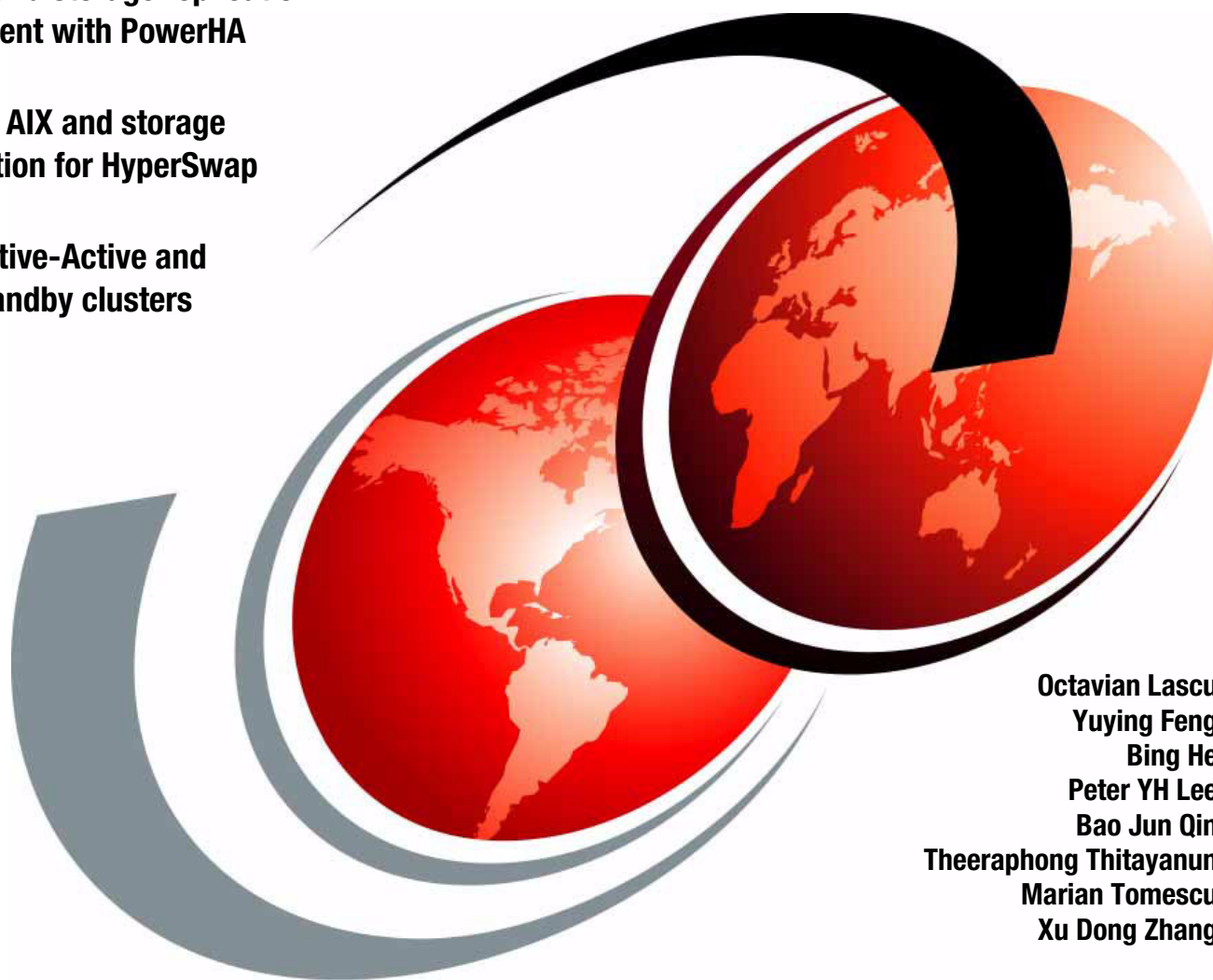


Deploying PowerHA Solution with AIX HyperSwap

Uses in-band storage replication
management with PowerHA

Describes AIX and storage
configuration for HyperSwap

Shows Active-Active and
Active-Standby clusters



Octavian Lascu
Yuying Feng
Bing He
Peter YH Lee
Bao Jun Qin
Theeraphong Thitayanun
Marian Tomescu
Xu Dong Zhang



International Technical Support Organization

Deploying PowerHA Solution with AIX HyperSwap

September 2014

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

First Edition (September 2014)

This edition applies to Version 7, Release 1, Modification 2 of IBM PowerHA Enterprise Edition (product number 5765-H24).

© Copyright International Business Machines Corporation 2014. All rights reserved.

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

Contents

Notices	vii
Trademarks	viii
Preface	ix
Authors	ix
Now you can become a published author, too!	x
Comments welcome	xi
Stay connected to IBM Redbooks	xi
Chapter 1. Introduction to PowerHA HyperSwap with the IBM DS8800	1
1.1 Overview and short history	2
1.2 Concepts and objectives	3
1.2.1 Basic configuration using PowerHA with HyperSwap	3
1.2.2 Basic workflow of PowerHA with HyperSwap	4
1.2.3 Typical solution for PowerHA with HyperSwap	4
1.2.4 The HyperSwap concept	5
1.3 Terminology and considerations	7
1.3.1 Planned/unplanned HyperSwap	7
1.3.2 Out-of-band versus in-band	8
1.3.3 Consistency Groups	9
1.3.4 PowerHA cross-site cluster	9
1.3.5 WAN considerations	9
1.3.6 Distance considerations	10
Chapter 2. PowerHA HyperSwap cluster planning	11
2.1 Introduction	12
2.2 HyperSwap architecture	13
2.3 Prerequisites	15
2.3.1 Hardware	15
2.3.2 Software requirements	15
2.4 HyperSwap for database applications	16
2.5 Testing environment description	17
2.6 Typical HyperSwap scenarios	18
2.6.1 Planned HyperSwap behavior details	19
2.6.2 Unplanned HyperSwap: Primary storage failure	20
2.6.3 Unplanned HyperSwap: All links to primary storage fail	21
2.6.4 Unplanned HyperSwap: Site down	22
2.6.5 Unplanned HyperSwap: Site partition	22
Chapter 3. PowerHA cluster with AIX HyperSwap Active-Standby for applications using a shared file system	25
3.1 Cluster description and diagrams	26
3.2 Installing a new configuration	28
3.2.1 Identifying the storage	28
3.2.2 Identifying the systems' HBA configurations	30
3.2.3 Zoning configuration	31
3.2.4 Configuring the storage	36
3.2.5 Enabling HyperSwap: Storage level	42
3.2.6 AIX configuration	43

3.2.7	PowerHA cluster configuration	50
3.2.8	Planned tests: Storage maintenance	64
3.2.9	Planned tests: Site maintenance	72
3.3	Migrating PowerHA cluster to HyperSwap enabled storage	74
3.3.1	Planning the cluster	75
3.3.2	Identifying the nodes and sites	75
3.3.3	Identifying and configuring the storage	75
3.3.4	Enabling HyperSwap in AIX	75
3.3.5	Reconfiguring the cluster for HyperSwap	76
3.4	Two-node cluster to four-node cluster with HyperSwap	83
Chapter 4. PowerHA HyperSwap cluster, Oracle stand-alone database, and ASM		103
4.1	Cluster description and diagrams	104
4.2	Storage configuration	105
4.2.1	LUN and mapping configuration	106
4.2.2	Zoning configuration	109
4.2.3	AIX disks information	110
4.3	Node configuration	113
4.3.1	AIX disk device driver and HBA attributes	113
4.3.2	Disk configuration	115
4.3.3	Time synchronization	117
4.4	Oracle installation and configuration on cluster nodes	119
4.4.1	Environment checking and configuration	119
4.4.2	Installing grid (Oracle Cluster Ready Services) and database software	123
4.4.3	Create a database instance on PS5n01base	124
4.4.4	Register the database instance on other nodes	125
4.4.5	Change the ASM disk group to the spfile	127
4.4.6	Test the database start-up and shutdown scripts	128
4.5	PowerHA configuration	130
4.5.1	Cluster topology	130
4.5.2	Cluster resources	133
4.5.3	Resource group configuration	135
4.6	Test scenarios	136
4.6.1	Node maintenance (planned)	136
4.6.2	Primary storage maintenance (planned)	139
4.6.3	Primary site maintenance (planned)	144
4.6.4	Node failure (unplanned)	149
4.6.5	Primary storage failure (unplanned)	151
4.6.6	Primary site failure (unplanned)	155
4.6.7	PPRC replication path failure (unplanned)	159
Chapter 5. PowerHA cluster with AIX HyperSwap Active-Active for applications using Oracle RAC		169
5.1	Cluster description and diagrams	170
5.1.1	Prerequisites	171
5.1.2	Implementation planning	171
5.2	Configuring the environment	172
5.2.1	Storage configuration	172
5.2.2	Storage area network configuration	174
5.2.3	LUN configuration in AIX and enabling HyperSwap	174
5.2.4	Oracle RAC cluster installation and configuration	177
5.2.5	PowerHA cluster installation and configuration	181
5.3	Test scenarios	188
5.3.1	Test method description	188

5.3.2 Primary storage maintenance (planned).....	189
5.3.3 Node failure (unplanned)	193
5.3.4 Primary storage failure (unplanned)	196
5.3.5 Site failure (unplanned).....	199
Related publications	203
IBM Redbooks	203
Online resources	203
Help from IBM	204

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

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

AIX®	IBM®	Redpaper™
DB2®	Parallel Sysplex®	Redbooks (logo)  ®
DS8000®	POWER®	System p®
eServer™	Power Systems™	System p5®
Geographically Dispersed Parallel Sysplex™	POWER6®	System Storage®
Global Technology Services®	POWER7®	System z®
GPFST™	PowerHA®	SystemMirror®
HACMP™	PowerVM®	Tivoli®
HyperSwap®	PureFlex®	
	Redbooks®	

The following terms are trademarks of other companies:

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

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

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

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

Preface

This IBM® Redpaper™ publication will help you plan, install, tailor, and configure the new IBM PowerHA® with IBM HyperSwap® clustering solution.

PowerHA with HyperSwap adds transparent storage protection for replicated storage, improving overall system availability by masking storage failures.

The PowerHA cluster is an Extended Distance cluster with two sites. It manages, in principle, the replicated storage infrastructure through HyperSwap functionality.

The storage is provided by two DS8800s configured to replicate each other using Metro Mirror Peer-to-Peer Remote Copy (PPRC) synchronous replication. DS8800 supports in-band (SCSI commands) communication, which is used to manage (and automate) the replication using IBM AIX® HyperSwap framework and PowerHA automation and management capabilities.

Authors

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

Octavian Lascu is a Project Leader at the International Technical Support Organization, Poughkeepsie Center. He writes extensively and teaches IBM classes worldwide on all areas of AIX, IBM Power Systems™, Linux, and Clustering. Before joining the ITSO 12 years ago, Octavian worked in IBM Integrated Technology Services Romania as an IT Infrastructure Consultant.

Yuying Feng is a Senior IT Specialist with the IBM Systems and Technology Group, IBM Greater China Group organization. Mr. Feng is responsible for providing technical support to IBM telecommunications industry sellers and clients in the South China region for IBM POWER® system-related solution design. Mr. Feng has over 11 years experience in supporting POWER products and IBM storage products. He has participated in many key projects for the telecom clients in the South China region.

Bing He is a Consulting I/T Specialist of the IBM Advanced Technical Skills (ATS) team in China. He has 14 years of experience with IBM Power Systems. He has worked at IBM for over seven years. His areas of expertise include PowerHA, PowerVM, and performance tuning on AIX.

Peter YH Lee is a Senior Certified IT Specialist of Systems and Technology Group in the IBM Greater China Group organization. Mr. Lee is in the Architect role of the High End Center of Competency currently to provide pre-sales technical support and advanced technical solution design based on high-end systems. Mr. Lee has over 19 years experience in supporting UNIX products and has participated in many large scale projects in the Greater China region.

Bao Jun Qin works with IBM Global Technology Services as a Senior IT Architect. He joined IBM in 2001. Currently, he is the Client Technical Architect (CTA) for Industrial and Commercial Bank of China Limited (“ICBC”). He is a Senior Certified IT Specialist in the IBM Greater China Group organization. Before the CTA role, he provided pre-sales technical support in the IBM China Advanced Technical Support team. He has 15 years experience in supporting UNIX. His skills include AIX, Power Virtualization, PowerHA, performance tuning, and application support.

Theeraphong Thitayanun is a Certified Consulting IT Specialist for IBM Thailand. His main responsibility is to provide services and support in all areas of the System p product set. His areas of expertise include IBM AIX/IBM Parallel System Support Program (PSSP), logical partitioning (LPAR)/Hardware Management Console (HMC), Product Lifecycle Management (PLM), GPFS/Andrew File System (AFS), HACMP, HACMP/XD for Metro Mirror, and IBM DB2® Universal Database. He holds a Bachelors degree in Computer Engineering from Chulalongkorn University and, as a Monbusho student, a Masters degree in Information Technology from Nagoya Institute of Technology, Japan.

Marian Tomescu has 15 years experience as an IT Specialist and currently works for IBM Global Technologies Services in Romania. Marian has nine years of experience in Power Systems. He is a certified specialist for IBM System p® Administration, High Availability Cluster Multi-Processing (HACMP™) for AIX, IBM Tivoli® Storage Management Administration Implementation, Oracle Certified Associated, IBM eServer™, Storage Technical Solutions Certified Specialist, and Cisco Information Security Specialist. His areas of expertise include Tivoli Storage Manager, PowerHA, IBM PowerVM®, IBM System Storage®, AIX, IBM General Parallel File System (GPFS™), VMware, Linux, and Windows. Marian has a Masters degree in Electronics Images, Shapes and Artificial Intelligence, from Polytechnic University - Bucharest, and Electronics and Telecommunications, Romania.

Xu Dong Zhang is an IBM Certified Specialist for Power Systems. He holds a Masters degree in Computer Science specializing in Information Technology. His areas of expertise include Power Systems and IBM PureFlex® System solutions, including Power Cloud, IBM Systems Director, PowerVM Virtualization Technology, and PowerHA High Availability Technology. He has been with IBM China Systems and Technology Group for 10 years and has a total 16 years of IT experience. Previously, he was a Senior Field Technical Sales Support (FTSS) Specialist and a Systems Architect in the IBM Power System Pre-Sales team.

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

William G. White, Dino Quintero
International Technical Support Organization, Poughkeepsie Center

Ravi A. Shankar
IBM Austin

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

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

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

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

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:

<http://www.facebook.com/IBMRedbooks>

- ▶ Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- ▶ Look for us on LinkedIn:

<http://www.linkedin.com/groups?home=&gid=2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

<http://www.redbooks.ibm.com/rss.html>



Introduction to PowerHA HyperSwap with the IBM DS8800

This chapter describes the concept of the HyperSwap feature introduced with IBM PowerHA SystemMirror® Enterprise Edition 7.1.2¹ and designed to work with IBM DS8800² storage. In this chapter, we present an overview of the solution, basic workflow, terminology, and considerations for building a high availability solution based on PowerHA HyperSwap with IBM DS8800.

¹ PowerHA 7.1.3 is available.

² For the latest supported storage model and type, check with your IBM representative.

1.1 Overview and short history

Available for some time, Open HyperSwap has been delivered as an IBM Tivoli Total Storage Productivity Center feature in combination with IBM DS8000® storage. Recently, the HyperSwap capability was announced also on IBM Power Systems running AIX as a new feature in conjunction with IBM PowerHA SystemMirror 7.1 Enterprise Edition (announced October, 2012). The PowerHA with HyperSwap solution created a lot of interest from enterprise clients because it can address a number of issues in the IT market today:

- ▶ Twenty-four x 7 x 365 operation causing difficulties to schedule downtime (for maintenance)
- ▶ Continuous availability with protection against storage failures
- ▶ Further improvement in applications' serviceability levels
- ▶ Standardization and automation management of data replication and data center failover
- ▶ Roadmap to support Active-Active data center workloads

At announcement time, PowerHA with HyperSwap supports the IBM DS8800 storage subsystem. For the latest list of hardware and software supported with HyperSwap, read the IBM PowerHA SystemMirror Enterprise Edition V7 announcement letter:

http://www-01.ibm.com/common/ssi/ShowDoc.wss?docURL=/common/ssi/rep_ca/4/760/ENUSJ12-0364/index.html&lang=en&request_locale=en

PowerHA with HyperSwap with DS8800 is an advanced technology that allows mission critical applications, such as database environments, to be run in a cluster within a campus or across two distant³ sites to achieve high availability (through automation). The PowerHA with HyperSwap solution improves data and application serviceability compared to traditional clustering technology in the following ways:

- ▶ Mission critical data is replicated synchronously from primary storage to secondary storage for better data resiliency.
- ▶ Server to SAN storage connections can automatically be switched to secondary storage in the event of a primary storage failure. The storage swap is transparent to applications.
- ▶ Application continuity is maintained due to the extremely fast failover time of SAN connections to secondary storage via the AIX Path Control Module (PCM) device driver.

The HyperSwap technology was first introduced on the IBM System z® platform together with the IBM Geographically Dispersed Parallel Sysplex™/ Peer-to-Peer Remote Copy (GDPS/PPRC) offering in 2002.

GDPS/PPRC has been widely adopted by clients who require the highest level of failure resiliency on IBM System z. HyperSwap allows continuous application availability by protecting against storage outages. GDPS/PPRC Multi-Site Workload can further extend the solution to geographically dispersed data centers tens of kilometers apart to achieve disaster recovery automation.

In the event of storage failure, scheduled maintenance, or site failure, HyperSwap controls failover to secondary storage to achieve continuous data accessibility. Existing server to storage connections will be swapped to new primary storage quickly to allow application continuity.

³ At the time of this writing, only synchronous replication was supported with PowerHA and HyperSwap. Check the latest announcements for the supported distance between sites and the types of storage replication solutions.

The HyperSwap solution has been available for 10 years on the IBM System z platform. Metro Mirror synchronous replication technology is also one of the most popular storage resiliency solutions in the market.

IBM PowerHA with HyperSwap offers similar architecture on the Power System and the AIX platform. This is an enterprise class technology that further enhances the reliability, availability, and serviceability (RAS) level of the overall system.

1.2 Concepts and objectives

In this section, we cover the concept and the workflow of the PowerHA HyperSwap solution with the IBM DS8800.

1.2.1 Basic configuration using PowerHA with HyperSwap

To use PowerHA HyperSwap, you need to create a cluster with at least two IBM Power System Servers and two⁴ IBM DS8800 storage subsystems. AIX 6.1 TL08 or AIX 7.1 TL02 (or later) and PowerHA 7.1.2 Enterprise Edition (or later) must be installed on IBM Power System servers. IBM Metro Mirror needs to be configured across two DS8800 storage subsystems.

Software and microcode/firmware levels: Always check the latest versions of AIX, PowerHA, and Storage microcode for HyperSwap support.

The basic configuration is illustrated in Figure 1-1.

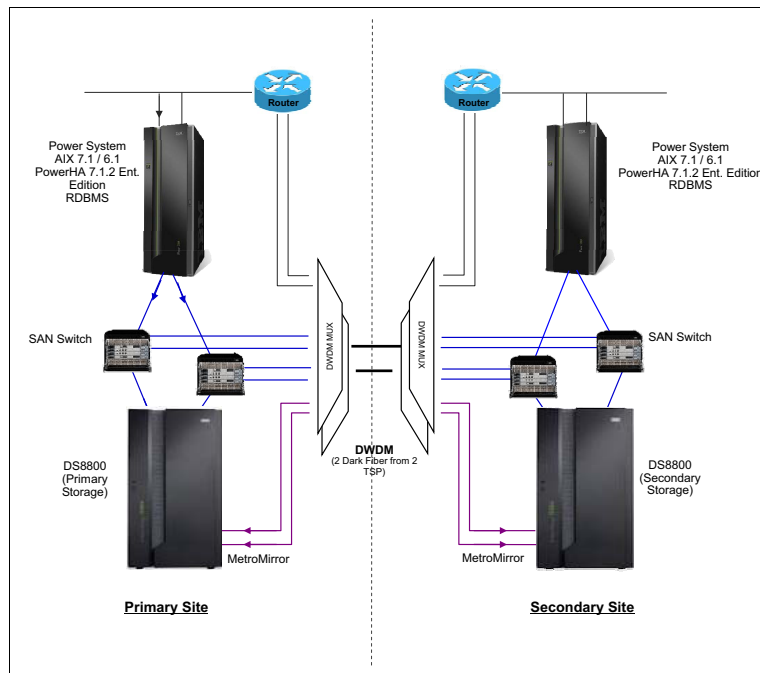


Figure 1-1 PowerHA HyperSwap reference architecture

⁴ As of PowerHA SystemMirror for AIX 7.1.3, single compute node HyperSwap deployment is also supported.

1.2.2 Basic workflow of PowerHA with HyperSwap

PowerHA with HyperSwap is an advanced feature that allows continuous application availability by masking storage outages. In Figure 1-1 on page 3, applications running on IBM Power System servers will perform data updates to the primary DS8800 storage. Updates will be replicated to the secondary DS8800 storage through Metro Mirror remote copy services, which is also known as Peer-to-Peer Remote Copy (PPRC). Because updates to primary storage are replicated synchronously to secondary storage, data integrity is maintained to ensure a recovery point objective (RPO⁵) equal to zero.

PowerHA HyperSwap takes advantages of in-band (SCSI) command capability to establish communication between the IBM Power System servers and the IBM DS8800 storage subsystems. As a result, designing the PowerHA with HyperSwap solution is greatly simplified, facilitating the storage management of the cluster via the integrated PowerHA interface, further improving cluster reliability and application availability.

In the event of a storage subsystem failure, SAN switch failure, or host bus adapter (HBA) failure, Cluster Aware AIX (CAA) will detect the failures. CAA will then generate cluster events that will trigger PowerHA events to send in-band (SCSI) commands to the storage subsystems to perform the necessary steps to promote secondary storage (target) to become primary storage (source).

CAA will also trigger AIX Path Control Module (PCM) to switch the SAN connections to the new primary storage subsystem with all devices, file systems, logical volumes, volume groups, and raw disks remaining unchanged from the application perspective. Because the storage failover is completed within SAN timeout period, and all device names remain unchanged, the application will not be interrupted by the underlying storage subsystem failure.

For other failure scenarios, such as server failure, network adapter failure, site failure, and so on, the behavior of PowerHA HyperSwap cluster will be the same as a standard PowerHA cluster. CAA will detect the hardware failure and will trigger PowerHA to perform the failover of resources to the backup server. The server to SAN storage connections remain unchanged (no storage swap occurs if the primary storage remains available).

When the PowerHA resource group is configured in Active-Standby mode, applications need to be restarted on the backup server. When the PowerHA resource group is configured in concurrent mode (for example, parallel database), user applications will re-establish connections to another server with minimal interruption.

1.2.3 Typical solution for PowerHA with HyperSwap

PowerHA with HyperSwap is suitable for the following usage scenarios:

- ▶ Storage continuity solution: Install the PowerHA HyperSwap cluster within a campus that allows applications to achieve continuous availability by protecting against storage errors. In addition, this solution also facilitates the scheduled (planned) downtime of the storage subsystem for upgrade or maintenance work in mission critical environments.

⁵ The open platform typically measures the recovery time objective (RTO)/RPO of the infrastructure alone. The open platform does *not* consider the impact to user transactions. System integration also needs to be done by the client to maintain a minimal RTO or RPO.

- ▶ **Active-Active data center solution:** Install the PowerHA HyperSwap cluster spanning across two sites with a specific application (for example, parallel database) that can run across two sites concurrently with shared disk access. The user application can be load-balanced between the two sites with access to shared data, while application data integrity is maintained cross-site through Metro Mirror and application data access is maintained cross-site through HyperSwap technology.

In the event of a cluster component or site failure, the user application can always establish connection to the same data via a different site. As a result, PowerHA HyperSwap extends the capability to build an Active-Active data center solution with the appropriate application installed.

- ▶ **Disaster recovery solution with improved serviceability:** Install the PowerHA HyperSwap stretched cluster across two sites with Power System servers and DS8800 storage subsystems cross-coupled to allow the automatic takeover to the secondary site in the event of a server or site failure. With HyperSwap enabled, increased application resiliency enhances the overall serviceability of the stretched cluster.

1.2.4 The HyperSwap concept

HyperSwap for AIX provides the framework to manage (transparent to the application) replicated storage (DS8800 Metro Mirror environment). However, automating storage operations can be achieved through a clustering mechanism capable of identifying various failures and providing the logic capable of handling the events.

Comparison between HyperSwap and PowerHA Extended Distance

PowerHA with HyperSwap offers an additional advantage over traditional PowerHA Extended Distance in the way that it allows the transparent swap of the underlying disk device, logical volumes, or file systems to secondary storage without the need to modify the application.

Traditional PowerHA Extended Distance (no HyperSwap)

Traditional PowerHA Extended Distance is implemented in the way that data from the primary site can be replicated to the remote site via synchronous or asynchronous copy services.

PowerHA Extended Distance supports IBM DS8000 series storage, as well as storage from other vendors. Check the IBM PowerHA SystemMirror Enterprise Edition V7 announcement letter for the latest list of hardware and copy services supported.

The data integrity is maintained by copy services (vendor-specific), while PowerHA handles the event management and resource takeover to the secondary site. Due to the fact that two different devices need to be maintained on primary storage and secondary storage servers, user scripts will need to include the steps to adjust the device name after failover or fallback. As a result, more customization work is required and some applications, such as parallel database (sensitive to underlying device name changes), might not function correctly in such an environment.

For an illustration of the traditional PowerHA Extended Distance configuration using disk replication technology and using different device names on the primary storage and secondary storage servers, see Figure 1-2 on page 6.

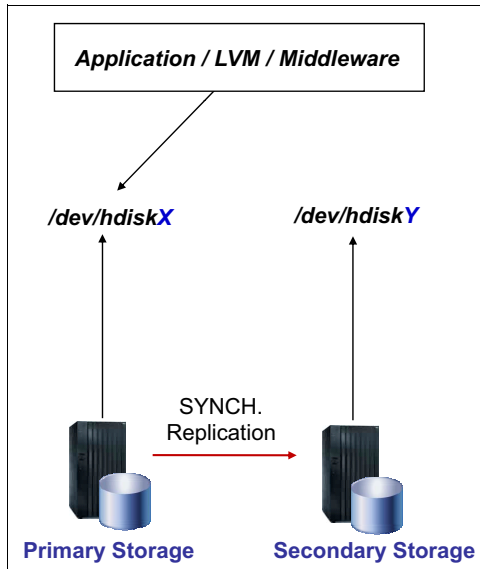


Figure 1-2 Traditional PowerHA Extended Distance with different device names

PowerHA Extended Distance with HyperSwap

PowerHA with HyperSwap, however, relies on Metro Mirror copy services for synchronous data replication from primary storage to secondary storage to ensure data integrity, and adds in-band replication control.

PowerHA with HyperSwap allows both primary storage and secondary storage to share the same disk name, logical volume, or file systems. The AIX path control module (PCM⁶) works together with the Power HyperSwap feature to perform the automatic switch of the SAN connection (path) to the secondary storage. The AIX PCM also works with the Power HyperSwap feature to send in-band commands to the secondary storage (PPRC target) to promote the target to become the primary storage and to ensure continuous disk access and application continuity. Because HyperSwap is transparent to applications, the system (AIX) device names are identical for both primary and secondary storage. The HyperSwap solution allows the application (including a parallel database environment) to run without changes or special customization (see Figure 1-3 on page 7).

⁶ The path control module is a component of the AIX disk device driver.

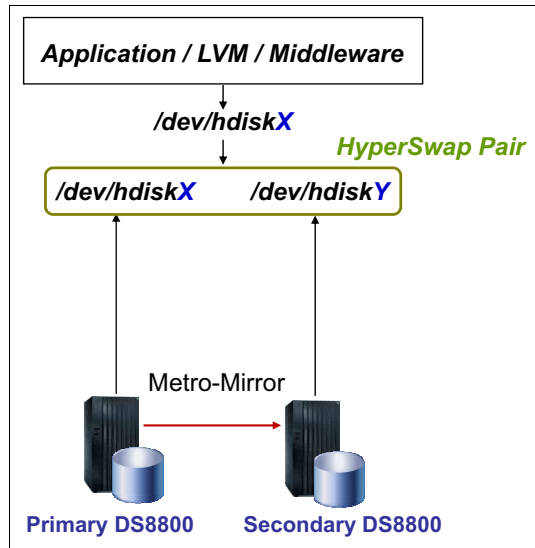


Figure 1-3 PowerHA with HyperSwap identical device names

1.3 Terminology and considerations

In this section, we describe some of the terminology and considerations used in building an enhanced high availability solution using PowerHA HyperSwap with the IBM DS8800.

1.3.1 Planned/unplanned HyperSwap

PowerHA supports two ways to trigger HyperSwap:

- ▶ **Planned HyperSwap:** With PowerHA HyperSwap, an administrator can schedule downtime in a much easier way. PowerHA with HyperSwap offers an option that allows an administrator to trigger HyperSwap on demand, which is known as *Planned HyperSwap*. Because this is a planned activity, the Fibre Channel (FC) connection does not need to wait for any timeout. The FC connection and Metro Mirror pair will be swapped quickly within seconds to access the secondary storage. With Planned HyperSwap, it becomes possible to schedule downtime for the storage subsystem to perform disk maintenance, a microcode upgrade, and reconfiguration without requiring user applications to be stopped.
- ▶ **Unplanned HyperSwap:** In the event of an unexpected failure of the primary storage subsystem, SAN switches, or FC connections, where servers cannot perform any read/write operation to primary storage, CAA will detect the failure and will trigger PowerHA to perform *unplanned HyperSwap*. Due to the fact that *unplanned HyperSwap* needs to wait for FC timeout, the failover time will be slightly longer than for a *planned HyperSwap*. *Unplanned HyperSwap* is best fit to protect mission critical applications to achieve near-continuous application availability.

1.3.2 Out-of-band versus in-band

PowerHA HyperSwap requires a FC communication channel between servers and storage subsystems so that storage management commands can be sent to storage subsystems to manage replication direction and storage access. The traditional way of managing replicated storage is achieved by using external elements via LAN (based on TCP/IP) connections. The traditional mechanism uses out-of-band communication (shown in Figure 1-4).

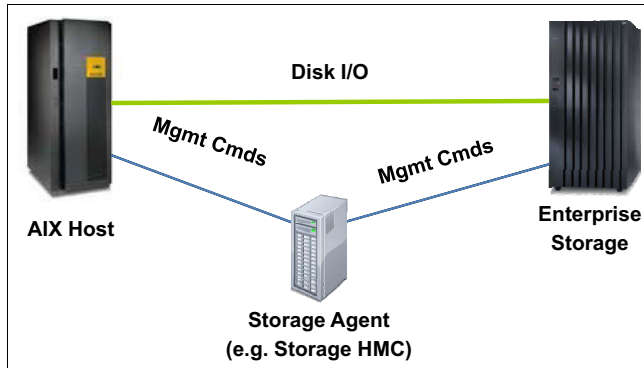


Figure 1-4 Out-of-band command

PowerHA HyperSwap supports IBM DS8800 Metro Mirror in-band communication where storage management commands are sent over the SAN (FC traffic and SCSI commands) using the same communication path as host disk I/O, as illustrated in Figure 1-5.

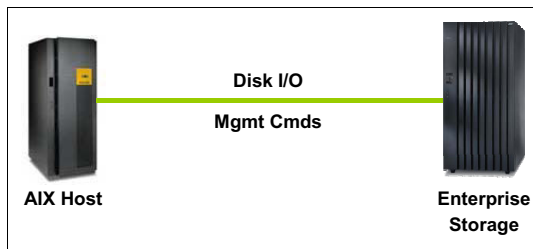


Figure 1-5 In-band command

In-band communication has the following advantages over out-of-band communication:

- ▶ Simplified network (LAN) and SAN infrastructure. This is important for complex data center environments or for stringent security control.
- ▶ Simplified command chain (eliminating “the middleman”)
- ▶ Better integration between the storage and server environment for ease of management. Communication via SAN is usually faster than LAN and can improve response and reconfiguration time in reaction to failures.
- ▶ Reliability and performance: Because fewer components are involved, in-band communication is more reliable and faster.

1.3.3 Consistency Groups

PowerHA with HyperSwap also takes advantage of the IBM DS8800 Consistency Group feature. User data can span across multiple IBM DS8800 storage subsystems. HyperSwap switches all logical unit numbers (LUNs) at the same time to secondary storage subsystems, therefore, maintaining data integrity. This feature is especially important for applications running across multiple storage subsystems. HyperSwap with Consistency Group support greatly simplifies the overall data management across multiple storage subsystems and multiple sites.

1.3.4 PowerHA cross-site cluster

The PowerHA SystemMirror Enterprise Edition 7.1 cluster can be configured across sites (two sites at the time of writing this paper). PowerHA offers two types of configurations:

- ▶ Stretched cluster: The PowerHA cluster is defined as single cluster spanning across two sites with a single PowerHA repository disk, Cluster Aware AIX (CAA)⁷, configured. PowerHA 7.1 uses multicast⁸ traffic for heartbeat communication among all member nodes within a stretched cluster.

Network infrastructure must support multicast traffic (switches, routers, and firewalls). Some applications that require concurrent data access from multiple nodes also use multicast communication; in this case, plan to configure a stretched cluster. The test clusters presented later in this document are configured as stretched clusters.

- ▶ Linked cluster: PowerHA is configured across two sites with a PowerHA cluster repository (CAA) configured in each site. Therefore, the cluster in each site can be operated separately with minimum inter-cluster communication required. This configuration does not require multicast traffic between sites, although multicast traffic will still require nodes in the same site. A linked cluster does not support concurrent resource groups. As a consequence, a linked cluster configuration cannot be used to deploy Active-Active solutions with concurrent storage access using HyperSwap across two sites.

1.3.5 WAN considerations

PowerHA with HyperSwap is designed to work in a single data center environment to provide a storage continuity solution or across data centers for improved disaster recovery and an Active-Active data center solution. For the Active-Active across data center solution, consider the following information:

- ▶ Redundant WAN: It is always important to ensure redundancy for connections between sites. Two sets of dense wavelength division multiplexing (DWDM) channels from separate providers are advised. For deploying PowerHA with HyperSwap solutions, it is also advised to isolate the communication channels used for different purposes, that is, Metro Mirror (PPRC) paths, SAN data access communication (FC and inter-switch links (ISLs)), and LAN/WAN communication. This isolation is suggested to ensure an adequate quality of service for each type of communication and to ease communications debugging, if communication issues occur.
- ▶ Inter-site bandwidth: This inter-site bandwidth is determined by the application transaction volume. Inter-site bandwidth determines the time taken by the initial storage synchronization sequence and subsequent recovery events.

⁷ AIX clustering infrastructure

⁸ With the latest PowerHA 7.1.3, unicast for cluster communication is also available as an option.

Network isolation (complete site isolation)

PowerHA provides a solution to determine which site will survive when all network connections between the two sites, primary and disaster recovery (DR), fail. We advise you to configure the `/usr/es/sbin/cluster/netmon.cf` file to define an IP address for a piece of external equipment that facilitates a PowerHA decision in the case of site isolation. You can select an IP address of one type of reliable equipment in the primary site (where the primary storage is installed). Or, if you can rely on the WAN infrastructure and if it is available, you can configure an IP address for one piece of equipment in a third site, which is different from both the primary and DR sites.

1.3.6 Distance considerations

When PowerHA HyperSwap is deployed in a stretched cluster, the distance between the two data centers needs to be considered carefully in these areas:

- ▶ **Network latency:** Due to light speed limitation (196,000 km/sec over fiber), a round trip for every 100 km (62 miles), including equipment latency and travel latency, will be approximately 1ms for each packet. A user transaction can consist of multiple transfers between the sites. Therefore, the longer the distance, the slower the application's response time.
- ▶ **Read/write ratio of application:** The read operation does not affect data integrity because the data is read from the primary storage directly. Write operations require an update to primary storage, which must be replicated to remote storage to ensure data integrity across sites. The larger the number of write operations, the longer the application's response time. If the distance causes an unacceptable performance penalty, consider a shorter distance between the two data centers.
- ▶ **Metro Mirror supported distance:** IBM DS8800 Metro Mirror (synchronous replication) currently supports a maximum distance of 300 km (186.4 miles). Considering the previously listed factors and application performance requirements, typical Metro Mirror solutions are implemented within a 100 km distance between sites. Active-Active across data centers solutions (for example, a parallel database) will generate more communication overhead; therefore, consider a shorter distance between the two data centers.

Distance between sites: The distance considerations listed previously apply to all storage vendors' solutions. Each specific implementation must be thoroughly tested to measure the actual performance impact to your applications.



PowerHA HyperSwap cluster planning

This chapter provides planning information for the IBM PowerHA HyperSwap solution. HyperSwap provides the ability to nondisruptively switch from using the primary volume of a mirrored pair (Peer-to-Peer Remote Copy (PPRC)) to using the PPRC target. The storage side of HyperSwap is based on IBM Metro Mirror. This advanced technology provides the capability for a cluster to span two sites with the storage and servers cross-coupled in a manner that keeps the application resilient through either a storage subsystem outage or server outage. In disk errors, the primary disk subsystem will be transparently switched over to the secondary disk subsystem, which makes the solution suitable for a variety of workloads, including Active-Active database deployments.

2.1 Introduction

HyperSwap is introduced as a facility of PowerHA SystemMirror for AIX Enterprise Edition in combination with select storage subsystems. This facility supports stretched cluster and linked cluster configurations.

HyperSwap can help to achieve these functions:

- ▶ Multisite PowerHA cluster with continuous storage availability
- ▶ Nondisruptive storage swap for application continuity in the event of one storage outage
- ▶ Storage maintenance without application downtime

A typical PowerHA HyperSwap cluster topology (shown in Figure 2-1) consists of the following components:

- ▶ Two nodes in each site, configured as a two-site stretched cluster
- ▶ Two DS8800¹ storage subsystems, one in each site, configured for Metro Mirror
- ▶ Highly available SAN infrastructure spanning over both sites

The network between the two sites can be connected with or without an IP router over dense wavelength division multiplexing (DWDM).

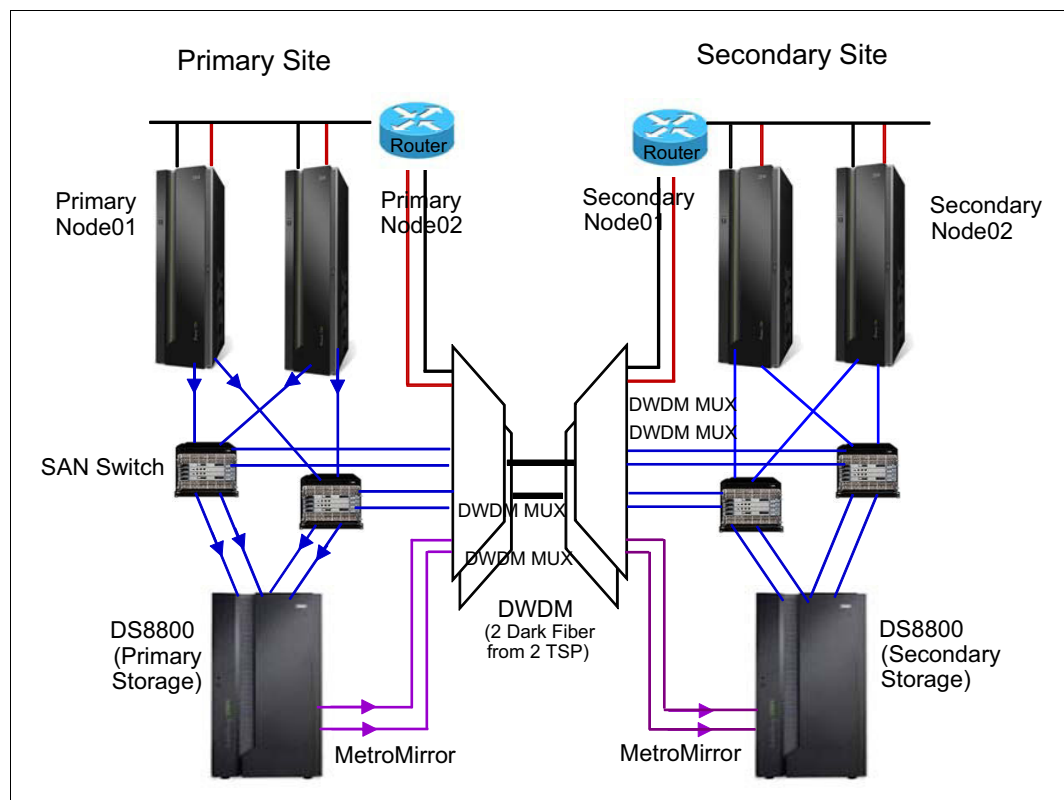


Figure 2-1 A typical topology of a HyperSwap solution

¹ DS8800 firmware tested for this paper: 6.3 or higher, microcode: 86.30.49.0 or higher

2.2 HyperSwap architecture

The HyperSwap solution consists of several software components (layers):

- ▶ AIX: HyperSwap is enabled in AIX Path Control Module (PCM)
- ▶ Firmware of DS8800 disk subsystem with in-band communication capability
- ▶ PowerHA Enterprise Edition 7.1.2

The basic function (see the diagram in Figure 2-2) is provided by AIX HyperSwap, which is integrated in the AIX PCM. HyperSwap manages the paths connecting to both the primary DS8800 and the secondary DS8800, and combines the paths into one path group. Only the primary side of the path group (connected to the PPRC source storage subsystem) is active and available for performing the I/O workload.

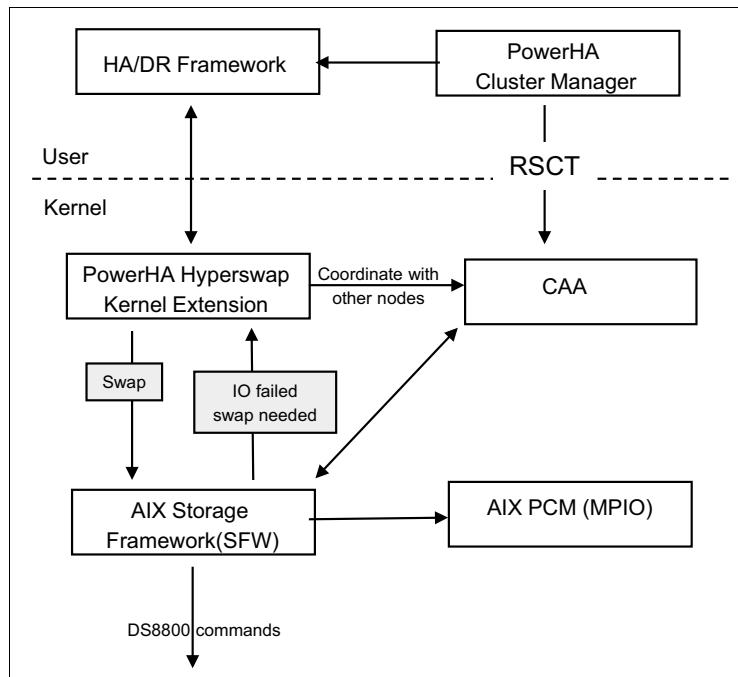


Figure 2-2 HyperSwap architecture overview

The diagram shown in Figure 2-3 describes the common view of the PPRC source disk and target disk as they are mapped to the host, *before enabling HyperSwap*.

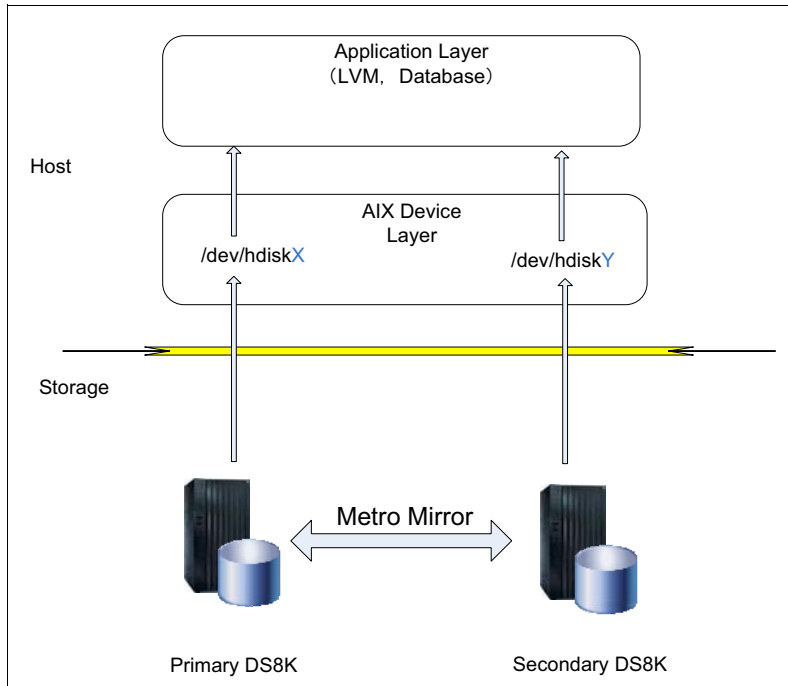


Figure 2-3 Device view of PPRC source and target within a host without HyperSwap

Both source and target logical unit numbers (LUNs) are *configured* on the same host. However, only the source LUN can be *accessed* for I/O, while the target LUN is blocked for I/O operations.

After enabling *HyperSwap* in AIX, a new, “composite” device is presented to the application layer, as seen in Figure 2-4.

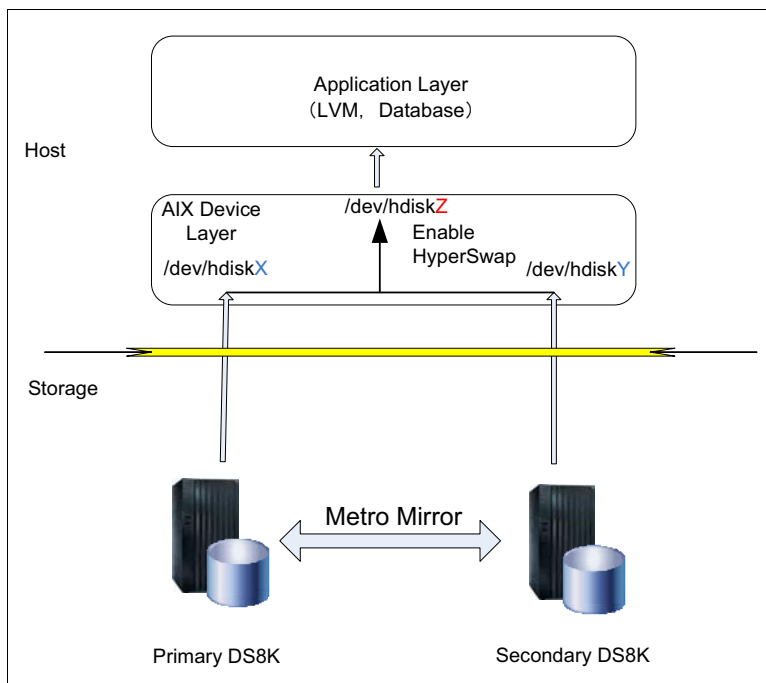


Figure 2-4 AIX disk device view with HyperSwap enabled

After enabling AIX HyperSwap, the source LUN (hdiskX) and target LUN (hdiskY) are combined into one logical device. The actual name² of the hdisk depends on the method used to name the HyperSwap enabled device:

- ▶ If using the `migrate_disk` option, the logical name of the source LUN will be kept (hdiskX).
- ▶ If using the `new` or `new_and_existing` options, a new logical name will be generated (hdiskZ in our example).

2.3 Prerequisites

This section introduces the prerequisites for building a PowerHA HyperSwap solution.

2.3.1 Hardware

The hardware required for the HyperSwap configuration consists of servers, a storage subsystem, and local and storage area network connectivity (LAN and SAN).

Servers

The following IBM systems run a supported level of AIX: IBM POWER5, POWER6®, or POWER7® technology-based processors, including the IBM Power Systems, System p, System p5®, eServer™ p5, and eServer pSeries server product lines.

Storage subsystems

IBM DS8800 is supported with firmware R6.3sp4 (86.xx.xx.x) or higher. Check the following URL:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1003740>

Important: HyperSwap requires you to specify a host profile to `pSeriesPowerswap` (“IBM pSeries - AIX with Powerswap support”).

If existing LUNs will be used for HyperSwap, the following *DSCLI commands* can help you to check and change to a supported profile:

- ▶ Check with `lshostconnect`
- ▶ Change with `chhostconnect`

Network and SAN

No specific requirements (normal connectivity support) exist for the network and SAN.

Range extenders: Communication range extenders (for example, Wavelength Division Multiplexing (WDM) equipment) are not described in this document. Consult your communication infrastructure provider for long-distance connectivity solutions design and implementation.

2.3.2 Software requirements

This section describes AIX and PowerHA required levels and additional considerations.

² You can change the device name with a name of your choice by using the AIX `rendev` command.

PowerHA SystemMirror

PowerHA SystemMirror Enterprise Edition V7.1.2 Service Pack 3 or higher is required for HyperSwap support.

Operating system

PowerHA SystemMirror V7 is supported on AIX V6.1 and AIX V7.1. However, for HyperSwap, the following minimum AIX levels are required:

- ▶ AIX Version 6, Release 1, Technology Level 8, Service Pack 2 (AIX 6.1TL08SP2)
- ▶ AIX Version 7, Release 1, Technology Level 2, Service Pack 2 (AIX 7.1TL02SP2)

Additional considerations

In addition to AIX and PowerHA requirements, you also need to understand the following considerations when planning for HyperSwap:

- ▶ In-band communication (HyperSwap requirement) is supported with either Fibre Channel (FC) (host bus adapter (HBA) FC adapter or virtual FC adapter (VFC)) or FC over Ethernet (FCoE).
- ▶ *HyperSwap is not supported with virtual SCSI (regardless of the Virtual I/O Server version).*
- ▶ At the time of the initial version of this paper (March 2013), HyperSwap was only supported with AIX default PCM. *The DS8800 IBM Subsystem Device Driver Path Control Module (SDDPCM) is NOT supported³. If SDDPCM is installed, it must be uninstalled before configuring HyperSwap.*
- ▶ Live Partition Mobility requires HyperSwap to be disabled for all the affected mirror groups.
- ▶ Before enabling HyperSwap, you need to configure PPRC paths and pairs (using the DSCLI). HyperSwap in-band commands can only change the replication status (no configuration changes).

2.4 HyperSwap for database applications

In our test environment, we deployed and tested the following configurations:

- ▶ Stand-alone application running on a single node (stand-alone Oracle database using journaled file system 2 (JFS2) storage), in an Active-Passive cluster configuration.
- ▶ Stand-alone application running on a single node, using application storage management (PowerHA manages storage as raw disks), that is, a stand-alone Oracle database using Automatic Storage Management (ASM) in an Active-Passive cluster configuration.
- ▶ Clustered application on a four-node cluster, using application-provided storage management (with concurrent access). We have tested a cluster running Oracle Real Application Clusters (RAC) 11gR2 (11.2.03) with ASM. The software requirements for Oracle are the same as for standard (non-HyperSwap) Oracle RAC configurations. The licensing requirements must be met according to the proposed configuration.

Application support: By design, HyperSwap is transparent to applications. Other applications can be deployed in the same environment, as well.

³ This might change in future microcode. Check the latest Release Notes.

2.5 Testing environment description

The goal of our exercise was to test the HyperSwap functionality in various cluster configurations. Some of the scenarios we have tested are briefly described here:

- ▶ Migrating a stand-alone Oracle database clustered with PowerHA from a non-HyperSwap environment to a HyperSwap enabled environment.
- ▶ Verifying the function of HyperSwap using an Oracle stand-alone instance (JFS2 and ASM storage).
- ▶ Deploying an Oracle 11gR2 clustered database (RAC using Oracle Cluster Ready Services (CRS) and ASM) on a PowerHA HyperSwap cluster.
- ▶ Testing for various failures, such as storage, node, or site.

Figure 2-5 on page 18 presents a diagram of our testing environment.

Restriction: The test environment that we used was not been designed to eliminate all possible single points of failure. Design your environment according to your availability requirements.

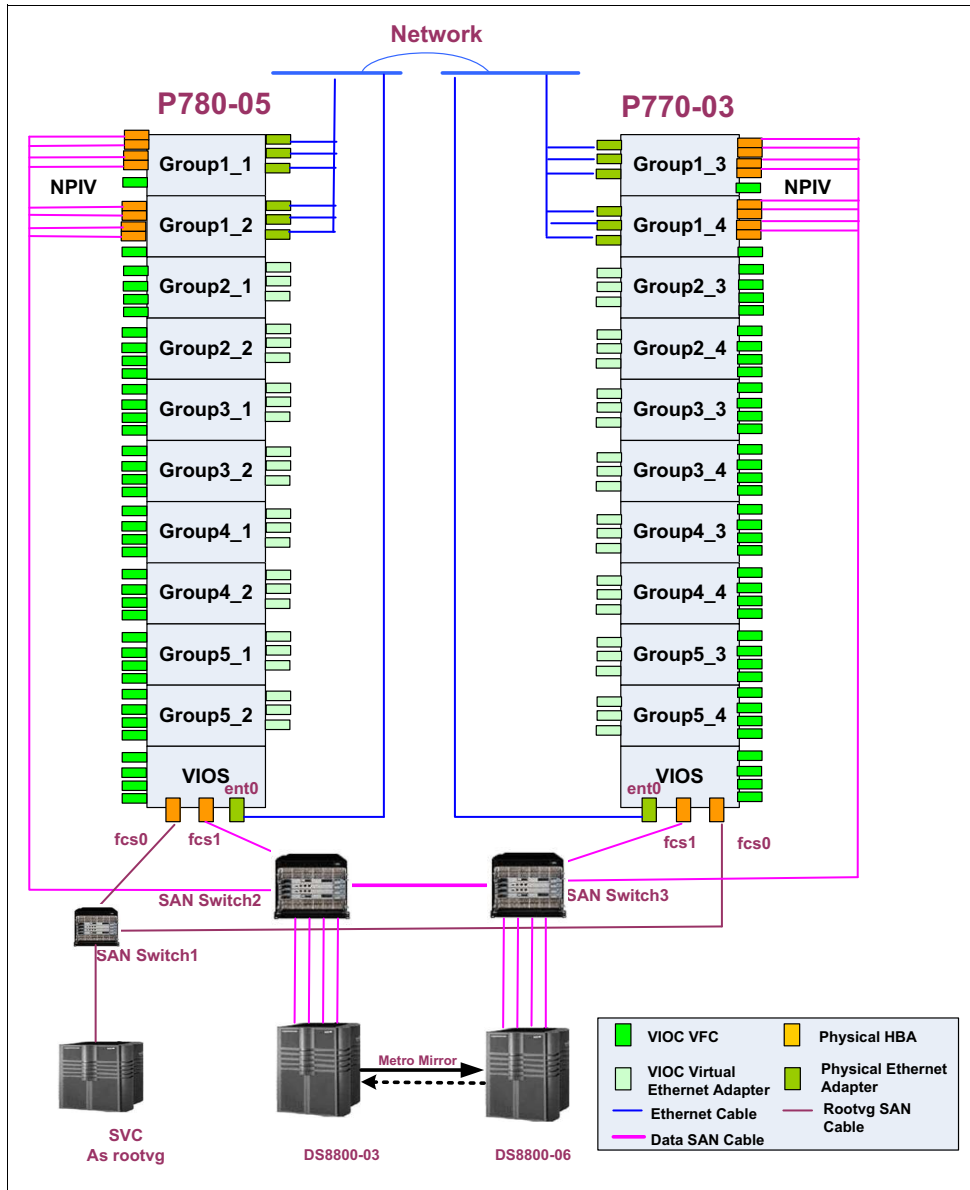


Figure 2-5 Test environment diagram

2.6 Typical HyperSwap scenarios

This section provides information about a PowerHA cluster and the expected HyperSwap behavior for various failures.

Unplanned HyperSwap

When primary storage fails, the OS that hosts the application detects the failure and, under the control of PowerHA, reacts by performing an automated PPRC failover. The application I/Os are transparently redirected to the secondary storage subsystem, therefore allowing the applications to continue running without any interruption.

Storage I/O timeout: I/O errors are detected by the operating system's SCSI driver layer. The decision to swap storage is made across multiple hosts to switch over to the secondary storage subsystem for all replicated LUNs defined in the PowerHA configuration, in a coordinated way.

For the duration of the HyperSwap swapping process, disk I/O is temporarily frozen. During this time, the applications only experience a delay that is shorter than the SCSI I/O timeout.

Unplanned HyperSwap recovery: After an unplanned HyperSwap, manually reverting to initial HyperSwap configuration is required. The `1spprc` command in AIX can be used for checking the replication status.

Metro Mirror recovery steps are required to revert the replication relationship to its original configuration.

Planned HyperSwap

The administrator can manually initiate a HyperSwap from the primary to the secondary storage subsystem. When the administrator has requested a planned HyperSwap, disk I/O activity is frozen (for a short while) and coordinated across the hosts in the cluster. The HyperSwap is performed, and then, I/O operations are resumed. Planned HyperSwap is helpful for maintenance on the primary storage and also for migrating from older storage.

Planned HyperSwap recovery: PowerHA provides the means (System Management Interface Tool (SMIT) menus) to revert to the original configuration, without the need to intervene at the storage level.

2.6.1 Planned HyperSwap behavior details

Figure 2-6 on page 20 describes the planned HyperSwap operation based on a PowerHA for AIX Enterprise Edition (EE) cluster with four nodes spanning two sites. A concurrent application is active on two nodes in the active site (Site_A). Node 1 and Node 2 have access to both storage subsystems (in Site_A and Site_B).

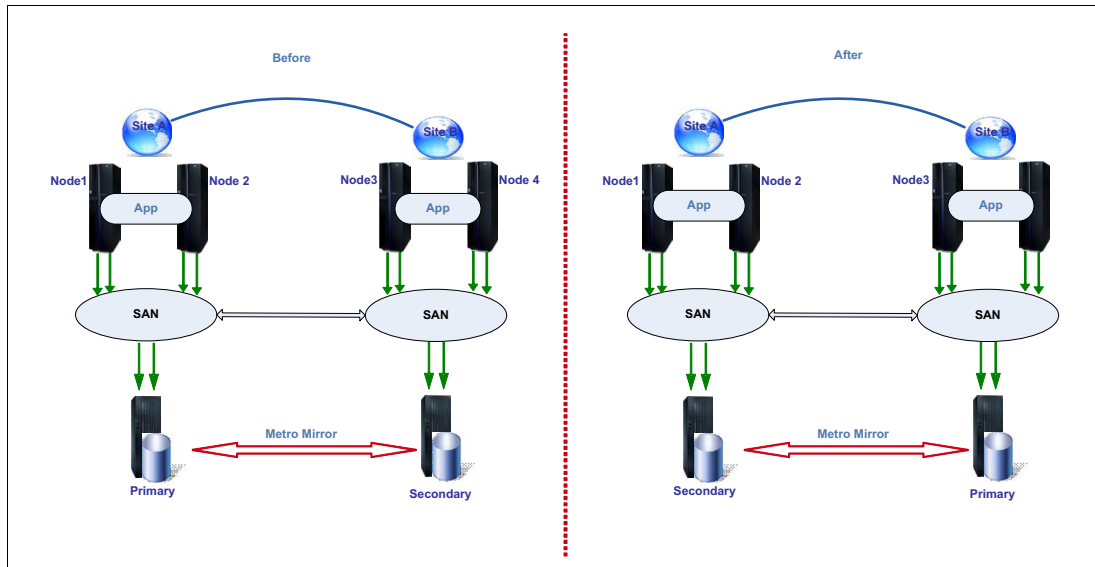


Figure 2-6 Planned HyperSwap diagram

When a planned HyperSwap operation is initiated, the applications using disks from Site_A storage are not affected. PowerHA validates to see whether all nodes that host the application (in this case, Node 1 and Node 2) can access the corresponding Site_B (PPRC target) disks. If one of the nodes (for example, Node 2) cannot access the Site_B storage, the operation is stopped and you are informed about the reason.

However, if the configurations are correct, a coordinated HyperSwap operation is performed so that both Node 1 and Node 2, together, start redirecting the application I/O to Site_B storage. This operation is performed with the cooperation of the AIX disk driver layer (on Node 1 and Node 2), and therefore, is completely transparent to the applications.

The planned HyperSwap reverts the replication direction. The storage in Site_A becomes the PPRC target.

If a planned swap completed successfully, the application I/O is now directed to Site_B disks, and Site_A storage can be taken offline for maintenance without affecting the application.

2.6.2 Unplanned HyperSwap: Primary storage failure

Figure 2-7 on page 21 describes the unplanned HyperSwap configuration in a PowerHA EE cluster with four nodes on two sites. The concurrent application is active on the primary site nodes. Node 1 and Node 2 both have access to storage subsystems in Site_A and Site_B.

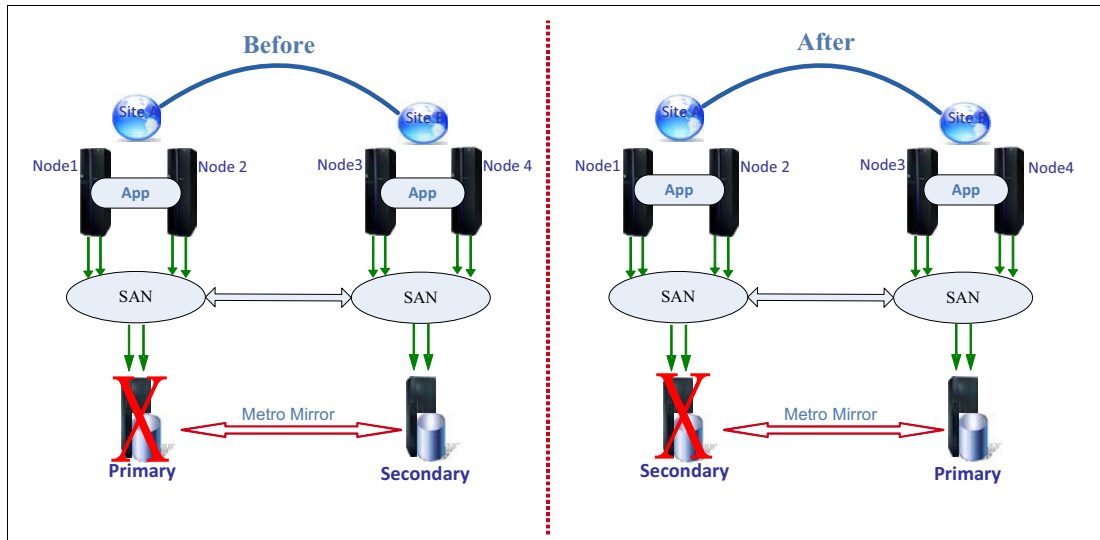


Figure 2-7 Primary storage failure

When the primary storage subsystem fails (out of service), an unplanned HyperSwap takes place under the control of PowerHA. This is a typical situation for HyperSwap: PowerHA receives the events generated by the Cluster Aware AIX (CAA) infrastructure (as a result of the primary storage failure). Then, PowerHA initiates a HyperSwap action, swapping paths and performing PPRC failover using in-band commands. The I/O workload switches to the secondary storage subsystem (in Site_B).

The application on Node 1 and Node 2 in Site_A keeps running, but it now performs disk I/O to the secondary storage in Site_B.

2.6.3 Unplanned HyperSwap: All links to primary storage fail

In the scenario illustrated in Figure 2-8, when Site_A Node 1 loses the access to the Site_A storage, the PowerHA unplanned HyperSwap occurs. PowerHA checks whether all nodes that run the application (in this case, Node 1 and Node 2) can access the corresponding Site_B disks.

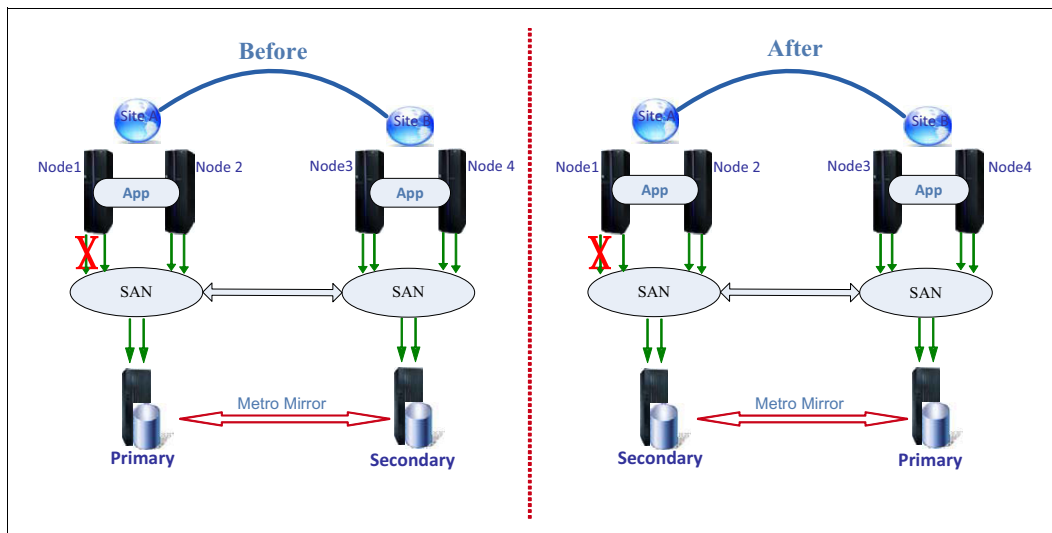


Figure 2-8 Links to the primary storage failure

If one of the nodes (for example, Node 2) cannot access the Site_B storage, the operation is stopped and you are informed about the reason.

However, if the configuration is correct, a coordinated HyperSwap operation is performed so that both Node 1 and Node 2, together, start redirecting the application I/O to the Site_B storage disks. This operation is performed with the cooperation of the AIX disk driver layer (on Node 1 and Node 2), and therefore, is completely transparent to the application.

If the unplanned swap completed successfully, the application I/O is now sent to storage in Site_B.

If HyperSwap fails for any reason, the I/O fails also, triggering resource group (RG) failover within the same site (based on RG policy).

2.6.4 Unplanned HyperSwap: Site down

The scenario shown in Figure 2-9 is a typical scenario in a disaster recovery (DR) situation. When a disaster strikes, Node 1, Node 2, and the primary storage subsystem (in Site_A) are out of service. Meanwhile, if Node 3 and Node 4 in Site_B are connected to the primary storage subsystem, they lose access to primary storage, as well.

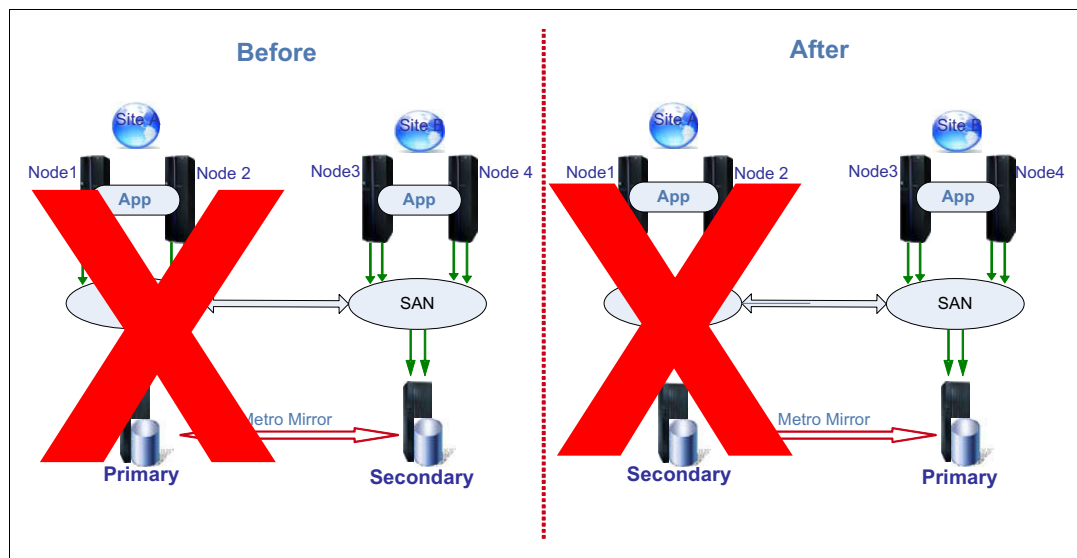


Figure 2-9 Site down

HyperSwap and RG takeover will take place. Depending on the PowerHA configuration, the application might be migrated from nodes in Site_A to nodes in Site_B. HyperSwap will swap the path from the primary storage in Site_A to the secondary storage in Site_B.

After these actions finish, the secondary storage subsystem in Site_B becomes the active storage subsystem (the PPRC source). The secondary (target) copy activation happens on DS8800 storage in Site_B and is initiated by PowerHA HyperSwap via in-band commands.

2.6.5 Unplanned HyperSwap: Site partition

In the scenario shown in Figure 2-10 on page 23, the workload continues to run on Site_A. Because both sites are partitioned, each site thinks it is the only surviving site, as such, the nodes in each site try to start the workload on each site.

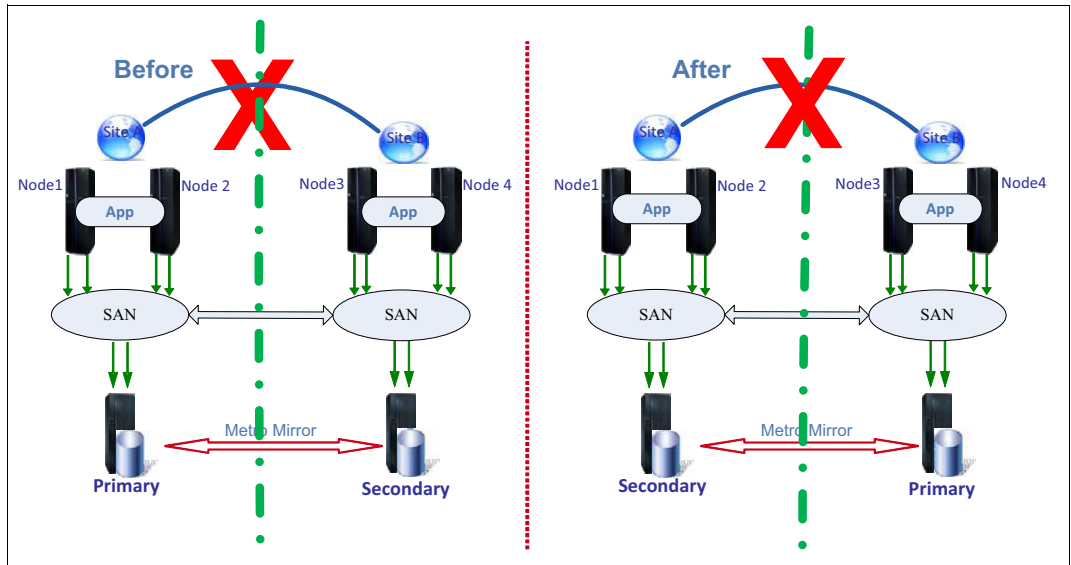


Figure 2-10 Site partition

Running the workload at the same time on both nodes results in data corruption. To maintain data integrity, PowerHA SystemMirror supports recovery mode for HyperSwap through manual workload activation. This option indicates that when the link between the sites is down (both sites are down), user intervention for manual recovery is needed, therefore maintaining data integrity.

When the site is down, because Auto Recovery Action is not supported, the resource groups (RGs) will remain in the ERROR state. User intervention is needed to correct the problem.

Expected user recovery action

The user has to shut down the cluster services on Site_B and fix the connectivity issues. When done, the user can start the cluster services on Site_B.



PowerHA cluster with AIX HyperSwap Active-Standby for applications using a shared file system

In this chapter, we describe two scenarios:

- ▶ Implementation of a two-site, Active-Standby stretched cluster configuration for an application using a file system for shared data and HyperSwap disks.
- ▶ Migration of an existing PowerHA cluster, which is an Active-Standby cluster configuration using file system as shared data to a two-site stretched cluster using AIX HyperSwap for providing protection against a single storage/site failure.

We describe how to configure storage, AIX, and PowerHA step-by-step for achieving higher application availability through enhanced storage availability. The last part of this chapter provides information about testing the configuration implemented for different types of failures.

The following tasks are described in this chapter:

- ▶ Cluster description and diagrams
- ▶ Installing a new configuration
- ▶ Migrating PowerHA cluster to HyperSwap enabled storage
- ▶ Two-node cluster to four-node cluster with HyperSwap

3.1 Cluster description and diagrams

Our configuration consists of four nodes (logical partitions (LPARs)) with virtual I/O resources (Shared Ethernet Adapter (SEA) and N_Port ID Virtualization (NPIV)) running in four different physical servers. Depending on the size of the servers and application requirements, you might also consider using dedicated (physical) I/O adapters for the storage (Fibre Channel (FC)) and network (Ethernet).

In our test environment, we use LPARs with virtual resources:

- ▶ NPIV for access to the shared storage
- ▶ Shared Ethernet Adapter using two physical Ethernet adapters on the Virtual I/O Server (VIOS)

The configuration of the I/O resources for the LPARs is beyond the scope of this document. For reference, see the Virtual I/O Server documentation:

<http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=%2Fiphb1%2Fiphb1kickoff.htm>

Important: See the application documentation for a supported configuration using virtual resources.

Systems diagrams and configuration data

For convenience, we provide the following diagrams and configuration data to help you understand the configuration steps:

- ▶ Storage and SAN diagram (Figure 3-1 on page 27)
- ▶ Networking diagram (Figure 3-2 on page 27)

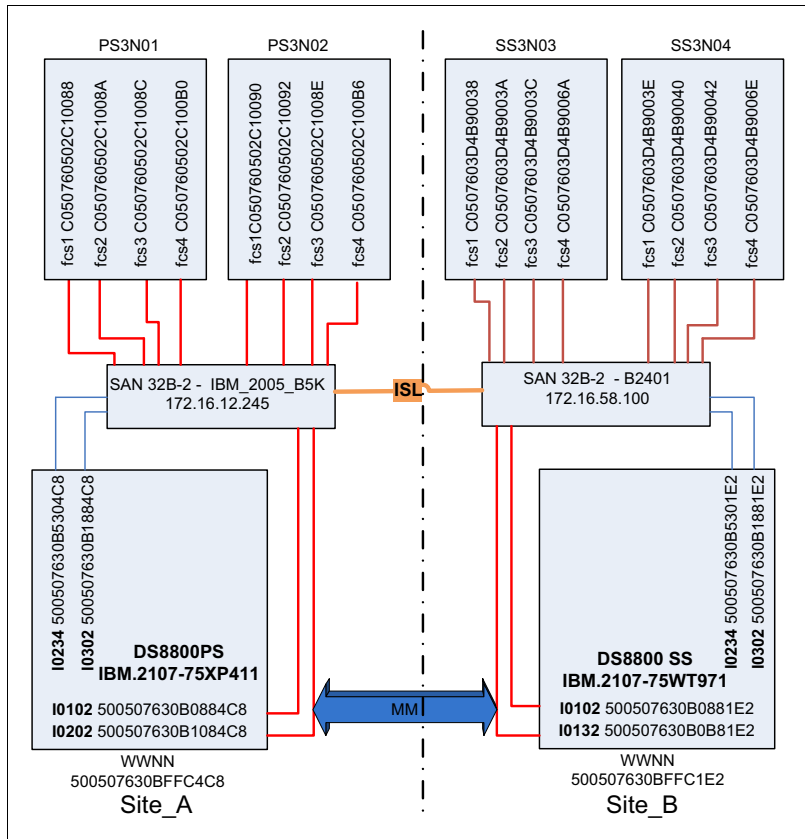


Figure 3-1 Storage and SAN diagram

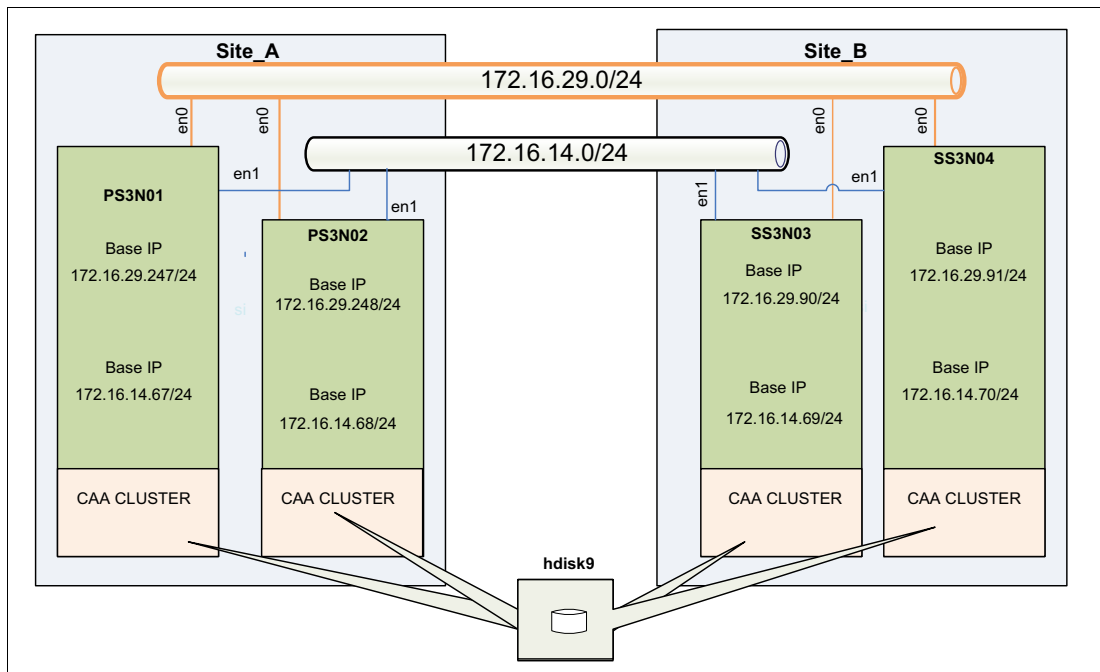


Figure 3-2 Networking diagram (logical)

3.2 Installing a new configuration

We provide a step-by-step infrastructure configuration for the storage, systems, and PowerHA cluster.

3.2.1 Identifying the storage

We provide an example of how to configure the storage for HyperSwap. We use the DS8000 storage command-line interface (**dscli**).

For details about the **dscli** interface, see the following manual:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S7002620>

We identify the two storage subsystems as shown in Example 3-1.

Example 3-1 Primary and secondary site storage subsystems

```
STORAGE_A
dscli> lssi
Name ID                Storage Unit      Model WWNN                State ESSNet
=====
-   IBM.2107-75XP411 IBM.2107-75XP410 951   500507630BFFC4C8 Online Enabled
dscli>

*****

STORAGE_B
dscli> lssi
Name ID                Storage Unit      Model WWNN                State ESSNet
=====
DS8803 IBM.2107-75WT971 IBM.2107-75WT970 951   500507630BFFC1E2 Online Enabled
dscli>
```

Identify the storage ports

We need to identify the storage ports used for host connectivity and for replication.

Ports used for host connectivity

The storage ports used for host connectivity are shown in Example 3-2.

Example 3-2 Storage ports used for host connectivity to our nodes

```
STORAGE in Site_A
dscli> lsioport
ID   WWPNN                State  Type                topo  portgrp
=====
.....<< Snippet >>.....
I0234 500507630B5304C8 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
I0302 500507630B1884C8 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
dscli>

*****
```

```

STORAGE in Site_B
dscli> lsioport
ID      WWPN              State  Type                topo    portgrp
=====
.....<< Snippet >>.....
I0200 500507630B1001E2 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
I0230 500507630B1301E2 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
dscli>

```

Although the storage is configured to allow host access through all available I/O ports, we will restrict the ports used by our systems through the zoning configuration (SAN fabric) to the following ports:

- ▶ Storage_A:
 - I0234
 - I0302
- ▶ Storage_B:
 - I0200
 - I0230

Ports used for replication

Also, in the zoning configuration, we need to configure the following physical ports on each storage subsystem for copy services as shown in Example 3-3.

Example 3-3 I/O ports used for copy services (Metro Mirror)

```

STORAGE in Site_A
dscli> lsioport
ID      WWPN              State  Type                topo    portgrp
=====
I0030 500507630B0304C8 Offline Fibre Channel-SW -      0
.....<< Snippet >>.....
I0100 500507630B0804C8 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
I0102 500507630B0884C8 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
I0202 500507630B1084C8 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....

I0333 500507630B1BC4C8 Online Fibre Channel-SW SCSI-FCP 0
dscli>

*****

STORAGE in Site_B
dscli> lsioport
ID      WWPN              State  Type                topo    portgrp
=====
.....<< Snippet >>.....
I0003 500507630B00C1E2 Online Fibre Channel-SW SCSI-FCP 0
I0030 500507630B0301E2 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
I0102 500507630B0881E2 Online Fibre Channel-SW SCSI-FCP 0

```

```

.....<< Snippet >>.....
I0132 500507630B0B81E2 Online Fibre Channel-SW SCSI-FCP 0
.....<< Snippet >>.....
I0333 500507630B1BC1E2 Online Fibre Channel-SW SCSI-FCP 0
dscli>

```

Although three ports on each storage subsystem were configured for replication, for our scenario, we use only the following two ports:

- ▶ Storage_A:
 - I0102
 - I0202
- ▶ Storage_B:
 - I0102
 - I0132

3.2.2 Identifying the systems' HBA configurations

We need to identify the system host bus adapter (HBA) Fibre Channel (FC) adapters and ports. Because we are using NPIV, we need to retrieve the worldwide port name (WWPN) information for the virtual FC adapters defined to the LPARs that we plan to configure.

We use the Hardware Management Console (HMC) command-line interface, as shown in Example 3-4.

Example 3-4 Virtual FC information for our LPARs

```

hyperswap@HMC58:~> lshwres -r virtualio --rsubtype fc -m SVRP7770-03-SN06F8DE6
--level lpar -F lpar_name,wwpns --filter "lpar_names=HSP77003N7"
HSP77003N7,"c050760502c10088,c050760502c10089"
HSP77003N7,"c050760502c10068,c050760502c10069"
HSP77003N7,"c050760502c100b0,c050760502c100b1"
HSP77003N7,"c050760502c1008c,c050760502c1008d"
HSP77003N7,"c050760502c1008a,c050760502c1008b"

```

```

hyperswap@HMC58:~> lshwres -r virtualio --rsubtype fc -m SVRP7770-03-SN06F8DE6
--level lpar -F lpar_name,wwpns --filter "lpar_names=HSP77003N8"
HSP77003N8,"c050760502c1006a,c050760502c1006b"
HSP77003N8,"c050760502c1008e,c050760502c1008f"
HSP77003N8,"c050760502c10092,c050760502c10093"
HSP77003N8,"c050760502c100b6,c050760502c100b7"
HSP77003N8,"c050760502c10090,c050760502c10091"

```

```

hyperswap@HMC58:~> lshwres -r virtualio --rsubtype fc -m SVRP7780-05-SN0681F3P
--level lpar -F lpar_name,wwpns --filter "lpar_names=HSP78005N7"
HSP78005N7,"c0507603d4b9003a,c0507603d4b9003b"
HSP78005N7,"c0507603d4b90038,c0507603d4b90039"
HSP78005N7,"c0507603d4b9001c,c0507603d4b9001d"
HSP78005N7,"c0507603d4b9006a,c0507603d4b9006b"
HSP78005N7,"c0507603d4b9003c,c0507603d4b9003d"

```

```

hyperswap@HMC58:~> lshwres -r virtualio --rsubtype fc -m SVRP7780-05-SN0681F3P
--level lpar -F lpar_name,wwpns --filter "lpar_names=HSP78005N8"
HSP78005N8,"c0507603d4b90042,c0507603d4b90043"

```

```
HSP78005N8,"c0507603d4b9006e,c0507603d4b9006f"
HSP78005N8,"c0507603d4b90040,c0507603d4b90041"
HSP78005N8,"c0507603d4b9001e,c0507603d4b9001f"
HSP78005N8,"c0507603d4b9003e,c0507603d4b9003f"
```

In our configuration, to access the shared storage, we use only the virtual adapters' WWPNs highlighted in Example 3-4 on page 30. For each node, the WWPNs that are not highlighted on the left column are used for accessing the node's rootvg.

Virtual HBA WWPNs: Live Partition Mobility (LPM) requires that each virtual HBA (NPIV) is generated with two worldwide port names (WWPNs). However, in our context, LPM is not used; therefore, the second WWPN of each virtual HBA is not used.

3.2.3 Zoning configuration

We use the SAN configuration presented in Figure 3-1 on page 27. Follow these steps for the zoning configuration:

1. Note that the switches are part of the same fabric, as shown in Example 3-5.

Example 3-5 Identifying the fabric information

```
IBM_2005_B5K:admin> fabricshow
Switch ID   Worldwide Name           Enet IP Addr   FC IP Addr     Name
-----
6: fffc06 10:00:00:05:1e:90:43:8a 172.16.12.245  0.0.0.0        >"IBM_2005_B5K"
10: fffc0a 10:00:00:05:33:6b:a1:3f 172.16.58.180  192.168.1.111 "B2401"
```

The Fabric has 2 switches

```
IBM_2005_B5K:admin>
```

Note: The SAN switch named IBM_2005_B5K is Switch #1 (in Site_A) and B2401 is Switch #2 (in Site_B). See Figure 3-1 on page 27 for the naming convention.

2. Example 3-6 shows the WWPNs of the ports logged in to the fabric via Switch #1.

Example 3-6 Port status on Switch #1

```
SWITCH #1
IBM_2005_B5K:admin> switchshow
switchName:   IBM_2005_B5K
switchType:   58.2
switchState:  Online
switchMode:   Native
switchRole:   Principal
switchDomain:  6
switchId:     fffc06
switchWwn:    10:00:00:05:1e:90:43:8a
zoning:       ON (powerswap)
switchBeacon: OFF

Index Port Address Media Speed State      Proto
-----
0 0 060000 id N4 Online FC F-Port 10:00:00:00:c9:c8:30:c6
```

```

.....<< Snippet >>.....
15 15 060f00 id N4 No_Light FC
16 16 061000 id N4 Online FC F-Port 1 N Port + 32 NPIV public
17 17 061100 id N4 No_Light FC
18 18 061200 id N4 Online FC E-Port 10:00:00:05:33:6b:a1:3f
"B2401" (downstream)
19 19 061300 id N4 No_Light FC
.....<< Snippet >>.....
24 24 061800 id N4 Online FC F-Port 50:05:07:63:0b:08:84:c8
25 25 061900 id N4 Online FC F-Port 50:05:07:63:0b:53:04:c8
26 26 061a00 id N4 Online FC F-Port 50:05:07:63:0b:18:84:c8
27 27 061b00 id N4 Online FC F-Port 50:05:07:63:0b:10:84:c8
.....<< Snippet >>.....
31 31 061f00 id N4 No_Light FC
IBM_2005_B5K:admin>

```

We identify the Storage_A ports connected to Switch #1 (marked in bold text in Example 3-6 on page 31):

- Storage ports used for host access are connected to ports **25** and **26**.
 - Storage ports used for replication are ports **24** and **27**.
3. Because we are using NPIV, we need to identify the WWPNs of the virtual HBAs defined to the LPARs (connected into the switch via port 16). We use the command shown in Example 3-7.

Note: The systems must be up and running (AIX operational) for their WWPNs to be logged in to the fabric. Compare this information with Power Systems' controlling HMC data (see Example 3-4 on page 30).

Example 3-7 Virtual HBA ports logged in to the fabric via port 16 in Switch #1

```

SWITCH #1
IBM_2005_B5K:admin> portshow 16
portIndex: 16
portName:
portHealth: HEALTHY

Authentication: None
.....<< Snippet >>.....
state transition count: 0

portId: 061000
portIfId: 43020010
portWwn: 20:10:00:05:1e:90:43:8a
portWwn of device(s) connected:
.....<< Snippet >>.....
c0:50:76:05:02:c1:00:94
c0:50:76:05:02:c1:00:8a
c0:50:76:05:02:c1:00:b0
c0:50:76:05:02:c1:00:88
c0:50:76:05:02:c1:00:8c
.....<< Snippet >>.....
c0:50:76:05:02:c1:00:b6
c0:50:76:05:02:c1:00:a6
c0:50:76:05:02:c1:00:8e

```

```

c0:50:76:05:02:c1:00:92
c0:50:76:05:02:c1:00:90
.....<< Snippet >>.....
10:00:00:00:c9:d2:1a:c4
Distance: normal
portSpeed: N4Gbps
.....<< Snippet >>.....
IBM_2005_B5K:admin>

```

4. Example 3-8 shows the WWPNs of the virtual HBA ports logged in to the fabric via Switch #2.

Example 3-8 Virtual HBA ports on Switch #2

```

SWITCH #2
B2401:admin> switchshow
switchName:    B2401
switchType:    71.2
switchState:   Online
switchMode:    Native
switchRole:    Subordinate
switchDomain:  10
switchId:      fffc0a
switchWwn:     10:00:00:05:33:6b:a1:3f
zoning:        ON (powerswap)
switchBeacon:  OFF

Index Port Address Media Speed State      Proto
=====
.....<< Snippet >>.....
  3  3  0a0300  id  N4  Online    FC  E-Port  10:00:00:05:1e:90:43:8a
"IBM_2005_B5K" (upstream)
  4  4  0a0100  id  N8  Online    FC  F-Port  10:00:00:00:c9:b7:02:1e
  5  5  0a0400  id  N8  Online   FC F-Port 50:05:07:63:0b:0b:81:e2
  6  6  0a0500  id  N8  Online   FC F-Port 50:05:07:63:0b:10:01:e2
.....<< Snippet >>.....
 10 10 0a0a00  id  N8  Online    FC  F-Port  50:05:07:63:0b:13:01:e2
.....<< Snippet >>.....
 13 13 0a0d00  id  N8  Online    FC  F-Port  1 N Port + 32 NPIV public
 14 14 0a0e00  id  N8  Online    FC  F-Port  50:05:07:63:0b:08:81:e2
.....<< Snippet >>.....
B2401:admin>

```

We identify the Storage_B ports connected to Switch #2 (These ports are marked in bold text in Example 3-8):

- Storage ports used for host access are connected to ports **6** and **10**.
- Storage ports used for replication are switch ports **5** and **14**.

5. Node 3 and Node 4 (systems) are connected using NPIV into Switch #2 via port **13**. We identify their WWPNs by using the command shown in Example 3-9 on page 34. The systems must be up and running (AIX operational) for their WWPNs to be logged in to the fabric. Compare this information with Power Systems' controlling HMC data shown in Example 3-4 on page 30.

Example 3-9 Systems' WWPNs connected via port 13 into Switch #2

```
SWITCH #2
B2401:admin> portshow 13
portIndex: 13
portName:
portHealth: No Fabric Watch License

Authentication: None
.....<< Snippet >>.....
state transition count: 1

portId: 0a0d00
portIfId: 430200c
portWwn: 20:0d:00:05:33:6b:a1:3f
portWwn of device(s) connected:
    c0:50:76:03:d4:b9:00:32
.....<< Snippet >>.....
    c0:50:76:03:d4:b9:00:44
    c0:50:76:03:d4:b9:00:3c
    c0:50:76:03:d4:b9:00:3a
    c0:50:76:03:d4:b9:00:6a
    c0:50:76:03:d4:b9:00:38
.....<< Snippet >>.....
    c0:50:76:03:d4:b9:00:6e
    c0:50:76:03:d4:b9:00:40
    c0:50:76:03:d4:b9:00:42
    c0:50:76:03:d4:b9:00:3e
    c0:50:76:03:d4:b9:00:2a
.....<< Snippet >>.....
    10:00:00:00:c9:aa:ac:a2
Distance: normal
portSpeed: N8Gbps

.....<< Snippet >>.....
B2401:admin>
```

6. Next, we identify the zoning configuration relevant to our cluster. In our case, we have a total of 18 zones in the active zoning configuration:

- Eight zones for zoning the fcs1 and fcs4 HBAs of our four nodes with storage ports connected into Switch #1
- Eight zones for zoning the fcs2 and fcs3 HBAs of our four nodes with storage ports connected into Switch #2
- Two zones for the two storage ports in each switch

The zoning configuration is shown in Example 3-10 on page 35. Refer also to Figure 3-1 on page 27.

Node 1, Node 2, Node 3, and Node 4 (fcs1 and fcs4) to storage ports of Storage_A connected in Switch #1:

Node 1 to Storage_A

```
zone: P7703LP7_fcs1_DS8805_I0302
      c0:50:76:05:02:c1:00:88; 50:05:07:63:0b:18:84:c8
zone: P7703LP7_fcs4_DS8805_I0234
      c0:50:76:05:02:c1:00:b0; 50:05:07:63:0b:53:04:c8
```

Node 2 to Storage_A

```
zone: P7703LP8_fcs1_DS8805_I0302
      c0:50:76:05:02:c1:00:90; 50:05:07:63:0b:18:84:c8
zone: P7703LP8_fcs4_DS8805_I0234
      c0:50:76:05:02:c1:00:b6; 50:05:07:63:0b:53:04:c8
```

Node 3 to Storage_A

```
zone: P7805LP7_fcs1_DS8805_I0302
      c0:50:76:03:d4:b9:00:38; 50:05:07:63:0b:18:84:c8
zone: P7805LP7_fcs4_DS8805_I0234
      c0:50:76:03:d4:b9:00:6a; 50:05:07:63:0b:53:04:c8
```

Node 4 to Storage_A

```
zone: P7805LP8_fcs1_DS8805_I0302
      c0:50:76:03:d4:b9:00:3e; 50:05:07:63:0b:18:84:c8
zone: P7805LP8_fcs4_DS8805_I0234
      c0:50:76:03:d4:b9:00:6e; 50:05:07:63:0b:53:04:c8
```

Node 1, Node 2, Node 3, and Node 4 (fcs2 and fcs3) to storage ports of Storage_B connected in Switch #2:

Node 1 to Storage_B:

```
zone: P7703LP7_fcs2_DS8803_I0200
      c0:50:76:05:02:c1:00:8a; 50:05:07:63:0b:10:01:e2
zone: P7703LP7_fcs3_DS8803_I0230
      c0:50:76:05:02:c1:00:8c; 50:05:07:63:0b:13:01:e2
```

Node 2 to Storage_B

```
zone: P7703LP8_fcs2_DS8803_I0200
      c0:50:76:05:02:c1:00:92; 50:05:07:63:0b:10:01:e2
zone: P7703LP8_fcs3_DS8803_I0203
      c0:50:76:05:02:c1:00:8e; 50:05:07:63:0b:13:01:e2
```

Node 3 to Storage_B:

```
zone: P7805LP7_fcs2_DS8803_I0200
      c0:50:76:03:d4:b9:00:3a; 50:05:07:63:0b:10:01:e2
zone: P7805LP7_fcs3_DS8803_I0230
      c0:50:76:03:d4:b9:00:3c; 50:05:07:63:0b:13:01:e2
```

Node 4 to Storage_B:

```
zone: P7805LP8_fcs2_DS8803_I0200
      c0:50:76:03:d4:b9:00:40; 50:05:07:63:0b:10:01:e2
zone: P7805LP8_fcs3_DS8803_I0230
```

c0:50:76:03:d4:b9:00:42; 50:05:07:63:0b:13:01:e2

Storage to storage ports for replication:

zone: DS8K_PPRC_G01
50:05:07:63:0b:08:84:c8; 50:05:07:63:0b:08:81:e2
zone: DS8K_PPRC_G02
50:05:07:63:0b:10:84:c8; 50:05:07:63:0b:0b:81:e2

3.2.4 Configuring the storage

We explain the steps that we performed to configure the storage space for our test systems.

LUN definitions

First, ensure that you discuss the definition of the logical unit numbers (LUNs) with the storage administrator to determine the space availability and the connectivity configuration. We logged on to the storage subsystems in Site_A and Site_B (we use `dsc1i`). Follow the steps we performed:

1. We check the existing logical subsystems (LSSs) on both Storage_A and Storage_B, as shown in Example 3-11.

Example 3-11 LSS configuration

STORAGE in Site_A

```
dsc1i> lslss
ID Group addrgrp stgtype confgvols
=====
35    1      3 fb          1
36    0      3 fb          1
63    1      6 fb          2
.....<< Snippet >>.....
AB    1      A fb          2
AD    1      A fb          2
AF    1      A fb          2
dsc1i>
```

STORAGE in Site_B

```
dsc1i> lslss
ID Group addrgrp stgtype confgvols
=====
00    0      0 fb          26
01    1      0 fb          28
10    0      1 fb          2
.....<< Snippet >>.....
AB    1      A fb          2
AD    1      A fb          2
AF    1      A fb          2
dsc1i>
```

2. In our test case, we create three new disks (B1, B2, and B3) in three new LSSs, on both storage subsystems. We start by checking the available space as shown in Example 3-12 on page 37.

Suggestion: Use dedicated storage LSSs for the HyperSwap configuration. Also, see the IBM DS8000 Copy Services documentation:

<http://publib.boulder.ibm.com/infocenter/dsichelp/ds8000ic/index.jsp>

Example 3-12 Checking available space

STORAGE in Site_A

```
dscli> lsxtpool
Name          ID stgtype rankgrp status  availstor (2^30B) %allocated available reserved numvols
=====
Perf_Pool01  P1 fb          1 exceeded      75          90          75          0          7
Perf_Pool02  P2 fb          0 exceeded       4          99           4          0          7
Perf_Pool03 P3 fb          1 below        3641         11        3641          0         23
Perf_Pool04 P4 fb          0 below        3995          2        3995          0         17
dscli>
```

STORAGE in Site_B

```
dscli> lsxtpool
Name          ID stgtype rankgrp status  availstor (2^30B) %allocated available reserved numvols
=====
extpool_00   P0 fb          0 exceeded      267          91          267          0         37
extpool_01   P1 fb          1 exceeded      424          89          424          0         36
extpool_02   P2 fb          0 below         892          78          892          0         57
extpool_03  P3 fb          1 exceeded      583          86          583          0         58
extpool_04  P4 fb          0 below        2008          36        2008          0          4
extpool_05   P5 fb          1 below        2178          48        2178          0          5
extpool_06   P6 fb          0 below        2008          36        2008          0          4
extpool_07   P7 fb          1 below        2218          47        2218          0          6
extpool_08   P8 fb          0 below        2007          36        2007          0          4
extpool_09   P9 fb          1 below        2147          49        2147          0          6
extpool_10   P10 fb         0 below        2147          49        2147          0          5
extpool_11   P11 fb         1 below        2147          49        2147          0          5
fujun_test   P12 fb         0 below        1573          0        1573          0          1
dscli>
```

3. Identify the storage volume group (VG) that allows access from your systems and add the newly created LUNs to that volume group. In our case, we use volume group V7 on Storage_A and volume group V16 on Storage_B.

Example 3-13 on page 38 shows the host connectivity for our systems to Storage_A.

Example 3-13 Host connectivity for our nodes to Storage_A

```
dscli> lshostconnect -volgrp V7
Name          ID   WWPN          HostType Profile          portgrp volgrpID ESSIOport
=====
G4_P7805LP7_fcs4 001F C0507603D4B9006A pSeries  IBM pSeries - AIX      0 V7      all
G4_P7805LP7_fcs1 0020 C0507603D4B9003B pSeries  IBM pSeries - AIX      0 V7      all
G4_P7805LP8_fcs4 0021 C0507603D4B9006E pSeries  IBM pSeries - AIX      0 V7      all
G4_P7805LP8_fcs1 0022 C0507603D4B9003E pSeries  IBM pSeries - AIX      0 V7      all
G4_P7805LP7_fcs4 0033 C050760502C10080 pSeries  IBM pSeries - AIX      0 V7      all
G4_P7703LP7_fcs1 0034 C050760502C10088 pSeries  IBM pSeries - AIX      0 V7      all
G4_P7703LP8_fcs4 0035 C050760502C10086 pSeries  IBM pSeries - AIX      0 V7      all
G4_P7703LP8_fcs1 0036 C050760502C10090 pSeries  IBM pSeries - AIX      0 V7      all
dscli>
```

4. Example 3-14 shows the host connectivity for our systems to Storage_B.

Example 3-14 Host connectivity for our nodes to Storage_B

```
dscli> lshostconnect -volgrp v16
Name          ID   WWPN          HostType Profile          portgrp volgrpID ESSIOport
=====
G4_P7805LP7_fcs2 0026 C0507603D4B9003A pSeries  IBM pSeries - AIX      0 V16     all
G4_P7805LP7_fcs3 0027 C0507603D4B9003C pSeries  IBM pSeries - AIX      0 V16     all
G4_P7805LP8_fcs2 0028 C0507603D4B90040 pSeries  IBM pSeries - AIX      0 V16     all
G4_P7805LP8_fcs2 0029 C0507603D4B90042 pSeries  IBM pSeries - AIX      0 V16     all
G4_P7703LP7_fcs2 003A C050760502C1008A pSeries  IBM pSeries - AIX      0 V16     all
G4_P7703LP7_fcs3 003B C050760502C1008C pSeries  IBM pSeries - AIX      0 V16     all
G4_P7703LP8_fcs2 003C C050760502C10092 pSeries  IBM pSeries - AIX      0 V16     all
G4_P7703LP8_fcs3 003D C050760502C1008E pSeries  IBM pSeries - AIX      0 V16     all
dscli>
```

5. We create the following LUNs on both storage subsystems:
- One 10 GB LUN for the Cluster Aware AIX disk (CAA)
 - Two LUNs (30 GB each) for the VG that will hold the application shared storage space

We chose to allocate space from Perf_Pool03 (P3) and Perf_Pool04 (P4) on Storage_A and extpool_03 (P3) and extpool_04 (P4) on Storage_B. Example 3-15 shows the LUN creation.

Example 3-15 LUN creation

STORAGE in Site_A

```
dscli> mkfbvol -extpool P3 -cap 10 -name CAA_rep01 -volgrp V7 -sam ese B100
CMUC00025I mkfbvol: FB volume B100 successfully created.
```

```
dscli> mkfbvol -extpool P4 -cap 30 -name ps3_data_v0 -volgrp V7 -sam ese B201
CMUC00025I mkfbvol: FB volume B101 successfully created.
```

```
dscli> mkfbvol -extpool P3 -cap 30 -name ps3_data_v1 -volgrp V7 -sam ese B301
CMUC00025I mkfbvol: FB volume B201 successfully created.
```

STORAGE in Site_B

```
dscli> mkfbvol -extpool P3 -cap 10 -name CAA_rep01 -volgrp V16 -sam ese B100
CMUC00025I mkfbvol: FB volume B100 successfully created.
```

```
dscli> mkfbvol -extpool P4 -cap 30 -name ps3_data_v0 -volgrp V16 -sam ese B201
CMUC00025I mkfbvol: FB volume B101 successfully created.
```

```
dscli> mkfbvol -extpool P3 -cap 30 -name ps3_data_v1 -volgrp V16 -sam ese B301
CMUC00025I mkfbvol: FB volume B201 successfully created.
```

6. We start configuring the replication for the previously created LUNs. We identify the available replication paths, which are based on our previous configuration, on our systems (see Example 3-16):

- Storage_A (WWNN: 500507630BFFC4C8):
 - I0102
 - I0202
- Storage_B (WWNN: 500507630BFFC1E2):
 - I0102
 - I0132

Example 3-16 Identifying available ports for replication on Storage_A

```
dscli> lsavailpprcport -l -remotewwnn 500507630BFFC1E2 B1:B1
Local Port Attached Port Type Switch ID Switch Port
=====
I0100      I0030          FCP NA      NA
I0102    I0102          FCP NA      NA
I0202    I0132          FCP NA      NA
dscli>
```

7. We create the replication relationship from Storage_A to Storage_B for all three LUNs that were previously created, as shown in Example 3-17.

Example 3-17 Creating replication relationship from Storage_A to Storage_B

```
dscli> mkpprcpath -remotewwnn 500507630BFFC1E2 -src1ss B1 -tgt1ss B1 -consistgrp I0102:I0102
I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path B1:B1 successfully established.
```

```
dscli> mkpprcpath -remotewwnn 500507630BFFC1E2 -src1ss B2 -tgt1ss B2 -consistgrp I0102:I0102
I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path B2:B2 successfully established.
```

```
dscli> mkpprcpath -remotewwnn 500507630BFFC1E2 -src1ss B3 -tgt1ss B3 -consistgrp I0102:I0102
I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path B3:B3 successfully established.
```

```
dscli> lsprrcpath -l B1
Src Tgt State  SS  Port  Attached Port Tgt WWNN      Failed Reason PPRC CG
=====
B1 B1  Success FFB1 I0102 I0102      500507630BFFC1E2 -      Enabled
B1 B1  Success FFB1 I0202 I0132      500507630BFFC1E2 -      Enabled
dscli>
```

8. Next, we identify the available replication paths from the Storage_B side, as shown in Example 3-18 on page 40.

Example 3-18 Identifying the available replication paths on Storage_B

```
dscli> lsavailpprcport -l -remotewwnn 500507630BFFC4C8 B1:B1
Local Port Attached Port Type Switch ID Switch Port
=====
I0030      I0100          FCP NA      NA
I0102      I0102          FCP NA      NA
I0132      I0202          FCP NA      NA
```

9. We create the replication relationship from Storage_B to Storage_A for all three LUNs that were previously created, as shown in Example 3-19.

Example 3-19 Creating replication relationship from Storage_B to Storage_A

```
dscli> mkpprcpath -remotewwnn 500507630BFFC4C8 -srclss B1 -tgtlss B1 -consistgrp I0102:I0102
I0132:I0202
CMUC00149I mkpprcpath: Remote Mirror and Copy path B1:B1 successfully established.
dscli>mkpprcpath -remotewwnn 500507630BFFC4C8 -srclss B2 -tgtlss B2 -consistgrp I0102:I0102
I0132:I0202
CMUC00149I mkpprcpath: Remote Mirror and Copy path B2:B2 successfully established.
dscli>mkpprcpath -remotewwnn 500507630BFFC4C8 -srclss B3 -tgtlss B3 -consistgrp I0102:I0102
I0132:I0202
CMUC00149I mkpprcpath: Remote Mirror and Copy path B3:B3 successfully established.
```

```
dscli> lsprrcpath -l B1
Src Tgt State SS Port Attached Port Tgt WNN Failed Reason PPRC CG
=====
B1 B1 Success FFB1 I0102 I0102 500507630BFFC4C8 - Enabled
B1 B1 Success FFB1 I0132 I0202 500507630BFFC4C8 - Enabled
dscli>
```

10. We enable the replication relationship, as shown in Example 3-20.

Example 3-20 Enabling replication

```
dscli> mkpprc -remotedev IBM.2107-75WT971 -type mmir -mode full -tgtse B100:B100 B101:B101 B201:B201
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship B201:B201 successfully created.
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship B100:B100 successfully created.
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship B101:B101 successfully created.
dscli>
```

11. We check the status of the replication, as shown in Example 3-21.

Example 3-21 Replication status for LUN B100

```
dscli> lsprrc -l B100:B100
CMUC00096E lsprrc: No remote storage image ID is specified.
dscli> lsprrc -l -fmt stanza -remotedev IBM.2107-75WT971 B100:B100
ID B100:B100
State Full Duplex
Reason -
Type Metro Mirror
Out Of Sync Tracks 0
Tgt Read Disabled
Src Cascade Disabled
Tgt Cascade Invalid
Date Suspended -
SourceLSS B1
```

```

Timeout (secs)      60
Critical Mode       Disabled
First Pass Status   Invalid
Incremental Resync  Disabled
Tgt Write           Disabled
GMIR CG             N/A
PPRC CG             Enabled
isTgtSE             Unknown
DisableAutoResync   -
dscli>

```

12. For our system, we also need to change the Peer-to-Peer Remote Copy (PPRC) (**pprconsistgrp**) and the *extend long busy* timeout (**xtndlbztimeout**) parameters on *both* Storage_A and Storage_B, as shown in Example 3-22 for Storage_A.

Example 3-22 Enabling the PPRC consistency group and changing the extlongbusy timeout

```

dscli> showlss B1
ID          B1
Group       1
addrgrp     B
stgtype     fb
confgvols   2
subsys      0xFFB1
pprconsistgrp Disabled
xtndlbztimeout 60 secs
resgrp      RG0
dscli>

```

```

dscli> chlss -pprconsistgrp enable -extlongbusy 5 B1
CMUC00029I chlss: LSS B1 successfully modified.
dscli> chlss -pprconsistgrp enable -extlongbusy 5 B2
CMUC00029I chlss: LSS B2 successfully modified.

```

13. Finally, check the source and the target LUNs to verify the replication relationship, as shown in Example 3-23 for Storage_A, LUN B100.

Example 3-23 Checking replication status for LUN B100

```

dscli> lsprrc -l -fmt stanza B100
ID          B100:B100
State      Full Duplex
Reason      -
Type        Metro Mirror
Out Of Sync Tracks 0
Tgt Read    Disabled
Src Cascade Disabled
Tgt Cascade Invalid
Date Suspended -
SourceLSS B1
Timeout (secs) 5
Critical Mode Disabled
First Pass Status Invalid
Incremental Resync Disabled
Tgt Write   Disabled
GMIR CG     N/A
PPRC CG    Enabled

```

```
isTgtSE           Unknown
DisableAutoResync -
dscli>
```

3.2.5 Enabling HyperSwap: Storage level

At this time, as shown in Example 3-13 on page 38 and Example 3-14 on page 38, the host connections are not enabled for in-band communication (required by HyperSwap). We change the host connection characteristics for our cluster nodes on both Storage_A and Storage_B. We followed these steps:

1. We determine whether our storage subsystems support HyperSwap by using the **lshosttype -type scsi** command as shown in Example 3-24.

Example 3-24 Checking communication protocols available for our storage subsystem

```
dscli> lshosttype -type scsi
```

HostType	Profile	AddrDiscovery	LBS
AMDLinuxRHEL	AMD - Linux RHEL	LUNPolling	512
.....<< Snippet >>.....			
pLinux	IBM pSeries - pLinux	LUNPolling	512
pSeries	IBM pSeries - AIX	reportLUN	512
pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	reportLUN	512
zLinux	IBM zSeries - zLinux	reportLUN	512

```
dscli>
```

2. If host type pSeriesPowerswap is listed, the storage subsystem supports HyperSwap. You can change the host type for all worldwide port names (WWPNs) used for your cluster as shown in Example 3-25.

Example 3-25 Enabling host connections for in-band communication (HyperSwap)

STORAGE in Site_A (See Example 3-13 on page 38 for ID)

```
dscli> chhostconnect -hosttype pSeriesPowerswap 001F
chhostconnect -hosttype pSeriesPowerswap 0020
chhostconnect -hosttype pSeriesPowerswap 0021
chhostconnect -hosttype pSeriesPowerswap 0022
chhostconnect -hosttype pSeriesPowerswap 0033
chhostconnect -hosttype pSeriesPowerswap 0034
chhostconnect -hosttype pSeriesPowerswap 0035
chhostconnect -hosttype pSeriesPowerswap 0036
CMUC00013I chhostconnect: Host connection 001F successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0020 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0021 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0022 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0033 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0034 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0035 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0036 successfully modified.
dscli>
```

STORAGE in Site_B (See Example 3-14 on page 38 for ID)

```
dscli> chhostconnect -hosttype pSeriesPowerswap 0026
chhostconnect -hosttype pSeriesPowerswap 0027
chhostconnect -hosttype pSeriesPowerswap 0028
```



```

chhostconnect -hosttype pSeriesPowerswap 0029
chhostconnect -hosttype pSeriesPowerswap 003A
chhostconnect -hosttype pSeriesPowerswap 003B
chhostconnect -hosttype pSeriesPowerswap 003C
chhostconnect -hosttype pSeriesPowerswap 003D
CMUC00013I chhostconnect: Host connection 0026 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0027 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0028 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 0029 successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 003A successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 003B successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 003C successfully modified.
dscli> CMUC00013I chhostconnect: Host connection 003D successfully modified.
dscli>

```

3. We verify that HyperSwap is enabled for the hosts, as shown in Example 3-26.

Example 3-26 HyperSwap enabled

STORAGE in Site_A

```
dscli> lshostconnect -volgrp V7
```

Name	ID	WWPN	HostType	Profile	portgrp	volgrpID	ESSIOport
G4_P7805LP7_fcs4	001F	C0507603D4B9006A	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7805LP7_fcs1	0020	C0507603D4B90038	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7805LP8_fcs4	0021	C0507603D4B9006E	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7805LP8_fcs1	0022	C0507603D4B9003E	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7805LP7_fcs4	0033	C050760502C100B0	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7703LP7_fcs1	0034	C050760502C10088	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7703LP8_fcs4	0035	C050760502C100B6	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all
G4_P7703LP8_fcs1	0036	C050760502C10090	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V7		all

STORAGE in Site_B

```
dscli> lshostconnect -volgrp v16
```

Name	ID	WWPN	HostType	Profile	portgrp	volgrpID	ESSIOport
G4_P7805LP7_fcs2	0026	C0507603D4B9003A	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7805LP7_fcs3	0027	C0507603D4B9003C	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7805LP8_fcs2	0028	C0507603D4B90040	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7805LP8_fcs3	0029	C0507603D4B90042	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7703LP7_fcs2	003A	C050760502C1008A	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7703LP7_fcs3	003B	C050760502C1008C	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7703LP8_fcs2	003C	C050760502C10092	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all
G4_P7703LP8_fcs3	003D	C050760502C1008E	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0 V16		all

3.2.6 AIX configuration

We verify the network and storage configuration on our systems. We assume that the correct level of AIX is installed and that the storage device drivers are at a supported level. (See 2.3, “Prerequisites” on page 15.)

For the AIX configuration, the following tasks are required:

- ▶ Network configuration
- ▶ AIX disk device driver configuration

Network configuration

We use two network interface cards (NICs) for each node for high availability. Although these NICs are virtual, they are bridged (SEA) to two distinct Shared Ethernet Adapter (SEA) devices. The network configuration diagram is shown in Figure 3-2 on page 27. The detailed information is shown in Example 3-27.

Example 3-27 Network interfaces configuration

```
ps3n01base:
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 2a.5c.2f.24.6f.a 8014608 0 45045 0 0
en0 1500 172.16.29 172.16.29.247 8014608 0 45045 0 0
en1 1500 link#3 2a.5c.2f.24.6f.b 7983683 0 3972 0 0
en1 1500 172.16.14 172.16.14.67 7983683 0 3972 0 0
lo0 16896 link#1 102009 0 102009 0 0
lo0 16896 127 127.0.0.1 102009 0 102009 0 0
lo0 16896 ::1%1 102009 0 102009 0 0

ps3n02base:
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 2a.5c.28.36.14.a 4309191 0 5431 0 0
en0 1500 172.16.29 172.16.29.248 4309191 0 5431 0 0
en1 1500 link#3 2a.5c.28.36.14.b 4127401 0 16 0 0
en1 1500 172.16.14 172.16.14.68 4127401 0 16 0 0
lo0 16896 link#1 8463 0 8463 0 0
lo0 16896 127 127.0.0.1 8463 0 8463 0 0
lo0 16896 ::1%1 8463 0 8463 0 0

ss3n03base:
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 4a.ca.b8.cf.88.a 20672578 0 105510 0 0
en0 1500 172.16.29 172.16.29.90 20672578 0 105510 0 0
en1 1500 link#3 4a.ca.b8.cf.88.b 20564118 0 2506 0 0
en1 1500 172.16.14 172.16.14.77 20564118 0 2506 0 0
lo0 16896 link#1 110553 0 110553 0 0
lo0 16896 127 127.0.0.1 110553 0 110553 0 0
lo0 16896 ::1%1 110553 0 110553 0 0

ss3n04base:
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 4a.ca.b4.5c.93.a 20596328 0 49855 0 0
en0 1500 172.16.29 172.16.29.91 20596328 0 49855 0 0
en1 1500 link#3 4a.ca.b4.5c.93.b 20560826 0 2374 0 0
en1 1500 172.16.14 172.16.14.78 20560826 0 2374 0 0
lo0 16896 link#1 101933 0 101933 0 0
lo0 16896 127 127.0.0.1 101933 0 101933 0 0
lo0 16896 ::1%1 101933 0 101933 0 0
```

Important: Check your nodes for the correct name resolution and connectivity. The nodes in the cluster must follow these rules:

- ▶ The name resolutions must follow the same pattern (short names or Fully Qualified Domain Name (FQDN), for example, name.domain.com).
- ▶ *The order of the name resolution methods must be the same on all nodes.* (See the /etc/netsvc.conf file.)

AIX disk device driver configuration

Tip: Always check the latest product documentation and AIX release notes for the up-to-date procedure.

Follow the steps that we performed:

1. On all nodes, determine whether the AIX Multipath I/O (MPIO) HyperSwap disk device driver code is installed on your system by using the `manage_disk_drivers` command, as shown in Example 3-28.

Important: *Unless otherwise specified, the following set of steps must be run on all nodes.* For clarity, we only show information retrieved from Node 1 in our configuration.

Example 3-28 Checking AIX device driver options

```
root@ps3n01base: /> manage_disk_drivers -l
Device                Present Driver      Driver Options
2810XIV               AIX_AAPCM          AIX_AAPCM,AIX_non_MPIO
DS4100                AIX_APPCM          AIX_APPCM,AIX_fcarray
DS4200                AIX_APPCM          AIX_APPCM,AIX_fcarray
DS4300                AIX_APPCM          AIX_APPCM,AIX_fcarray
DS4500                AIX_APPCM          AIX_APPCM,AIX_fcarray
DS4700                AIX_APPCM          AIX_APPCM,AIX_fcarray
DS4800                AIX_APPCM          AIX_APPCM,AIX_fcarray
DS3950                AIX_APPCM          AIX_APPCM
DS5020                AIX_APPCM          AIX_APPCM
DCS3700               AIX_APPCM          AIX_APPCM
DS5100/DS5300         AIX_APPCM          AIX_APPCM
DS3500                AIX_APPCM          AIX_APPCM
XIVCTRL               MPIO_XIVCTRL       MPIO_XIVCTRL,nonMPIO_XIVCTRL,MPIO_XIVCTRL,nonMPIO_XIVCTRL
2107DS8K             NO_OVERRIDE       NO_OVERRIDE,AIX_AAPCM,NO_OVERRIDE
root@ps3n01base: />
```

2. At this time, AIX_AAPCM is available but not used to access the shared storage subsystem. If you have another device driver installed on your system, for example, Subsystem Device Driver Path Control Module (SDDPCM), you might need to remove the driver from your systems. Follow the instructions from your IBM support representative.
3. *Change the device driver for accessing the shared storage to use in-band communication (used for HyperSwap) on all nodes.* See Example 3-29 on page 46.

Example 3-29 Enabling HyperSwap driver

```

root@ps3n01base: /> manage_disk_drivers -d 2107DS8K -o AIX_AAPCM
***** ATTENTION *****

For the change to take effect the system must be rebooted
root@ps3n01base: />

root@ps3n01base: /> lsdev -C |grep fscsi
fscsi0    Available C2-T1-01    FC SCSI I/O Controller Protocol Device
fscsi1    Available 42-T1-01    FC SCSI I/O Controller Protocol Device
fscsi2    Available 43-T1-01    FC SCSI I/O Controller Protocol Device
fscsi3    Available 44-T1-01    FC SCSI I/O Controller Protocol Device
fscsi4    Available 67-T1-01    FC SCSI I/O Controller Protocol Device
root@ps3n01base: />

```

4. For all HBAs that will be used to access the HyperSwap disks, we also change the FC SCSI I/O Controller Protocol Device attributes, as shown in Example 3-30. In our configuration, we use fscsi1, fscsi2, fscsi3, and fscsi4.

Example 3-30 FC SCSI I/O Controller Protocol Device attributes

```

root@ps3n01base: /> lsattr -El fscsi1
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     0xa6a30  Adapter SCSI ID                          False
sw_fc_class 3         FC Class for Fabric                       True
root@ps3n01base: />

```

Reboot all nodes in the cluster at this time.

5. Reboot the systems to activate the AIX HyperSwap driver (AIX-AAPCM).

Important: If SAN paths and zoning are configured correctly, the LUNs from Storage_B (PPRC target) are also configured as *available* on your system (`lsdev -Cc disk`). However, the disks from Storage_B are not accessible at this time (`lsquerypv -h /dev/hdisk*`).

6. To check the accessible disks, use the command shown in Example 3-31. These accessible disks can be identified by the selected attribute (**s**) in the third column of the results displayed by the `lspprc -Ao` command.

Example 3-31 Identifying the accessible (PPRC source) disks on your systems

```

Command for all nodes: "lspprc -Ao|tail +4|sort -nk1.6,7"
HOSTS -----
ps3n01base, ps3n02base, ss3n03base, ss3n04base
-----
hdisk1    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk2    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk3    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk4    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk5    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk6    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk7    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk8    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk9    Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk10   Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2
hdisk11   Active    0(s)      -1          500507630bffc4c8  500507630bffc1e2

```

hdisk12	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk13	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk14	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk15	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk16	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk17	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk18	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk19	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk20	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk21	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk22	Active	-1	0	500507630bffc4c8	500507630bffc1e2

7. Change the reservation policy to **no_reserve** and ensure that, in Object Data Manager (ODM), the reservation policy is also set to NO RESERVE *for all shared disks on all systems that are part of the cluster*, as shown in Example 3-32. (We only show hdisk1 on each node.)

Example 3-32 Changing the disk reservation policy

```
root@ps3n01base: /> chdev -l hdisk1 -a reserve_policy=no_reserve
ps3n01base: hdisk1 changed
```

```
root@ps3n02base: /> chdev -l hdisk1 -a reserve_policy=no_reserve
ps3n02base: hdisk1 changed
```

```
root@ss3n03base: /> chdev -l hdisk1 -a reserve_policy=no_reserve
ss3n03base: hdisk1 changed
```

```
root@ps3n01base: /> devrsrv -c query -l hdisk1
=====
Device Reservation State Information
Device Name                : hdisk1
Device Open On Current Host? : NO
ODM Reservation Policy      : NO RESERVE
Device Reservation State    : NO RESERVE
```

```
root@ps3n02base: /> devrsrv -c query -l hdisk1
=====
Device Reservation State Information
Device Name                : hdisk1
Device Open On Current Host? : NO
ODM Reservation Policy      : NO RESERVE
Device Reservation State    : NO RESERVE
```

```
root@ss3n03base: /> devrsrv -c query -l hdisk1
=====
Device Reservation State Information
Device Name                : hdisk1
Device Open On Current Host? : NO
ODM Reservation Policy      : NO RESERVE
Device Reservation State    : NO RESERVE
```

```
root@ss3n04base: /> devrsrv -c query -l hdisk1
Device Reservation State Information
=====
Device Name                : hdisk1
```

```
Device Open On Current Host?      : NO
ODM Reservation Policy           : NO RESERVE
Device Reservation State         : NO RESERVE
```

8. Because we have already enabled replication on our logical unit numbers (LUNs), we can verify the device (hdisk) replication capability at this time. We use the command shown in Example 3-33.

Example 3-33 Checking the replication capability of the disks

```
root@ps3n01base: /> lsattr -El hdisk1 | grep rep
san_rep_cfg      none      SAN Replication Device Configuration Policy True+
san_rep_device detected SAN Replication Device                False
root@ps3n01base: />
```

Tip: The `san_rep_device` attribute shows the hdisk's HyperSwap configuration state:

- ▶ no: (Default) This value does not support PPRC SCSI in-band communication; therefore, it is ineligible for HyperSwap.
- ▶ supported: The device supports PPRC SCSI in-band but it is not a PPRC disk.
- ▶ detected: The device supports in-band communication and it is a PPRC disk, but HyperSwap has not been enabled.
- ▶ yes: The device is a PPRC-configured disk with HyperSwap enabled. (However, this does not guarantee that the AIX host has access to both storage subsystems in the PPRC pair. Check the SAN zoning and LUN masking definitions.)

When `san_rep_device` is set to yes, the hdisk is HyperSwap ready.

9. For better disk identification, we suggest that you assign a `pvid` for each disk that will be used in your cluster. Creating a `pvid` for a disk is shown in Example 3-34. Repeat for all accessible disks on all nodes that you plan to use in the cluster.

Tip: You only need to change the PVID on the accessible disks (PPRC source). The PVID will be automatically replicated to the PPRC target.

Example 3-34 Changing the pvid for the accessible disks

```
root@ps3n01base: /> lspv | grep -w hdisk1
hdisk1          none                                     None
root@ps3n01base: /> chdev -l hdisk1 -a pv=yes
hdisk1 changed
root@ps3n01base: />
```

10. Before activating the HyperSwap function for a disk, check the following information for *all disks* and save it for future reference, as shown in Example 3-35, for `hdisk1` on each system. We save the PVID, Unique Device Identifier (UDID), and IEEE Universally Unique Identifier (UUID) in a text file.

Example 3-35 PVID, UDID, and UUID information

```
root@ps3n01base: /> lspv -u | grep -w hdisk1
hdisk1          00f681f3697a80d1                                     None
200B75XP411A80007210790003IBMfcp -----> UDID
5e722cb5-4e32-1690-3c1c-6718c06b55d3 -----> UUID
```

```
root@ps3n02base: /> lspv -u |grep -w hdisk1
hdisk1          00f681f3697a80d1          None
200B75XP411A80007210790003IBMfc
5e722cb5-4e32-1690-3c1c-6718c06b55d3
```

```
root@ss3n03base: /> lspv -u |grep -w hdisk1
hdisk1          00f681f3697a80d1          None
200B75XP411A80007210790003IBMfc
5e722cb5-4e32-1690-3c1c-6718c06b55d3
```

11. Before we activate the disk to use the HyperSwap (in-band path migration capability) capability, we check the disk availability for hdisk1 and hdisk12. *Note that hdisk1 and hdisk12 are both in the Available state, as shown in Example 3-36.*

Example 3-36 Checking disk availability before activating migration capability

```
root@ps3n01base: /> lsdev -Cc disk
hdisk0 Available C2-T1-01 MPIO IBM 2145 FC Disk
hdisk1 Available 42-T1-01 MPIO IBM 2107 FC Disk
hdisk2 Available 42-T1-01 MPIO IBM 2107 FC Disk
.....<< Snippet >>.....
hdisk11 Available 42-T1-01 MPIO IBM 2107 FC Disk
hdisk12 Available 43-T1-01 MPIO IBM 2107 FC Disk
hdisk13 Available 43-T1-01 MPIO IBM 2107 FC Disk
.....<< Snippet >>.....
hdisk22 Available 43-T1-01 MPIO IBM 2107 FC Disk
root@ps3n01base: />
```

12. We activate the HyperSwap capability for hdisk1 on *all* nodes in the cluster, as shown in Example 3-37.

Example 3-37 Activating HyperSwap for hdisk1

```
root@ps3n01base: /> chdev -l hdisk1 -a san_rep_cfg=migrate_disk -U
hdisk1 changed
root@ps3n02base: /> chdev -l hdisk1 -a san_rep_cfg=migrate_disk -U'
hdisk1 changed
root@ss3n03base: /> chdev -l hdisk1 -a san_rep_cfg=migrate_disk -U'
hdisk1 changed
```

13. We check the disk availability again and observe that the PPRC target (hdisk12) has changed to the **Defined** state, as shown in Example 3-38.

Example 3-38 Checking disk status after activating HyperSwap

```
root@ps3n01base: /> lsdev -Cc disk
hdisk0 Available C2-T1-01 MPIO IBM 2145 FC Disk
hdisk1 Available 43-T1-01 MPIO IBM 2107 FC Disk
hdisk2 Available 42-T1-01 MPIO IBM 2107 FC Disk
hdisk3 Available 42-T1-01 MPIO IBM 2107 FC Disk
.....<< Snippet >>.....
hdisk11 Available 42-T1-01 MPIO IBM 2107 FC Disk
hdisk12 Defined 43-T1-01 MPIO IBM 2107 FC Disk
hdisk13 Available 43-T1-01 MPIO IBM 2107 FC Disk
.....<< Snippet >>.....
hdisk22 Available 43-T1-01 MPIO IBM 2107 FC Disk
root@ps3n01base: />
```

14. Verify the replication attributes of the disk as shown in Example 3-39. The attributes must be **yes** on all nodes in the cluster.

Example 3-39 SAN replication parameters

```
root@ps3n01base: /> lsattr -El hdisk1 | grep san
san_rep_cfg      migrate_disk     SAN Replication Device Configuration Policy True+
san_rep_device   yes              SAN Replication Device                               False
root@ps3n01base: />
```

15. We also verify the PPRC status as shown in Example 3-40. Observe that the secondary path group has changed from **-1** to **1**, which means that the path to access the disk can be swapped to the secondary storage by using HyperSwap. Also, the PPRC target of **hdisk1** (which is **hdisk12** in this case) is removed from this display.

Example 3-40 Checking information about PPRC replicated disks

```
root@ps3n01base: /> lsprrc -Ao | tail +4 | sort -nk1.6,7
hdisk1   Active   0(s)         1           500507630bffc4c8 500507630bffc1e2
.....<< Snippet >>.....
hdisk11   Active   0(s)         -1           500507630bffc4c8 500507630bffc1e2
hdisk13   Active  -1           0            500507630bffc4c8 500507630bffc1e2
.....<< Snippet >>.....
hdisk22   Active  -1           0            500507630bffc4c8 500507630bffc1e2
```

16. Observe that the PVID has not changed. However, the UDID and UUID have changed for the replicated device, as shown in Example 3-41.

Example 3-41 Checking the PVID, UDID, and UUID after activating HyperSwap

```
root@ps3n01base: /> lspv -u | grep -w hdisk1
hdisk1      00f681f3697a80d1      None
352037355850343131413830300050a4cc1307210790003IBMfcp
f7a8da24-4da6-e3c4-434a-d10fca47b0a9

root@ps3n02base: /> lspv -u | grep -w hdisk1
hdisk1      00f681f3697a80d1      None
352037355850343131413830300050a4cc1307210790003IBMfcp
f7a8da24-4da6-e3c4-434a-d10fca47b0a9

root@ss3n03base: /> lspv -u | grep -w hdisk1
hdisk1      00f681f3697a80d1      None
352037355850343131413830300050a4cc1307210790003IBMfcp
f7a8da24-4da6-e3c4-434a-d10fca47b0a9
```

At this point, you can proceed to install and configure PowerHA on your nodes.

3.2.7 PowerHA cluster configuration

We use the following cluster topology configuration:

- ▶ Cluster name: ps3n01base_cluster
- ▶ Cluster sites:
 - Site_A
 - Site_B

- ▶ Cluster nodes:
 - ps3n01base
 - ps3n02base
 - ss3n03base
 - ss3n04base
- ▶ Cluster networks: net_ether_01 (172.16.15.0/24 172.16.29.0/24 172.16.14.0/24)

We used the following steps:

1. We check the disks to use for the cluster configuration, as shown in Example 3-42.

Example 3-42 Disk information shown from all the nodes

```
Command to be run on all nodes: lspv | egrep "None"
HOSTS -----
ps3n01base, ps3n02base, ss3n03base, ss3n04base
-----
hdisk1      00f681f3697a80d1      None
hdisk2      00f681f3697a80fb      None
hdisk3      00f681f3697a8134      None
hdisk4      00f681f3697a816e      None
hdisk5      00f681f3697a81a3      None
hdisk6      00f681f3697a81da      None
hdisk7      00f681f3697a8213      None
hdisk8      00f681f3697a8249      None
hdisk9      00f681f3697a827f      None
hdisk10     00f681f3697a82ed      None
hdisk11     00f681f36a720446     None
hdisk12     none                    None
hdisk13     none                    None
hdisk14     none                    None
hdisk15     none                    None
hdisk16     none                    None
hdisk17     none                    None
hdisk18     none                    None
hdisk19     none                    None
hdisk20     none                    None
hdisk21     none                    None
hdisk22     none                    None
```

2. For our cluster configuration, we chose hdisk9, hdisk10, and hdisk11. Determine whether HyperSwap has been activated for these disks, as shown in Example 3-43.

Example 3-43 Checking whether HyperSwap has been enabled

```
Command to be run on all nodes: lsprrc -Ao | egrep "hdisk9|hdisk10|hdisk11"
HOSTS -----
ps3n01base, ps3n02base, ss3n03base, ss3n04base
-----
hdisk9  Active  0(s)      1          500507630bffc4c8  500507630bffc1e2
hdisk10 Active  0(s)      1          500507630bffc4c8  500507630bffc1e2
hdisk11 Active  0(s)      1          500507630bffc4c8  500507630bffc1e2
```

3. We also check the size of these disks, as shown in Example 3-44 on page 52.

Example 3-44 Checking the size of the disks that we use for our cluster

```
Command to be used on all nodes: bootinfo -s hdisk9 (also for hdisk10 and hdisk11)
HOSTS -----
ps3n01base, ps3n02base, ss3n03base
-----
10240
30720
30720
```

4. We use hdisk9 for the Cluster Aware AIX (CAA) cluster repository, and hdisk10 and hdisk11 to store application data.

Checking prerequisites

Install and check the PowerHA for AIX Enterprise Edition packages on your nodes. Example 3-45 shows an extract of the **lslpp** command executed on each node in the cluster.

Important: You must install PowerHA for AIX Enterprise Edition (EE) V 7.1.2 SP1 or higher. Also, check with your IBM representative for any required or recommended fixes for the AIX and HyperSwap device driver before you configure the PowerHA cluster.

Example 3-45 Checking for PowerHA installed packages

```
Command to be run on all nodes: lslpp -L cluster.*
HOSTS -----
ps3n01base, ps3n02base, ss3n03base, ss3n04base
-----
Fileset                Level  State  Type  Description (Uninstaller)
-----
cluster.adt.es.client.include
                        7.1.2.0  C    F    PowerHA SystemMirror Client
.....<< Snippet >>.....

cluster.es.cfs.rte      7.1.2.0  C    F    Cluster File System Support
cluster.es.cgpprc.cmds  7.1.2.0  C    F    PowerHA SystemMirror
.....<< Snippet >>.....

cluster.xd.license    7.1.2.0  C    F    PowerHA SystemMirror
                        Enterprise Edition License
                        Agreement Files
.....<< Snippet >>.....
```

Although PowerHA cluster verification checks the consistency of installed packages on nodes defined in a cluster, we suggest that you also check manually before you start the cluster configuration.

Cluster topology

Follow these steps:

1. Define the cluster topology: cluster name, nodes, and networks. Use the SMIT fast path **smitty cm_setup_sites_menu** command for multisite configuration as shown in Example 3-46 on page 53.

Example 3-46 SMIT configuration entry point for multisite deployment

Multi Site Cluster Deployment

Move cursor to desired item and press Enter.

Setup a Cluster, Nodes and Networks
Define Repository Disk and Cluster IP Address

Learn more about repository disk and cluster IP address

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

2. Select **Setup a Cluster, Nodes, and Networks**. In the next menu, enter the required information, as shown in Example 3-47.

Example 3-47 Cluster definition menu

Setup Cluster, Sites, Nodes and Networks

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

* Cluster Name	[Entry Fields]
	[ps3n01base_cluster]
* Site 1 Name	[Site_A]
* New Nodes (via selected communication paths)	[ps3n01base ps3n02base] +
* Site 2 Name	[Site_B]
* New Nodes (via selected communication paths)	[ss3n03base ss3n04base] +
Cluster Type	[Stretched Cluster] +

F1=Help	F2=Refresh	F3=Cancel	F4=List	Esc+5=Reset	F6=Command
F7=Edit	F8=Image	F9=Shell	F10=Exit	Enter=Do	

The command execution is successful, as shown in Example 3-48.

Example 3-48 Cluster configuration command status

COMMAND STATUS

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

```
[TOP]
Cluster Name: ps3n01base_cluster
Cluster Connection Authentication Mode: Standard
.....<< Snippet >>.....
```

3. Define a cluster repository disk (CAA) by using the menu shown in Example 3-49.

Example 3-49 Defining the CAA repository disk

```

Multi Site Cluster Deployment

Move cursor to desired item and press Enter.

  Setup a Cluster, Nodes and Networks
  Define Repository Disk and Cluster IP Address

  Learn more about repository disk and cluster IP address

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

```

Example 3-50 shows disk selection for the CAA repository.

Example 3-50 Choosing the CAA repository disk

```

Define Repository Disk and Cluster IP Address

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

* Cluster Name          [Entry Fields]
* Repository Disk      ps3n01base_cluster
Cluster IP Address     [None]
                       []

.....
.                      Repository Disk                      .
.                                                                .
. Move cursor to desired item and press Enter.                .
.                                                                .
.  hdisk1 (00f681f3697a80d1) on all cluster nodes              .
.  hdisk2 (00f681f3697a80fb) on all cluster nodes              .
.  hdisk3 (00f681f3697a8134) on all cluster nodes              .
.  hdisk4 (00f681f3697a816e) on all cluster nodes              .
.  hdisk5 (00f681f3697a81a3) on all cluster nodes              .
.  hdisk6 (00f681f3697a81da) on all cluster nodes              .
.  hdisk7 (00f681f3697a8213) on all cluster nodes              .
.  hdisk8 (00f681f3697a8249) on all cluster nodes              .
.  hdisk9 (00f681f3697a827f) on all cluster nodes              .
.  hdisk10 (00f681f3697a82b5) on all cluster nodes            .
.  hdisk11 (00f681f3697a82ed) on all cluster nodes            .
.                                                                .
.  F1=Help           F2=Refresh           F3=Cancel           .
F1=H. F8=Image      F10=Exit           Enter=Do             .
Esc+. /=Find        n=Find Next           .
F9=S.....

```

4. Verify and synchronize the configuration by using the standard SMIT menu as shown in Example 3-51 on page 55.

Example 3-51 Cluster verification and synchronization menu

Cluster Nodes and Networks

Move cursor to desired item and press Enter.

Standard Cluster Deployment
Multi Site Cluster Deployment

Manage the Cluster
Manage Nodes
Manage Sites
Manage Networks and Network Interfaces
Manage Repository Disks

Discover Network Interfaces and Disks

Verify and Synchronize Cluster Configuration

The verification and synchronization are successful as shown in Example 3-52.

Example 3-52 Cluster verification command status

COMMAND STATUS

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

[TOP]

Verification to be performed on the following:
Cluster Topology
Cluster Resources

Verification will interactively correct verification errors.

Retrieving data from available cluster nodes. This could take a few minutes.

Start data collection on node ss3n03base
Start data collection on node ps3n01base
Start data collection on node ps3n02base
Collector on node ps3n01base completed
Collector on node ps3n02base completed
Collector on node ss3n03base completed
Data collection complete

Verifying Cluster Topology...
.....<< Snippet >>.....

5. Verify the cluster topology by using the **clmgr** command, as shown in Example 3-53 on page 56.

Example 3-53 Verifying the cluster topology

```
root@ps3n01base: /> clmgr q site
Site_A
Site_B
root@ps3n01base: /> clmgr -v q site
NAME="Site_A"
GID="15991788271"
STATE="STABLE"
NODES="ps3n01base ps3n02base"
SITE_IP=""
RECOVERY_PRIORITY="1"

NAME="Site_B"
GID="15991788272"
STATE="STABLE"
NODES="ss3n03base ss3n04base"
SITE_IP=""
RECOVERY_PRIORITY="2"
root@ps3n01base: />
```

6. Start cluster services on all nodes and verify the status as shown in Example 3-54.

Example 3-54 Start cluster services on all nodes and verify the service status

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        [ps3n01base]                  +
* 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?
.....
.                Start Cluster Services on these nodes                .
.
. Move cursor to desired item and press F7.                               .
.   ONE OR MORE items can be selected.                                   .
. Press Enter AFTER making all selections.                               .
.
. > ps3n01base                                                            .
. > ps3n02base                                                            .
. > ss3n03base                                                            .
. > ss3n04base                                                            .
.
. F1=Help          F2=Refresh          F3=Cancel                          .
F1=Help  . F7=Select          F8=Image          F10=Exit                          .
Esc+5=Res. Enter=Do  ./=Find          n=Find Next                          .
F9=Shell .....

```

7. On every node, we check the status of the PowerHA services as shown in Example 3-55 on page 57.

Example 3-55 Service status in node ps3n01

```

root@ps3n01base: /> /usr/es/sbin/cluster/utilities/clshsrv -v
Status of the RSCT subsystems used by HACMP:
Subsystem      Group      PID      Status
cthags         cthags    7798900  active
ctrmc         rsct      11075740 active

Status of the HACMP subsystems:
Subsystem      Group      PID      Status
clstrmgrES    cluster   7667826  active
clcomd        caa       6291658  active

Status of the optional HACMP subsystems:
Subsystem      Group      PID      Status
clinforES     cluster   7995458  active

```

Cluster resources

The cluster resources and resource group (RG) definition are described.

Resource groups

Follow these steps:

1. We define one resource group (Example 3-56) using the following SMIT fast path:

smitty cm_add_resource_group

Example 3-56 Defining resource group using smit -C cm_add_resource_group

```

                Add a Resource Group (extended)

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

                                [Entry Fields]
* Resource Group Name           [R1]

Inter-Site Management Policy    [Online On Either Site]      +
* Participating Nodes from Primary Site [ps3n01base ps3n02base]      +
  Participating Nodes from Secondary Site [ss3n03base ss3n04base]      +

Startup Policy                  Online On Home Node Only      +
Fallover Policy                 Fallover To Next Priority Node In Th> +
Fallback Policy                 Never Fallback                 +

F1=Help      F2=Refresh      F3=Cancel    F4=List
Esc+5=Reset  F6=Command   F7=Edit     F8=Image
F9=Shell     F10=Exit     Enter=Do

```

We use the previously HyperSwap enabled disks in a volume group named datavg. The datavg group is part of RG1, as shown in Example 3-57 on page 58.

We also define IP service addresses for Site_A and Site_B, and add them in the resource group configuration.

Example 3-57 Adding a volume group in the resource group and IP service addresses

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                  RG1
Inter-site Management Policy         Online On Either Site
Participating Nodes from Primary Site ps3n01base ps3n02base
Participating Nodes from Secondary Site ss3n03base ss3n04base

Startup Policy                       Online On Home Node Only
Failover Policy                      Fallover To Next Priority Node In Th>
Fallback Policy                      Never Fallback

Service IP Labels/Addresses          [ps3n01svc ss3n03svc]      +
Application Controller Name          []                          +

Volume Groups                        [datavg]                  +
Use forced varyon of volume groups, if necessary false             +
Automatically Import Volume Groups   false                       +

Allow varyon with missing data updates? true                       +
  (Asynchronous GLVM Mirroring Only)
Default choice for data divergence recovery ignore                   +
[MORE...30]
```

.....<<<<<<Snippet>>>>>>.....

```
Primary Workload Manager Class      []                          +
Secondary Workload Manager Class     []                          +

Miscellaneous Data                  []
WPAR Name                           []                          +
User Defined Resources               []                          +
SVC PPRC Replicated Resources       []                          +
EMC SRDF(R) Replicated Resources    []                          +
DS8000 Global Mirror Replicated Resources []                          +
XIV Replicated Resources             []                          +
TRUFCOPY Replicated Resources       []                          +
DS8000-Metro Mirror (In-band) Resources []                          +
```

```
F1=Help          F2=Refresh      F3=Cancel       F4=List
Esc+5=Reset     F6=Command     F7=Edit         F8=Image
F9=Shell        F10=Exit       Enter=Do
```

2. After the resource group configuration, we verify and synchronize the cluster configuration.

The resource group status is shown in Example 3-58 on page 59.

Example 3-58 Resource group status

```
root@ps3n01base: /> clRGinfo
-----
Group Name      State                Node
-----
RG1             ONLINE              ps3n01base@Sit
                OFFLINE             ps3n02base@Sit
                ONLINE SECONDARY   ss3n03base@Sit
                OFFLINE             ss3n04base@Sit
-----
```

The volume group datavg uses hdisk10 and hdisk11 as shown in Example 3-59.

Example 3-59 Hdisks used in datavg volume group

```
root@ps3n01base: /> lsvg -p datavg
datavg:
PV_NAME          PV STATE            TOTAL PPs   FREE PPs   FREE DISTRIBUTION
hdisk10          active              239         199        48..08..47..48..48
hdisk11          active              239         239        48..48..47..48..48

root@ps3n01base: /> lsvg -l datavg
datavg:
LV NAME          TYPE               LPs         PPs         PVs  LV STATE      MOUNT POINT
lv01             jfs2               40          40          1   open/syncd    /data1
-----
```

3. The **lsprrc** command is used to determine the hdisk status as shown in Example 3-60.

Example 3-60 Replication status

```
root@ps3n01base: /> lsprrc -p hdisk10
path          WWNN              LSS  VOL  path
group id      group status
=====
0(s)          500507630bffc4c8 0xb2 0x01 PRIMARY
1             500507630bffc1e2 0xb2 0x01 SECONDARY

path          path path          parent connection
group id     id  status
=====
0             0   Enabled  fscsi1 500507630b1884c8,40b2400100000000
0             1   Enabled  fscsi4 500507630b5304c8,40b2400100000000
1             2   Enabled  fscsi2 500507630b1001e2,40b2400100000000
1             3   Enabled  fscsi3 500507630b1301e2,40b2400100000000

root@ps3n01base: /> lsprrc -p hdisk11
path          WWNN              LSS  VOL  path
group id      group status
=====
0(s)          500507630bffc4c8 0xb3 0x01 PRIMARY
1             500507630bffc1e2 0xb3 0x01 SECONDARY

path          path path          parent connection
group id     id  status
=====
0             0   Enabled  fscsi1 500507630b1884c8,40b3400100000000
0             1   Enabled  fscsi4 500507630b5304c8,40b3400100000000
```

1	2	Enabled	fscsi2	500507630b1001e2,40b3400100000000
1	3	Enabled	fscsi3	500507630b1301e2,40b3400100000000

Storage systems

Follow these steps:

1. To enable PowerHA control over the HyperSwap facility, we need to define the mirror groups (MGs) in the PowerHA SystemMirror cluster configuration. Before we define the mirror groups, we need to define the DS8000 Metro Mirror resources, as shown in Example 3-61.

Tip: PowerHA uses the mirror group (MG) to control the paths to the active storage. Three types of mirror groups can be defined:

- ▶ User: For user-application disks
- ▶ System: Disks that are owned by the AIX system/node, for example, rootvg and paging devices
- ▶ Cluster repository: CAA repository disk associated with PowerHA

2. We use the `smit cm_cfg_ds8k_mm_in_band_resource` fast path command. Alternately, you can use the following SMIT command and menu selections:

smit hacmp → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources

Example 3-61 Configuring the storage systems (one for each site)

Configure DS8000 Metro Mirror (In-Band) Resources

Move cursor to desired item and press Enter.

```
Configure Storage Systems
Configure Mirror Groups
```

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

.....

Move cursor to desired item and press Enter.

```
Add a Storage System
Change/Show a Storage System
Remove a Storage System
```

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

3. We add both storage systems in PowerHA for each site as shown in Example 3-62 on page 61.

Tip: You can add a storage subsystem *after* enabling the AIX_AAPCM storage driver in AIX.

Example 3-62 Defining the storage subsystem for both sites

For SITE_A

Add a Storage System

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

	[Entry Fields]	
* Storage System Name	[STORAGE_A]	
* Site Association	Site_A	+
* Vendor Specific Identifier	IBM.2107-00000XP411	+
* WWNN	500507630BFFC4C8	+

F1=Help	F2=Refresh	F3=Cancel	F4=List
Esc+5=Reset	F6=Command	F7=Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

.....

For SITE_B

Add a Storage System

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

	[Entry Fields]	
* Storage System Name	[STORAGE_B]	
* Site Association	Site_B	+
* Vendor Specific Identifier	IBM.2107-00000WT971	+
* WWNN	500507630BFFC1E2	+

F1=Help	F2=Refresh	F3=Cancel	F4=List
Esc+5=Reset	F6=Command	F7=Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Mirror groups

Follow these steps:

1. Because we have one volume group designated for user data and one CAA repository disk, we define the corresponding MGs for these resources.

We create the CAA Repository - cluster repository mirror group using the SMIT fast path, **smit cm_cfg_mirr_gps** as shown in Example 3-63 on page 62.

Example 3-63 Defining the CAA MG

Configure Mirror Groups

Move cursor to desired item and press Enter.

Add a Mirror Group

- Change/Show a Mirror Group
- Remove a Mirror Group

```

.....
.           Select the type of Mirror Group to Add           .
.                                                                 .
. Move cursor to desired item and press Enter.                .
.                                                                 .
.  User                                                         .
.  System                                                         .
.  Cluster_Repository                                         .
.                                                                 .
.  F1=Help           F2=Refresh           F3=Cancel           .
.  F8=Image          F10=Exit             Enter=Do           .
F1=H. /=Find        n=Find Next          .
F9=S.....

```

2. We specify the details required by the cluster repository MG definition as shown in Example 3-64.

Example 3-64 Defining the Cluster Repository MG

Add cluster Repository Mirror Group

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

```

                                     [Entry Fields]
* Mirror Group Name                 [CAA_MG]
* Site Name                          Site_A Site_B          +
* Non Hyperswap Disk                hdisk8:84b85e42-7eec-4786> +
* Hyperswap Disk                    hdisk9:64652ce3-3fe9-d40b> +
  Hyperswap                          Enabled                +
  Consistency Group                  Enabled                +
  Unplanned HyperSwap Timeout (in sec) [60]                    #
  Hyperswap Priority                  High                    +

F1=Help           F2=Refresh           F3=Cancel           F4=List
Esc+5=Reset       F6=Command           F7=Edit             F8=Image
F9=Shell          F10=Exit             Enter=Do

```

3. We create the User MG using the datavg disks as shown in Example 3-65.

Example 3-65 Defining the User MG

Configure Mirror Groups

Move cursor to desired item and press Enter.

Add a Mirror Group

Change/Show a Mirror Group
 Remove a Mirror Group

```

.....
.           Select the type of Mirror Group to Add           .
.                                                                 .
. Move cursor to desired item and press Enter.                .
.                                                                 .
. User                                                         .
. System                                                         .
. Cluster_Repository                                           .
.                                                                 .
. F1=Help                F2=Refresh                F3=Cancel    .
. F8=Image                F10=Exit                 Enter=Do     .
F1=H· /|=Find            n=Find Next                .
F9=S.....
  
```

.....

Add a User Mirror Group

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

	[Entry Fields]	
* Mirror Group Name	[DATA_MG]	
Volume Group(s)	datavg	+
Raw Disk(s)		+
Hyperswap	Enabled	+
Consistency Group	Enabled	+
Unplanned HyperSwap Timeout (in sec)	[60]	#
Hyperswap Priority	Medium	+
Recovery Action	Manual	+

F1=Help	F2=Refresh	F3=Cancel	F4=List
Esc+5=Reset	F6=Command	F7=Edit	F8=Image

Tip: After a user MG is defined, it must be added to the corresponding cluster resource group.

- We add the User MG to the resource group, as shown in Example 3-66, by choosing the DS8000-Metro Mirror (In-band) Resources.

Example 3-66 Adding a mirror group to existing resource group

Change/Show All Resources and Attributes for a Resource Group

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

[TOP]	[Entry Fields]
Resource Group Name	RG1
Inter-site Management Policy	Online On Either Site
Participating Nodes from Primary Site	ps3n01base ps3n02base

```

Participating Nodes from Secondary Site          ss3n03base ss3n04base

Startup Policy                                 Online On Home Node Only
Failover Policy                               Fallback To Next Priority>
Fallback Policy                               Never Fallback

Service IP Labels/Addresses                   [ps3n01svc ss3n03svc]      +
Application Controller Name                   []                          +

Volume Groups                                 [datavg ]                  +
Use forced varyon of volume groups, if necessary  false                       +
Automatically Import Volume Groups            false                       +

Allow varyon with missing data updates?
  (Asynchronous GLVM Mirroring Only)          true                         +
Default choice for data divergence recovery     ignore

+.....<<Snippet>>.....
Tape Resources                                []                          +
Raw Disk PVIDs                                []                          +
Raw Disk UUIDs/hdisks                        []                          +
Disk Error Management?                       no                           +

Primary Workload Manager Class                 []                          +
Secondary Workload Manager Class               []                          +
DS8000 Global Mirror Replicated Resources     []                          +
XIV Replicated Resources                      []                          +
TRUECOPY Replicated Resources                 []                          +
DS8000-Metro Mirror (In-band) Resources      DATA_MG                   +
[BOTTOM]
F1=Help          F2=Refresh          F3=Cancel          F4=List
Esc+5=Reset      F6=Command          F7=Edit            F8=Image

```

Stop the cluster services before synchronizing cluster configuration.

5. After the mirror groups configuration is complete, we verify and synchronize the PowerHA cluster.

Important: *After the MGs are defined, the cluster configuration must be verified and synchronized with the PowerHA services stopped.*

3.2.8 Planned tests: Storage maintenance

This scenario describes storage maintenance. Swap the active storage and check the application status:

- ▶ Check the PowerHA log files by using the `lspprc -Ao/-p/v` command.
- ▶ PowerHA swaps the MGs: cluster repository (CAA), and user, one by one.
- ▶ Check the PowerHA log files again by using the `lspprc -Ao/-p/v` command.

Storage maintenance of Storage_A

We swap the disks to Storage_B and verify that the applications are still running.

Starting point

Follow these steps:

1. In the beginning of the test, the disks are configured as shown in Example 3-67. All the disks are accessed from the primary storage (FC path points to Storage_ A). The secondary copy is in Storage_ B.

Example 3-67 Disk configuration before storage maintenance test

```
root@ps3n01base: /> lsprrc -p hdisk9
path      WWNN          LSS  VOL  path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xb1 0x00 PRIMARY
1         500507630bffc1e2 0xb1 0x00 SECONDARY

path      path path      parent connection
group id  id  status
=====
0         0   Enabled fscsi1 500507630b1884c8,40b1400000000000
0         1   Enabled fscsi4 500507630b5304c8,40b1400000000000
1         2   Enabled fscsi2 500507630b1001e2,40b1400000000000
1         3   Enabled fscsi3 500507630b1301e2,40b1400000000000

root@ps3n01base: /> lsprrc -p hdisk10
path      WWNN          LSS  VOL  path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xb2 0x01 PRIMARY
1         500507630bffc1e2 0xb2 0x01 SECONDARY

path      path path      parent connection
group id  id  status
=====
0         0   Enabled fscsi1 500507630b1884c8,40b2400100000000
0         1   Enabled fscsi4 500507630b5304c8,40b2400100000000
1         2   Enabled fscsi2 500507630b1001e2,40b2400100000000
1         3   Enabled fscsi3 500507630b1301e2,40b2400100000000

root@ps3n01base: /> lsprrc -p hdisk11
path      WWNN          LSS  VOL  path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xb3 0x01 PRIMARY
1         500507630bffc1e2 0xb3 0x01 SECONDARY

path      path path      parent connection
group id  id  status
=====
0         0   Enabled fscsi1 500507630b1884c8,40b3400100000000
0         1   Enabled fscsi4 500507630b5304c8,40b3400100000000
1         2   Enabled fscsi2 500507630b1001e2,40b3400100000000
1         3   Enabled fscsi3 500507630b1301e2,40b3400100000000
root@ps3n01base: />
```

The Metro Mirror status for the corresponding disks before the planned swap operation is shown in Example 3-68.

Example 3-68 Metro Mirror status

Storage_A

```

dscli> lsprrc -l b100-b3ff
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
=====
=====
B100:B100 Full Duplex - Metro Mirror 0          Disabled Disabled
Invalid -          B1      5          Disabled Invalid
Disabled          Disabled N/A      Enabled Unknown -
B101:B101 Full Duplex - Metro Mirror 0          Disabled Disabled
Invalid -          B1      5          Disabled Invalid
Disabled          Disabled N/A      Enabled Unknown -
B201:B201 Full Duplex - Metro Mirror 0          Disabled Disabled
Invalid -          B2      5          Disabled Invalid
Disabled          Disabled N/A      Enabled Unknown -
B301:B301 Full Duplex - Metro Mirror 0          Disabled Disabled
Invalid -          B3      5          Disabled Invalid
Disabled          Disabled N/A      Enabled Unknown -
dscli>

```

Storage_B

```

dscli> lsprrc -l b100-b3ff
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
=====
=====
B100:B100 Target Full Duplex - Metro Mirror 0          Disabled
Invalid Disabled -          B1      unknown Disabled
Invalid          Disabled Disabled N/A      N/A      Unknown -
B101:B101 Target Full Duplex - Metro Mirror 0          Disabled
Invalid Disabled -          B1      unknown Disabled
Invalid          Disabled Disabled N/A      N/A      Unknown -
B201:B201 Target Full Duplex - Metro Mirror 0          Disabled
Invalid Disabled -          B2      unknown Disabled
Invalid          Disabled Disabled N/A      N/A      Unknown -
B301:B301 Target Full Duplex - Metro Mirror 0          Disabled
Invalid Disabled -          B3      unknown Disabled
Invalid          Disabled Disabled N/A      N/A      Unknown -
=====

```

2. We perform the swap operation by using the `smit cm_user_mirr_gp` fast path command, as shown in Example 3-69 on page 67, only for the user MG named DATA_MG.

Example 3-69 Manage the user MG

Manage User Mirror Group(s)

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

			[Entry Fields]	
* Mirror Group(s)			DATA_MG	+
* Operation			Swap	+
F1=Help	F2=Refresh	F3=Cancel	F4=List	
Esc+5=Reset	F6=Command	F7=Edit	F8=Image	
F9=Shell	F10=Exit	Enter=Do		

3. By using the `lspprc -p` command, we check the disks' status (shown in Example 3-70).

Example 3-70 Swapping paths to Storage_B

```
root@ps3n01base: /> lspprc -p hdisk9
path      WWNN          LSS  VOL    path
group id  group status
=====
0(s)      500507630bffc4c8  0xb1 0x00  PRIMARY
1         500507630bffc1e2  0xb1 0x00  SECONDARY

path      path path      parent connection
group id  id  status
=====
0 0 Enabled fscsi1 500507630b1884c8,40b1400000000000
0 1 Enabled fscsi4 500507630b5304c8,40b1400000000000
1 2 Enabled fscsi2 500507630b1001e2,40b1400000000000
1 3 Enabled fscsi3 500507630b1301e2,40b1400000000000

root@ps3n01base: /> lspprc -p hdisk10
path      WWNN          LSS  VOL    path
group id  group status
=====
0         500507630bffc4c8  0xb2 0x01  SECONDARY
1(s)     500507630bffc1e2  0xb2 0x01  PRIMARY

path      path path      parent connection
group id  id  status
=====
0 0 Enabled fscsi1 500507630b1884c8,40b2400100000000
0 1 Enabled fscsi4 500507630b5304c8,40b2400100000000
1 2 Enabled fscsi2 500507630b1001e2,40b2400100000000
1 3 Enabled fscsi3 500507630b1301e2,40b2400100000000

root@ps3n01base: /> lspprc -p hdisk11
path      WWNN          LSS  VOL    path
group id  group status
=====
0         500507630bffc4c8  0xb3 0x01  SECONDARY
1(s)     500507630bffc1e2  0xb3 0x01  PRIMARY
```

path group	path id	path status	parent	connection
0	0	Enabled	fscsi1	500507630b1884c8,40b3400100000000
0	1	Enabled	fscsi4	500507630b5304c8,40b3400100000000
1	2	Enabled	fscsi2	500507630b1001e2,40b3400100000000
1	3	Enabled	fscsi3	500507630b1301e2,40b3400100000000

- On the storage side, on both storage subsystems, the replication paths for the corresponding LSSs are defined, and the PPRC relationships were changed, as shown in Example 3-71.

Example 3-71 Path status on both Storage_A and Storage_B after the HyperSwap operation

Storage_A

```
dscli> lsprrcpath -l b1-b3
```

Src	Tgt	State	SS	Port	Attached	Port	Tgt	WWNN	Failed	Reason	PPRC	CG
B1	B1	Success	FFB1	I0102	I0102		500507630BFFC1E2	-	-	Enabled		
B1	B1	Success	FFB1	I0202	I0132		500507630BFFC1E2	-	-	Enabled		
B2	B2	Success	FFB2	I0102	I0102		500507630BFFC1E2	-	-	Enabled		
B2	B2	Success	FFB2	I0202	I0132		500507630BFFC1E2	-	-	Enabled		
B3	B3	Success	FFB3	I0102	I0102		500507630BFFC1E2	-	-	Enabled		
B3	B3	Success	FFB3	I0202	I0132		500507630BFFC1E2	-	-	Enabled		

```
dscli> lsprrc -l b100-b3ff
```

ID	State	Reason	Type	Out Of Sync	Tracks	Tgt	Read	Src	Cascade	Tgt
Cascade	Date	Suspended	SourceLSS	Timeout	(secs)	Critical	Mode	First	Pass	Status
Resync	Tgt	Write	GMIR	CG	PPRC	CG	isTgtSE	Disable	AutoResync	
B100:B100	Full Duplex	-	Metro Mirror	0		Disabled	Disabled	Invalid		
-	B1	unknown	Disabled	Invalid		Disabled	Disabled			
Disabled	N/A	Enabled	Unknown	-						
B101:B101	Full Duplex	-	Metro Mirror	0		Disabled	Disabled	Invalid		
-	B1	unknown	Disabled	Invalid		Disabled	Disabled			
Disabled	N/A	Enabled	Unknown	-						
B201:B201	Target Full Duplex	-	Metro Mirror	0		Disabled	Invalid			
Disabled	-	B2	unknown	Disabled	Invalid		Disabled			
Disabled	N/A	N/A	Unknown	-						
B301:B301	Target Full Duplex	-	Metro Mirror	0		Disabled	Invalid			
Disabled	-	B3	unknown	Disabled	Invalid		Disabled			
Disabled	N/A	N/A	Unknown	-						

Storage_B

```
dscli> lsprrcpath -l b1-b3
```

Src	Tgt	State	SS	Port	Attached	Port	Tgt	WWNN	Failed	Reason	PPRC	CG
B1	B1	Success	FFB1	I0102	I0102		500507630BFFC4C8	-	-	Enabled		
B1	B1	Success	FFB1	I0132	I0202		500507630BFFC4C8	-	-	Enabled		
B2	B2	Success	FFB2	I0102	I0102		500507630BFFC4C8	-	-	Enabled		
B2	B2	Success	FFB2	I0132	I0202		500507630BFFC4C8	-	-	Enabled		
B3	B3	Success	FFB3	I0102	I0102		500507630BFFC4C8	-	-	Enabled		

B3 B3 Success FFB3 I0132 I0202 500507630BFFC4C8 - Enabled

dscli> lsprrc -l b100-b3ff

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
=====					
=====					
B100:B100	Target Full Duplex -	Metro Mirror 0		Disabled Invalid	
Disabled -	B1	unknown	Disabled	Invalid	Disabled
Disabled N/A	N/A	Unknown -			
B101:B101	Target Full Duplex -	Metro Mirror 0		Disabled Invalid	
Disabled -	B1	unknown	Disabled	Invalid	Disabled
Disabled N/A	N/A	Unknown -			
B201:B201	Full Duplex -	Metro Mirror 0		Disabled Disabled Invalid	
-	B2	unknown	Disabled	Invalid	Disabled
Disabled N/A	Enabled	Unknown -			
B301:B301	Full Duplex -	Metro Mirror 0		Disabled Disabled Invalid	
-	B3	unknown	Disabled	Invalid	Disabled
Disabled N/A	Enabled	Unknown -			

Tip: The PPRC direction must be maintained by PowerHA. However, if you have manually reverted the replication direction, the operations are recognized by PowerHA and the replication is not automatically reverted.

- 5. We pause the PPRC relationships for B201 and B301, as shown in Example 3-72, and we also check the status of the disks at the operating system level (Example 3-73 on page 70).

Example 3-72 The disk status after pausepprc operation

```
dscli> pausepprc -remotedev IBM.2107-75WT971 b201:b201 b301:b301
CMUC00157I pausepprc: Remote Mirror and Copy volume pair B201:B201 relationship successfully
paused.
CMUC00157I pausepprc: Remote Mirror and Copy volume pair B301:B301 relationship successfully
paused.
```

.....

The disk status:

```
dscli> lsprrc -l b100-b3ff
```

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
=====					
=====					
B100:B100	Full Duplex -	Metro Mirror 0		Disabled Disabled Invalid	
-	B1	5	Disabled	Invalid	Disabled
Disabled N/A	Enabled	Unknown -			

B101:B101	Full Duplex -	Metro Mirror 0	Disabled	Disabled	Invalid	Disabled	Invalid
-	B1	5	Disabled	Invalid	Disabled	Disabled	
Disabled	N/A	Enabled Unknown -					
B201:B201	Suspended	Host Source Metro Mirror 1	Disabled	Invalid	Disabled	Disabled	Invalid
-	B2	5	Disabled	Invalid	Disabled	Disabled	
Disabled	N/A	Enabled Unknown -					
B301:B301	Suspended	Host Source Metro Mirror 0	Disabled	Invalid	Disabled	Disabled	Invalid
-	B3	5	Disabled	Invalid	Disabled	Disabled	
Disabled	N/A	Enabled Unknown -					

After the suspend operation, the disks appear as shown in Example 3-73.

Example 3-73 Disks' status after pausepprc command

```

root@ps3n01base:/var/hacmp/xd/log> lsprrc -p hdisk10
path      WWNN          LSS  VOL  path
group id
=====
0(s)      500507630bffc4c8  0xb2  0x01  PRIMARY,
          500507630bffc4c8  0xb2  0x01  SUSPENDED,
          500507630bffc4c8  0xb2  0x01  OOS
1         500507630bffc1e2  0xb2  0x01  SECONDARY,
          500507630bffc1e2  0xb2  0x01  SUSPENDED

path      path  path      parent  connection
group id  id    status
=====
0         0     Enabled   fscsi1  500507630b1884c8,40b2400100000000
0         1     Enabled   fscsi4  500507630b5304c8,40b2400100000000
1         2     Enabled   fscsi2  500507630b1001e2,40b2400100000000
1         3     Enabled   fscsi3  500507630b1301e2,40b2400100000000

root@ps3n01base:/var/hacmp/xd/log> lsprrc -p hdisk11
path      WWNN          LSS  VOL  path
group id
=====
0(s)      500507630bffc4c8  0xb3  0x01  PRIMARY,
          500507630bffc4c8  0xb3  0x01  SUSPENDED
1         500507630bffc1e2  0xb3  0x01  SECONDARY,
          500507630bffc1e2  0xb3  0x01  SUSPENDED

path      path  path      parent  connection
group id  id    status
=====
0         0     Enabled   fscsi1  500507630b1884c8,40b3400100000000
0         1     Enabled   fscsi4  500507630b5304c8,40b3400100000000
1         2     Enabled   fscsi2  500507630b1001e2,40b3400100000000
1         3     Enabled   fscsi3  500507630b1301e2,40b3400100000000

```

Tip: DS8800 PPRC operations:

- ▶ Failover = Issued to a secondary PPRC LUN. This operation moves the PPRC LUN to primary, suspended. This operation does not immediately alter the state of the previous primary PPRC LUN. Mirroring is stopped.
- ▶ Failback = Issued to a primary, suspended PPRC LUN, which has valid replication path to peer. This operation moves this PPRC LUN to primary, and the alternate LUN to secondary. Mirroring is resumed.
- ▶ Suspend = Moves a primary PPRC LUN to primary, suspended. Mirroring is stopped.
- ▶ Resume = Moves a primary, suspended PPRC LUN to primary. Mirroring is resumed.
- ▶ Freeze = This operation destroys the replication paths for the LSS, which stops the mirroring communication, and moves LUNs on this LSS into a 60-second long busy state.
- ▶ UnFreeze = This operation immediately exits the 60-second long busy state caused by a freeze. This operation does not resume mirroring.

6. The HyperSwap status of the cluster repository (CAA) disk is shown in Example 3-74.

Operational logs: If the HyperSwap operations are performed through PowerHA SMIT menus, the corresponding events are logged in the `hacmp.out` file. Otherwise, all events are logged through the `syslog` subsystem.

Example 3-74 HyperSwap CAA disk status

```
root@ss3n03base: /> lsprrc -p hdisk9
path      WWNN          LSS  VOL   path
group id
=====
0(s)      500507630bffc4c8 0xb1 0x00  PRIMARY
1         500507630bffc1e2 0xb1 0x00  SECONDARY

path      path path      parent connection
group id id   status
=====
0         0   Enabled  fscsi1  500507630b1884c8,40b1400000000000
0         1   Enabled  fscsi4  500507630b5304c8,40b1400000000000
1         2   Enabled  fscsi2  500507630b1001e2,40b1400000000000
1         3   Enabled  fscsi3  500507630b1301e2,40b1400000000000
```

7. The swap for the CAA Repository disk is performed by using the `smit cm_cluser_repos_mirr_gp` fast path command as shown in Example 3-75 on page 72.

Example 3-75 CAA Repository swapping paths

Manage Cluster Repository Mirror Group

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

			[Entry Fields]	
* Mirror Group Name			CAA_MG	+
* Operation			Swap	+
F1=Help	F2=Refresh	F3=Cancel	F4=List	
Esc+5=Reset	F6=Command	F7=Edit	F8=Image	
F9=Shell	F10=Exit	Enter=Do		

8. After swapping the cluster repository (CAA) mirror group, the status of the CAA repository disk is shown in Example 3-76. The CAA repository is now accessing the disk from Storage_B as the primary.

Example 3-76 Cluster repository MG after swap

```

root@ps3n01base:/usr/es/sbin/cluster/worksheets> lspprc -p hdisk9
path          WWNN              LSS  VOL  path
group id      group status
=====
0             500507630bffc4c8  0xb1 0x00 SECONDARY
1(s)         500507630bffc1e2  0xb1 0x00 PRIMARY

path  path  path  parent  connection
group id  id  status
=====
0  0  Enabled  fscsi1  500507630b1884c8,40b1400000000000
0  1  Enabled  fscsi4  500507630b5304c8,40b1400000000000
1  2  Enabled  fscsi2  500507630b1001e2,40b1400000000000
1  3  Enabled  fscsi3  500507630b1301e2,40b1400000000000

```

The previous operations are logged in the hacmp.out file, as shown in Example 3-77.

Example 3-77 Swap operation logs in hacmp.out log

```

ERROR: rep_disk_notify : Thu Dec 13 14:33:04 BEIST 2012 : Node ss3n04base on
Cluster ps3n01base_cluster has lost access to repository disk hdisk9. Please
recover from this error or replace the repository disk using smitty.
.....<<Snippet>>.....
rep_disk_notify: Thu Dec 13 14:33:30 BEIST 2012 : Access to repository disk has
been restored on Node ss3n04base

```

9. After the application is verified, we have swapped the user MG and cluster repository MG back to Storage_A.

3.2.9 Planned tests: Site maintenance

In this scenario, we move all resources from Site_A to Site_B (storage resources, as well as resource groups).

The steps are almost identical to the steps for storage maintenance. The only difference is that the resource groups will also be moved to Site_B.

Sequential order: The HyperSwap operations and application failover can be done only in sequential order.

Starting point

Follow these steps:

1. We start by swapping the user and cluster repository MGs as we did in the previous test. (See 3.2.8, “Planned tests: Storage maintenance” on page 64.)

After the MGs have been swapped to Storage_B (active FC paths pointing to secondary storage and nodes accessing the LUNs in Storage_B in addition to reverting the Metro Mirror replication direction), we proceed to the resource group movement to one of the nodes in Site_B.

2. In the next step, we move our application from Site_A to Site_B, using cluster single point of control (C-SPOC) as shown in Example 3-78.

Example 3-78 Moving a resource group from Site_A to Site_B

Resource Group and Applications

Move cursor to desired item and press Enter.

```
Show the Current State of Applications and Resource Groups
Bring a Resource Group Online
Bring a Resource Group Offline
Move Resource Groups to Another Node
Move Resource Groups to Another Site

Suspend/Resume Application Monitoring
Applica.....
.                               Select a Resource Group                               .
.                                                                           .
. Move cursor to desired item and press Enter. Use arrow keys to scroll. .
.                                                                           .
. #                                                                           .
. # Resource Group                  State                  Node(s) / Site .
. #                                                                           .
. RG2                                ONLINE                ss3n03base / Si .
.                                                                           .
. #                                                                           .
. # 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=Shell .....
```

Moving applications on the secondary site implies a short downtime for the applications during the failover on the secondary site.

3.3 Migrating PowerHA cluster to HyperSwap enabled storage

The migration of an existing cluster to HyperSwap for PowerHA SystemMirror must be carefully planned. The configuration steps are detailed in the PowerHA *Storage-based high availability and disaster recovery manual*:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/hacmp_pprc_pdf.pdf

We verify that our environment meets the requirements before starting the configuration of HyperSwap for PowerHA SystemMirror:

- ▶ DS8800 configuration (including microcode level).
- ▶ A PowerHA SystemMirror cluster is defined.
- ▶ All PowerHA SystemMirror nodes are defined.
- ▶ All PowerHA SystemMirror sites are defined.
- ▶ All PowerHA SystemMirror resource groups and associated resources are configured and working correctly.

The HyperSwap configuration requires the PowerHA Enterprise Edition environment. In this scenario, two nodes are located onsite named Site_A (ps3n01 and ps3n02), and the third node (ss3n03) is in Site_B.

The nodes meet all of the following requirements for HyperSwap enablement:

- ▶ The latest fixes for AIX operating system, Path Control Module (PCM) driver subsystem, and PowerHA SystemMirror have been applied.
- ▶ Storage has the required firmware for HyperSwap in-band communication.
- ▶ The disks are correctly distributed across the logical subsystems (LSSs) on both storage subsystems, following the PowerHA SystemMirror recommendations:
http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/hacmp_pprc_pdf.pdf
- ▶ SAN zoning is configured as required by the configuration. In the HyperSwap environment, the target disks from both storage subsystems must be configured to the logical partition (LPAR) (node) that will be added to the cluster.

Planning for the maintenance window: This scenario has two steps that require downtime:

- ▶ Enabling HyperSwap in AIX (changing the PCM for HyperSwap disks)
- ▶ Cluster verification and synchronization after defining the storage resources, cluster repository, and user mirror groups

We advise that you plan for a single maintenance window that covers the entire migration procedure. *Ensure that you create a cluster snapshot before and after the migration.*

3.3.1 Planning the cluster

Because HyperSwap is a feature of PowerHA SystemMirror Enterprise Edition, we have already configured the cluster planning:

- ▶ Naming convention for the cluster nodes
- ▶ Networks used in the configuration and the IP addressing plan, considering also the network requirements for the applications that will be highly available in the cluster
- ▶ Storage requirements for the CAA repository disk and also the requirements for the HyperSwap environment
- ▶ SAN intersite configuration, bandwidth between sites, zoning configuration for storage subsystems, and host communications

3.3.2 Identifying the nodes and sites

The HyperSwap feature in PowerHA SystemMirror requires an existing (configured) cluster, based on PowerHA SystemMirror Enterprise Edition, with an extended distance cluster configuration (sites defined).

At the time of writing this paper, the HyperSwap operations are limited to DS8800 Metro Mirror replication. Therefore, we configured a cluster with two sites, Site_A and Site_B. Although the two storage subsystems can be in the same physical site, site definition in PowerHA is mandatory.

The PowerHA SystemMirror software packages are already installed, as shown in Example 3-45 on page 52, and we have defined a cluster with two sites, Site_A and Site_B.

3.3.3 Identifying and configuring the storage

The storage (DS8800) must have the minimum microcode bundle level 86.31.70.0 (R6.3.SP4) and each storage subsystem must be at the same microcode level. Consider the following information:

- ▶ The HyperSwap operation supports only Metro Mirror (synchronous) replications.
- ▶ The DS8800 Storage Subsystem must have the license for Peer-to-Peer Remote Copy (PPRC) and SCSI in-band command support.
- ▶ The AIX hosts must be set to AIX HyperSwap on the DS8800 storage subsystems.

The storage configuration is described in detail in “Storage systems” on page 60.

3.3.4 Enabling HyperSwap in AIX

The activation of HyperSwap requires a maintenance window because the OS storage drivers' reconfiguration requires systems to be rebooted for activation. The Path Control Module (PCM) part of the device driver is one of the key components in the HyperSwap environment, and it must be enabled. For details about how to enable HyperSwap PCM, see “AIX disk device driver configuration” on page 45.

The (replicated) disk pairs that will be enabled for HyperSwap also need to be configured as described in 3.3.4, “Enabling HyperSwap in AIX” on page 75.

3.3.5 Reconfiguring the cluster for HyperSwap

We describe how to migrate cluster shared storage to HyperSwap. We cover the Cluster Repository and user data disks.

Migrating the Cluster Repository to the HyperSwap disk

The existing PowerHA SystemMirror has the cluster repository based on a standard shared disk (non-HyperSwap enabled), as shown in Example 3-79.

Example 3-79 CAA Repository disk based on non-HyperSwap enabled disk

```
root@ps3n01base: /> lsccluster -d
Storage Interface Query

Cluster Name: ps3n01base_cluster
Cluster UUID: 9b2d80ec-3f7e-11e2-b210-2a5c2f246f0a
Number of nodes reporting = 4
Number of nodes expected = 4

Node ps3n01base
Node UUID = 9b3414a2-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
  hdisk3:
    State : UP
    uDid : 200B75XP411A80207210790003IBMfcp
    uUid : b78c2d40-f6ac-0713-e34c-a0beaa69150d
    Site uUid : 51735173-5173-5173-5173-517351735173
    Type : REPDISK

Node ps3n02base
Node UUID = 9b3970dc-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
  hdisk3:
    State : UP
    uDid : 200B75XP411A80207210790003IBMfcp
    uUid : b78c2d40-f6ac-0713-e34c-a0beaa69150d
    Site uUid : 51735173-5173-5173-5173-517351735173
    Type : REPDISK

Node ss3n04base
Node UUID = 9b399166-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
  hdisk3:
    State : UP
    uDid : 200B75XP411A80207210790003IBMfcp
    uUid : b78c2d40-f6ac-0713-e34c-a0beaa69150d
    Site uUid : 51735173-5173-5173-5173-517351735173
    Type : REPDISK

Node ss3n03base
Node UUID = 9b39839c-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
  hdisk3:
    State : UP
    uDid : 200B75XP411A80207210790003IBMfcp
```

```

uId : b78c2d40-f6ac-0713-e34c-a0beaa69150d
Site uId : 51735173-5173-5173-5173-517351735173
Type : REPDISK

```

Unique identifier: It is not possible to transform the repository disk into a HyperSwap device because of the IEEE unique ID (UUID) (<http://www.ietf.org/rfc/rfc4122.txt>) of the CAA repository disk. Instead, a new HyperSwap enabled disk must be added to all nodes in the cluster, and the CAA repository disk must be migrated from the non-HyperSwap disk (PowerHA provides the tool for this action).

To migrate the existing (non-HyperSwap) CAA repository disk (hdisk3) to HyperSwap, we must configure another disk to be HyperSwap enabled by using the standard procedure shown in Example 3-37 on page 49.

PowerHA SystemMirror provides the function to replace the CAA repository disk with another disk. In Example 3-80, we replace the existing CAA disk hdisk3 with the HyperSwap enabled disk hdisk9. We use the following SMIT menu options:

smit hacmp → Problem Determination Tools → Replace the Primary Repository Disk

Alternately, you can use the following SMIT fast path:

smit cl_replace_repository_nm

Important: The PowerHA SystemMirror services must be stopped on all nodes before migrating (replacing) the CAA repository disk.

Example 3-80 Replacing the CAA disk

Select a new Cluster repository disk

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

```

                                     [Entry Fields]
* Cluster Name                       ps3n01base_cluster
* Repository Disk                     [hdisk9]
+

F1=Help           F2=Refresh       F3=Cancel         F4=List
Esc+5=Reset      F6=Command       F7=Edit          F8=Image
F9=Shell         F10=Exit         Enter=Do

```

The hdisk9 attributes are shown in Example 3-81.

Example 3-81 CAA HyperSwap enabled disk attributes

```

root@ps3n01base: /> lsattr -El hdisk9
PCM                PCM/friend/aixmpiots8k      Path Control Module          False
PR_key_value       none                        Persistent Reserve Key Value True
algorithm          fail_over                  Algorithm                     True
clr_q              no                          Device CLEARS its Queue on error True
dist_err_pcmt      0                          Distributed Error Percentage  True
dist_tw_width      50                         Distributed Error Sample Time True
hcheck_cmd         test_unit_rdy              Health Check Command          True
hcheck_interval    60                         Health Check Interval         True
hcheck_mode        nonactive                   Health Check Mode             True

```

location		Location Label	True
lun_id	0x40b1400000000000	Logical Unit Number ID	False
lun_reset_spt	yes	LUN Reset Supported	True
max_coalesce	0x40000	Maximum Coalesce Size	True
max_retry_delay	60	Maximum Quiesce Time	True
max_transfer	0x80000	Maximum TRANSFER Size	True
node_name	0x500507630bffc1e2	FC Node Name	False
pvid	00f681f3697a827f0000000000000000	Physical volume identifier	False
q_err	yes	Use QERR bit	True
q_type	simple	Queueing TYPE	True
queue_depth	20	Queue DEPTH	True
reassign_to	120	REASSIGN time out value	True
reserve_policy	no_reserve	Reserve Policy	True
rw_timeout	30	READ/WRITE time out value	True
san_rep_cfg	migrate_disk	SAN Replication Device Configuration Policy	True+
san_rep_device	yes	SAN Replication Device	False
scsi_id	0xa0500	SCSI ID	False
start_timeout	60	START unit time out value	True
timeout_policy	fail_path	Timeout Policy	True
unique_id	352037355850343131423130300050be162a07210790003IBMfcp	Unique device identifier	False
ww_name	0x500507630b1001e2	FC World Wide Name	False

After the operation is complete, we verify the status of the CAA cluster by observing that hdisk9 is used as the repository disk, as shown in Example 3-82.

Example 3-82 New (migrated) CAA repository

```

root@ps3n01base: /> lscluster -d
Storage Interface Query

Cluster Name: ps3n01base_cluster
Cluster UUID: 9b2d80ec-3f7e-11e2-b210-2a5c2f246f0a
Number of nodes reporting = 4
Number of nodes expected = 4

Node ps3n01base
Node UUID = 9b3414a2-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
    hdisk9:
        State : UP
        uDid : 352037355850343131423130300050be162a07210790003IBMfcp
        uUID : 64652ce3-3fe9-d40b-72b6-07bd285a47d4
        Site uUID : 51735173-5173-5173-5173-517351735173
        Type : REPDISK

Node ps3n02base
Node UUID = 9b3970dc-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
    hdisk9:
        State : UP
        uDid : 352037355850343131423130300050be162a07210790003IBMfcp
        uUID : 64652ce3-3fe9-d40b-72b6-07bd285a47d4
        Site uUID : 51735173-5173-5173-5173-517351735173
        Type : REPDISK

Node ss3n04base
Node UUID = 9b399166-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
    hdisk9:

```

```

State : UP
uDid : 352037355850343131423130300050be162a07210790003IBMfcp
uUId : 64652ce3-3fe9-d40b-72b6-07bd285a47d4
Site uUId : 51735173-5173-5173-5173-517351735173
Type : REPDISK

```

```

Node ss3n03base
Node UUID = 9b39839c-3f7e-11e2-b210-2a5c2f246f0a
Number of disks discovered = 1
  hdisk9:
    State : UP
    uDid : 352037355850343131423130300050be162a07210790003IBMfcp
    uUId : 64652ce3-3fe9-d40b-72b6-07bd285a47d4
    Site uUId : 51735173-5173-5173-5173-517351735173
    Type : REPDISK

```

Tip: The procedure for the CAA repository disk replacement can be used to replace the CAA repository disk either with HyperSwap or a non-HyperSwap enabled disk (backing out).

The next configuration step is to define the Cluster Repository mirror group (MG), as shown in Example 3-83. We use the following SMIT fast path:

```
smit cm_cfg_mirr_gps
```

We use the Add cluster Repository Mirror Group window.

Example 3-83 Configuring the Cluster Repository MG

Add cluster Repository Mirror Group

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

	[Entry Fields]	
* Mirror Group Name	[CAA_MG]	
* Site Name	Site_A Site_B	+
* Non Hyperswap Disk	hdisk9:64652ce3-3fe9-d40b-72b6-07bd>	+
* Hyperswap Disk	hdisk9:64652ce3-3fe9-d40b-72b6-07bd>	+
Hyperswap	Enabled	+
Consistency Group	Enabled	+
Unplanned HyperSwap Timeout (in sec)	[60]	#
Hyperswap Priority	High	+
F1=Help	F2=Refresh	F3=Cancel
Esc+5=Reset	F6=Command	F7=Edit
F9=Shell	F10=Exit	Enter=Do
		F4=List
		F8=Image

Cluster verification: *You need to verify and synchronize the cluster configuration with the PowerHA services stopped on all nodes.*

Configuring existing user disks for HyperSwap

After a planned migration to HyperSwap, the existing user (data) disks used in the cluster must be also configured to HyperSwap functionality.

To convert the user disks to the HyperSwap configuration, all disks that are part of the volume groups that are part of the resource groups must already have configured their disk pairs in the secondary storage and also the Metro Mirror PPRC relationship established.

The zoning for every LPAR is configured and activated to access the target disk pairs from both storage subsystems. At the end of this operation, we run the `cfgmgr` command on all required nodes. Follow these steps:

1. We identify the corresponding disk pairs on the AIX operating system used for HyperSwap and change the `reserve_policy` attribute of the disk to `no_reserve` for all pairs as shown Example 3-32 on page 47.

We change the `san_rep_cfg` attribute of the disks as shown in Example 3-37 on page 49.

Disk attributes: The `san_rep_cfg` disk attribute is changed to `migrate_disk`. The `chdev` command is used with the `-U` flag only for the disk acting as a source in the PPRC relationship. The secondary (target) disk will change to the Defined state. Do not change the `san_rep_cfg` attribute on the PPRC target disk.

We use disks belonging to RG1 (`hdisk11(datavg)` and `hdisk12(datavg)`), and the disk belonging to RG2 (`hdisk2(oravg)`), as shown in Example 3-84.

Example 3-84 Disks' status before you enable HyperSwap

```

HOSTS -----
ps3n01base, ps3n02base, ss3n03base, ss3n04base
-----
hdisk1      00cf8de691f6b487      None
hdisk2    00cf8de6746ac060      oravg
hdisk3      00f681f3697a8134      None
.....<< Snippet >>.....hdisk8
00f681f3697a8249      None
hdisk9      00f681f3697a827f      caavg_private  active
hdisk10     00cf8de68d114d09      data2vg
hdisk11   00f681f3697a82ed      datavg
hdisk12   00f681f36a720446      datavg
hdisk13     none                    None
.....<< Snippet >>.....

```

The disks in AIX are paired by Metro Mirror (PPRC) in the storage subsystems as shown in Example 3-85. We maintain the same volume ID across the storage subsystems for the LUNs. Storage_A in Site_A (primary site) has the serial number 75XP411A and Storage_B in Site_B (secondary storage) has the serial number 75WT971B.

Example 3-85 Volumes' IDs

```

hdisk0
Serial Number.....2145
hdisk1
Serial Number.....75XP411A
Device Specific.(Z7).....A800
hdisk2
Serial Number.....75XP411A
Device Specific.(Z7).....A801      oravg ( Storage_A )
.....<<Snippet>>.....
hdisk11
Serial Number.....75XP411A

```

```

Device Specific.(Z7).....B201                datavg ( Storage_A )
hdisk12
Serial Number.....75XP411A
Device Specific.(Z7).....B301                datavg ( Storage_A)
hdisk13
Serial Number.....75WT971A
Device Specific.(Z7).....A800
hdisk14
Serial Number.....75WT971B
Device Specific.(Z7).....A801                oravg ( Storage_B)
.....<<Snippet>>.....
hdisk21
Serial Number.....75WT971B
Device Specific.(Z7).....B201                datavg ( Storage_B)
hdisk22
Serial Number.....75WT971B
Device Specific.(Z7).....B301                datavg ( Storage_B)

```

The PPRC relationships are shown in Example 3-86.

Example 3-86 PPRC relationships

```

dscli> lssi
Name ID                Storage Unit      Model WNN          State  ESSNet
=====
- IBM.2107-75XP411 IBM.2107-75XP410 951  500507630BFFC4C8 Online Enabled
dscli>

dscli> lsprrc -l a800
ID      State      Reason Type      Out Of Sync Tracks Tgt Read Src Cascade Tgt Cascade Date Suspended SourceLSS T
imeout (secs) Critical Mode First Pass Status Incremental Resync Tgt Write GMIR CG PPRC CG isTgtSE DisableAutoResync
=====
A800:A800 Full Duplex - Metro Mirror 0 Disabled Disabled Invalid - A8 5
          Disabled Invalid Disabled Disabled N/A Enabled Unknown -
dscli> lsprrc -l b200-b3ff
ID      State      Reason Type      Out Of Sync Tracks Tgt Read Src Cascade Tgt Cascade Date Suspended SourceLSS T
imeout (secs) Critical Mode First Pass Status Incremental Resync Tgt Write GMIR CG PPRC CG isTgtSE DisableAutoResync
=====
B201:B201 Full Duplex - Metro Mirror 0 Disabled Disabled Invalid - B2 5
          Disabled Invalid Disabled Disabled N/A Enabled Unknown -
B301:B301 Full Duplex - Metro Mirror 0 Disabled Disabled Invalid - B3 5
          Disabled Invalid Disabled Disabled N/A Enabled Unknown -

```

2. We continue the disk reconfiguration by using the **chdev** command as shown in Example 3-87. We change the SCSI reservation policy for hdisk2, hdisk11, hdisk12, hdisk14, hdisk21, and hdisk22 on all nodes.

Example 3-87 Changing the disk attributes

```

for i in 2 11 12 14 21 22 ; do chdev -l hdisk$i -a reserve_policy=no_reserve;done
ps3n01base: hdisk2 changed
ps3n01base: hdisk11 changed
ps3n01base: hdisk12 changed
ps3n01base: hdisk14 changed
ps3n01base: hdisk21 changed
ps3n01base: hdisk22 changed

```

3. We change the disk attribute `san_rep_cfg` to enable HyperSwap only on the disks that are in the primary storage (Storage_A) on all nodes as shown in Example 3-88.

Example 3-88 Activating HyperSwap on disks

```
for i in 2 11 12 ; do chdev -l hdisk$i -a san_rep_cfg=migrate_disk;done
ps3n01base: hdisk2 changed
ps3n01base: hdisk11 changed
ps3n01base: hdisk12 changed
```

After changing the `san_rep_cfg` attribute, the secondary disk for each pair will change to the Defined state, as shown in Example 3-89,

Example 3-89 Hdisks form the secondary storage status

```
root@ps3n01base:/bubu> lsdev -Cc disk
hdisk0 Available 27-T1-01 MPIO IBM 2145 FC Disk
.....<< Snippet >>.....
hdisk13 Available 45-T1-01 MPIO IBM 2107 FC Disk
hdisk14 Defined 44-T1-01 MPIO IBM 2107 FC Disk
hdisk15 Available 45-T1-01 MPIO IBM 2107 FC Disk
.....<< Snippet >>.....
hdisk20 Available 45-T1-01 MPIO IBM 2107 FC Disk
hdisk21 Defined 44-T1-01 MPIO IBM 2107 FC Disk
hdisk22 Defined 44-T1-01 MPIO IBM 2107 FC Disk
```

Because all resource groups are already defined in the PowerHA SystemMirror cluster, now we have all the disks activated for HyperSwap.

4. We verify the PPRC relationships for disks by using the `lspprc` command as shown in Example 3-90.

Example 3-90 Status of the pprc relationships as shown by the lspprc command

```
root@ps3n01base: /> lspprc -Ao |sort -k1n6,7
      ID      ID
state path group path group WNN WNN
PPRC Primary Secondary Primary Storage Secondary Storage
hdisk# PPRC
hdisk1 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk2 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk3 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk4 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk5 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk6 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk7 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk8 Active 0(s) -1 500507630bffc4c8 500507630bffc1e2
hdisk9 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk10 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk11 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk12 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk13 Active -1 0 500507630bffc4c8 500507630bffc1e2
hdisk15 Active -1 0 500507630bffc4c8 500507630bffc1e2
hdisk16 Active -1 0 500507630bffc4c8 500507630bffc1e2
hdisk17 Active -1 0 500507630bffc4c8 500507630bffc1e2
hdisk18 Active -1 0 500507630bffc4c8 500507630bffc1e2
hdisk19 Active -1 0 500507630bffc4c8 500507630bffc1e2
hdisk20 Active -1 0 500507630bffc4c8 500507630bffc1e2
```

5. We configure a mirror group for every volume group that is part of a resource group and add the corresponding mirror group at its related resource group as shown in “Mirror groups” on page 61.
6. We verify and synchronize the cluster configuration. This operation requires the PowerHA services to be down on all nodes.

3.4 Two-node cluster to four-node cluster with HyperSwap

We illustrate how to extend an existing two-node PowerHA single-site cluster to a four-node (two nodes on each site) PowerHA EE cluster using HyperSwap to achieve the enhanced storage availability.

The initial configuration is a two-node (ps2n01 and ps2n02) configuration with access to the following LUNs: A600, A601, A700, and A701. The PowerHA installed is 7.1.2. We extend this configuration to a four-node cluster (ps2n01 and ps2n02 in Site_A and ss2n03 and ss2n04 in Site_B (Figure 3-3 on page 84) with PPRC pairs for A600, A601, A700, and A701 enabled for HyperSwap.

License information: To implement this configuration, you need to upgrade the PowerHA license from PowerHA Standard Edition to PowerHA Enterprise Edition.

Table 3-1 Base IP addresses and IP labels

Node	en0		en1	
	IP address	IP label	IP address	IP label
ps2n01	172.16.29.247	ps2n01base	172.16.14.67	ps2n01en1
ps2n02	172.16.29.248	ps2n02base	172.16.14.68	ps2n02en1
ss2n03	172.16.29.90	ss2n03base	172.16.14.69	ss2n03en1
ss2n04	172.16.29.91	ss2n04base	172.16.14.70	ss2n04en1

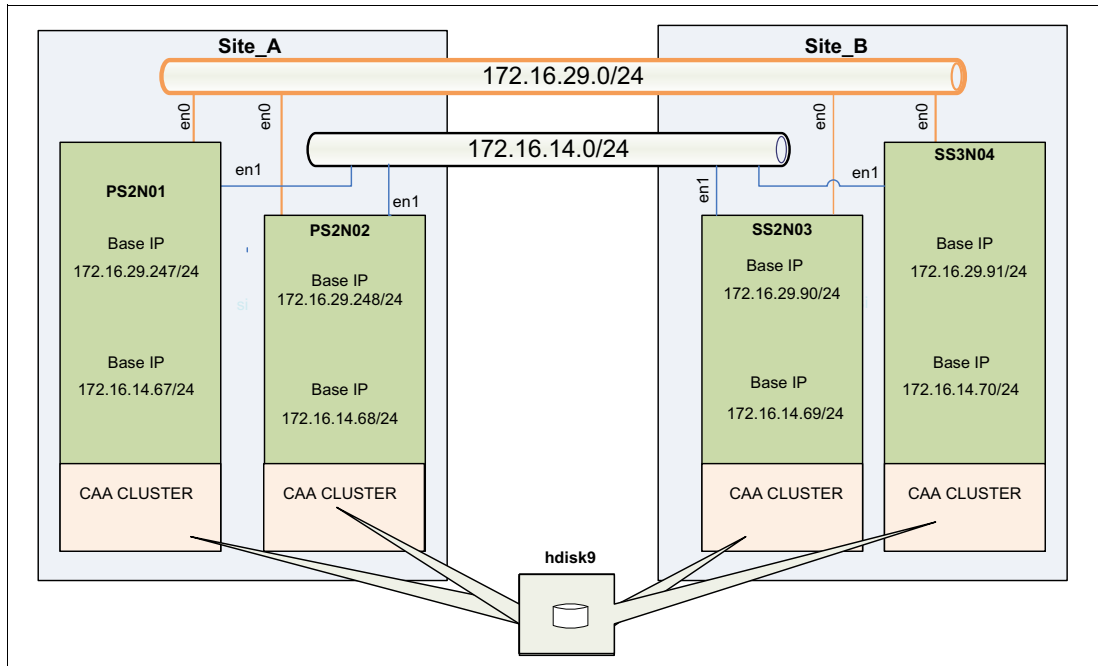


Figure 3-3 Target cluster diagram (logical)

Planning information:

- ▶ We assume that the existing cluster is upgraded to PowerHA 7.1.2 SP2 and the AIX level that supports HyperSwap and the existing DS8800 firmware is also updated to support HyperSwap.
- ▶ In this scenario, we assume that the existing cluster is in production so our goal is to minimize the application downtime during the cluster extension.

Implementation steps

Follow these steps:

1. Prepare the new nodes and DS8800 on Site_B:
 - Install the same level of AIX and PowerHA on nodes ss2n03base and ss2n04base (DR site).
 - Create the following LUNs, A600-A601 and A700-A701, on the DR site storage (DS8800-03). (Make them the same size as the LUNs on the primary site.)
2. Modify the SAN zoning and prepare the disks:
 - Check the SAN connectivity between sites and modify the zoning so that all four nodes can access all LUNs on both sites.
 - Make the LUNS known to AIX. (Run the `cfgmgr` command on *all* nodes to detect new LUNs.)
 - Import the shared storage to the nodes in the secondary (DR) site by using the `importvg` command.
3. Extend the PowerHA cluster to two sites (but using the storage on Site_A only):
 - Add nodes ss2n03base and ss2n04base to the PowerHA cluster.
 - Test the RG movement back and forth (optional because this action requires downtime).

- Add sites A and B to the PowerHA cluster.
 - Test the RG movement back and forth (optional because this action requires downtime).
4. Prepare the disk replication
- Create the PPRC paths and create the PPRC pairs (for both DS8800 storage subsystems).
5. Implement PowerHA EE with HyperSwap (*this step requires downtime*):
- a. Install the PowerHA `cluster.genxd.*` packages on all nodes.
 - b. Change the `hostconnect` host type to `pSeriesPowerswap` on *both* DS8800s.
 - c. In AIX, follow these steps:
 - i. Set `manage_disk_drivers` to `AIX_AAPCM`.
 - ii. Install any recommended PTFs and EFIXs for PowerHA.
 - iii. Reboot all nodes.
 - iv. Set the disk attribute `reserve_policy` to `no_reserve`.
 - v. Change the disk attribute `san_rep_cfg` to `migrate_disk`.
 - vi. Reimport the volume user groups.
 - d. For the PowerHA configuration, follow these steps:
 - i. Define the storage resources, mirror groups, add the MGs to the RGs, and synchronize.
 - ii. Start the cluster services.
 - iii. Test the RG movement back and forth.
 - iv. Test the planned storage HyperSwap.

Current cluster configuration

Example 3-91 shows the current cluster topology.

Example 3-91 Cluster topology

```

root@ps2n02base: /> cltopinfo
Cluster Name: g2_c1
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: hdisk2
Cluster IP Address: 228.16.29.89
There are 2 node(s) and 1 network(s) defined
NODE ps2n01base:
    Network net_ether_01
        ps2n02svc      172.16.15.140
        ps2n01base    172.16.29.88
NODE ps2n02base:
    Network net_ether_01
        ps2n02svc      172.16.15.140
        ps2n02base    172.16.29.89

Resource Group g2_rg
Startup Policy Online On First Available Node

```

```

Fallover Policy Fallover To Next Priority Node In The List
Fallback Policy Never Fallback
Participating Nodes ps2n02base ps2n01base
Service IP Label ps2n02svc

```

The cluster resource groups are shown in Example 3-92.

Example 3-92 Cluster RGs

```
root@ps2n02base: /> clRGinfo -v
```

```
Cluster Name: g2_c1
```

```
Resource Group Name: g2_rg
Startup Policy: Online On First Available Node
Fallover Policy: Fallover To Next Priority Node In The List
Fallback Policy: Never Fallback
Site Policy: ignore
```

Node	Group State
ps2n02base	ONLINE
ps2n01base	OFFLINE

Extending the cluster

Follow these steps to extend the cluster.

Step 1. Prepare the new nodes and DS8800 in Site_B

Follow these steps:

1. Install the same level of AIX/PowerHA on ss2n03base and ss2n04base (Site_B).
2. Create LUNs A600-A601 and A700-A701 on DS8800-03 (Storage_B) in the DR site. (Make them the same size as the LUNs in Site_A.)

Step 2. Modify SAN zoning and prepare disks

Follow these steps:

1. Modify the zoning so that the four nodes can access all shared LUNs in both sites.
2. Run the **cfgmgr** command on all nodes to detect the new LUNs.
3. Import the shared volume groups to the new nodes (ss2n03 and ss2n04). See Example 3-93.

Example 3-93 Importing VGs on nodes in Site_B

```

tt@hsp78005n4:/home/tt> ssh root@n23 importvg -V 51 -y g2vg hdisk3
g2vg
0516-783 importvg: This imported volume group is concurrent capable.
Therefore, the volume group must be varied on manually.
tt@hsp78005n4:/home/tt> ssh root@n24 importvg -V 51 -y g2vg hdisk3
g2vg
0516-783 importvg: This imported volume group is concurrent capable.
Therefore, the volume group must be varied on manually.

```

Tip: This step can be done without any downtime because PowerHA 7.1.2 converts any VG defined into a resource group to be a concurrent-capable volume group. Therefore, we can run the `importvg` command, even if, on other nodes, the VG is already varied on.

Step 3. Extend PowerHA to two sites (but using storage on Site_A only)

Follow these steps:

1. Add the `ss2n03base` and `ss2n04base` nodes to the cluster. Use the following SMIT menu options to add a node:

`smitty hacmp` → **Cluster Nodes and Networks** → **Manage Nodes** → **Add a Node**

Tip: After adding a node, run the `cltopinfo` command to see whether PowerHA has recorded the IP address for the node.

If not, use the following SMIT menu options to add the IP address manually:

`smitty hacmp` → **Cluster Nodes and Networks** → **Manage Networks and Network Interfaces** → **Network Interfaces** → **Add a Network Interface**

2. After adding nodes, verify and synchronize the PowerHA cluster configuration and start the new nodes.
3. Modify the resource group to add the new nodes to the resource group, and synchronize the PowerHA cluster configuration again. The volume groups will be varied on in concurrent mode on all nodes, as shown in Example 3-94.

Example 3-94 VG information

Command run on all nodes:

```

-----Running lspv on ps2n01n-----
hdisk0      00f681f3d99b7740      rootvg      active
hdisk1      00f681f3b69ca761      None
hdisk2      00f681f303641e8c      caavg_private active
hdisk3     00f681f303641ee0     g2vg      concurrent
hdisk4      00f681f308a1d561      None
hdisk5      none                    None
hdisk6      none                    None
hdisk7      none                    None
hdisk8      none                    None

-----Running lspv on ps2n02-----
hdisk0      00f681f3d99b7740      rootvg      active
hdisk1      00f681f3b69ca761      None
hdisk2      00f681f303641e8c      caavg_private active
hdisk3     00f681f303641ee0     g2vg      concurrent
hdisk4      00f681f308a1d561      None
hdisk5      none                    None
hdisk6      none                    None
hdisk7      none                    None
hdisk8      none                    None

-----Running lspv on ss2n03-----
hdisk0      00cf8de6e476d1fa      rootvg      active
hdisk1      00f681f3b69ca761      None
hdisk2      00f681f303641e8c      caavg_private active
hdisk3     00f681f303641ee0     g2vg      concurrent
hdisk4      00f681f308a1d561      None

```

```

hdisk5          none          None
hdisk6          none          None
hdisk7          none          None
hdisk8          none          None
-----Running lspv on ss2n04-----
hdisk0          00cf8de6e47cdb4a      rootvg          active
hdisk1          00f681f3b69ca761      None
hdisk2          00f681f303641e8c      caavg_private  active
hdisk3         00f681f303641ee0      g2vg          concurrent
hdisk4          00f681f308a1d561      None
hdisk5          none          None
hdisk6          none          None
hdisk7          none          None
hdisk8          none          None

```

Optional: This task results in a shorter application downtime. Test the RG movement back and forth.

4. Define the sites to the PowerHA cluster: Add Site_A and Site_B. Use the following SMIT menu options to add the sites (Example 3-95):

smitty hacmp → **Cluster Nodes and Networks** → **Manage Sites** → **Add a Site**

Example 3-95 Adding sites to cluster configuration

Add a Site

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

```

* Site Name          [Entry Fields]
* Site Nodes        [Site_A]
Cluster Type        ps2n01base ps2n02base
                   Stretched Cluster

```

```

F1=Help             F2=Refresh         F3=Cancel         F4=List
Esc+5=Reset        F6=Command        F7=Edit           F8=Image
F9=Shell           F10=Exit          Enter=Do

```

5. Repeat the task to add Site_B. Verify and synchronize the cluster configuration.

Optional: This task results in a shorter application downtime. Test the RG movement back and forth.

Step 4. Prepare disk replication

Assuming that the connectivity between the two storage subsystems is already configured, create the PPRC path between Storage_A and Storage_B. Follow these steps:

1. On DS8800-05 (Storage_A in Site_A), run the commands shown in Example 3-96 on page 89.

Example 3-96 PPRC path definition on primary storage

```
dscli> mkpprcpath -remotewwn 500507630BFFC1E2 -srclss A6 -tgtlss A6 -consistgrp
I0102:I0102 I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path B1:B1 successfully established.
dscli> mkpprcpath -remotewwn 500507630BFFC1E2 -srclss A7 -tgtlss A7 -consistgrp
I0102:I0102 I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path B2:B2 successfully established.
```

2. On DS8800-03 (Storage_B in Site_B), run the commands shown in Example 3-97.

Example 3-97 PPRC path definition on secondary storage

```
dscli> mkpprcpath -remotewwn 500507630BFFC4C8 -srclss A6 -tgtlss A6 -consistgrp
I0102:I0102 I0132:I0202
CMUC00149I mkpprcpath: Remote Mirror and Copy path B1:B1 successfully established.
dscli> mkpprcpath -remotewwn 500507630BFFC4C8 -srclss A7 -tgtlss A7 -consistgrp
I0102:I0102 I0132:I0202
CMUC00149I mkpprcpath: Remote Mirror and Copy path B2:B2 successfully established.
```

3. Verify the PPRC path on both storage subsystems, as shown in Example 3-98.

Example 3-98 Verifying PPRC path definition

On Storage_A:

```
dscli> lsprrcpath a6
Src Tgt State  SS  Port  Attached Port Tgt WNN
=====
A6 A6  Success FFA6 I0102 I0102          500507630BFFC1E2
A6 A6  Success FFA6 I0202 I0132          500507630BFFC1E2
dscli> lsprrcpath a7
Src Tgt State  SS  Port  Attached Port Tgt WNN
=====
A7 A7  Success FFA7 I0102 I0102          500507630BFFC1E2
A7 A7  Success FFA7 I0202 I0132          500507630BFFC1E2
```

On Storage_B:

```
dscli> lsprrcpath a6
Src Tgt State  SS  Port  Attached Port Tgt WNN
=====
A6 A6  Success FFA6 I0102 I0102          500507630BFFC4C8
A6 A6  Success FFA6 I0132 I0202          500507630BFFC4C8
dscli> lsprrcpath a7
Src Tgt State  SS  Port  Attached Port Tgt WNN
=====
A7 A7  Success FFA7 I0102 I0102          500507630BFFC4C8
A7 A7  Success FFA7 I0132 I0202          500507630BFFC4C8
```

4. Create the PPRC pairs on Storage_A and verify, as shown in Example 3-99.

Example 3-99 PPRC pairs on Storage_A

Create:

```
dscli> mkpprc -type mmir -tgtse a600:a600
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship A600:A600 successfully created.
dscli> mkpprc -type mmir -tgtse a601:a601
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship A601:A601 successfully created.
dscli> mkpprc -type mmir -tgtse a700:a700
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship A700:A700 successfully created.
```

```
dscli> mkpprc -type mmir -tgtse a701:a701
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship A701:A701 successfully created.
```

Verify:

```
dscli> lsprrc a600-a7ff
ID          State      Reason Type          SourceLSS Timeout (secs) Critical Mode First Pass Status
=====
A600:A600 Full Duplex - Metro Mirror A6      5          Disabled          Invalid
A601:A601 Full Duplex - Metro Mirror A6      5          Disabled          Invalid
A700:A700 Full Duplex - Metro Mirror A7      5          Disabled          Invalid
A701:A701 Full Duplex - Metro Mirror A7      5          Disabled          Invalid
dscli>
```

5. Verify the PPRC pairs on Storage_B, as shown in Example 3-100.

Example 3-100 PPRC pairs on Storage_B

```
dscli> lsprrc a600-a7ff
ID          State      Reason Type          SourceLSS Timeout (secs) Critical Mode First Pass Status
=====
A600:A600 Target Full Duplex - Metro Mirror A6      unknown      Disabled          Invalid
A601:A601 Target Full Duplex - Metro Mirror A6      unknown      Disabled          Invalid
A700:A700 Target Full Duplex - Metro Mirror A7      unknown      Disabled          Invalid
A701:A701 Target Full Duplex - Metro Mirror A7      unknown      Disabled          Invalid
dscli>
```

Step 5. Implement PowerHA for AIX Enterprise Edition (EE) with HyperSwap

Downtime required: This step requires downtime.

Follow these steps:

1. Stop the PowerHA cluster and install the `cluster.genxd.*` package on all nodes, as shown in Example 3-101.

Example 3-101 Cluster genxd packages installed

```
.....<< Snippet >>.....
SUCSESSES
-----
Filesets listed in this section passed pre-installation verification
and will be installed.

Selected Filesets
-----
cluster.es.genxd.cmds 7.1.2.0          # PowerHA SystemMirror Enterpr...
cluster.es.genxd.cmds 7.1.2.1          # PowerHA SystemMirror Enterpr...
cluster.es.genxd.rte 7.1.2.0          # PowerHA SystemMirror Enterpr...
cluster.es.genxd.rte 7.1.2.1          # PowerHA SystemMirror Enterpr...

Requisites
-----
(being installed automatically; required by filesets listed above)
cluster.xd.license 7.1.2.0            # PowerHA SystemMirror Enterpr...

<< End of Success Section >>
```


2. Collect the following AIX disk information before making any more changes and save it for reference:

- Physical volume identifier (PVID)
- Unique Device Identifier (UDID)
- Universally Unique Identifier (UUID)

In Example 3-102, we show ps2n02 only.

Example 3-102 AIX information

```

root@ps2n02base: /> lspv -u
hdisk0          00f681f3d99b7740          rootvg          active
33213600507680185057370000000000015904214503IBMfcp
79e8fa61-f988-4b25-4953-19799a33ec4d
hdisk1          00f681f3b69ca761          None
200B75XP411A60007210790003IBMfcp
5c706998-3fe4-bc9d-b869-4915e524882f
hdisk2          00f681f303641e8c          caavg_private  active
200B75XP411A60107210790003IBMfcp
6aeb8cf4-85cb-67c4-18ff-4edfbe807783
hdisk3          00f681f303641ee0          g2vg           concurrent
200B75XP411A70007210790003IBMfcp
8af707e9-b531-2864-46db-8ea730df2b09
hdisk4          00f681f308a1d561          None
200B75XP411A70107210790003IBMfcp
f351e848-67eb-4f71-f691-45842cfe3cc1
hdisk5          none                          None
200B75WT971A60007210790003IBMfcp
f9c1a08e-444d-40fd-8b34-41116db10b6b
hdisk6          none                          None
200B75WT971A60107210790003IBMfcp
d0d3865f-1492-06cf-e793-8e505154f272
hdisk7          none                          None
200B75WT971A70007210790003IBMfcp
194e171f-7c5e-237a-56c4-7f21e2e9b497
hdisk8          none                          None
200B75WT971A70107210790003IBMfcp
d1803584-520a-2803-8ef6-72f00411ff64

```

3. On both storage subsystems, change the **hostconnect** host type to pSeriesPowerswap. Example 3-103 shows the commands used to change the **hostconnect** information.

Example 3-103 Changing host type

Storage_A:

```

dscli> chostconnect -hosttype pseriespowerswap 1b
CMUC00013I chostconnect: Host connection 001B successfully modified.
dscli> chostconnect -hosttype pseriespowerswap 1c
CMUC00013I chostconnect: Host connection 001C successfully modified.
dscli> chostconnect -hosttype pseriespowerswap 1d
CMUC00013I chostconnect: Host connection 001D successfully modified.
dscli> chostconnect -hosttype pseriespowerswap 1e
CMUC00013I chostconnect: Host connection 001E successfully modified.
dscli> chostconnect -hosttype pseriespowerswap 2f
CMUC00013I chostconnect: Host connection 002F successfully modified.
dscli> chostconnect -hosttype pseriespowerswap 30
CMUC00013I chostconnect: Host connection 0030 successfully modified.

```

```

dscli> chhostconnect -hosttype pseriespowerswap 31
CMUC00013I chhostconnect: Host connection 0031 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 32
CMUC00013I chhostconnect: Host connection 0032 successfully modified.

```

Storage_B:

```

dscli> chhostconnect -hosttype pseriespowerswap 22
CMUC00013I chhostconnect: Host connection 0022 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 23
CMUC00013I chhostconnect: Host connection 0023 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 24
CMUC00013I chhostconnect: Host connection 0024 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 25
CMUC00013I chhostconnect: Host connection 0025 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 36
CMUC00013I chhostconnect: Host connection 0036 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 37
CMUC00013I chhostconnect: Host connection 0037 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 38
CMUC00013I chhostconnect: Host connection 0038 successfully modified.
dscli> chhostconnect -hosttype pseriespowerswap 39
CMUC00013I chhostconnect: Host connection 0039 successfully modified.

```

4. Example 3-104 shows the **hostconnect** host type changed to AIX HyperSwap.

Example 3-104 Host type changed to HyperSwap on both storage subsystems

Storage_A:

```

dscli> lshostconnect -volgrp v6
Name          ID    WWPN          HostType          Profile portgrp volgrpID ESSIOport
=====
G3_P7805LP5_fcs4 001B C0507603D4B90062 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7805LP5_fcs1 001C C0507603D4B9002C pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7805LP6_fcs4 001D C0507603D4B90066 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7805LP6_fcs1 001E C0507603D4B90032 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7703LP5_fcs4 002F C050760502C100AA pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7703LP5_fcs1 0030 C050760502C1007C pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7703LP6_fcs4 0031 C050760502C100AE pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
G3_P7703LP6_fcs1 0032 C050760502C10082 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V6      all
dscli>

```

Storage_B:

```

dscli> lshostconnect -volgrp v15
Name          ID    WWPN          HostType          Profile
portgrp volgrpID ESSIOport
=====
G3_P7805LP5 0022 C0507603D4B9002E pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7805LP5 0023 C0507603D4B90030 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7805LP6 0024 C0507603D4B90034 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7805LP6 0025 C0507603D4B90036 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7703LP5 0036 C050760502C1007E pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7703LP5 0037 C050760502C10080 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7703LP6 0038 C050760502C10084 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
G3_P7703LP6 0039 C050760502C10086 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V15     all
dscli>

```

5. Next, change the AIX Path Control Module (PCM) to AIX_AAPCM by using the **manage_disk_drivers** command on *all* nodes, as shown in Example 3-105 on page 93.

Example 3-105 Changing PCM

Command to run on ALL nodes: 'manage_disk_drivers -d 2107DS8K -o AIX_AAPCM'

```
-----Running manage_disk_drivers -d 2107DS8K -o AIX_AAPCM on ps2n01base-----
***** ATTENTION *****
For the change to take effect the system must be rebooted
-----Running manage_disk_drivers -d 2107DS8K -o AIX_AAPCM on ps2n02base-----
***** ATTENTION *****
For the change to take effect the system must be rebooted
-----Running manage_disk_drivers -d 2107DS8K -o AIX_AAPCM on ss2n03base-----
***** ATTENTION *****
For the change to take effect the system must be rebooted
-----Running manage_disk_drivers -d 2107DS8K -o AIX_AAPCM on ss2n04base-----
***** ATTENTION *****
For the change to take effect the system must be rebooted
```

6. Verify the changes as shown in Example 3-106.

Example 3-106 Verifying PCM

Command to execute on all nodes: 'manage_disk_drivers -l|grep DS8K'

```
-----Running manage_disk_drivers -l|grep DS8K on ps2n01base-----
2107DS8K      AIX_AAPCM      NO_OVERRIDE,AIX_AAPCM,NO_OVERRIDE
-----Running manage_disk_drivers -l|grep DS8K on ps2n02base-----
2107DS8K      AIX_AAPCM      NO_OVERRIDE,AIX_AAPCM,NO_OVERRIDE
-----Running manage_disk_drivers -l|grep DS8K on ss2n03base-----
2107DS8K      AIX_AAPCM      NO_OVERRIDE,AIX_AAPCM,NO_OVERRIDE
-----Running manage_disk_drivers -l|grep DS8K on ss2n04base-----
2107DS8K      AIX_AAPCM      NO_OVERRIDE,AIX_AAPCM,NO_OVERRID
```

SDDPCM tip: We discovered that even after setting the current PCM to AIX_AAPCM, if DS8000 SDDPCM drivers are installed, issues might occur after rebooting the node (disks might not be recognized during the **cfgmgr** phases).

Therefore, we suggest to uninstall SDDPCM from all nodes before rebooting.

7. Install any suggested PowerHA fixes (PTF/EFIX). At the time that we tested this configuration, there were two efixes that had to be applied.
8. Reboot all nodes to activate the AIX_AAPCM.
9. After the reboot, the disk attribute **san_rep_device** is changed to *detected*. This means that AIX detected that the LUN is replicated but not yet configured for HyperSwap in-band commands.

The **lspprc** command output is shown in Example 3-107.

Example 3-107 PPRC information in AIX

```
-----Running lspprc -Ao on ss2n04base-----
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk1	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk2	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk5	Active	-1	0	500507630bffc4c8	500507630bffc1e2

hdisk6	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk7	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk8	Active	-1	0	500507630bffc4c8	500507630bffc1e2

Tip: The `lsprrc` command does not show all PPRC pairs. It only shows PPRC pairs that are HyperSwap enabled. There are three possible values for the path group ID:

- -1 represents no path group (HyperSwap is not enabled yet.)
- 0 represents the primary path group
- 1 represents the secondary path group

The (s) after the path group ID represents the current selected path group.

10. Change the AIX disk attribute `reserve_policy` to `no_reserve` on all shared disks and nodes in the cluster:

```
chdev -l <hdisk#> -a reserve_policy=no_reserve
```

If you cannot modify the attribute on some of the disks because it is still in use, you can specify the `-P` flag (to change only ODM) and reboot the nodes later.

11. To enable HyperSwap, change the disk attribute `san_rep_cfg` to `migrate_disk`.

We change the attribute for `hdisk1` (planned for the repository mirror group) and `hdisk3` (data volume group `g2vg`) as shown in Example 3-108.

Tip: HyperSwap for the current CAA repository `hdisk2` (non-HyperSwap) must not be activated because we will migrate the CAA repository to a new (HyperSwap enabled) disk later.

When you enable HyperSwap, the disk UDID and UUID change to reflect the new “composite” device, which makes the repository disk unusable for CAA.

Example 3-108 Enabling HyperSwap for hdisk1

Command to run on all nodes: `chdev -a san_rep_cfg=migrate_disk -l hdisk1 -U`

```

-----Running chdev -a san_rep_cfg=migrate_disk -l hdisk1 -U on
ps2n01base-----
hdisk1 changed
-----Running chdev -a san_rep_cfg=migrate_disk -l hdisk1 -U on
ps2n02base-----
hdisk1 changed
-----Running chdev -a san_rep_cfg=migrate_disk -l hdisk1 -U on
ss2n03base-----
hdisk1 changed
-----Running chdev -a san_rep_cfg=migrate_disk -l hdisk1 -U on
ss2n04base-----
hdisk1 changed

```

Example 3-109 shows that after enabling HyperSwap, `hdisk5` changes to the Defined state (only `ps2n01` nodes shown here).

Example 3-109 HyperSwap enabled

```

-----Running lsdev -Cc disk on ps2n01base-----
hdisk0 Available C2-T1-01 MPIIO IBM 2145 FC Disk
hdisk1 Available 38-T1-01 MPIIO IBM 2107 FC Disk
hdisk2 Available 36-T1-01 MPIIO IBM 2107 FC Disk

```

```

hdisk3 Available 36-T1-01 MPIO IBM 2107 FC Disk
hdisk4 Available 36-T1-01 MPIO IBM 2107 FC Disk
hdisk5 Defined 38-T1-01 MPIO IBM 2107 FC Disk
hdisk6 Available 38-T1-01 MPIO IBM 2107 FC Disk
hdisk7 Available 38-T1-01 MPIO IBM 2107 FC Disk
hdisk8 Available 38-T1-01 MPIO IBM 2107 FC Disk

```

UDID and
UUID change

Also, the UDID and UUID of hdisk1 changed (check Example 3-102 on page 91), as shown in Example 3-110.

Example 3-110 New UDID and UUID for HyperSwap enabled disk

```

-----Running lspv -u|grep k1 on ss2n04base-----
hdisk1          00f681f3b69ca761          None
352037355850343131413630300050d7d40307210790003IBMfcp
aeff57ba-362e-24a9-56a9-8daf2bd9a73c

```

Checking the PPRC status in AIX (by using the `lsprrc` command) reveals that the secondary path group ID of the migrated LUN changed from -1 to 1, as shown in Example 3-111 (ps2n01 node only).

Example 3-111 PPRC information in AIX

```

-----Running lsprrc -Ao on ps2n01base-----
hdisk#   PPRC   Primary   Secondary   Primary Storage   Secondary Storage
         state  path group path group   WWNN              WWNN
hdisk1   Active  0(s)     1           500507630bffc4c8  500507630bffc1e2
hdisk6   Active  -1       0           500507630bffc4c8  500507630bffc1e2
hdisk7   Active  -1       0           500507630bffc4c8  500507630bffc1e2
hdisk2   Active  0(s)    -1           500507630bffc4c8  500507630bffc1e2
hdisk8   Active  -1       0           500507630bffc4c8  500507630bffc1e2
hdisk4   Active  0(s)    -1           500507630bffc4c8  500507630bffc1e2
hdisk3   Active  0(s)    -1           500507630bffc4c8  500507630bffc1e2

```

Also, observe that hdisk5 is not shown in the `lsprrc` command output any longer (because it changed to the Defined state after enabling HyperSwap).

12. Export and reimport the VGs on all nodes, as shown in Example 3-112.

Example 3-112 Export and import VGs

Export VGs

Command to be executed on all nodes: exportvg g2vg

```

-----Running exportvg g2vg on ps2n01base-----
-----Running exportvg g2vg on ps2n02base-----
-----Running exportvg g2vg on ss2n03base-----
-----Running exportvg g2vg on ss2n04base-----

```

Import VGs

Command to be executed on all nodes: importvg -V 51 -y g2vg hdisk3

```

-----Running importvg -V 51 -y g2vg hdisk3 on ps2n01base-----
g2vg
0516-783 importvg: This imported volume group is concurrent capable.
Therefore, the volume group must be varied on manually.
-----Running importvg -V 51 -y g2vg hdisk3 on ps2n02base-----
g2vg
0516-783 importvg: This imported volume group is concurrent capable.

```

Therefore, the volume group must be varied on manually.

-----Running importvg -V 51 -y g2vg hdisk3 on ss2n03base-----

g2vg

0516-783 importvg: This imported volume group is concurrent capable.

Therefore, the volume group must be varied on manually.

-----Running importvg -V 51 -y g2vg hdisk3 on ss2n04base-----

g2vg

0516-783 importvg: This imported volume group is concurrent capable.

Therefore, the volume group must be varied on manually.

13.Reboot all nodes to activate changes that needed to be specified with the -P flag (if any).

14.Define the DS8800 storage resources on **both sites** in the PowerHA cluster. We used the following SMIT menu options (see Example 3-113):

smitty hacmp → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Storage Systems → Add a Storage System

Example 3-113 Defining storage subsystems

Add a Storage System

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

	[Entry Fields]	
* Storage System Name	[Storage_siteA]	
* Site Association	Site_A	+
* Vendor Specific Identifier	IBM.2107-00000XP411	+
* WWNN	500507630BFFC4C8	+

F1=Help	F2=Refresh	F3=Cancel	F4=List
Esc+5=Reset	F6=Command	F7=Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Add a Storage System

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

	[Entry Fields]	
* Storage System Name	[Storage_siteB]	
* Site Association	Site_B	+
* Vendor Specific Identifier	IBM.2107-00000WT971	+
* WWNN	500507630BFFC1E2	+

F1=Help	F2=Refresh	F3=Cancel	F4=List
Esc+5=Reset	F6=Command	F7=Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Tip: For Site Association, Vendor Specific Identifier, and WWNN, you can press F4 to choose from the list.

15. Define the user mirror group by using the following SMIT menu options (see Example 3-114):

smitty hacmp → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Mirror Groups → Add a Mirror Group

Example 3-114 Defining the user MG

Add a User Mirror Group

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

			[Entry Fields]	
* Mirror Group Name			[g2_mg]	
Volume Group(s)			g2vg	+
Raw Disk(s)				+
Hyperswap			Enabled	+
Consistency Group			Enabled	+
Unplanned HyperSwap Timeout (in sec)			[60]	#
Hyperswap Priority			Medium	+
Recovery Action			Manual	+
F1=Help	F2=Refresh	F3=Cancel	F4=List	
Esc+5=Reset	F6=Command	F7=Edit	F8=Image	
F9=Shell	F10=Exit	Enter=Do		

16. Add the previously defined user MG to the existing resource group by using the following SMIT menu options (see Example 3-115):

smitty hacmp → Cluster Applications and Resources → Resource Groups → Change/Show Resources and Attributes for a Resource Group

Example 3-115 Adding an MG to an RG

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.

[MORE...26]			[Entry Fields]	
Raw Disk PVIDs			[]	+
Raw Disk UUIDs/hdisks			[]	+
Disk Error Management?			no	+
Primary Workload Manager Class			[]	+
Secondary Workload Manager Class			[]	+
Miscellaneous Data			[]	
WPAR Name			[]	+
User Defined Resources			[]	+
DS8000(GM)/XIV Replicated Resources			[]	+

```

XIV Replicated Resources +
DS8000-Metro Mirror (In-band) Resources g2_mg +
[BOTTOM]

```

```

F1=Help          F2=Refresh      F3=Cancel      F4=List
Esc+5=Reset     F6=Command     F7=Edit       F8=Image
F9=Shell        F10=Exit       Enter=Do

```

17. Verify and synchronize the cluster configuration and start the PowerHA service.

Testing the final configuration

This section presents the results of two basic tests: resource group movement and planned storage swap (HyperSwap). More detailed testing is described in 4.6, “Test scenarios” on page 136.

RG movement

Follow these steps:

1. Test the RG movement back and forth as shown in Example 3-116.

Example 3-116 Moving the RG to another site

```

root@ss2n04base: /> clmgr move rg g2_rg site=Site_B
Attempting to move resource group g2_rg to site Site_B.

```

Waiting for the cluster to process the resource group movement request....

Waiting for the cluster to stabilize.....

Resource group g2_rg is online on site Site_B.

Cluster Name: g2_c1

```

Resource Group Name: g2_rg
Node                Group State
-----
ps2n01base         OFFLINE
ps2n02base         OFFLINE
ss2n03base         ONLINE
ss2n04base         OFFLINE

```

```

=====
root@ss2n04base: /> clmgr move rg g2_rg site=Site_A
Attempting to move resource group g2_rg to site Site_A.

```

Waiting for the cluster to process the resource group movement request....

Waiting for the cluster to stabilize.....

Resource group g2_rg is online on site Site_A.

Cluster Name: g2_c1

Resource Group Name: g2_rg

Node	Group State
ps2n01base	ONLINE
ps2n02base	OFFLINE
ss2n03base	OFFLINE
ss2n04base	OFFLINE

2. Test the planned storage HyperSwap.

The PPRC status before the planned swap is shown in Example 3-117.

Example 3-117 PPRC status before the planned swap

Command to run on all nodes: lsprrc -Ao

```
-----Running lsprrc -Ao on ps2n01base-----
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk1	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk6	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk2	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2

```
-----Running lsprrc -Ao on ps2n02base-----
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk1	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk2	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk6	Active	-1	0	500507630bffc4c8	500507630bffc1e2

```
-----Running lsprrc -Ao on ss2n03base-----
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk1	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk2	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk6	Active	-1	0	500507630bffc4c8	500507630bffc1e2

```
-----Running lsprrc -Ao on ss2n04base-----
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk1	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk6	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk2	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2

3. Use the following PowerHA SMIT menu options to swap the storage (see Example 3-118):

smitty hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups

Example 3-118 Swapping the storage subsystems

```

Manage User Mirror Group(s)

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

                                     [Entry Fields]
* Mirror Group(s)                    g2_mg                    +
* Operation                          Swap                      +

F1=Help          F2=Refresh          F3=Cancel          F4=List
Esc+5=Reset      F6=Command          F7=Edit           F8=Image
F9=Shell         F10=Exit            Enter=Do
  
```

Observe that the hdisk3 primary path ID changed from 0 to 1 and the primary storage changed also (by using the **lspprc -Ao** output, as shown in Example 3-119).

Example 3-119 PPRC information after the swap

```

Command to run on all nodes: lspprc -Ao
-----Running lspprc -Ao on ps2n01base-----
hdisk#  PPRC      Primary      Secondary      Primary Storage      Secondary Storage
        state    path group    path group    WWNN                WWNN
        ID      ID
hdisk1  Active    0(s)         1              500507630bffc4c8    500507630bffc1e2
hdisk6  Active    -1           0              500507630bffc4c8    500507630bffc1e2
hdisk2  Active    0(s)         -1             500507630bffc4c8    500507630bffc1e2
hdisk4  Active    0(s)         1              500507630bffc4c8    500507630bffc1e2
hdisk3  Active    1(s)         0              500507630bffc1e2    500507630bffc4c8
  
```

The path status can be revealed by using the **lspprc -p hdiskx** command, as shown in Example 3-120.

Example 3-120 PPRC path information

```

Command to run on all nodes: lspprc -p hdisk3
-----Running lspprc -p hdisk3 on ps2n01base-----
path      WWNN                LSS  VOL  path
group id
=====
0          500507630bffc4c8    0xa7 0x00  SECONDARY
1(s)      500507630bffc1e2    0xa7 0x00  PRIMARY

path      path path      parent connection
group id  id  status
=====
0         0   Enabled  fscsi1 500507630b1884c8,40a7400000000000
1         1   Enabled  fscsi3 500507630b1301e2,40a7400000000000
  
```




PowerHA HyperSwap cluster, Oracle stand-alone database, and ASM

This chapter describes the deployment of the PowerHA HyperSwap solution with the Oracle stand-alone database. This database uses Oracle's Automatic Storage Management (ASM) technology to manage the disks that store the database files. We also provide several test scenarios to verify this solution.

This chapter contains the following topics:

- ▶ Cluster description and diagrams
- ▶ Storage configuration
- ▶ Node configuration
- ▶ Oracle installation and configuration on cluster nodes
- ▶ PowerHA configuration
- ▶ Test scenarios

4.1 Cluster description and diagrams

In our test scenario, we configure a PowerHA *stretched cluster* with three nodes (two in Site_A and one in Site_B), two DS8800 storage subsystems, and two SAN switches. Each site has one storage subsystem and one SAN switch.

Clustering infrastructure: The PowerHA clustering infrastructure provides automated storage handling, masking storage subsystem failures by sending the (in-band) commands to switch to the available copy of the data. This method is transparent to the application because storage failures are handled at the device driver level and automated through the PowerHA AIX kernel extension.

The diagram of our test environment is shown in Figure 4-1. The cluster configuration consists of three nodes. Two are in the primary site (PS) with the primary storage. One node is in the secondary site (SS) with the secondary storage, which holds the storage replica (Metro Mirror).

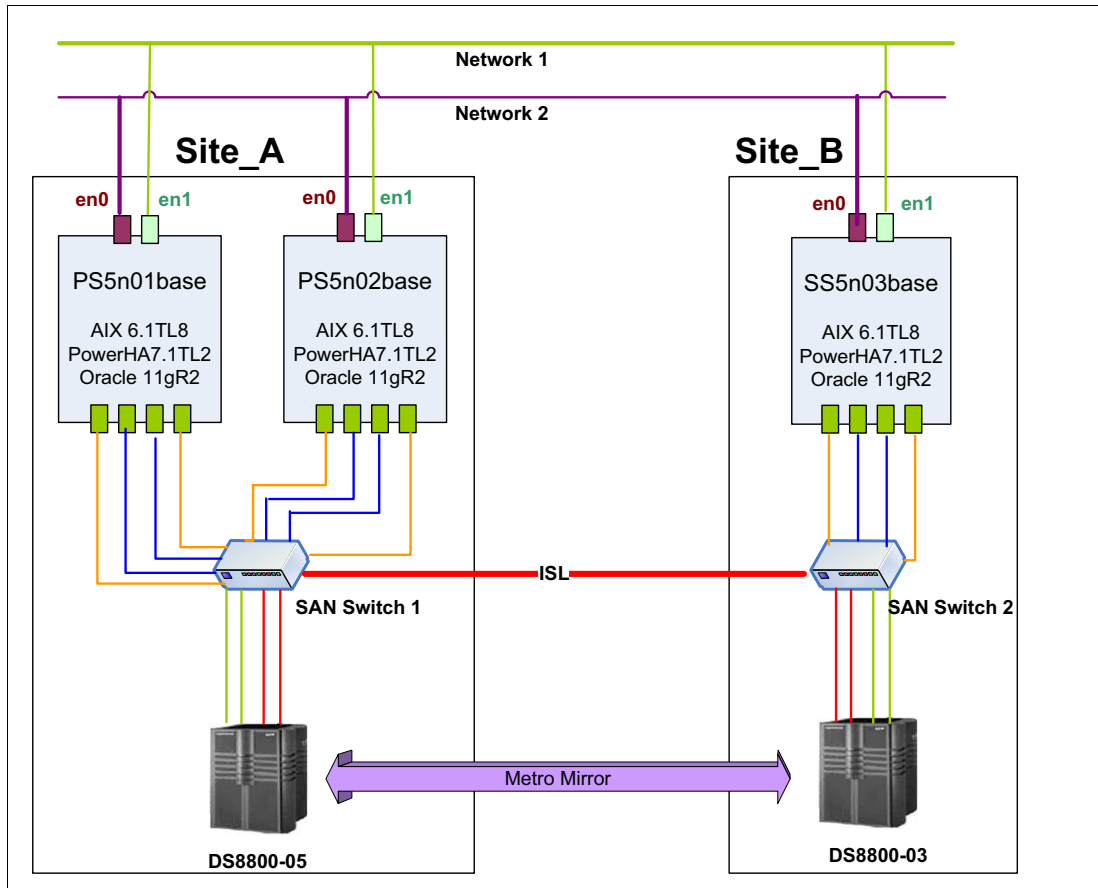


Figure 4-1 Cluster diagram

Before configuring the PowerHA SystemMirror, the following steps are needed:

1. Configure the storage with replicated resources (DS8000 Metro Mirror) and SAN connectivity among systems, storage, and sites.
2. Configure AIX (OS, networking, and storage device drivers).
3. Install Oracle (Grid software, ASM configuration, and database).

4. Install PowerHA SystemMirror For AIX Enterprise Edition 7.1.2¹ code.
5. Create the Oracle database on one node.
6. Register the database on the other cluster nodes.

4.2 Storage configuration

We describe the storage configuration that we used in our test environment. The storage diagram is shown in Figure 4-2. For detailed steps about how to configure the storage and SAN, see Chapter 3, “PowerHA cluster with AIX HyperSwap Active-Standby for applications using a shared file system” on page 25.

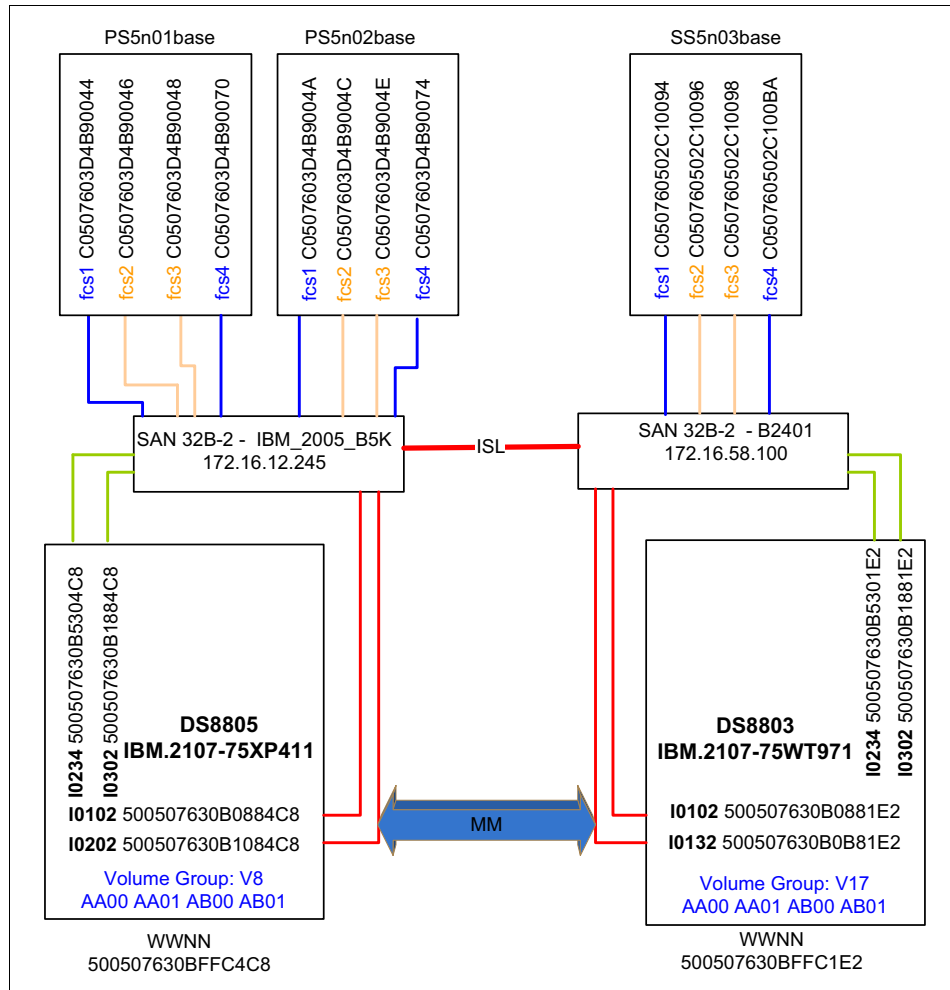


Figure 4-2 Storage and SAN diagram

¹ Always check the latest AIX and PowerHA fixes, and the DS8800 microcode levels required for HyperSwap.

4.2.1 LUN and mapping configuration

Storage command-line interface (CLI): Throughout this chapter, we use the DS8800 command-line interface (DSCLI). Setting up the DSCLI is not covered in this material. See the IBM System Storage DS8000 Information Center:

<http://publib.boulder.ibm.com/infocenter/dsichelp/ds8000ic/index.jsp>

Primary storage configuration

We describe the primary storage configuration. The information presented is relevant for cluster configuration and will be used subsequently to describe testing results.

In our test scenario, DS8800-05 is the primary storage and DS8800-03 is the secondary storage.

Storage identification

Example 4-1 shows basic information for the primary storage (using the DSCLI).

Example 4-1 Display primary storage basic information

```
dscli> lssi
Name ID                Storage Unit      Model WNN          State  ESSNet
=====
-   IBM.2107-75XP411  IBM.2107-75XP410  951   500507630BFFC4C8 Online Enabled
```

Storage volume group configuration

Example 4-2 shows the four logical unit numbers (LUNs) for this testing and that they belong to Volume Group V8.

Example 4-2 Displaying volume group information on the primary storage (DS8800-05)

```
dscli> lsvolgrp V8
Name          ID Type
=====
OpenSwap_G5VG V8 SCSI Mask
```

```
dscli> showvolgrp V8
Name OpenSwap_G5VG
ID   V8
Type SCSI Mask
Vols AA00 AA01 AB00 AB01
```

Metro Mirror information (Peer-to-Peer Remote Copy)

Example 4-3, shows the four LUNs' Peer-to-Peer Remote Copy (PPRC) volume relationships on the primary storage.

Example 4-3 Display PPRC volume relationship on the primary storage

```

dsccli> lspprc AA00-AB01
ID          State      Reason Type      SourceLSS Timeout (secs) Critical Mode First Pass
Status
=====
AA00:AA00 Full Duplex - Metro Mirror AA      5          Disabled      Invalid
AA01:AA01 Full Duplex - Metro Mirror AA      5          Disabled      Invalid
AB00:AB00 Full Duplex - Metro Mirror AB      5          Disabled      Invalid
AB01:AB01 Full Duplex - Metro Mirror AB      5          Disabled      Invalid

```

PPRC path information

Example 4-4 shows the PPRC paths of the four LUNs from the primary storage to the secondary storage.

Example 4-4 Display the pprcpath information on the primary storage

```

dsccli> lspprcpath AA
Src Tgt State  SS  Port  Attached Port Tgt WWNN
=====
AA  AA  Success FFAA I0102 I0102      500507630BFFC1E2
AA  AA  Success FFAA I0202 I0132      500507630BFFC1E2

```

```

dsccli> lspprcpath AB
Src Tgt State  SS  Port  Attached Port Tgt WWNN
=====
AB  AB  Success FFAB I0102 I0102      500507630BFFC1E2
AB  AB  Success FFAB I0202 I0132      500507630BFFC1E2

```

LUN masking

Example 4-5 shows host connection (LUN masking) configuration of Volume Group V8.

Storage mapping: The four LUNs in the primary storage are mapped to all three nodes of the test environment.

Example 4-5 Display the hostconnect information on the primary storage

```

dsccli> lshostconnect -volgrp V8
Name          ID  WWPN          HostType      Profile          portgrp volgrpID
ESSIOport
=====
====
G5_P7805LP9_fcs4 0023 C0507603D4B90070 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V8      all
G5_P7805LP9_fcs1 0024 C0507603D4B90044 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V8      all
G5_P7805LP10_fcs4 0025 C0507603D4B90074 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V8      all
G5_P7805LP10_fcs1 0026 C0507603D4B9004A pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V8      all
G5_P7703LP9_fcs4 0037 C050760502C100BA pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V8      all
G5_P7703LP9_fcs1 0038 C050760502C10094 pSeriesPowerswap IBM pSeries - AIX with Powerswap support      0 V8      all

```

Secondary storage configuration

We describe the secondary storage configuration. The information presented is relevant for cluster configuration and will be used to describe the testing.

Secondary storage identification

In this testing, DS8800-03 is the secondary storage. We use the DSCCLI commands that are shown in Example 4-6 on page 108 to identify the secondary storage.

Example 4-6 Display the secondary storage's identification information

```
dscli> lssi
Name      ID              Storage Unit      Model WWNN              State  ESSNet
=====
DS8803   IBM.2107-75WT971  IBM.2107-75WT970  951   500507630BFFC1E2  Online Enabled
```

Storage volume group information

Example 4-7 shows that the four PPRC target LUNs belong to Volume Group V17.

Example 4-7 Display the volume group information on the secondary storage

```
dscli> lsvolgrp V17
Name      ID  Type
=====
OpenSwap_G5VG V17 SCSI Mask
```

```
dscli> showvolgrp V17
Name OpenSwap_G5VG
ID   V17
Type SCSI Mask
Vols AA00 AA01 AB00 AB01
```

Metro Mirror information (PPRC)

Example 4-8 shows the four replicated LUNs on the secondary storage.

Example 4-8 PPRC relationship on the secondary storage

```
dscli> lspprc AA00-AB01
ID      State      Reason Type      SourceLSS Timeout (secs) Critical Mode First Pass
Status
=====
AA00:AA00 Target Full Duplex - Metro Mirror AA unknown Disabled Invalid
AA01:AA01 Target Full Duplex - Metro Mirror AA unknown Disabled Invalid
AB00:AB00 Target Full Duplex - Metro Mirror AB unknown Disabled Invalid
AB01:AB01 Target Full Duplex - Metro Mirror AB unknown Disabled Invalid
```

PPRC path information

Example 4-9 shows the PPRC paths for the four LUNs from the secondary storage to the primary storage.

Example 4-9 PPRC path information on the secondary storage

```
dscli> lspprcpath AA
Src Tgt State  SS  Port  Attached Port Tgt WWNN
=====
AA  AA  Success FFAA I0102 I0102      500507630BFFC4C8
AA  AA  Success FFAA I0132 I0202      500507630BFFC4C8
```

```
dscli> lspprcpath AB
Src Tgt State  SS  Port  Attached Port Tgt WWNN
=====
AB  AB  Success FFAB I0102 I0102      500507630BFFC4C8
AB  AB  Success FFAB I0132 I0202      500507630BFFC4C8
```

LUN masking

Example 4-10 shows the host connection (LUN masking) configuration for Volume Group V17. The four LUNs in the secondary storage are also mapped to all three cluster nodes.

Example 4-10 Display the host connectivity information on the secondary storage

```
dscli> lshostconnect -volgrp V17
```

Name	ID	WWPN	HostType	Profile	portgrp	volgrpID	ESSIOport
G5_P7805LP9	002A	C0507603D4B90046	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0	V17	all
G5_P7805LP9	002B	C0507603D4B90048	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0	V17	all
G5_P7805LP10	002C	C0507603D4B9004C	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0	V17	all
G5_P7805LP10	002D	C0507603D4B9004E	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0	V17	all
G5_P7703LP9	003E	C050760502C10096	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0	V17	all
G5_P7703LP9	003F	C050760502C10098	pSeriesPowerswap	IBM pSeries - AIX with Powerswap support	0	V17	all

4.2.2 Zoning configuration

In our testing scenario, there are four Virtual Fibre Channel (VFC) adapters² in each logical partition (LPAR) (cluster node).

Two VFC adapters are configured to access the primary storage, and the other two VFC adapters access the secondary storage. Example 4-11 shows the zoning configuration for the nodes to the primary storage.

Example 4-11 Display the nodes' VFC adapters (fcs1 and fcs4) to DS8800-05 (primary storage)

PS5n01base to Primary Storage:

```
zone: P7805LP9_fcs1_DS8805_I0302
      c0:50:76:03:d4:b9:00:44; 50:05:07:63:0b:18:84:c8
zone: P7805LP9_fcs4_DS8805_I0234
      c0:50:76:03:d4:b9:00:70; 50:05:07:63:0b:53:04:c8
```

PS5n02base to Primary Storage:

```
zone: P7805LP10_fcs1_DS8805_I0302
      c0:50:76:03:d4:b9:00:4a; 50:05:07:63:0b:18:84:c8
zone: P7805LP10_fcs4_DS8805_I0234
      c0:50:76:03:d4:b9:00:74; 50:05:07:63:0b:53:04:c8
```

SS5n03base to Primary Storage:

```
zone: P7703LP9_fcs1_DS8805_I0302
      c0:50:76:05:02:c1:00:94; 50:05:07:63:0b:18:84:c8
zone: P7703LP9_fcs4_DS8805_I0234
      c0:50:76:05:02:c1:00:ba; 50:05:07:63:0b:53:04:c8
```

Example 4-12 shows the zoning configuration for the nodes to the secondary storage.

Example 4-12 Zoning configuration for our test scenario for the nodes to the secondary storage

PS5n01base to Secondary Storage:

```
zone: P7805LP9_fcs2_DS8803_I0200
      c0:50:76:03:d4:b9:00:46; 50:05:07:63:0b:10:01:e2
zone: P7805LP9_fcs3_DS8803_I0203
      c0:50:76:03:d4:b9:00:48; 50:05:07:63:0b:13:01:e2
```

² Physical Fibre Channel host bus adapters (HBAs) can be used as well.

PS5n02base to Secondary Storage:

```
zone: P7805LP10_fcs2_DS8803_I0200
      c0:50:76:03:d4:b9:00:4e; 50:05:07:63:0b:10:01:e2
zone: P7805LP10_fcs3_DS8803_I0203
      c0:50:76:03:d4:b9:00:4e; 50:05:07:63:0b:13:01:e2
```

SS5n03base to Secondary Storage:

```
zone: P7703LP9_fcs2_DS8803_I0200
      c0:50:76:05:02:c1:00:96; 50:05:07:63:0b:10:01:e2
zone: P7703LP9_fcs3_DS8803_I0203
      c0:50:76:05:02:c1:00:98; 50:05:07:63:0b:13:01:e2
```

SAN configuration: SAN design is beyond the purpose of this document. The configuration that we used is for exemplification only and does not guarantee resiliency to all possible types of SAN failures. Consult with your SAN administrators and intersite communication providers (for dark fiber, Wave Division Multiplexing (WDM), and so on) to understand the SAN resiliency to various types of failures and plan accordingly.

4.2.3 AIX disks information

After the storage and SAN configuration are in place, AIX recognizes the allocated LUNs. Example 4-13 shows that eight disks are available as configured in DS8800 storage (hdisk2 to hdisk9). Four LUNs are from the primary storage, and the other four LUNs are from the secondary storage.

Disk accessibility: Although AIX on the cluster node reports that eight disks are available (`lsdev -Cc disk`), only four of them are accessible for I/O data operations (read/write). The four LUNs are the LUNs identified in the storage as PPRC source. The PPRC target LUNs are reported as available, but they are not accessible for I/O (`lsquerypv -h /dev/hdisk*`).

Example 4-13 Disk information on all nodes (PS5n01base, PS5n02base, and SS5n03base)

```
# lsdev -Cc disk
hdisk0 Available C2-T1-01 MPIO IBM 2145 FC Disk
hdisk1 Available C2-T1-01 MPIO IBM 2145 FC Disk
hdisk2 Available 48-T1-01 MPIO IBM 2107 FC Disk
hdisk3 Available 48-T1-01 MPIO IBM 2107 FC Disk
hdisk4 Available 48-T1-01 MPIO IBM 2107 FC Disk
hdisk5 Available 48-T1-01 MPIO IBM 2107 FC Disk
hdisk6 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk7 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk8 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk9 Available 49-T1-01 MPIO IBM 2107 FC Disk

# lspath|egrep "fscsi1|fscsi2|fscsi3|fscsi4"
Enabled hdisk2 fscsi1
Enabled hdisk3 fscsi1
Enabled hdisk4 fscsi1
Enabled hdisk5 fscsi1
Enabled hdisk6 fscsi2
Enabled hdisk7 fscsi2
Enabled hdisk8 fscsi2
Enabled hdisk9 fscsi2
```

```

Enabled hdisk6 fscsi3
Enabled hdisk7 fscsi3
Enabled hdisk8 fscsi3
Enabled hdisk9 fscsi3
Enabled hdisk2 fscsi4
Enabled hdisk3 fscsi4
Enabled hdisk4 fscsi4
Enabled hdisk5 fscsi4

# for i in 2 3 4 5 6 7 8 9
>do
>echo "hdisk$i";lscfg -vpl hdisk$i|egrep "Serial|Z7"
>done
hdisk2
    Serial Number.....75XP411A
    Device Specific.(Z7).....AA00
hdisk3
    Serial Number.....75XP411A
    Device Specific.(Z7).....AA01
hdisk4
    Serial Number.....75XP411A
    Device Specific.(Z7).....AB00
hdisk5
    Serial Number.....75XP411A
    Device Specific.(Z7).....AB01
hdisk6
    Serial Number.....75WT971A
    Device Specific.(Z7).....AA00
hdisk7
    Serial Number.....75WT971A
    Device Specific.(Z7).....AA01
hdisk8
    Serial Number.....75WT971A
    Device Specific.(Z7).....AB00
hdisk9
    Serial Number.....75WT971A
    Device Specific.(Z7).....AB01

```

Example 4-14 shows the worldwide port name (WWPN) information of the VFC adapters on the nodes. The VFC adapters are configured into the SAN zoning, which is shown in 4.2.2, “Zoning configuration” on page 109.

Example 4-14 WWPN information on the nodes

```

Node1:
# for i in 1 2 3 4
>do
>echo "fcs$i"
>lscfg -vpl fcs$i|grep "Network"
>done
fcs1
    Network Address.....C0507603D4B90044
fcs2
    Network Address.....C0507603D4B90046
fcs3
    Network Address.....C0507603D4B90048

```

```
fcs4
    Network Address.....C0507603D4B90070
```

```
Node2:
# for i in 1 2 3 4
>do
>echo "fcs$i"
>lscfg -vpl fcs$i|grep "Network"
>done
```

```
fcs1
    Network Address.....C0507603D4B9004A
fcs2
    Network Address.....C0507603D4B9004C
fcs3
    Network Address.....C0507603D4B9004E
fcs4
    Network Address.....C0507603D4B90074
```

```
Node3:
# for i in 1 2 3 4
>do
>echo "fcs$i"
>lscfg -vpl fcs$i|grep "Network"
>done
fcs1
    Network Address.....C050760502C10094
fcs2
    Network Address.....C050760502C10096
fcs3
    Network Address.....C050760502C10098
fcs4
    Network Address.....C050760502C100BA
```

PPRC source and target information

To identify the available hdisks that can be accessed for I/O (read/write) operations, we use the `lsprrc -Ao` command (see Example 4-15 on page 113). The command output shows paths to the primary and secondary LUNs, grouped based on accessibility. The groups are numbered 0 and 1. A -1 for the path group ID indicates no active paths to the LUN. The (s) attribute indicates that the group is currently selected for disk I/O (read/write). For example, hdisk2 has a path group 0 to the primary LUN, and it is selected for I/O. It has no paths to its secondary group because it is not yet configured for HyperSwap.

Example 4-15 Disk information on all nodes (PS5n01base, PS5n02base, and SS5n03base)

```
# lspprc -Ao
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk2	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk5	Active	0(s)	-1	500507630bffc4c8	500507630bffc1e2
hdisk6	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk7	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk8	Active	-1	0	500507630bffc4c8	500507630bffc1e2
hdisk9	Active	-1	0	500507630bffc4c8	500507630bffc1e2

Example 4-16 lists the hdisk2 disk attributes that are important for HyperSwap.

Example 4-16 Default values of the important attributes of the hdisk

```
# lsattr -E1 hdisk2 | egrep "PCM|reserve_policy|san_rep_cfg|san_rep_device"
```

PCM	PCM/friend/aixmpi	Path Control Module	False
reserve_policy	single_path	Reserve Policy	True
san_rep_cfg	none	SAN Replication Device Configuration Policy	True+
san_rep_device	detected	SAN Replication Device	False

The following attributes are relevant for our configuration:

- ▶ Path Control Module
- ▶ Reserve policy
- ▶ SAN replication device

4.3 Node configuration

After the shared disks are configured on all nodes, we perform the following AIX configuration changes:

- ▶ AIX default disk device driver for DS8800 (to enable HyperSwap)
- ▶ Host bus adapter (HBA) attributes (to enable the secondary path)
- ▶ hdisk attributes to enable the replicated device
- ▶ Network Time Protocol (NTP) for time synchronization across nodes (for logging consistency)

4.3.1 AIX disk device driver and HBA attributes

AIX level: AIX 6.1TL8SP1 or AIX 7.1TL2SP1 is the minimum AIX requirement for PowerHA HyperSwap. In our environment, we use AIX 6.1TL8SP1.

We configure the AIX device driver and HBA attributes to enable the HyperSwap capability:

1. On all nodes, check the current disk device driver that is used for the DS8000 storage family (see Example 4-17).

Important: *Unless otherwise specified, the following steps must be performed on all nodes.* For clarity, we only show information from the ps5n01base node.

Example 4-17 Display the storage families and the driver

```
# manage_disk_drivers -l
Device          Present Driver Driver Options
2810XIV         AIX_AAPCM     AIX_AAPCM,AIX_non_MPIO
DS4100         AIX_APPCM     AIX_APPCM,AIX_fcarray
.....<< snippet >>.....
DS3500         AIX_APPCM     AIX_APPCM
XIVCTRL        MPIO_XIVCTRL  MPIO_XIVCTRL,nonMPIO_XIVCTRL,MPIO_XIVCTRL,nonMPIO_XIVCTRL
2107DS8K       NO_OVERRIDE   NO_OVERRIDE,AIX_AAPCM,NO_OVERRIDE
```

The AIX `manage_disk_drivers` command supports the following options:

- ▶ `NO_OVERRIDE` (default)
- ▶ `AIX_AAPCM`

By using the `NO_OVERRIDE` option, you can use storage vendor device driver software, such as Subsystem Device Driver Path Control Module (SDDPCM) for the IBM DS8000 series, to manage the storage systems. For the HyperSwap environment, the `AIX_AAPCM` option must be set. This option prevents any problems that might occur if the SDDPCM driver is installed later.

Example 4-18 shows the command used to change the DS8800 storage driver from the default to `AIX_AAPCM`.

Example 4-18 Set disk driver for DS8800

```
# manage_disk_drivers -d 2107DS8K -o AIX_AAPCM
***** ATTENTION *****
For the change to take effect the system must be rebooted
```

HyperSwap also requires that you change the HBA's attributes (`dyntrk` and `fc_err_recov`) as shown in Example 4-19.

Example 4-19 Enable fast recover for the HBA device

```
# chdev -l fcs1 -a dyntrk=yes fc_err_recov=fast_fail -P
# chdev -l fcs2 -a dyntrk=yes fc_err_recov=fast_fail -P
# chdev -l fcs3 -a dyntrk=yes fc_err_recov=fast_fail -P
# chdev -l fcs4 -a dyntrk=yes fc_err_recov=fast_fail -P
```

Reboot required: The systems must be rebooted at this time to enable the HyperSwap driver (`AIX-AAPCM`) and the HBA's attributes.

4.3.2 Disk configuration

We configure the disk attributes to enable the HyperSwap composite disk device (hdisk*). We change the disk reservation policy (**reserve_policy**) to `no_reserve`, path failover policy (**san_rep_cfg**) to `migrate_disk`, and PVID (**pv**) to *yes for all disks on all systems*, as shown in Example 4-20.

Example 4-20 Change a disk's attribute to enable HyperSwap

```
# for i in 2 3 4 5
>do
>chdev -l hdisk$i -a reserve_policy=no_reserve -a san_rep_cfg=migrate_disk -a
pv=yes
>done
hdisk2 changed
hdisk3 changed
hdisk4 changed
hdisk5 changed
```

We check the disk availability again and see that the PPRC targets (hdisk6 to hdisk9) changed to the Defined state, as shown in Example 4-21.

Example 4-21 Display the disk's status after enabling HyperSwap

```
# lsdev -Cc disk
hdisk0 Available C2-T1-01 MPIO IBM 2145 FC Disk
hdisk1 Available C2-T1-01 MPIO IBM 2145 FC Disk
hdisk2 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk3 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk4 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk5 Available 49-T1-01 MPIO IBM 2107 FC Disk
hdisk6 Defined 49-T1-01 MPIO IBM 2107 FC Disk
hdisk7 Defined 49-T1-01 MPIO IBM 2107 FC Disk
hdisk8 Defined 49-T1-01 MPIO IBM 2107 FC Disk
hdisk9 Defined 49-T1-01 MPIO IBM 2107 FC Disk
```

We also verify again the PPRC status as shown in Example 4-22. Observe that the secondary path group changed from -1 to 1, which means that the disk access can be swapped to secondary storage by using HyperSwap. Also, the PPRC targets (from hdisk6 to hdisk9) are removed from this display.

Example 4-22 Display the LUN's PPRC status with the `lspprc` command

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WNN	Secondary Storage WNN
hdisk2	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk3	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk4	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2
hdisk5	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2

We can use the `lspprc -p` or `lspprc -v` commands to display the detailed path group and vital product data (VPD) information of the individual disk, as shown in Example 4-23 on page 116.

Example 4-23 Display the disk's detailed path group and VPD information

lspprc -p hdisk2

path	WWNN	LSS	VOL	path
group id				group status

```
=====
0(s)      500507630bffc4c8 0xaa 0x00 PRIMARY
1         500507630bffc1e2 0xaa 0x00 SECONDARY
```

path	path	path	parent	connection
group id	id	status		

```
=====
0 0 Enabled fscsi1 500507630b1884c8,40aa400000000000
1 1 Enabled fscsi2 500507630b1001e2,40aa400000000000
1 2 Enabled fscsi3 500507630b1301e2,40aa400000000000
```

lspprc -v hdisk2

Hyperswap lun unique
identifier.....352037355850343131414130300050b75eb207210790003IBMfcp

hdisk2 Primary MPIO IBM 2107 FC Disk

Manufacturer.....IBM
Machine Type and Model.....2107900
ROS Level and ID.....2E313630
Serial Number.....75XP411A
Device Specific.(Z7).....AA00
Device Specific.(Z0).....000005329F101002
Device Specific.(Z1).....A00
Device Specific.(Z2).....075
Unique Device Identifier.....200B75XP411AA0007210790003IBMfcp
Logical Subsystem ID.....0xaa
Volume Identifier.....0x00
Subsystem Identifier(SS ID)...0xFFAA
Control Unit Sequence Number..00000XP411
Storage Subsystem WWNN.....500507630bffc4c8
Logical Unit Number ID.....40aa400000000000

hdisk2 Secondary MPIO IBM 2107 FC Disk

Manufacturer.....IBM
Machine Type and Model.....2107900
ROS Level and ID.....2E313630
Serial Number.....75WT971A
Device Specific.(Z7).....AA00
Device Specific.(Z0).....000005329F101002
Device Specific.(Z1).....A00
Device Specific.(Z2).....075
Unique Device Identifier.....200B75WT971AA0007210790003IBMfcp
Logical Subsystem ID.....0xaa
Volume Identifier.....0x00
Subsystem Identifier(SS ID)...0xFFAA
Control Unit Sequence Number..00000WT971
Storage Subsystem WWNN.....500507630bffc1e2
Logical Unit Number ID.....40aa400000000000

4.3.3 Time synchronization

We suggest that you enable time synchronization for the PowerHA cluster nodes. Time synchronization is useful for log management and analysis. We use Network Time Protocol (NTP) to achieve time synchronization and describe how to configure time synchronization with NTP.

NTP server service configuration

In our tests, we designated our Network Installation Management (NIM) server (ITSONIM at IP address 172.16.66.122) to act as the NTP server. To enable the NTP server, we add one line to the `/etc/ntp.conf` file as shown in Example 4-24.

Example 4-24 The /etc/ntp.conf file on NTP server

```
broadcastclient
driftfile /etc/ntp.drift
logfile /etc/ntp.trace
server 127.127.1.0 prefer
```

We start the NTP service as shown in Example 4-25.

Example 4-25 Starting the xntpd daemon on the NTP server

```
# startsrc -s xntpd -a -x
0513-059 The xntpd Subsystem has been started. Subsystem PID is 5505328.
```

NTP client configuration

We need to configure the NTP client on all nodes. For example, on the ps5n01base node, we add one line to the `/etc/ntp.conf` file, as shown in Example 4-26.

Example 4-26 The /etc/ntp.conf file on the NTP client

```
broadcastclient
driftfile /etc/ntp.drift
logfile /etc/ntp.trace
server 172.16.66.122
```

We start the NTP service as shown in Example 4-27.

Example 4-27 Starting the xntpd daemon on the NTP client

```
# startsrc -s xntpd -a -x
0513-059 The xntpd Subsystem has been started. Subsystem PID is 5609393.
```

Depending on the time difference (drift) between the NTP server and its clients, it might take several minutes for the clients to complete the time synchronization. Example 4-28 shows the unsynchronized status.

Example 4-28 Time not synchronized

```
#lssrc -ls xntpd
Program name: /usr/sbin/xntpd
Version: 3
Leap indicator: 11 (Leap indicator is insane.)
Sys peer: no peer, system is insane
Sys stratum: 16
```

```

Sys precision: -18
Debug/Tracing: DISABLED
Root distance: 0.000000
Root dispersion: 0.000000
Reference ID: no refid, system is insane
Reference time: no reftime, system is insane
Broadcast delay: 0.003906 (sec)
Auth delay: 0.000122 (sec)
System flags: bclient pll monitor filegen
System uptime: 13 (sec)
Clock stability: 0.000000 (sec)
Clock frequency: 0.000000 (sec)
Peer: 172.16.66.122
    flags: (configured)
    stratum: 16, version: 3
    our mode: client, his mode: unspecified
Subsystem      Group      PID      Status
xntpd          tcpip     2949250  active

```

Example 4-29 shows the synchronized status of the clients.

Example 4-29 Time synchronized on the clients

```

# lssrc -ls xntpd
Program name: /usr/sbin/xntpd
Version: 3
Leap indicator: 00 (No leap second today.)
Sys peer: 172.16.66.122
Sys stratum: 5
Sys precision: -18
Debug/Tracing: DISABLED
Root distance: 0.000412
Root dispersion: 0.011932
Reference ID: 172.16.66.122
Reference time: d4698d99.5aa24000 Wed, Dec 5 2012 17:08:41.354
Broadcast delay: 0.003906 (sec)
Auth delay: 0.000122 (sec)
System flags: bclient pll monitor filegen
System uptime: 3504 (sec)
Clock stability: 12.683578 (sec)
Clock frequency: 0.000000 (sec)
Peer: 172.16.66.122
    flags: (configured)(sys peer)
    stratum: 4, version: 3
    our mode: client, his mode: server
Subsystem      Group      PID      Status
xntpd          tcpip     2949250  active

```

Modify the /etc/rc.tcpip file

Look for the following line in the /etc/rc.tcpip file and remove the comment from it to ensure that the **xntpd** daemon starts automatically at the system reboot:

```
#start /usr/sbin/xntpd "$src_running"
```

4.4 Oracle installation and configuration on cluster nodes

We describe how to install and configure the Oracle stand-alone database by using ASM to manage the disks:

- ▶ Environment checking and configuration
- ▶ Installing grid (Oracle Cluster Ready Services) and database software
- ▶ Create a database instance on PS5n01base
- ▶ Register the database instance on other nodes
- ▶ Change the ASM disk group to the spfile
- ▶ Test the database start-up and shutdown scripts

Important: We use the Oracle 11gR2 11.2.0.3 version in our test environment.

The parameters and environment variables that we use in our testing are only for demonstration purposes. You must adjust these parameters according to your deployment rules.

4.4.1 Environment checking and configuration

Follow the installation guidelines provided by the Oracle documentation:

http://www.oracle.com/pls/db112/portal.portal_db?selected=11

For information related to Oracle support for environments using virtualized hardware resources, see this website:

<http://www.oracle.com/technetwork/database/virtualizationmatrix-172995.html>

AIX filesets

Example 4-30 shows how to check the filesets required for Oracle installation. For simplicity, the output of the command is not shown here. Verify the output against the minimum requirements provided with the Oracle 11gR2 installation documentation.

Example 4-30 Checking the AIX filesets

```
#1slpp -l |egrep "(opens|bos.adt|bos.perf|rsct.basic|rsct.compat|x1C.aix61)"
```

AIX parameters

The various required AIX parameters are shown in Example 4-31.

Example 4-31 AIX parameters

```
no -p -o rfc1323=1
no -p -o tcp_recvspace=262144
no -p -o tcp_sendspace=262144
no -p -o udp_sendspace=262144
no -p -o udp_recvspace=655360
chdev -l sys0 -a maxuproc=16384
```

Group and user

The group (dba) and users (grid and oracle) are required for the Oracle grid and database software installation. Example 4-32 on page 120 shows the details.

Storage requirement: In our testing scenario, we use Oracle Automatic Storage Management (ASM) to manage disk space used for Oracle database files. The user grid is required for ASM.

Example 4-32 Creating the dba group, oracle user, and grid user

Create dba group

```
mkgroup -'a' id='300' admin=false projects='System' dba
```

Create grid user

```
mkuser id='311' admin=true pgrp='dba' groups='dba' admgroups='dba'  
home='/home/grid' grid  
chuser  
capabilities=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE,CAP_NUMA_ATTACH,CAP_BYPASS_RAC_VMM  
grid  
chown -R grid:dba /home/grid
```

Create oracle user

```
mkdir /oracle  
mkuser id=301 admin=true pgrp=dba groups=dba admgroups=dba home=/oracle  
shell=/usr/bin/ksh oracle  
chuser  
capabilities=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE,CAP_NUMA_ATTACH,CAP_BYPASS_RAC_VMM  
oracle
```

Then, use the **passwd** command to set the password for the grid and oracle users.

File system configuration

The file system shown in Example 4-33 is needed for storing the Oracle code.

Example 4-33 File system configuration for Oracle code installation

```
# mklv -e x -t jfs2log -y lvoralog -U oracle -G dba rootvg 1  
# crfs -v jfs2 -g rootvg -a logname=lvoralog -a agblksize=4096 -a ea=v2 -A yes  
-asize=40G -m /oracle -p rw -t no  
# mount /oracle  
# chown -R oracle:dba /oracle  
# chmod -R 777 /oracle  
  
# chfs -a size=5G /tmp
```

Prerequisite: If the PowerHA code is already installed, you need to apply the Oracle patch 1384060.1 for **rootpre.sh**.

Check the Oracle Metalink for detailed information:

<http://bit.ly/YEgCBq>

Profiles for grid and oracle users

Example 4-34 shows the grid user's .profile file.

Example 4-34 The .profile file for the grid user

```
export PS1="\~/usr/bin/hostname~-> "  
export PATH=/usr/bin:/etc:/usr/sbin:/usr/ucb:$HOME/bin:/usr/bin/X11:/sbin:.  
export ORACLE_BASE=/oracle/11g  
export ORACLE_HOME=/oracle/grid  
export PATH=$ORACLE_HOME/bin:$PATH:.  
export ORACLE_SID=+ASM1 #+ASM2 on node2 #+ASM3 on node3  
export TEMP=$ORACLE_BASE/tmp  
export TMPDIR=$ORACLE_BASE/tmp  
export LIBPATH=${ORACLE_HOME}/lib:.$LIBPATH
```

Example 4-35 shows the oracle user's .profile file.

Example 4-35 The .profile file for the oracle user

```
PATH=/usr/bin:/etc:/usr/sbin:/usr/ucb:$HOME/bin:/usr/bin/X11:/sbin:.  
export PATH  
if [ -s "$MAIL" ]           # This is at Shell startup. In normal  
then echo "$MAILMSG"       # operation, the Shell checks  
fi                           # periodically.  
export PS1="\~/usr/bin/hostname~-> "  
export ORACLE_BASE=/oracle/11g  
export ORACLE_HOME=/oracle/product  
export PATH=$ORACLE_HOME/bin:$PATH:.  
export LIBPATH=${ORACLE_HOME}/lib:.$LIBPATH  
export ORACLE_SID=testdb  
export ORA_GRID_HOME=$ORACLE_BASE/grid  
export ORACLE_OWNER=oracle  
export ORACLE_CRS=$ORACLE_BASE  
export ORA_CRS_HOME=/oracle/grid/product/11.2.0/grid  
export PATH=$ORA_CRS_HOME/bin:$PATH  
export LIBPATH=$ORA_CRS_HOME/lib:$LIBPATH  
export NLS_LANG=AMERICAN_AMERICA.ZHS16GBK  
export NLS_DATE_FORMAT='YYYY-MM-DD HH24:MI:SS'  
export LDR_CNTRL=TEXTSIZE=64K@STACKSIZE=64K@DATASIZE=64K@SHMPSIZE=64K  
export  
CLASSPATH=$CLASSPATH:$ORACLE_HOME/jre:$ORACLE_HOME/jlib:$ORACLE_HOME/rdbms/jlib:$O  
RACLE_HOME/network/jlib  
export  
CLASSPATH=$CLASSPATH:$ORA_CRS_HOME/jre:$ORA_CRS_HOME/jlib:$ORA_CRS_HOME/rdbms/jlib  
export TEMP=$ORACLE_BASE/tmp  
export TMPDIR=$ORACLE_BASE/tmp  
umask 022
```

AIX networking configuration

In this cluster, there are two network interfaces on each node, which are configured with base IP addresses. Our cluster configuration will handle one resource group with one service IP address. The IP address list is shown in Table 4-1 on page 122.

Table 4-1 IP address information

	Node1	Node2	Node3
Base1 IP address	172.16.29.92	172.16.29.93	172.16.29.249
Base2 IP address	172.16.14.69	172.16.14.70	172.16.14.79
Service IP address	172.16.15.147		

We use the local IP address name resolution. Example 4-36 shows the content of the /etc/hosts file on each node.

Example 4-36 The /etc/hosts file

```

127.0.0.1          loopback localhost      # loopback (1o0) name/address
172.16.29.92     PS5n01base
172.16.14.69     PS5n01std

172.16.29.93     PS5n02base
172.16.14.70     PS5n02std

172.16.29.249   SS5n03base
172.16.14.79    SS5n03std

172.16.15.147   PS5n01svc
    
```

Disks used for the Oracle installation

Table 4-2 shows the disk usage in this testing.

Table 4-2 Disk usage information

Disk	Logical subsystem (LSS) and volume ID	Purpose
hdisk1	Internal disk	ASM diskgroup during the installation. It will be removed after the installation.
hdisk2	AA00	CAA repository (reserved, not used for Oracle).
hdisk3	AA01	Future use (spare capacity).
hdisk4	AB00	ASM diskgroup for database.
hdisk5	AB01	ASM diskgroup for database.

Disk reservation policy: As required by Oracle ASM, you need to ensure that the reservation policy (`reserve_policy`) for the hdisk devices is set to `no_reserve`.

Before the installation, we need to change the ownership and access mode (permissions) for the disks that are used for Oracle (see Example 4-37 on page 123).

Example 4-37 Change the disk's owner and mode

```
# chmod 660 /dev/rhdisk1 /dev/rhdisk3 /dev/rhdisk4
# chown grid:dba /dev/rhdisk1 /dev/rhdisk3 /dev/rhdisk4

# ls -l /dev/rhdisk*
crw----- 2 root    system    19, 6 Dec 05 11:23 /dev/rhdisk0
crw-rw---- 1 grid    dba       19, 0 Dec 06 18:02 /dev/rhdisk1
crw----- 1 root    system    19, 1 Dec 06 16:28 /dev/rhdisk2
crw----- 1 root    system    19, 1 Dec 06 16:28 /dev/rhdisk3
crw-rw---- 1 grid    dba       19, 5 Dec 06 16:28 /dev/rhdisk4
crw-rw---- 1 grid    dba       19, 7 Dec 06 16:28 /dev/rhdisk5
```

4.4.2 Installing grid (Oracle Cluster Ready Services) and database software

We install the Oracle grid and database software using the Oracle graphical user interface (GUI) on the three cluster nodes. We demonstrate this process on node PS5n01base. For the detailed installation guide, see the following Oracle website:

http://www.oracle.com/pls/db112/portal.portal_db?selected=11&frame=#aix_installation_guides

Installing the Oracle grid software

Follow these steps:

1. Log in as the grid user and execute the `.profile`. Then, run `runInstaller` in the grid software directory.
2. Select **Configure Oracle Grid Infrastructure for a Standalone Server**.
3. Select `/dev/rhdisk1` as the ASM disk group named ASMDG. Select the **External** redundancy policy.
4. Select `dba` for the Oracle ASM operator group.
5. Keep the default values for the Oracle base directory (`/oracle/grid`) and the software location (`/oracle/grid/product/11.2.0/grid`).
6. Keep the default value for the inventory directory (`/oracle/grid/oraInventory`).
7. After the installation finishes, run the following scripts as the root user:
 - `/oracle/grid/oraInventory/orainstRoot.sh`
 - `/oracle/grid/product/11.2.0/grid/root.sh`
8. The ASM daemons start automatically as shown in Example 4-38.

Example 4-38 Checking the ASM instance

```
# ps -ef | grep ASM
grid 2752702      1  0 20:09:25      -  0:00 asm_smon_+ASM
grid 2949228      1  0 20:09:25      -  0:00 asm_lgwr_+ASM
....
grid 7798784      1  0 20:09:25      -  0:00 asm_gmon_+ASM
grid 8323122      1  0 20:09:24      -  0:01 asm_pmon_+ASM
grid 8388608      1  0 20:09:25      -  0:01 asm_mmon_+ASM
grid 8454146      1  0 20:09:25      -  0:01 asm_mnnl_+ASM
```

Installing the database software

Follow these steps:

1. Log in as the `oracle` user and execute the `.profile`. Then, run the `runInstaller` script in the database software directory.
2. Select **Skip software update**.
3. Select **Install database software only**.
4. Select **Single instance database installation**.
5. Select **Enterprise Edition**.
6. Keep the default value for the Oracle Base (`/oracle/11gdb`) and Software Location (`/oracle/product`).
7. Select **dba** for the Database Operator Group.
8. After the installation completes, run `/oracle/product/root.sh` as the root user.

4.4.3 Create a database instance on PS5n01base

We use the Oracle GUI to create a database disk group and the database instance, which is named `testdb`.

Create the disk group for the database

Follow these steps:

1. Log in as the `grid` user and run `asmca`.
2. Create a new disk group, which is named `TESTDG`, and select `/dev/rhdisk4` and `/dev/rhdisk5` as members.

Create the database instance

Follow these steps:

1. Log in as the `oracle` user and run `dbca`.
2. Select **Customer Database**.
3. Set `testdb` for the Global Database Name and SID.
4. Select **ASM** for the Storage Type and select **TESTDG**.
5. After the installation completes, the instance is started, as shown in Example 4-39.

Example 4-39 Checking the database instance

```
# ps -ef|grep testdb
grid 5243058      1  0 20:03:13    -  0:00 ora_ocf0_testdb
grid 6684700      1  0 20:03:26    -  0:00 ora_j000_testdb
grid 6750268      1  0 20:03:21    -  0:00 ora_qmnc_testdb
grid 6946896      1  0 20:03:12    -  0:00 ora_diag_testdb
...
grid 15466496     1  0 20:03:13    -  0:00 ora_mmn1_testdb
grid 15532264     1  0 20:03:26    -  0:00 ora_j001_testdb
grid 15597808     1  0 20:03:31    -  0:00 ora_q000_testdb
```

Copy the Oracle code to other nodes in the cluster

The copy source is the node, `PS5n01base`, that we installed in previous step.

There are two directories that we need to copy from PS5n01base to the other two nodes in the cluster (see Example 4-40). Log in on PS5n01base as the `oracle` user and use the `tar` command to archive the following directories:

- ▶ `$ORACLE_BASE/admin`
- ▶ `$ORACLE_HOME/dbs`

Create the users and groups on the other cluster nodes. We suggest that you use the same Group IDs (GIDs) and unique identifier (UID). Then, copy the archives to the remaining two nodes (PS5n02base and SS5n01base) and unpack them in the corresponding directories.

Changing ownership and permissions for ASM disks

Change the raw disk (`/dev/rhdiskX`) ownership and permissions on the other cluster nodes as shown in Example 4-37 on page 123.

4.4.4 Register the database instance on other nodes

In this environment, the database instance will be activated on the cluster node under the control of PowerHA as part of a resource group, using an application server (start and stop scripts). Therefore, there is no need to start the database instance automatically after the ASM instance is started. To accomplish this task, we perform the following configuration on the cluster nodes.

On node PS5n01base

Follow these steps:

1. Log in as the `oracle` user and check the current grid management policy for the database, as shown in Example 4-40.

Example 4-40 Checking current management policy for database

```
$srvctl status database -d testdb
```

```
Database is running.
```

```
$srvctl config database -d testdb
```

```
Database unique name: testdb
```

```
Database name: testdb
```

```
Oracle home: /oracle/product
```

```
Oracle user: oracle
```

```
Spfile: +/testdb/spfiletestdb.ora
```

```
Domain:
```

```
Start options: open
```

```
Stop options: immediate
```

```
Database role: PRIMARY
```

```
Management policy: AUTOMATIC
```

```
Database instance: testdb
```

```
Disk Groups: TESTDG
```

```
Services:
```

2. Change the management policy from AUTOMATIC to MANUAL:

```
$srvctl modify database -d testdb -y MANUAL
```

3. Verify the policy as shown in Example 4-41.

Example 4-41 Checking the management policy after the change

```
$ srvctl config database -d testdb
```

Database unique name: testdb
 Database name: testdb
 Oracle home: /oracle/product
 Oracle user: oracle
 Spfile: +TESTDG/testdb/spfiletestdb.ora
 Domain:
 Start options: open
 Stop options: immediate
 Database role: PRIMARY
Management policy: MANUAL
 Database instance: testdb
 Disk Groups: TESTDG
 Services:

On node PS5n02base

Follow these steps:

1. Register the database instance in ASM as shown in Example 4-42.

Example 4-42 Register the database on the secondary node

```

$srvctl add database -d testdb -n testdb -o $ORACLE_HOME -p
+TESTDG/testdb/spfiletdb.ora -s OPEN -y MANUAL -a TESTDG -t IMMEDIATE"
  
```

2. Verify that the instance (testdb) is registered, as shown in Example 4-43.

Example 4-43 Check whether the database is registered

```

# crsctl stat res -t
-----
NAME                TARGET  STATE        SERVER                     STATE_DETAILS
-----
Local Resources
-----
ora.ASM DG
ora.LISTENER.lsnr   ONLINE  ONLINE       ps5n02base
ora.TESTDG.dg       OFFLINE OFFLINE       ps5n02base
ora.asm              ONLINE  ONLINE       ps5n02base                Started
ora.ons              OFFLINE OFFLINE       ps5n02base
-----
Cluster Resources
-----
ora.cssd
ora.diskmon
ora.evmd
ora.testdb.db
  
```

NAME	TARGET	STATE	SERVER	STATE_DETAILS
ora.ASM DG				
ora.LISTENER.lsnr	ONLINE	ONLINE	ps5n02base	
ora.TESTDG.dg	OFFLINE	OFFLINE	ps5n02base	
ora.asm	ONLINE	ONLINE	ps5n02base	Started
ora.ons	OFFLINE	OFFLINE	ps5n02base	
ora.cssd				
ora.cssd 1	ONLINE	ONLINE	ps5n02base	
ora.diskmon 1	OFFLINE	OFFLINE		
ora.evmd 1	ONLINE	ONLINE	ps5n02base	
ora.testdb.db 1	OFFLINE	OFFLINE		Instance Shutdown

On node SS5n03base

Repeat the steps that we performed on the PS5n02base node.

4.4.5 Change the ASM disk group to the spfile

In our environment, each node has a separate ASM instance and the ASM daemons are started at AIX start-up. During the Oracle grid software installation, we chose one internal disk for the ASM disk group. In this step, we replace the internal disk with a parameter file (spfile) pointing to the ASM instance. We show the configuration on the PS5n01base node.

Create the parameter file (spfile) for the ASM instance

Example 4-44 shows the creation of the spfile for the ASM instance.

Example 4-44 Create the spfile for the ASM instance

```
PS5n01base-> sqlplus /nolog
SQL*Plus: Release 11.2.0.3.0 Production on Mon Dec 10 19:50:42 2012
Copyright (c) 1982, 2011, Oracle. All rights reserved.

SQL> conn / as sysasm;
Connected.

SQL> create pfile='/oracle/grid/product/11.2.0/grid/dbs/asmfile.ora' from spfile;
File created.

SQL> create spfile='/oracle/grid/product/11.2.0/grid/dbs/spfile+ASM.ora' from
pfile='/oracle/grid/product/11.2.0/grid/dbs/asmfile.ora';
File created.
```

Removing the ASMDG resource

Example 4-45 shows how to remove the ASMDG resource.

Example 4-45 Remove the ASMDG resource

```
PS5n01base-> crsctl stat res -t
-----
NAME                TARGET  STATE        SERVER                     STATE_DETAILS
-----
Local Resources
-----
ora.ASM DG          ONLINE  ONLINE      ps5n01base
...

PS5n01base-> crsctl stop res ora.ASM DG
CRS-2673: Attempting to stop 'ora.ASM DG' on 'ps5n01base'
CRS-2677: Stop of 'ora.ASM DG' on 'ps5n01base' succeeded

PS5n01base-> crsctl stat res -t
-----
NAME                TARGET  STATE        SERVER                     STATE_DETAILS
-----
Local Resources
-----
```

```

ora.ASM DG      OFFLINE OFFLINE      ps5n01base
...

PS5n01base-> crsctl delete res ora.ASM.dg

```

Repeat this step on the remaining nodes (PS5n02base and SS5n01base).

4.4.6 Test the database start-up and shutdown scripts

We describe the manual database takeover using two scripts to start and stop the Oracle instance. The start and stop scripts handle both the Oracle database instance and the associated ASM disk group.

Important: The application start and stop scripts are for demonstration purposes only. The ITSO does not provide any support or warranty for using these scripts. You can develop your own application scripts based on the needs of your environment.

The main actions of the startdb.sh script

Follow these steps:

1. Checking the Oracle service, if it is not started, starts the service.
2. After the ASM daemon is successfully started, the script mounts the database disk group (TESTDG).
3. After the disk group is mounted, the script starts the database instance.
4. After the instance start-up, the script starts the database listener.

The `startdb.sh` script is shown in Example 4-46.

Example 4-46 The startdb.sh script

```

#####start has####
GRID_HOME=/oracle/grid/product/11.2.0/grid
ORACLE_HOME=/oracle/product

$GRID_HOME/bin/crsctl stat res -t
if [ ! $? -eq 0 ];then
    echo "start has " >/tmp/start.out
    $GRID_HOME/bin/crsctl start has
fi
echo "sleep 5..." >>/tmp/start.out
sleep 5

while [ 1 ]
do
    if $GRID_HOME/bin/crsctl stat res ora.asm |grep STATE|grep ONLINE
    then
        sleep 1
        break
    else
        sleep 1
    fi
done

```

```

su - grid -c "sqlplus / as sysasm << EOF
alter diskgroup TESTDG mount;
exit
EOF"

$ORACLE_HOME -p +TESTDG/testdb/spfiletdb.ora -s OPEN
-y MANUAL -a TESTDG -t IMMEDIATE"

###start db###
su - oracle -c "$ORACLE_HOME/bin/srvctl start database -d testdb"
if [ ! $? -eq 0 ];then
    echo "start db failed!!!" >>/tmp/start.out
    exit 1
fi

####start lsnr###
$GRID_HOME/bin/crsctl stat res ora.LISTENER.lsnr | grep STATE | grep ONLINE
if [ ! $? -eq 0 ];then
    echo "start lsnr" >>/tmp/start.out
    $GRID_HOME/bin/crsctl start res ora.LISTENER.lsnr
fi

```

The main actions of the stopdb.sh script

Follow these steps:

1. Stop the database listener.
2. Stop the Oracle database instance.
3. Dismount the database disk group (TESTDG).

The **stopdb.sh** script is shown in Example 4-47.

Example 4-47 The stopdb.sh script

```

GRID_HOME=/oracle/grid/product/11.2.0/grid
ORACLE_HOME=/oracle/product
lv_dbret=0
lv_dgret=0

###shutdown lsnr###
$GRID_HOME/bin/crsctl stop res ora.LISTENER.lsnr

###stop db###
su - oracle -c "$ORACLE_HOME/bin/srvctl stop database -d testdb"
if [ ! $? -eq 0 ];then
    echo "stop db failed!!!" >>/tmp/start.out
    lv_dbret=-1
fi

###dismount diskgroup###
su - grid -c "sqlplus / as sysasm << EOF
alter diskgroup TESTDG dismount;
exit
EOF"

```

Service IP address: During testing, we do not need to add the service IP address to listener.ora because the PowerHA service is not yet started. We will add the service IP address to the Oracle database listener configuration after we finalize the PowerHA configuration.

4.5 PowerHA configuration

We describe the PowerHA configuration in our environment. Figure 4-3 shows the network and mirror group configuration.

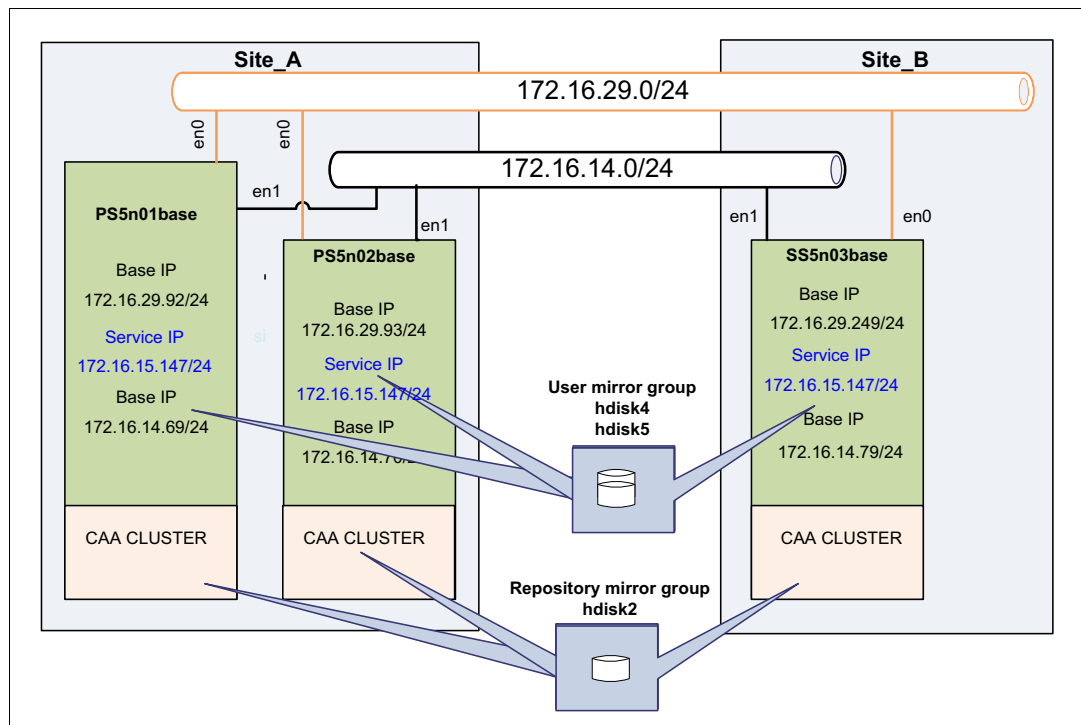


Figure 4-3 Cluster diagram (service IP address will move along with its associated resource group)

We describe the following topics:

- ▶ Configuring cluster topology
- ▶ Configuring cluster resources
- ▶ Configuring the resource groups

4.5.1 Cluster topology

We explain how to configure the cluster topology. The data is shown in Table 4-3 on page 131.

Table 4-3 Attributes of cluster topology

	Site_A		Site_B
Node name	PS5n01base	PS5n02base	SS5n03base
Network interface	en0 PS5n01base 172.16.29.92	en0 PS5n02base 172.16.29.93	en0 SS5n03base 172.16.29.249
Network interface	en1 PS5n01std 172.16.14.69	en1 PS5n02std 172.16.14.70	en1 SS5n03std 172.16.14.79
Network	net_ether_01(172.16.15.0/24 172.16.29.0/24 172.16.14.0/24)		
Service IP address	PS5n01svc 172.16.15.147		
CAA disk	hdisk2 (PVID: 00cf8de64b045b47)		
CAA multicast IP	228.16.29.92 (default)		

Follow these steps:

1. Edit the `/etc/cluster/rhosts` file on each node and add the node information as shown in Example 4-48.

Example 4-48 Content of the `/etc/cluster/rhosts` file

```
# cat /etc/cluster/rhosts
PS5n01base
PS5n02base
SS5n03base
```

2. Restart the `clcomd` daemon on each node with following command:

```
stopsrc -s clcomd;sleep2;startsrc -s clcomd
```

3. Add the cluster information with the following SMIT menu options:

```
smitty hacmp → Cluster Nodes and Networks → Multi Site Cluster Deployment → Setup a Cluster, Nodes and Networks
```

4. Define the CAA Repository Disk and Cluster IP Address with the following menu options:

```
smitty hacmp → Cluster Nodes and Networks → Multi Site Cluster Deployment → Define Repository Disk and Cluster IP Address
```

5. Run “Verify and Synchronize Cluster Configuration”.

During this process, PowerHA creates the CAA configuration among the cluster nodes. You can use the `lsccluster` and `lspv` commands to check the CAA configuration (see Example 4-49).

6. Add the service IP label with the following SMIT menu options:

```
smitty hacmp → Cluster Applications and Resources → Resources → Configure Service IP Labels/Addresses → Add a Service IP Label/Address
```

Example 4-49 Checking the CAA cluster status

```
# lsccluster -m
Calling node query for all nodes...
Node query number of nodes examined: 3
```

```
Node name: PS5n01base
```

Cluster shorthand id for node: 1
 UUID for node: 70efd1d4-4421-11e2-abf3-4acabba5d00b
State of node: UP NODE_LOCAL
 Smoothed rtt to node: 0
 Mean Deviation in network rtt to node: 0
 Number of clusters node is a member in: 1

CLUSTER NAME	SHID	UUID
ASMCluster	0	70e9af3e-4421-11e2-abf3-4acabba5d00b
SITE NAME	SHID	UUID
LOCAL	1	51735173-5173-5173-5173-517351735173

Points of contact for node: 0

Node name: PS5n02base

Cluster shorthand id for node: 2
 UUID for node: 70efc4aa-4421-11e2-abf3-4acabba5d00b
State of node: UP
 Smoothed rtt to node: 7
 Mean Deviation in network rtt to node: 3
 Number of clusters node is a member in: 1

CLUSTER NAME	SHID	UUID
ASMCluster	0	70e9af3e-4421-11e2-abf3-4acabba5d00b
SITE NAME	SHID	UUID
LOCAL	1	51735173-5173-5173-5173-517351735173

Points of contact for node: 3

Interface	State	Protocol	Status
dpcom	DOWN	none	RESTRICTED
en0	UP	IPv4	none
en1	UP	IPv4	none

Node name: SS5n03base

Cluster shorthand id for node: 3
 UUID for node: 70efb280-4421-11e2-abf3-4acabba5d00b
State of node: UP
 Smoothed rtt to node: 7
 Mean Deviation in network rtt to node: 3
 Number of clusters node is a member in: 1

CLUSTER NAME	SHID	UUID
ASMCluster	0	70e9af3e-4421-11e2-abf3-4acabba5d00b
SITE NAME	SHID	UUID
LOCAL	1	51735173-5173-5173-5173-517351735173

Points of contact for node: 3

Interface	State	Protocol	Status
dpcom	DOWN	none	RESTRICTED
en0	UP	IPv4	none
en1	UP	IPv4	none

```

# lspv
hdisk0          00f681f3d99b7740          rootvg          active
hdisk1          00f681f3839d94c8          None
hdisk2          00cf8de64b045b47          caavg_private  active
hdisk3          00f681f36bf9b43e          None
hdisk4          00f681f36bf9b47e          None
hdisk5          00f681f36bf9b4b4          None

```

4.5.2 Cluster resources

The following examples are cluster resources:

- ▶ Storage (shared disks)
- ▶ Application controllers (applications' start and stop scripts)
- ▶ Service IP addresses

These resources will be kept highly available by PowerHA when they are grouped into resource groups.

Important: Our test configuration is an Extended Distance cluster, based on PowerHA SystemMirror Enterprise Edition for AIX. There is specific data that needs to be entered in the cluster resource configuration:

- ▶ Storage definition
- ▶ Mirror groups definition
- ▶ Application controller definition (formerly "Application server")

Storage definition

In our scenario, we use two DS8800 storage subsystems. We add these DS8800 storage subsystems as storage resources into the PowerHA configuration by using the following SMIT menu options, based on the information listed in Table 4-4:

smitty hacmp → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Storage Systems → Add a Storage System

Table 4-4 Attribute of storage Metro Mirror (in-band) resources

	Primary storage	Secondary storage
Storage system name	DS8805	DS8803
Site association	Site_A	Site_B
Vendor-specific identifier	IBM.2107-00000XP411	IBM.2107-00000WT971
Worldwide node name (WNN)	500507630BFFC4C8	500507630BFFC1E2

Mirror groups definition

Shared disks that are used for CAA, the user application, or the rootvg need to be configured into PowerHA mirror groups. To add a mirror group, use the following SMIT menu options:

smitty hacmp → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Mirror Groups → Add a Mirror Group

In our testing, we use one *user mirror group* and one *cluster repository mirror group*. The attributes are shown in Table 4-5 and Table 4-6.

Table 4-5 Attributes of the cluster user mirror group

	User mirror group
Mirror group name	dbmg
Raw disks	702f1177-cb29-4ca6-66b6-081eab05e21d (hdisk4) 5e4e0bae-1892-6c0c-ed47-49375d03782e (hdisk5)
HyperSwap	Enabled
Consistency Group	Enabled
Unplanned HyperSwap timeout (in seconds)	60s (default)
HyperSwap priority	Medium
Recovery action	Automatic

Table 4-6 Attributes of the cluster repository mirror group

	Cluster_Repository mirror group
Mirror group name	repmg
Site name	Site_A Site_B
NonHyperSwap disk	381fac72-bf73-407a-fb62-c9c178ebfa14 (hdisk2)
HyperSwap disk	381fac72-bf73-407a-fb62-c9c178ebfa14 (hdisk2)
HyperSwap	Enabled
Consistency Group	Enabled
Unplanned HyperSwap timeout (in seconds)	60s (default)
HyperSwap priority	High

Application controller definition

In our test environment, we use one application controller to handle the Oracle database start and stop. To configure the scripts in the application controller, use the following SMIT menu options:

smitty hacmp → Cluster Applications and Resources → Resources → Configure User Applications (Scripts and Monitors) → Application Controller Scripts → Add Application Controller Scripts

The attributes of the application controller are listed in Table 4-7 on page 135. The actual content of the start script is shown in Example 4-46 on page 128 and the content of the stop script is shown in Example 4-47 on page 129.

Table 4-7 Attributes of the cluster application controller

	Application controller attributes
Application controller name	dbcontrol
Start script	/home/startdb.sh
Stop script	/home/stopdb.sh
Application monitor names	None
Application startup mode	Background (default)

4.5.3 Resource group configuration

We configure one resource group (RG) named dborarg. The attributes of the RG are listed in Table 4-8. We define the RG by using the following SMIT fast path:

smitty hacmp → Cluster Applications and Resources → Resource Groups → Add a Resource Group

After the RG is defined, we must populate the RG with the corresponding resources. We change the RG attributes by using the “Change/Show Resources and Attributes for a Resource Group” menu.

Storage resources: In this scenario, Oracle ASM uses raw disks to store database files. We need to configure the storage as Raw Disk Universally Unique Identifiers (UUIDs)/hdisks that belong to the user mirror group that we previously defined.

Table 4-8 Attributes of the cluster resource group

	Resource group attributes
Resource group name	dborarg
Intersite management policy	Prefer Primary Site
Participating nodes from primary site	PS5n01base PS5n02base
Participating nodes from secondary site	SS5n03base
Startup policy	Online On Home Node Only
Fallover policy	Fallover To Next Priority Node In The List
Fallback policy	Never Fallback
Service IP labels/addresses	PS5n01svc
Application controller name	dbcontrol
Raw disk UUIDs/hdisks	5e4e0bae-1892-6c0c-ed47-49375d03782e 702f1177-cb29-4ca6-66b6-081eab05e21d
DS8000 Metro Mirror (In-band) resources	dbmg

After the RG configuration is complete, run **Verify and Synchronize Cluster Configuration**.

Cluster synchronization: During the cluster configuration synchronization process, you might encounter some error and warning messages.

It is important that you read these messages carefully and take appropriate actions to fix the causes of these issues. Contact IBM PowerHA technical support if necessary.

4.6 Test scenarios

Restriction: The test results that are presented in this section are specific to our test environment. Your test results might vary based on your specific configuration.

We cover the following test scenarios:

- ▶ Node maintenance (planned)
- ▶ Primary storage maintenance (planned)
- ▶ Primary site maintenance (planned)
- ▶ Node failure (unplanned)
- ▶ Primary storage failure (unplanned)
- ▶ Primary site failure (unplanned)
- ▶ PPRC replication path failure (unplanned)

For each scenario, we provide the detailed testing procedure and results.

4.6.1 Node maintenance (planned)

In this test scenario, we shut down one node for maintenance and move the application onto another (available) node. The scope of this action is to improve the application availability by reducing the interrupt time (downtime).

Testing behavior expectation

We perform a “Move Resource Groups to Another Node” from the PowerHA SMIT menu. During the RG movement, the application will be stopped on the current node and started on the target node, so the application cannot provide service for a brief period. After we finalize the RG movement, the (application) service is recovered.

Display the current status

Follow these steps:

1. Display the PowerHA resource group status.

In the initial state, the RG (dborarg) is online on node PS5n01base as shown in Example 4-50 on page 137. The PowerHA service is in ST_STABLE status on all nodes.

Example 4-50 Current resource group status of planned node maintenance scenario

```
# clRGinfo
```

Group Name	State	Node
dborarg	ONLINE	PS5n01base@Sit
	OFFLINE	PS5n02base@Sit
	ONLINE SECONDARY	SS5n03base@Sit

2. Display the application status.

The Oracle database is running on node PS5n01base. We use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations. Example 4-51 shows the I/O throughput with the application running.

Example 4-51 Current I/O running status of application

```
PS5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk
```

...						
hdisk4	0.0	320.0	20.0	320	0	11:19:29
hdisk5	7.0	8170.0	97.0	777	7393	11:19:30
hdisk4	4.0	7657.0	63.0	408	7249	11:19:30
hdisk5	1.0	4336.0	16.0	0	4336	11:19:31
...						

RG manual move from PS5n01base to PS5n02base

Use the following SMIT path *or* execute the command that is shown in Example 4-52:

smitty hacmp → System Management (C-SPOC) → Resource Group and Applications → Move Resource Groups to Another Node

Example 4-52 Command to perform the RG movement

```
PS5n01base#/usr/es/sbin/cluster/utilities/clRGmove -s 'false' -m -i -g 'dborarg' -n 'PS5n02base'
```

Test results

Follow these steps:

1. Display the PowerHA cluster.log file information.

Example 4-53 shows the PowerHA actions in this testing scenario. The testing begins at **11:20:29** and ends at **11:20:44**. The total RG move time is 15 seconds.

Tip: Depending on the workload conditions and the number of configured disks, the RG move time varies.

Example 4-53 PowerHA cluster.log for planned node maintenance

```
Dec 13 11:20:29 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT START: rg_move_fence PS5n01base 1
Dec 13 11:20:30 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: rg_move_fence PS5n01base 1 0
Dec 13 11:20:32 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT START: rg_move_fence PS5n01base 1
```

```
Dec 13 11:20:32 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
rg_move_fence PS5n01base 1 0
...
Dec 13 11:20:44 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT START:
external_resource_state_change_complete PS5n02base
Dec 13 11:20:44 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
external_resource_state_change_complete PS5n02base 0
```

2. Display the PowerHA resource group movement result.

After the RG move ends, the RG is online on the PS5n02base node, as shown in Example 4-54.

Example 4-54 Resource group status as a result of the planned node maintenance scenario

```
# clRGinfo
-----
Group Name      State                Node
-----
dborarg          OFFLINE              PS5n01base@Sit
                 ONLINE              PS5n02base@Sit
                 ONLINE SECONDARY    SS5n03base@Sit
```

3. Check the application.

During the RG move, the Oracle database is stopped (shut down) on node PS5n01base and restarted on node PS5n02base. In our testing, the application did not provide service from **11:20:15** to **11:21:02**, for a total time of 57 seconds (see Example 4-55).

Example 4-55 Application running result of planned node maintenance scenario

```
On PS5n01base node:
PS5n01base# iostat -T hdisk4 hdisk5|grep hdisk
...
hdisk5          2.0      4188.0      69.0        768      3420  11:20:14
hdisk4          1.0      1616.0      44.0        720       896  11:20:14
hdisk5          1.0       208.0      19.0         84       124  11:20:15
hdisk4          0.0       100.0       7.0          48         52 11:20:15
hdisk5          0.0         0.0         0.0           0           0  11:20:16
..
On PS5n02base node:
PS5n02base# iostat -T hdisk4 hdisk5|grep hdisk
...
hdisk5          0.0         0.0         0.0           0           0  11:21:01
hdisk4          0.0         4.0         1.0           0           4  11:21:01
hdisk5          9.0      5690.0     471.0       5064       626 11:21:02
hdisk4          7.0      3910.0     328.0       3597       313  11:21:02
hdisk5         13.0     4913.0     561.0       4840         73  11:21:03
```

Testing scenario summary

Based on the testing results, we see that the resource group can be moved in a HyperSwap environment in the usual manner. (HyperSwap does not influence the behavior in this case.) This is the normal resource group takeover scenario.

4.6.2 Primary storage maintenance (planned)

This scenario tests primary storage maintenance in a HyperSwap environment and shows how HyperSwap can help you avoid application interruption.

Testing behavior expectation

The basic function of Power HyperSwap allows applications to continue running without any interruption when one of the replicated storage subsystems cannot provide service. The expectation is that the application is not affected in this scenario.

Display current status

Follow these steps:

1. Display the PowerHA resource group status.

Example 4-56 shows that the resource group is ONLINE on node PS5n02base.

Example 4-56 Current resource group status of planned primary storage maintenance scenario

```
# clRGinfo
-----
Group Name      State                Node
-----
dborarg         OFFLINE             PS5n01base@Sit
                ONLINE             PS5n02base@Sit
                ONLINE SECONDARY   SS5n03base@Sit
```

2. Display the application status.

The Oracle database is running on node PS5n02base. We use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations. Example 4-57 shows the I/O throughput with the application running.

Example 4-57 Current I/O running status of application

```
PS5n02base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5          0.0    4505.0    84.0      825    3680 12:09:32
hdisk4          0.0    4323.0    46.0      296    4027 12:09:32
hdisk5          0.0   13688.0   509.0      48    13640 12:09:33
hdisk4          0.0   13719.0   602.0      24    13695 12:09:33
...
```

3. Display the Metro Mirror (PPRC) status.

Example 4-58 shows the PPRC status and path group with the AIX `lspprc` command. The primary path of the LUNs is pointing to the primary storage (DS8805). Two HBAs (fcs1 and fcs4) are used to access the storage. The other two adapters (fcs2 and fcs3) are configured to access the secondary storage (DS8803) and are, therefore, not active at this time.

Example 4-58 Current PPRC status of planned primary storage maintenance scenario

```
# lspprc -Ao
hdisk#  PPRC      Primary      Secondary      Primary Storage      Secondary Storage
         state    path group   path group      WWNN                WWNN
         ID      ID
hdisk3  Active    0(s)        1                500507630bffc4c8    500507630bffc1e2
hdisk5  Active    0(s)        1                500507630bffc4c8    500507630bffc1e2
```

```

hdisk4   Active  0(s)      1          500507630bffc4c8 500507630bffc1e2
hdisk2   Active  0(s)      1          500507630bffc4c8 500507630bffc1e2

```

```
# lsprrc -p hdisk4
```

```

path      WWNN          LSS  VOL    path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xab 0x00  PRIMARY
1         500507630bffc1e2 0xab 0x00  SECONDARY

```

```

path      path path      parent connection
group id id   status
=====
0         0   Enabled fscsi1 500507630b1884c8,40ab400000000000
0         1   Enabled fscsi4 500507630b5304c8,40ab400000000000
1         2   Enabled fscsi2 500507630b1001e2,40ab400000000000
1         3   Enabled fscsi3 500507630b1301e2,40ab400000000000

```

Swap user mirror group and repository mirror group manually

We perform the mirror groups (MG) swap by using the following SMIT menu options *or* by running the command that is shown in Example 4-59.

User mirror group swap

Use this command and these menu selections:

```
smitty hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage User Mirror Group(s)
```

Repository mirror group swap

Use this command and these menu selections:

```
smitty hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage Cluster Repository Mirror Group
```

Example 4-59 Command to perform mirror group swap

```

PS5n01base# /usr/es/sbin/cluster/xd_generic/xd_cli/cl_clxd_manage_mg_smit -t 'user'
-m 'dbmg' -o 'swap'
PS5n01base# /usr/es/sbin/cluster/xd_generic/xd_cli/cl_clxd_manage_mg_smit -t
'repository' -m 'repmg' -o 'swap'

```

Test results for swapping the user mirror group

In this testing, we swap the user mirror group and the repository mirror group one at a time (in sequence).

1. Display the PowerHA `c1xd` daemon log file information.

The PowerHA `c1xd` daemon log shown in Example 4-60 on page 141 indicates that the swap procedure for the user MG finished within seconds.

Example 4-60 PowerHA clxd daemon log of the planned swap of the user mirror group

```

INFO      |2012-12-13T12:12:15.095565|MG Name='dbmg'
INFO      |2012-12-13T12:12:15.095581|MG Mode='Synchronous'
INFO      |2012-12-13T12:12:15.095597|CG Enabled = 'Yes'
INFO      |2012-12-13T12:12:15.095611|Recovery Action = 'Automatic'
INFO      |2012-12-13T12:12:15.095627|Vendor's unique ID =
INFO      |2012-12-13T12:12:15.095642|Printing Storage System Set @(0x200d4710)
INFO      |2012-12-13T12:12:15.095658|Num Storage System: '2'
INFO      |2012-12-13T12:12:15.095674|Storage System Name = 'DS8805'
INFO      |2012-12-13T12:12:15.095689|Storage System Name = 'DS8803'
INFO      |2012-12-13T12:12:15.095705|Printing Opaque Attribute Value Set ...
@(0x2014d788)
INFO      |2012-12-13T12:12:15.095720|Num Opaque Attributes Values = '0'
INFO      |2012-12-13T12:12:15.095736|Hyperswap Policy = Enabled
INFO      |2012-12-13T12:12:15.095752|MG Type = user
INFO      |2012-12-13T12:12:15.095768|Hyperswap Priority = medium
INFO      |2012-12-13T12:12:15.095783|Unplanned Hyperswap timeout = 20
INFO      |2012-12-13T12:12:15.095799|RawDisks =
fe7e3fe2-e418-7de4-0206-adbb5f6a2a51
INFO      |2012-12-13T12:12:15.095824|RawDisks =
fb4e7243-a7e7-ed0e-c6db-e384bcf74631
INFO      |2012-12-13T12:12:15.094742|Received XD CLI request = '' (0x1d)
INFO      |2012-12-13T12:12:17.094930|Received XD CLI request = 'Swap Mirror
Group' (0x1c)
INFO      |2012-12-13T12:12:17.094950|Request to Swap Mirror Group 'dbmg',
Direction 'siteB', Outfile ''
WARNING   |2012-12-13T12:12:17.097689|Not able to find any VG disks for MG=dbmg
INFO      |2012-12-13T12:12:17.171791|err_num = 0, retval = 0, errno=10
INFO      |2012-12-13T12:12:17.172148|Swap Mirror Group 'dbmg'
completed.

```

2. Display the PowerHA resource group status.

During the storage swap, the resource group is not affected. The resource group remains ONLINE on the PS5n02base node as shown in Example 4-61.

Example 4-61 Resource group status result of the planned swap of the user mirror group

```

# clRGinfo
-----
Group Name      State                Node
-----
dborarg         OFFLINE              PS5n01base@Sit
                 ONLINE              PS5n02base@Sit
                 ONLINE SECONDARY    SS5n03base@Sit

```

3. Display the application status.

During the storage swap, the `ioostat` command output shows that the Oracle database is not affected, as shown in Example 4-62 on page 142.

Example 4-62 Application running result of the planned swap of the user mirror group

```
PS5n02base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5          0.0    1600.0    40.0      576    1024  12:12:16
hdisk4          0.0    1510.0    22.0      288    1222  12:12:16
hdisk5          0.0    15426.0   113.0     296    15130  12:12:17
hdisk4          0.0    12513.0   107.0     249    12264  12:12:17
hdisk5          0.0     5349.0    70.0      48     5301  12:12:18
hdisk4          0.0     7007.0    71.0      24     6983  12:12:18
hdisk5          0.0    12106.0   357.0      0    12106  12:12:19
hdisk4          0.0    11693.0   353.0      0    11693  12:12:19
...
```

4. Display the Metro Mirror (PPRC) status.

After the user mirror group (MG) swap, the `lsprrc` command output shows that the primary path for `hdisk4` and `hdisk5` (the ASM disk group, which is used for the Oracle database) is changed from `DS8805 (Storage_A)` to `DS8803 (Storage_B)`, as shown in Example 4-63.

Example 4-63 PPRC result of the planned swap of the user MG

```
# lsprrc -Ao
hdisk#    PPRC    Primary    Secondary    Primary Storage    Secondary Storage
          state   path group  path group  WWNN              WWNN
          ID     ID
hdisk5    Active  1(s)      0           500507630bffc1e2  500507630bffc4c8
hdisk4    Active  1(s)      0           500507630bffc1e2  500507630bffc4c8
hdisk2    Active  0(s)      1           500507630bffc4c8  500507630bffc1e2
hdisk3    Active  0(s)      1           500507630bffc4c8  500507630bffc1e2

# lsprrc -p hdisk4
path      WWNN              LSS  VOL  path
group id  WWNN              LSS  VOL  group status
=====
0         500507630bffc4c8 0xab 0x00 SECONDARY
1(s)     500507630bffc1e2 0xab 0x00 PRIMARY

path      path path      parent connection
group id  id  status
=====
0 0 Enabled fscsi1 500507630b1884c8,40ab400000000000
0 1 Enabled fscsi4 500507630b5304c8,40ab400000000000
1 2 Enabled fscsi2 500507630b1001e2,40ab400000000000
1 3 Enabled fscsi3 500507630b1301e2,40ab400000000000
```

Test results for swapping the repository mirror group

Follow these steps:

1. Display the PowerHA `c1xd` daemon's log.

Example 4-64 on page 143 shows the PowerHA HyperSwap repository mirror group swap.

Example 4-64 PowerHA clxd daemon's log of the planned swap of the repository mirror group

```

INFO      |2012-12-13T12:18:33.368218|Received XD CLI request = 'List Mirror Group' (0xc)
INFO      |2012-12-13T12:18:33.368705|MG Name='repmg'
INFO      |2012-12-13T12:18:33.368722|MG Mode='Synchronous'
INFO      |2012-12-13T12:18:33.368739|CG Enabled = 'Yes'
INFO      |2012-12-13T12:18:33.368754|Recovery Action = 'Manual'
INFO      |2012-12-13T12:18:33.368771|Vendor's unique ID =
INFO      |2012-12-13T12:18:33.368787|Printing Storage System Set @(0x200f0710)
INFO      |2012-12-13T12:18:33.368807|Num Storage System: '2'
INFO      |2012-12-13T12:18:33.368824|Storage System Name = 'DS8805'
INFO      |2012-12-13T12:18:33.368839|Storage System Name = 'DS8803'
INFO      |2012-12-13T12:18:33.368855|Printing Opaque Attribute Value Set ... @(0x2012a308)
...
INFO      |2012-12-13T12:18:33.063286|Received XD CLI request = '' (0x1d)
INFO      |2012-12-13T12:18:33.067744|Received XD CLI request = 'Swap Mirror Group' (0x1c)
INFO      |2012-12-13T12:18:33.067765|Request to Swap Mirror Group 'repmg', Direction
'siteA', Outfile ''
ERROR     |2012-12-13T12:18:33.069046|Failed to get rg name record from ODM
'HACMPresource'. odmerrno=0 for mg repmg
WARNING   |2012-12-13T12:18:33.069067|Not able to find any RG for MG repmg
WARNING   |2012-12-13T12:18:33.070260|Not able to find any VG disks for MG=repmg
INFO      |2012-12-13T12:18:35.642922|err_num = 0, retval = 0, errno=10
INFO      |2012-12-13T12:18:35.643098|Swap Mirror Group 'repmg' completed.

```

2. Display the application status.

During the repository MG swap, the application is not affected, as shown in Example 4-65.

Example 4-65 Application running result of the planned swap of the repository mirror group

```

PS5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5      0.0    8631.0    35.0      64    8567 12:18:34
hdisk4      0.0   10105.0    67.0      8    10097 12:18:34
hdisk5      0.0    5976.0    26.0      0    5976 12:18:35
hdisk4      0.0    5447.0    21.0      0    5447 12:18:35
hdisk5      0.0   13216.0   354.0      0   13216 12:18:36
hdisk4      0.0   13975.0   404.0      8   13967 12:18:36
hdisk5      0.0    6019.0    32.0     64    5955 12:18:37
hdisk4      0.0    6061.0    22.0      0    6061 12:18:37
...

```

3. Display the Metro Mirror (PPRC) status.

After the repository MG swap, the primary path of the repository disk changed from DS8805 (Storage_A) to DS8803 (Storage_B), as shown in Example 4-66.

Example 4-66 PPRC result of the planned swap of the repository mirror group

```

# lsprrc -Ao
hdisk#    PPRC      Primary      Secondary    Primary Storage    Secondary Storage
          state    path group   path group   WWNN              WWNN
          ID      ID
hdisk5    Active    1(s)        0            500507630bffc1e2  500507630bffc4c8
hdisk4    Active    1(s)        0            500507630bffc1e2  500507630bffc4c8
hdisk2    Active    1(s)        0            500507630bffc1e2  500507630bffc4c8

# lsprrc -p hdisk2
path      WWNN              LSS  VOL  path

```

group id					group status
0	500507630bffc4c8	0xaa	0x00		SECONDARY
1(s)	500507630bffc1e2	0xaa	0x00		PRIMARY

path group id	path id	path status	parent	connection
0	0	Enabled	fscsi1	500507630b1884c8,40aa400000000000
0	1	Enabled	fscsi4	500507630b5304c8,40aa400000000000
1	2	Enabled	fscsi2	500507630b1001e2,40aa400000000000
1	3	Enabled	fscsi3	500507630b1301e2,40aa400000000000

Testing scenario summary

Based on the test results, we can see that there is no impact to the application when swapping the user mirror group and the repository mirror group. After the swap, the primary storage (DS8805) can be taken offline for maintenance.

Important: Before the storage maintenance, pause PPRC for the replicated LUNs first. After the storage is recovered, resume PPRC manually.

4.6.3 Primary site maintenance (planned)

This scenario describes the actions for planned maintenance in the primary site. Both cluster nodes and the storage in the primary site will be taken offline. The goal of this scenario is to avoid or reduce application interruption during site maintenance.

Testing behavior expectation

This scenario consists of the resource group movement with the storage swap to the secondary site. Because the resource group needs to be moved to the target node on the secondary site, the application will be affected for a brief period. However, a storage swap typically does not affect the application that is running.

Important: Like with any clustering environment, events are handled in the sequence that they are detected by the clustering mechanisms. Do not try to issue multiple cluster changes in parallel, because they might affect the cluster recovery and stabilization time.

Display the current status

Follow these steps:

1. Display the PowerHA resource group status.

Example 4-67 on page 145 shows that the resource group (dborarg) is ONLINE on node PS5n02base.

Example 4-67 RG status before the planned primary site maintenance

```
# clRGinfo
```

Group Name	State	Node
dborarg	OFFLINE	PS5n01base@Sit
	ONLINE	PS5n02base@Sit
	ONLINE SECONDARY	SS5n03base@Sit

2. Display the application status.

The Oracle database is running on node PS5n02base. We use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations.

Example 4-68 shows the I/O throughput with the application running.

Example 4-68 Current I/O status for the application disks

```
PS5n02base#iostat -T hdisk4 hdisk5 1|grep hdisk
```

```
...
```

hdisk5	0.0	13096.0	168.0	0	13096	12:26:11
hdisk4	0.0	14131.0	149.0	8	14123	12:26:11
hdisk5	0.0	4416.0	35.0	320	4096	12:26:12
hdisk4	0.0	3826.0	27.0	160	3666	12:26:12
hdisk5	0.0	4971.0	78.0	600	4371	12:26:13

```
...
```

3. Display the Metro Mirror (PPRC) status.

Example 4-69 on page 146 shows the current PPRC status and path group. The primary path of the LUNs points to the primary storage (DS8805). The two HBAs (fcs1 and fcs4) are used for storage access. The other two adapters (fcs2 and fcs3) are used to access the secondary storage (DS8803).

Example 4-69 Current PPRC status of the planned primary site maintenance scenario

```
# lspprc -Ao
hdisk#    PPRC      Primary      Secondary    Primary Storage  Secondary Storage
          state    path group   path group   WWNN           WWNN
          ID      ID
hdisk3    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2
hdisk5    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2
hdisk4    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2
hdisk2    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2

# lspprc -p hdisk4
path      WWNN                LSS  VOL    path
group id                                group status
=====
0(s)      500507630bffc4c8  0xab 0x00   PRIMARY
1         500507630bffc1e2  0xab 0x00   SECONDARY

path      path path      parent connection
group id  id   status
=====
0  0    Enabled  fscsi1 500507630b1884c8,40ab400000000000
0  1    Enabled  fscsi4 500507630b5304c8,40ab400000000000
1  2    Enabled  fscsi2 500507630b1001e2,40ab400000000000
1  3    Enabled  fscsi3 500507630b1301e2,40ab400000000000
```

Perform the user mirror group and the repository mirror group swap

Perform the user MG and the repository MG swap using the following SMIT menu options, *or* execute the command that is shown in Example 4-70.

User MG swap

Use this command and these options:

smitty hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage User Mirror Group(s)

Repository MG swap

Use this command and these options:

smitty hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage Cluster Repository Mirror Group

Example 4-70 Command to perform the mirror group swap

```
PS5n01base# /usr/es/sbin/cluster/xd_generic/xd_cli/cl__manage_mg_smit -t 'user' -m
'dbmg' -o 'swap'
PS5n01base# /usr/es/sbin/cluster/xd_generic/xd_cli/cl_clxd_manage_mg_smit -t
'repository' -m 'repmg' -o 'swap'
```

During the MG swap, the application is not affected and continues running on the current node. For detailed information, see “Test results for swapping the repository mirror group” on page 142.

Move a resource group to another site

Switch the resource groups to another site by using the following SMIT path or by executing the command that is shown in Example 4-71:

smitty hacmp → **System Management (C-SPOC)** → **Resource Group and Applications** → **Move Resource Groups to Another Site**

Example 4-71 Command to move the resource group to another site

```
/usr/es/sbin/cluster/utilities/clRGmove -s 'false' -x -i -g 'dborarg' -n 'Site_B'
```

Testing result of move resource groups to another site

Follow these steps:

1. Display the PowerHA cluster.log file information.

Example 4-72 shows the PowerHA activities during this test scenario.

Example 4-72 PowerHA cluster.log of the planned site maintenance scenario

```
Dec 13 12:27:24 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT START:
external_resource_state_change PS5n01base
Dec 13 12:27:24 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
external_resource_state_change PS5n01base 0
...
Dec 13 12:28:09 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
start_server dbcontrol 0
Dec 13 12:28:10 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
rg_move_complete PS5n01base 1 0
Dec 13 12:28:12 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT START:
external_resource_state_change_complete PS5n01base
Dec 13 12:28:12 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
external_resource_state_change_complete PS5n01base 0
```

2. Display the PowerHA resource group movement result.

After the resource group movement ends, the RG becomes ONLINE on the SS5n03base node, which is in the secondary site (see Example 4-73).

Example 4-73 Resource group result of the planned primary site maintenance scenario

```
# clRGinfo
-----
Group Name      State                Node
-----
dborarg         ONLINE SECONDARY    PS5n01base@Sit
                OFFLINE              PS5n02base@Sit
                ONLINE               SS5n03base@Sit
```

3. Display the application status.

During the RG move, the Oracle database is stopped (shut down) on node PS5n02base and restarted on node SS5n03base. In our testing, the application did not provide service from 12:27:18 to 11:28:29, as shown in Example 4-74 on page 148.

Example 4-74 Application running result of the planned primary site maintenance scenario

```
PS5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5      0.0    3148.0    113.0      896      2252 12:27:16
hdisk4      0.0    1296.0     48.0      688       608 12:27:16
hdisk5      0.0    160.0     10.0       80        80 12:27:17
hdisk4      0.0    116.0     11.0       48         68 12:27:17
hdisk5      0.0     0.0       0.0        0          0 12:27:18
hdisk4      0.0     4.0       1.0         0           4 12:27:18
...

hdisk4      0.0     8.0       2.0         0            8 12:28:29
hdisk5      0.0     0.0       0.0         0            0 12:28:29
hdisk4      0.0   3078.0    229.0     2781       297 12:28:30
hdisk5      0.0   4417.0    343.0     3776       641 12:28:30
```

4. Display the Metro Mirror (PPRC) status.

After the user MG and the repository MG swap, the `lsprrc` command output shows that the primary path of hdisk4 and hdisk5 (used for the Oracle database in user mirror group) and hdisk2 (used for CAA in the repository mirror group) changed from DS8805 to DS8803, as shown in Example 4-75.

Example 4-75 PPRC status as a result of the planned primary site maintenance scenario

```
# lsprrc -Ao
hdisk#   PPRC   Primary   Secondary   Primary Storage   Secondary Storage
         state  path group path group   WWNN              WWNN
         ID    ID
hdisk5   Active  1(s)     0           500507630bffc1e2  500507630bffc4c8
hdisk4   Active  1(s)     0           500507630bffc1e2  500507630bffc4c8
hdisk2   Active  1(s)     0           500507630bffc1e2  500507630bffc4c8
hdisk3   Active  0(s)     1           500507630bffc4c8  500507630bffc1e2

# lsprrc -p hdisk2
path     WWNN              LSS  VOL  path
group id  ID                ID   ID   group status
=====
0        500507630bffc4c8  0xaa 0x00  SECONDARY
1(s)    500507630bffc1e2  0xaa 0x00  PRIMARY

path     path  path     parent  connection
group id id    status
=====
0        0     Enabled  fscsi1  500507630b1884c8,40aa400000000000
0        1     Enabled  fscsi4  500507630b5304c8,40aa400000000000
1        2     Enabled  fscsi2  500507630b1001e2,40aa400000000000
1        3     Enabled  fscsi3  500507630b1301e2,40aa400000000000
```

Testing scenario summary

If you want to perform site maintenance, there are two necessary steps. The first step is to move the resource groups from nodes in the primary site to nodes in the secondary site. This step will result in a short application outage. The second step is to swap both the user MG and the repository MG from the primary site storage to the secondary site storage. The application can provide continuous service during this process.

Important: Before you perform storage maintenance, pause PPRC for the replicated LUNs first. After the storage is recovered, resume PPRC manually.

4.6.4 Node failure (unplanned)

This scenario describes what happened in our test environment when the node with the ONLINE RG fails (unplanned).

Testing behavior expectation

In this scenario, PowerHA detects the node failure and brings up the resource group on another (available) node based on the resource group policy. During this process, the application cannot provide service until the resource group is brought back ONLINE.

Display the current status

Follow these steps:

1. Display the PowerHA resource group status.

Example 4-76 shows the resource group ONLINE on node PS5n01base.

Example 4-76 Current resource group status of the unplanned node failure scenario

```
# clRGinfo
-----
Group Name      State                Node
-----
dborarg         ONLINE              PS5n01base@Sit
                 OFFLINE             PS5n02base@Sit
                 ONLINE SECONDARY   SS5n03base@Sit
```

2. Display the application status.

The Oracle database is running on node PS5n01base, and we use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations. Example 4-77 shows the I/O throughput with the application running.

Example 4-77 Current I/O running status of the application

```
PS5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk4          0.0    8544.0    71.0      457    8087 12:40:48
hdisk5          0.0    4034.0    15.0       0    4034 12:40:49
hdisk4          0.0    4415.0    16.0       0    4415 12:40:51
hdisk5          0.0   19512.0    66.0       0   19512 12:40:52
...
```

Perform the halt -q command on node PS5n01base

Issuing the `halt -q` command on PS5n01base causes a quick halt, triggering a node failure detection by the surviving cluster nodes. This detection results in PowerHA activating the RG on a surviving node based on the RG policy.

Test results

Follow these steps:

1. Display the PowerHA cluster.log file information.

Example 4-78 shows a part of the cluster.log, after the PS5n01base node is halted. PowerHA detects the failure and triggers the RG takeover. In the end, the resource group is brought ONLINE on node PS5n02base.

Example 4-78 PowerHA cluster.log for the unplanned node failure scenario

```
Dec 13 12:42:10 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT START: node_down
PS5n01base
Dec 13 12:42:10 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: node_down
PS5n01base 0
...
Dec 13 12:42:26 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
rg_move_acquire PS5n02base 1 0
Dec 13 12:42:26 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT START:
rg_move_complete PS5n02base 1
Dec 13 12:42:27 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT START: start_server
dbcontrol
Dec 13 12:42:27 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
start_server dbcontrol 0
...
Dec 13 12:42:30 PS5n02base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED:
node_down_complete PS5n01base 0
```

2. Display the PowerHA resource group movement status.

Example 4-79 shows the resource group (dborarg) online on node PS5n02base.

Example 4-79 Resource group status result of the unplanned node failure scenario

```
# clRGinfo
-----
Group Name      State                Node
-----
dborarg         OFFLINE              PS5n01base@Sit
                 ONLINE              PS5n02base@Sit
                 ONLINE SECONDARY    SS5n03base@Sit
```

3. Display the application status.

Example 4-80 on page 151 shows the online resources on node PS5n02base.

Example 4-80 Oracle service status on PS5n02base

```
$ crsctl stat res -t
```

NAME	TARGET	STATE	SERVER	STATE_DETAILS

Local Resources				

ora.LISTENER.lsnr				
	ONLINE	ONLINE	ps5n02base	
ora.TESTDG.dg				
	ONLINE	ONLINE	ps5n02base	
...				
ora.testdb.db				
1	ONLINE	ONLINE	ps5n02base	Open

Example 4-81 shows that the application is interrupted at 12:42:01 on node P5n01base and continues at 12:42:50 on node P5n02base.

Example 4-81 Application running result of the unplanned node failure scenario

```
P5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk  
...  
hdisk5      0.0      816.0      51.0        800         16 12:42:00  
hdisk4      0.0      160.0       10.0         160          0 12:42:00  
hdisk5      0.0     10978.0      87.0         505       10473 12:42:01  
hdisk4      0.0     10837.0      66.0         264       10573 12:42:01  
Hal ted here
```

```
P5n02base#iostat -T hdisk4 hdisk5 1|grep hdisk  
...  
hdisk5      0.0         0.0         0.0          0          0 12:42:49  
hdisk4      0.0         4.0         1.0          0          4 12:42:49  
hdisk5      0.0      988.0        59.0         628         360 12:42:50  
hdisk4      0.0      772.0       108.0         540         232 12:42:50  
hdisk5      0.0     95546.0      328.0       95542          4 12:42:51  
hdisk4      0.0     95031.0      303.0       95031          0 12:42:51  
...
```

Testing scenario summary

Based on the test results, PowerHA behaves normally if one node goes down in the HyperSwap environment. PowerHA will start the resource group on another node (normal RG takeover); during the process, the application service is interrupted for a brief time.

4.6.5 Primary storage failure (unplanned)

This scenario describes the PowerHA cluster with HyperSwap behavior when the primary storage fails. In our testing, we simulate a primary storage failure by disabling all zones between the cluster nodes and the primary storage.

Testing behavior expectation

The basic function of Power HyperSwap to provide the applications with continuous storage access (so that applications can continue running without any interruption) when one of the storage subsystems (the primary storage in our scenario) fails. It is expected that the application is not affected in this scenario.

Display the current status

Follow these steps:

1. Display the PowerHA resource status.

Example 4-82 shows the resource group ONLINE on node PS5n02base.

Example 4-82 Current resource group status of the unplanned primary storage failure scenario

```
# clRGinfo
-----
Group Name      State                Node
-----
dborarg         OFFLINE              PS5n01base@Sit
                 ONLINE              PS5n02base@Sit
                 ONLINE SECONDARY   SS5n03base@Sit
```

2. Display the application status.

The Oracle database is running on node PS5n02base. We use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations.

Example 4-83 shows the I/O throughput with the application running.

Example 4-83 Current I/O running status of application

```
PS5n02base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5      0.0    4405.0    84.0      825     3680  11:40:32
hdisk4      0.0    4423.0    46.0      296     4027  11:40:32
hdisk5      0.0    12688.0   509.0     48     13640  11:40:33
hdisk4      0.0    12719.0   602.0     24     13695  11:40:33
...
```

3. Display the Metro Mirror (PPRC) path status.

Example 4-84 shows the current PPRC status and path group. The primary path to the LUNs points to the primary storage (DS8805). The two HBAs (fcs1 and fcs4) access this storage. The other two adapters (fcs2 and fcs3) access the secondary storage (DS8803).

Example 4-84 Current PPRC status of the unplanned primary storage failure scenario

```
# lsprrc -Ao
hdisk#    PPRC    Primary    Secondary    Primary Storage    Secondary Storage
          state   path group path group    WWNN              WWNN
          ID      ID
hdisk3    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2
hdisk5    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2
hdisk4    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2
hdisk2    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2

# lsprrc -p hdisk4
path      WWNN              LSS  VOL    path
group id  group status
```

```

=====
0(s)      500507630bffc4c8 0xab 0x00 PRIMARY
1         500507630bffc1e2 0xab 0x00 SECONDARY

path      path path      parent connection
group id  id  status
=====
0         0   Enabled fscsi1 500507630b1884c8,40ab400000000000
0         1   Enabled fscsi4 500507630b5304c8,40ab400000000000
1         2   Enabled fscsi2 500507630b1001e2,40ab400000000000
1         3   Enabled fscsi3 500507630b1301e2,40ab400000000000

```

Simulate a primary storage failure

There are six zones that we disable at the same time.

Important: Be familiar with your SAN infrastructure to evaluate the consequences of this action correctly. SAN switch management depends on the switch manufacturer and the firmware version.

The zones are listed in 4.2.2, “Zoning configuration” on page 109:

- ▶ P7805LP9_fcs1_DS8805_I0302
- ▶ P7805LP9_fcs4_DS8805_I0234
- ▶ P7805LP10_fcs1_DS8805_I0302
- ▶ P7805LP10_fcs4_DS8805_I0234
- ▶ P7703LP9_fcs1_DS8805_I0302
- ▶ P7703LP9_fcs4_DS8805_I0234

Example 4-85 shows the commands to be issued on the SAN switch to disable the zones.

Example 4-85 Commands to disable zones

```

cfgremove
"CSC_Base", "P7805LP9_fcs1_DS8805_I0302;P7805LP9_fcs4_DS8805_I0234;P7805LP10_fcs1_DS8805_I0302;P7805LP10_fcs4_DS8805_I0234;P7703LP9_fcs1_DS8805_I0302;P7703LP9_fcs4_DS8805_I0234"

cfgenable "CSC_Base"

```

Test results

Follow these steps:

1. Display the PowerHA resource group status.

Example 4-86 on page 154 shows that the resource group status does not change as a result of a primary storage failure.

Example 4-86 Resource group status result of the unplanned primary storage failure scenario

```
# clRGinfo
```

Group Name	State	Node
dborarg	OFFLINE	PS5n01base@Sit
	ONLINE	PS5n02base@Sit
	ONLINE SECONDARY	SS5n03base@Sit

2. Display the application status.

Example 4-87 shows that the application's I/O activity is suspended from **11:50:14** to **11:50:44**.

Example 4-87 Application (running) result of the unplanned primary storage failure scenario

```
PS5n02base#iostat -T hdisk4 hdisk5 1|grep hdisk
```

```
...
hdisk5      0.0    3693.0    42.0      480     3213  11:50:13
hdisk4      1.0    3617.0    22.0       96     3521  11:50:13
hdisk5     29.0    560.0    35.0      560        0  11:50:14
hdisk4      0.0    112.0     7.0      112        0  11:50:14
hdisk5     100.0     0.0     0.0        0        0  11:50:15
---skipped
hdisk5     100.0     0.0     0.0        0        0  11:50:43
hdisk4      0.0     0.0     0.0        0        0  11:50:43
hdisk5     92.0   1883.0    44.0      249     1634  11:50:44
hdisk4      0.0    283.0    23.0      200        83  11:50:44
hdisk5      2.0   3072.0    13.0        0     3072  11:50:45
hdisk4      0.0   3250.0     9.0        0     3250  11:50:45
...

```

3. Display the Metro Mirror (PPRC) status.

After you disable the zones, the `lspprc` command output reveals that the primary paths for `hdisk4`, `hdisk5`, and `hdisk2` (used for the Oracle database and the CAA repository respectively) change from DS8805 to DS8803, as shown in Example 4-88.

Example 4-88 PPRC result of the unplanned primary storage failure scenario

```
#lspprc -Ao
```

hdisk#	PPRC state	Primary path group ID	Secondary path group ID	Primary Storage WWNN	Secondary Storage WWNN
hdisk5	Active	1(s)	0	500507630bffc1e2	500507630bffc4c8
hdisk4	Active	1(s)	0	500507630bffc1e2	500507630bffc4c8
hdisk2	Active	0, 1(s)	-1	500507630bffc4c8, 500507630bffc1e2	
hdisk3	Active	0(s)	1	500507630bffc4c8	500507630bffc1e2


```
# lspprc -p hdisk4
```

path group id	WWNN	LSS	VOL	path group status
0	500507630bffc4c8	0xab	0x00	SECONDARY
1(s)	500507630bffc1e2	0xab	0x00	PRIMARY

path group	path id	path status	parent	connection
0	0	Failed	fscsi1	500507630b1884c8,40ab400000000000
0	1	Failed	fscsi4	500507630b5304c8,40ab400000000000
1	2	Enabled	fscsi2	500507630b1001e2,40ab400000000000
1	3	Enabled	fscsi3	500507630b1301e2,40ab400000000000

Testing scenario summary

If the primary storage fails, the application's I/O is suspended for a short period of time (in our testing, 30 seconds). Then, it will resume.

4.6.6 Primary site failure (unplanned)

This scenario describes the PowerHA HyperSwap behavior when the primary site fails. In our testing, we simulate a primary site failure by disabling all zones between the hosts and the primary storage. We shut down the PS5n01base and PS5n02base nodes by using the `halt -q` command at the same time.

Testing behavior expectation

When the primary site fails, the secondary site takes over the RG and runs the application. The resource group is moved to the SS5n03base node, and the DS8803 storage acts as the primary storage (PPRC source).

Display the current status

Follow these steps:

1. Display the PowerHA resource group status.

Example 4-89 shows the resource group (dborarg) online on node PS5n01base.

Example 4-89 Current resource group status of the unplanned primary site failure scenario

```
# clRGinfo
```

Group Name	State	Node
dborarg	ONLINE	PS5n01base@Sit
	OFFLINE	PS5n02base@Sit
	ONLINE SECONDARY	SS5n03base@Sit

2. Display the application status.

The Oracle database is running on node PS5n01base. We use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations. Example 4-90 on page 156 shows the I/O throughput with the application running.

Example 4-90 Current I/O status of the application that is running

```
PS5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5          0.0    13159.0    147.0         0    13159  14:37:40
hdisk4          0.0    12148.0     79.0         8    12140  14:37:40
hdisk5          0.0     3982.0     35.0        320    3662  14:37:41
hdisk4          0.0     4256.0     25.0        160    4096  14:37:41
...
```

3. Display the Metro Mirror (PPRC) status.

Example 4-91 shows the current PPRC status and path group. The primary path of the LUNs points to the primary storage (DS8805) and two HBAs (fcs1 and fcs4) are used to access this storage. The other two adapters (fcs2 and fcs3) access the secondary storage (DS8803).

Example 4-91 Current PPRC status of the unplanned primary site failure scenario

```
# lsprrc -Ao
hdisk#    PPRC    Primary    Secondary    Primary Storage    Secondary Storage
          state   path group path group    WWNN              WWNN
          ID     ID
hdisk5    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2
hdisk4    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2
hdisk2    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2
hdisk3    Active  0(s)      1            500507630bffc4c8  500507630bffc1e2

# lsprrc -p hdisk4
path      WWNN              LSS  VOL  path
group id
=====
0(s)      500507630bffc4c8  0xab 0x00  PRIMARY
1         500507630bffc1e2  0xab 0x00  SECONDARY

path      path path      parent connection
group id id  status
=====
0  0    Enabled  fscsi1  500507630b1884c8,40ab400000000000
0  1    Enabled  fscsi4  500507630b5304c8,40ab400000000000
1  2    Enabled  fscsi2  500507630b1001e2,40ab400000000000
1  3    Enabled  fscsi3  500507630b1301e2,40ab400000000000
```

Simulating the primary site failure

Important: Be familiar with your SAN infrastructure to evaluate the consequences of this action correctly. The SAN switch management depends on the switch manufacturer and firmware version.

There are six zones that need to be disabled at the same time. See 4.2.2, “Zoning configuration” on page 109.

- ▶ P7805LP9_fcs1_DS8805_I0302
- ▶ P7805LP9_fcs4_DS8805_I0234
- ▶ P7805LP10_fcs1_DS8805_I0302
- ▶ P7805LP10_fcs4_DS8805_I0234

- ▶ P7703LP9_fcs1_DS8805_I0302
- ▶ P7703LP9_fcs4_DS8805_I0234

Example 4-92 shows the command that is executed on the SAN switch to disable the zones.

Example 4-92 Commands to disable the zones

```

cfgremove
"CSC_Base", "P7805LP9_fcs1_DS8805_I0302;P7805LP9_fcs4_DS8805_I0234;P7805LP10_fcs1_DS8805_I0302;P7805LP10_fcs4_DS8805_I0234;P7703LP9_fcs1_DS8805_I0302;P7703LP9_fcs4_DS8805_I0234"

cfgenable "CSC_Base"

```

At the same time, run the `halt -q` command on both the PS5n01base node and the PS5n02base node.

Test results

Follow these steps:

1. Display the PowerHA `cluster.log` file information.

Example 4-93 shows the PowerHA activities in this testing scenario. A failure was detected at 14:38:12. The cluster reconfiguration ended at 14:38:57.

Example 4-93 PowerHA cluster.log of the unplanned primary site failure scenario

```

Dec 13 14:38:12 SS5n03base local0:crit clstrmgrES[12845164]: Thu Dec 13 14:38:12 Removing 2 from ml_idx
Dec 13 14:38:12 SS5n03base daemon:err|error ConfigRM[8126552]: (Recorded using libct_ffdc.a cv 2):::Error ID: :::Reference ID: :::Template ID: a098bf90:::Details File: :::Location: RSCT,PeerDomain.C,1.99.22.110,18997 :::CONFIGRM_PENDINGQUORUM_ER The operational quorum state of the active peer domain has changed to PENDING_QUORUM. This state usually indicates that exactly half of the nodes that are defined in the peer domain are online. In this state cluster resources cannot be recovered although none will be stopped explicitly.
Dec 13 14:38:13 SS5n03base local0:crit clstrmgrES[12845164]: Thu Dec 13 14:38:13 Removing 1 from ml_idx
...
Dec 13 14:38:52 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: start_server dbcontrol 0
...
Dec 13 14:38:57 SS5n03base user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: node_down_complete PS5n02base 0

```

2. Display the PowerHA resource group status.

After the primary site fails, the resource group is brought back online on SS5n03base, which is in the secondary site. Example 4-94 shows the RG status.

Example 4-94 Resource group status result of the unplanned primary site failure scenario

```

# clRGinfo
-----
Group Name      State           Node
-----
dborarg         OFFLINE        PS5n01base@Sit
                OFFLINE        PS5n02base@Sit
                ONLINE        SS5n03base@Sit

```

3. Display the application status.

Example 4-95 shows that the application was interrupted at 14:37:52 on node PS5n01base and resumed at 14:39:16 on node SS5n03base.

Example 4-95 Application running result of the unplanned primary site failure scenario

```
PS5n01base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk5          0.0      2850.0      10.0          0      2850  14:37:52
hdisk4          0.0      2666.0       9.0          0      2666  14:37:52
forced shutdown by 'halt -q'

SS5n03base#iostat -T hdisk4 hdisk5 1|grep hdisk
...
hdisk4          0.0         0.0         0.0           0         0  14:39:15
hdisk5          0.0         0.0         0.0           0         0  14:39:15
hdisk4          0.0      47888.0     4130.0     47812         76  14:39:16
hdisk5          0.0     43599.0     198.0     43463        136  14:39:16
...
```

4. Display the Metro Mirror (PPRC) status.

After we disabled the zones, the `lspprc` command output shows that the primary paths of `hdisk4`, `hdisk5`, and `hdisk2` (used for Oracle database and CAA respectively) are changed from DS8805 to DS8803, as shown in Example 4-96.

Example 4-96 PPRC result of the unplanned primary site failure scenario

```
# lspprc -Ao
hdisk#   PPRC   Primary   Secondary   Primary Storage   Secondary Storage
         state  path group path group   WWNN              WWNN
hdisk4 Active 1(s)    0         500507630bffc1e2 500507630bffc4c8
hdisk3   Active  0(s)     1          500507630bffc4c8 500507630bffc1e2
hdisk5 Active 1(s)    0         500507630bffc1e2 500507630bffc4c8
hdisk2 Active 1(s)    0         500507630bffc1e2 500507630bffc4c8

# lspprc -p hdisk2
path     WWNN              LSS  VOL   path
group id                               group status
=====
0        500507630bffc4c8 0xaa 0x00 SECONDARY
1(s)    500507630bffc1e2 0xaa 0x00 PRIMARY

path     path  path      parent  connection
group id id    status
=====
0 0 Failed fscsi1 500507630b1884c8,40aa400000000000
0 1 Failed fscsi4 500507630b5304c8,40aa400000000000
1 2 Enabled fscsi2 500507630b1001e2,40aa400000000000
1 3 Enabled fscsi3 500507630b1301e2,40aa400000000000
```

Testing scenario summary

If the primary site fails, the resource group is moved to the secondary site, which results in the application being unavailable for a brief period.

4.6.7 PPRC replication path failure (unplanned)

This scenario describes the PowerHA HyperSwap behavior when the PPRC path fails. In our testing, we forcefully remove the PPRC path to simulate this type of failure.

Testing behavior expectation

When the PPRC path fails, the application continues to access the storage in the primary site, but the data cannot be synchronized to the secondary site.

After the replication path is recovered, data synchronization can be achieved after we perform the `resumepprc` operation (using the DSCLI, manually). The application is not affected during the recovery.

Display the current status

Follow these steps:

1. Display the PowerHA resource group status.

Example 4-97 shows the resource group online on node PS5n01base.

Example 4-97 Current resource group status of the unplanned PPRC replication path failure

```
# clRGinfo
-----
Group Name      State                Node
-----
oradbrg         ONLINE              PS5n01base@Sit
                 OFFLINE             PS5n02base@Sit
                 ONLINE SECONDARY    SS5n03base@Sit
-----
```

2. Display the status of the application that is running.

The Oracle database is running on node PS5n01base. We use an application that sends continuous SQL requests to the database, resulting in a number of disk I/O operations.

Example 4-98 shows the I/O throughput with the application running.

Example 4-98 Current IO running status of the application

```
PS5n01base#iostat -T hdisk3 hdisk4 1|grep disk
...
hdisk4          0.0    21333.0    953.0         8    21325  15:10:42
hdisk3          0.0    24394.0    118.0        320   24074  15:10:43
hdisk4          0.0    24110.0    100.0        36   24074  15:10:43
hdisk3          0.0    12048.0     55.0         0   12048  15:10:44
...
```

3. Display the Metro Mirror (PPRC) status.

Example 4-99 on page 160 shows the current PPRC status and path group. The primary path of the LUNs points to the primary storage (DS8805), and the two HBAs (fcs1 and fcs4) are used to access it. The other two adapters (fcs2 and fcs3) access the secondary storage (DS8803).

Example 4-99 Current PPRC status of the unplanned PPRC replication path failure

```
# lsprrc -Ao
hdisk#    PPRC      Primary      Secondary    Primary Storage  Secondary Storage
          state    path group   path group   WWNN            WWNN
          ID      ID
hdisk2    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2
hdisk3    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2
hdisk4    Active    0(s)        1            500507630bffc4c8 500507630bffc1e2

# lsprrc -p hdisk2
path      WWNN            LSS  VOL    path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xaa 0x00   PRIMARY
1         500507630bffc1e2 0xaa 0x00   SECONDARY

path      path path      parent  connection
group id  id   status
=====
0  0    Enabled  fscsi1 500507630b1884c8,40aa400000000000
0  1    Enabled  fscsi4 500507630b5304c8,40aa400000000000
1  2    Enabled  fscsi2 500507630b1001e2,40aa400000000000
1  3    Enabled  fscsi3 500507630b1301e2,40aa400000000000
```

Note: In this testing scenario, we removed the AA01 LUN because this LUN was not used. So, there are only three LUNs (AA00, AB00, and AB01) in this scenario; hdisk2 is for the CAA repository, and hdisk3 and hdisk4 are for the Oracle database (ASM disk group).

4. Display the Metro Mirror (PPRC) status.

Example 4-100 shows the PPRC and PPRC path status on the primary storage (DS8805).

Example 4-100 Current PPRC status of the primary storage (DS8805) from the DSCLI

```
dscli> lssi
Name ID          Storage Unit      Model WWNN            State ESSNet
=====
-   IBM.2107-75XP411 IBM.2107-75XP410 951   500507630BFFC4C8 Online Enabled

dscli> lsprrc AA00-AB01
ID      State      Reason Type SourceLSS Timeout (secs) Critical Mode First Pass Status
=====
AA00:AA00 Full Duplex - Metro Mirror AA      5          Disabled Invalid
AB00:AB00 Full Duplex - Metro Mirror AB      5          Disabled Invalid
AB01:AB01 Full Duplex - Metro Mirror AB      5          Disabled Invalid

dscli> lsprrcpath AA AB
Src Tgt State  SS  Port Attached Port Tgt WWNN
=====
AA  AA  Success FFAA I0102 I0102      500507630BFFC1E2
AA  AA  Success FFAA I0202 I0132      500507630BFFC1E2
AB  AB  Success FFAB I0102 I0102      500507630BFFC1E2
AB  AB  Success FFAB I0202 I0132      500507630BFFC1E2
```

Example 4-101 lists the PPRC and PPRC path status on the secondary storage (DS8803).

Example 4-101 Current PPRC status of the secondary storage (DS8803) from the DSCLI

```
dscli> lssi
Name ID Storage Unit Model WNN State ESSNet
=====
DS8803 IBM.2107-75WT971 IBM.2107-75WT970 951 500507630BFFC1E2 Online Enabled

dscli> lsprrc AA00-AB01
ID State Reason Type SourceLSS Timeout (secs) Critical Mode First
Pass Status
=====
AA00:AA00 Target Full Duplex - Metro Mirror AA unknown Disabled Invalid
AB00:AB00 Target Full Duplex - Metro Mirror AB unknown Disabled Invalid
AB01:AB01 Target Full Duplex - Metro Mirror AB unknown Disabled Invalid

dscli> lsprrcpath AA AB
Src Tgt State SS Port Attached Port Tgt WNN
=====
AA AA Success FFAA I0102 I0102 500507630BFFC4C8
AA AA Success FFAA I0132 I0202 500507630BFFC4C8
AB AB Success FFAB I0102 I0102 500507630BFFC4C8
AB AB Success FFAB I0132 I0202 500507630BFFC4C8
```

Removing the PPRC replication path between the LUNs

We forcefully remove the replication path by using the DSCLI, as shown in Example 4-102.

Example 4-102 Command to remove the PPRC replication path

```
dscli> rmprrcpath -remotedev IBM.2107-75WT971 -remotewwnn 500507630BFFC1E2 -type fcp -force
AA:AA AB:AB
CMUC00152W rmprrcpath: Are you sure you want to remove the Remote Mirror and Copy path AA:AA:?
[y/n]:y
CMUC00150I rmprrcpath: Remote Mirror and Copy path AA:AA successfully removed.
CMUC00152W rmprrcpath: Are you sure you want to remove the Remote Mirror and Copy path AB:AB:?
[y/n]:y
CMUC00150I rmprrcpath: Remote Mirror and Copy path AB:AB successfully removed.
```

Testing result of removing the PPRC replication path

Follow these steps:

1. Display the PowerHA resource group result.

The resource group was not affected during this process, as shown in Example 4-103.

Example 4-103 Resource group result of removing the pprc path

```
# clRGinfo
-----
Group Name State Node
-----
oradbrg ONLINE PS5n01base@Sit
OFFLINE PS5n02base@Sit
ONLINE SECONDARY SS5n03base@Sit
```

2. Display the application status.

There is no effect on the application, as shown in Example 4-104.

Example 4-104 Application running result of removing the pprc path

```
PS5n01base#iostat -T hdisk3 hdisk4 1|grep hdisk
...
hdisk3          0.0    19700.0    132.0        176    19524  15:20:54
hdisk4          0.0    19605.0    128.0         81    19524  15:20:54
hdisk3          0.0    32544.0    114.0         0    32544  15:20:55
hdisk4          0.0    32712.0    115.0         0    32712  15:20:55
hdisk3          0.0     3164.0     10.0          0     3164  15:20:56
hdisk4          0.0     2996.0      9.0           0     2996  15:20:56
hdisk3          0.0    28053.0    116.0          8    28045  15:20:57
hdisk4          0.0    28053.0    116.0          8    28045  15:20:57
hdisk3          0.0    13289.0     54.0          0    13289  15:20:58
...
```

3. Display the PPRC result.

From the **lspprc** command output shown in Example 4-105, we can see that the group status of the primary storage is changed to PRIMARY, SUSPENDED, OOS. Removing the PPRC replication path results in LUN data that is Out-Of-Sync (OOS) between the two storage subsystems.

Example 4-105 PPRC result of removing the pprc path

```
# lspprc -p hdisk2
path          WWNN          LSS  VOL  path
group id
=====
0(s)          500507630bffc4c8 0xaa 0x00  PRIMARY,
              500507630bffc4c8 0xaa 0x00  SUSPENDED,
              500507630bffc1e2 0xaa 0x00  OOS
              500507630bffc1e2 0xaa 0x00  SECONDARY

path          path  path          parent  connection
group id      id   status
=====
0  0  Enabled  fscsi1  500507630b1884c8,40aa400000000000
0  1  Enabled  fscsi4  500507630b5304c8,40aa400000000000
1  2  Enabled  fscsi2  500507630b1001e2,40aa400000000000
1  3  Enabled  fscsi3  500507630b1301e2,40aa400000000000

# lspprc -p hdisk3
path          WWNN          LSS  VOL  path
group id
=====
0(s)          500507630bffc4c8 0xab 0x00  PRIMARY,
              500507630bffc4c8 0xab 0x00  SUSPENDED,
              500507630bffc1e2 0xab 0x00  OOS
              500507630bffc1e2 0xab 0x00  SECONDARY

path          path  path          parent  connection
group id      id   status
=====
0  0  Enabled  fscsi1  500507630b1884c8,40ab400000000000
0  1  Enabled  fscsi4  500507630b5304c8,40ab400000000000
1  2  Enabled  fscsi2  500507630b1001e2,40ab400000000000
```



```

1 3 Enabled fscsi3 500507630b1301e2,40ab400000000000

# lsprrc -p hdisk4
path WNN LSS VOL path
group id group status
=====
0(s) 500507630bffc4c8 0xab 0x01 PRIMARY,
SUSPENDED,
OOS
1 500507630bffc1e2 0xab 0x01 SECONDARY

path path path parent connection
group id id status
=====
0 0 Enabled fscsi1 500507630b1884c8,40ab400100000000
0 1 Enabled fscsi4 500507630b5304c8,40ab400100000000
1 2 Enabled fscsi2 500507630b1001e2,40ab400100000000
1 3 Enabled fscsi3 500507630b1301e2,40ab400100000000

```

4. Display the Metro Mirror (PPRC) status.

From the DSCLI on the primary storage, we can see that the replication path (between the primary storage and the secondary storage) changed to Failed. The PPRC status changed to Suspended Internal Conditions, as shown in Example 4-106.

Example 4-106 Primary storage's PPRC result of removing the pprc path (from the DSCLI)

```

dscli> lssi
Name ID Storage Unit Model WNN State ESSNet
=====
DS8800-05 IBM.2107-75XP411 IBM.2107-75XP410 951 500507630BFFC4C8 Online Enabled

dscli> lsprrc AA00-AB01
ID State Reason Type SourceLSS Timeout
(secs) Critical Mode First Pass Status
=====
AA00:AA00 Suspended Internal Conditions Target Metro Mirror AA 5
Disabled Invalid
AB00:AB00 Suspended Internal Conditions Target Metro Mirror AB 5
Disabled Invalid
AB01:AB01 Suspended Internal Conditions Target Metro Mirror AB 5
Disabled Invalid

dscli> lsprrcpath AA AB
Src Tgt State SS Port Attached Port Tgt WNN
=====
AA AA Failed FFAA - - 500507630BFFC1E2
AB AB Failed FFAB - - 500507630BFFC1E2

```

From the DSCLI on the secondary storage, we can see that the replication path (between the secondary storage and the primary storage) is not changed. The PPRC status is not changed either, as shown in Example 4-107 on page 163.

Example 4-107 Secondary storage's PPRC result of removing the pprc path (from the DSCLI)

```

dscli> lssi
Name ID Storage Unit Model WNN State ESSNet

```

```
=====
DS8803 IBM.2107-75WT971 IBM.2107-75WT970 951 500507630BFFC1E2 Online Enabled
```

```
dscli> lsprrc AA00-AB01
ID          State          Reason Type          SourceLSS Timeout (secs) Critical
Mode First Pass Status
=====
AA00:AA00 Target Full Duplex - Metro Mirror AA    unknown    Disabled
Invalid
AB00:AB00 Target Full Duplex - Metro Mirror AB    unknown    Disabled
Invalid
AB01:AB01 Target Full Duplex - Metro Mirror AB    unknown    Disabled
Invalid
```

```
dscli> lsprrcpath AA AB
Src Tgt State  SS  Port  Attached Port Tgt WWNN
=====
AA  AA  Success FFAA I0102 I0102      500507630BFFC4C8
AA  AA  Success FFAA I0132 I0202      500507630BFFC4C8
AB  AB  Success FFAB I0102 I0102      500507630BFFC4C8
AB  AB  Success FFAB I0132 I0202      500507630BFFC4C8
```

Re-creating the PPRC replication path

We re-create the PPRC path by using the `mkpprcpath` command (DSCLI). After the replication path is re-created, the PPRC status does not change, as shown in Example 4-108.

Example 4-108 Commands to re-create the pprc path

```
dscli> mkpprcpath -remotedev IBM.2107-75WT971 -remotewwnn 500507630BFFC1E2
-srclss AA -tgtlss AA I0102:I0102 I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path AA:AA successfully established.
dscli> mkpprcpath -remotedev IBM.2107-75WT971 -remotewwnn 500507630BFFC1E2
-srclss AB -tgtlss AB I0102:I0102 I0202:I0132
CMUC00149I mkpprcpath: Remote Mirror and Copy path AB:AB successfully established.
```

```
dscli> lssi
Name          ID          Storage Unit      Model WWNN          State ESSNet
=====
DS8800-05 IBM.2107-75XP411 IBM.2107-75XP410 951 500507630BFFC4C8 Online Enabled
```

```
dscli> lsprrc AA00-AB01
ID          State          Reason          Type          SourceLSS Timeout
(secs) Critical Mode First Pass Status
=====
AA00:AA00 Suspended Internal Conditions Target Metro Mirror AA    5
Disabled      Invalid
AB00:AB00 Suspended Internal Conditions Target Metro Mirror AB    5
Disabled      Invalid
AB01:AB01 Suspended Internal Conditions Target Metro Mirror AB    5
Disabled      Invalid
```

```
dscli> lsprrcpath AA AB
Src Tgt State  SS  Port  Attached Port Tgt WWNN
=====
```

```
AA AA Success FFAA I0102 I0102 500507630BFFC1E2
AA AA Success FFAA I0202 I0132 500507630BFFC1E2
AB AB Success FFAB I0102 I0102 500507630BFFC1E2
AB AB Success FFAB I0202 I0132 500507630BFFC1E2
```

Resuming PPRC

Finally, we resume PPRC on the primary storage (Example 4-109).

Example 4-109 Command to resume PPRC

```
dscli> lssi
Name ID                      Storage Unit      Model WNN                      State  ESSNet
-----
-   IBM.2107-75XP411 IBM.2107-75XP410 951  500507630BFFC4C8 Online Enabled

dscli> resumepprc -remotedev IBM.2107-75WT971 -type mmir -tgtse AA00:AA00
CMUC00158I resumepprc: Remote Mirror and Copy volume pair AA00:AA00 relationship
successfully resumed. This message is being returned before the copy completes.

dscli> resumepprc -remotedev IBM.2107-75WT971 -type mmir -tgtse AB00:AB00
AB01:AB01
CMUC00158I resumepprc: Remote Mirror and Copy volume pair AB00:AB00 relationship
successfully resumed. This message is being returned before the copy completes.
CMUC00158I resumepprc: Remote Mirror and Copy volume pair AB01:AB01 relationship
successfully resumed. This message is being returned before the copy completes.
```

Follow these steps:

1. Display the Metro Mirror (PPRC) status.

Example 4-110 PPRC status during resuming PPRC on the primary storage

```
dscli> lsprrc AA00-AB01
ID      State      Reason Type      SourceLSS Timeout (secs) Critical Mode
First Pass Status
=====
AA00:AA00 Full Duplex - Metro Mirror AA      5      Disabled
Invalid
AB00:AB00 Copy Pending - Metro Mirror AB      5      Disabled
Invalid
AB01:AB01 Copy Pending - Metro Mirror AB      5      Disabled
Invalid
```

#After several seconds, the data is synchronized successfully.

```
dscli> lsprrc AA00-AB01
ID      State      Reason Type      SourceLSS Timeout (secs) Critical Mode
First Pass Status
=====
AA00:AA00 Full Duplex - Metro Mirror AA      5      Disabled
Invalid
AB00:AB00 Full Duplex - Metro Mirror AB      5      Disabled
Invalid
AB01:AB01 Full Duplex - Metro Mirror AB      5      Disabled
Invalid
```

We check the PPRC status from the secondary storage and everything is returned to normal, as shown in Example 4-111.

Example 4-111 PPRC result on the secondary storage after resuming PPRC is completed

```

dscli> lssi
Name      ID              Storage Unit      Model WWNN              State  ESSNet
=====
DS8803 IBM.2107-75WT971 IBM.2107-75WT970 951  500507630BFFC1E2 Online Enabled

dscli> lsprrc AA00-AB01
ID      State          Reason Type          SourceLSS Timeout (secs) Critical
Mode First Pass Status
=====
AA00:AA00 Target Full Duplex - Metro Mirror AA    unknown    Disabled
Invalid
AB00:AB00 Target Full Duplex - Metro Mirror AB    unknown    Disabled
Invalid
AB01:AB01 Target Full Duplex - Metro Mirror AB    unknown    Disabled
Invalid

```

2. Display the Metro Mirror (PPRC) status in AIX.

After data synchronization completes, the output of the **lsprrc** command shows the replication status recovered, as shown in Example 4-112.

Example 4-112 PPRC status result on AIX after the data synchronization completes

```

# lsprrc -p hdisk2
path      WWNN              LSS  VOL  path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xaa 0x00 PRIMARY
1         500507630bffc1e2 0xaa 0x00 SECONDARY

path      path path      parent connection
group id id  status
=====
0 0 Enabled fscsi1 500507630b1884c8,40aa400000000000
0 1 Enabled fscsi4 500507630b5304c8,40aa400000000000
1 2 Enabled fscsi2 500507630b1001e2,40aa400000000000
1 3 Enabled fscsi3 500507630b1301e2,40aa400000000000

# lsprrc -p hdisk3
path      WWNN              LSS  VOL  path
group id                                     group status
=====
0(s)      500507630bffc4c8 0xab 0x00 PRIMARY
1         500507630bffc1e2 0xab 0x00 SECONDARY

path      path path      parent connection
group id id  status
=====
0 0 Enabled fscsi1 500507630b1884c8,40ab400000000000
0 1 Enabled fscsi4 500507630b5304c8,40ab400000000000
1 2 Enabled fscsi2 500507630b1001e2,40ab400000000000
1 3 Enabled fscsi3 500507630b1301e2,40ab400000000000

```

```
# lsprrc -p hdisk4
path          WNNN          LSS  VOL  path
group id      group status
=====
0(s)          500507630bffc4c8 0xab 0x01 PRIMARY
1             500507630bffc1e2 0xab 0x01 SECONDARY

path          path path          parent connection
group id     id  status
=====
0            0   Enabled  fscsi1 500507630b1884c8,40ab400100000000
0            1   Enabled  fscsi4 500507630b5304c8,40ab400100000000
1            2   Enabled  fscsi2 500507630b1001e2,40ab400100000000
1            3   Enabled  fscsi3 500507630b1301e2,40ab400100000000
```

3. Display the AIX error log.

From the AIX error report, we can see the PPRC suspend and resume events recorded as shown in Example 4-113.

Example 4-113 PPRC events in the AIX error report

IDENTIFIER	TIMESTAMP	T	C	RESOURCE_NAME	DESCRIPTION
63B1A1E6	0207154113	I	H	pha_1065451171	PPRC Replication Path Recovered
5011BAF4	0207154113	I	H	hdisk4	PPRC Device Resumed
5011BAF4	0207154113	I	H	hdisk3	PPRC Device Resumed
5011BAF4	0207154113	I	H	hdisk2	PPRC Device Resumed
BFCFD000	0207154013	T	H	hdisk2	PPRC Device Suspended
DCB47997	0207154013	T	H	hdisk4	DISK OPERATION ERROR
DCB47997	0207154013	T	H	hdisk3	DISK OPERATION ERROR
4BD7BBF6	0207154013	T	H	pha_1065451171	PPRC Replication Path Failed
BFCFD000	0207154013	T	H	hdisk4	PPRC Device Suspended
BFCFD000	0207154013	T	H	hdisk3	PPRC Device Suspended

4. Display the application status.

During the resume operation, the application I/O was only slightly affected (for about 2 seconds), as shown in Example 4-114.

Example 4-114 Application affected during the operation to resume PPRC

```
#iostat -T hdisk3 hdisk4 1|grep disk
...
hdisk4          0.0      28491.0      1580.0          16      28475  15:41:21
hdisk3          0.0      22005.0       105.0           0      22005  15:41:22
hdisk4          0.0      22013.0       106.0           8      22005  15:41:22
hdisk3          0.0      10406.0        56.0          128     10278  15:41:23
hdisk4          0.0      10342.0        52.0          64     10278  15:41:23
hdisk3          0.0       724.0         46.0          704       20  15:41:24
hdisk4          0.0       352.0         22.0          352         0  15:41:24
hdisk3          0.0      4462.0         87.0          824     3638  15:41:25
hdisk4          0.0      3801.0         58.0          360     3441  15:41:25
hdisk3          0.0          0.0          0.0           0         0  15:41:26
hdisk4          0.0          0.0          0.0           0         0  15:41:26
hdisk3          0.0          0.0          0.0           0         0  15:41:27
hdisk4          0.0          0.0          0.0           0         0  15:41:27
hdisk3          0.0     13659.0         83.0          144     13515  15:41:28
```

hdisk4	0.0	13749.0	78.0	17	13732	15:41:28
hdisk3	0.0	11905.0	60.0	0	11905	15:41:29
hdisk4	0.0	11905.0	60.0	0	11905	15:41:29
...						

Testing scenario summary

If the PPRC replication path between two storage subsystems is broken, or the PPRC replication path fails, there is no I/O pending for the application. When you recover the PPRC pair (by using the **resumepprc** command), the I/O is affected for about 2 seconds.



PowerHA cluster with AIX HyperSwap Active-Active for applications using Oracle RAC

In this chapter, we describe how to plan and configure a two-site, four-node cluster with PowerHA SystemMirror with AIX HyperSwap. This cluster provides the foundation for a four-node Oracle Database 11g Release 2 Real Application Cluster (RAC).

We provide a high-level description of system prerequisites and the configuration process. The second part of the chapter provides information about the various tests that we performed in our environment.

The following topics are described:

- ▶ Cluster description and diagrams:
 - Prerequisites
 - Implementation planning
- ▶ Configuring the environment:
 - Storage configuration
 - Storage area network configuration
 - LUN configuration in AIX and enabling HyperSwap
 - Oracle RAC cluster installation and configuration
 - PowerHA cluster installation and configuration
- ▶ Test scenarios:
 - Test method description
 - Primary storage maintenance (planned)
 - Node failure (unplanned)
 - Primary storage failure (unplanned)
 - Site failure (unplanned)

5.1 Cluster description and diagrams

The configuration tested in this scenario consists of a two-site, four-node PowerHA stretched cluster with two nodes in each site. Oracle Database RAC is deployed over all four nodes.

This deployment is complex because it involves two different cluster management solutions:

- ▶ Oracle GRID 11gR2 (11.2.0.3) manages the system resources to support the Oracle Database, including the storage infrastructure, Automatic Storage Management (ASM).
- ▶ PowerHA SystemMirror Enterprise Edition 7.1.2 with AIX HyperSwap manages the automated storage swap.

The two cluster management suites complement each other to improve application availability. In addition to the benefits and features of the Oracle GRID suite, PowerHA with HyperSwap adds transparent storage protection for replicated storage, improving overall system availability by masking storage failures.

The cluster management scope is different (there is no functional overlap) for each cluster suite. Oracle GRID manages the Oracle Database Cluster (RAC) and can be considered a standard four-node Oracle Database RAC deployment. The PowerHA cluster is an Extended Distance cluster (with two sites) that manages, in principle, the replicated storage infrastructure through HyperSwap functionality.

Based on the Oracle RAC requirements, each node has two network (Ethernet) interfaces and four Fibre Channel (FC) adapters. We use Virtual Fibre Channel (VFC) and a SAN infrastructure that supports N_Port ID Virtualization (NPIV).

The storage is provided by two DS8800 storage subsystems configured to replicate each other by using Metro Mirror Peer-to-Peer Remote Copy (PPRC) synchronous replication. The DS8800 supports in-band (SCSI commands) communication, which is used to manage (and automate) the replication using the AIX HyperSwap framework and the PowerHA automation and management capabilities.

To support all involved clustering layers, we allocated direct access shared storage:

- ▶ For the Oracle GRID and Oracle Database, logical unit numbers (LUNs) for the Oracle Cluster Repository (OCR) files and voting disk and database disks
- ▶ For PowerHA, the Cluster Aware AIX (CAA) repository disk

The diagram in Figure 5-1 on page 171 provides basic information for our test environment. The GRIDDG is a disk group managed by ASM, which is used for the voting disk and to store OCR files. The DATADG also is an ASM-managed disk group, which is used to store the Oracle database. The REPOSITORY disk is the CAA repository disk.

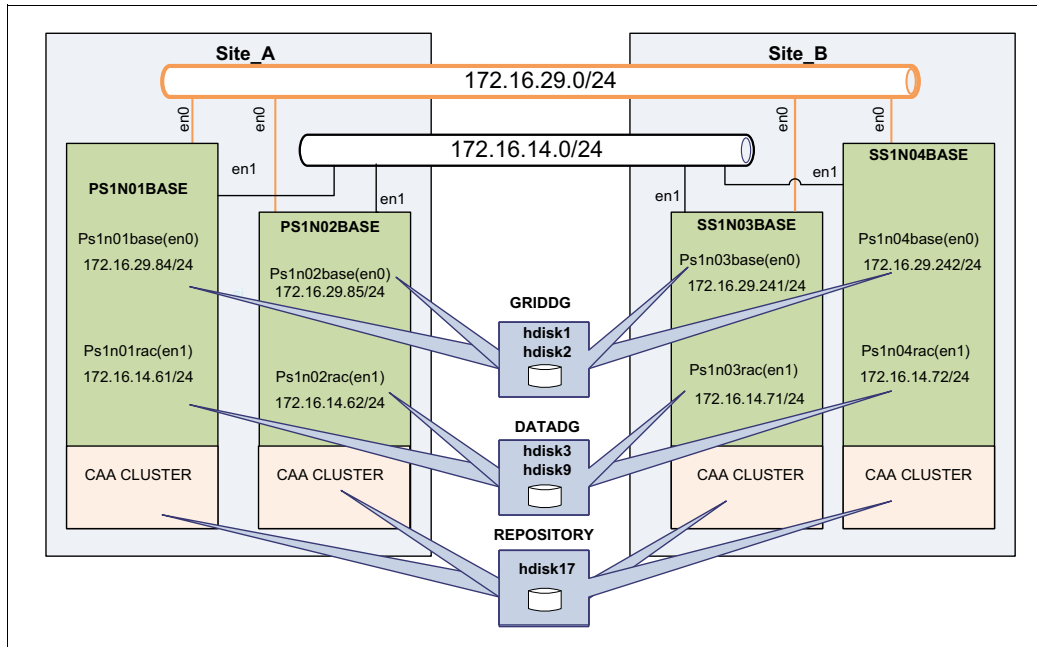


Figure 5-1 Diagram of the testing environment

5.1.1 Prerequisites

The cluster nodes are logical partitions (LPARs) that use virtual resources. All LPARs are hosted in two IBM Power 7 780 Servers (9179-MHB) using Micro-Partitioning. I/O virtualization is provided by Virtual I/O Servers running VIOS 2.2.2.

The operating system that is running on the LPARs is AIX 6.1 TL8 SP2.

The storage firmware is DS8800 microcode R6.3SP4.

The following Oracle software is used in our configuration:

- ▶ Oracle Grid Infrastructure 11g Release 2 (patch level 11.2.0.3)
- ▶ Oracle Database 11g Release 2 (patch level 11.2.0.3)

The PowerHA software is IBM PowerHA SystemMirror for AIX Enterprise Edition Release 7 Version 1.2 Service Pack 2 (7.1.2 SP2)

5.1.2 Implementation planning

We used the following steps to build the environment:

1. Configure the systems: VIOS, LPARs, and virtual resources.
2. Install AIX (including networking and VFC).
3. Configure the storage subsystems.
4. Configure the SAN (zoning, inter-switch link (ISL), and so on).
5. Configure the storage replication.
6. Configure the HyperSwap devices in AIX.
7. Install and configure the Oracle Clusterware and Database.
8. Install the PowerHA software and configure the PowerHA stretched cluster.
9. Test the configuration.

The order of the steps presented in the previous list is not precise. Occasionally, you might need to go back and forth between configuring the storage, SAN, and systems, depending on the actual environment.

5.2 Configuring the environment

This section contains a summary of the configuration tasks that we performed in our environment. We emphasize specific details.

5.2.1 Storage configuration

In this scenario, we allocate five LUN pairs from the two DS8800 storage subsystems (DS8800-05 in Site_A and DS8800-03 in Site_B). For management purposes, we allocate the LUNs in three logical subsystems (LSS):

- ▶ One LUN pair in an individual LSS for the CAA repository disk.
- ▶ Two LUN pairs are allocated in another LSS that is managed by Oracle ASM and used for the voting disk and OCR files.
- ▶ The last two LUN pairs are allocated in a third LSS and used for the oracle data disk group.

Each LUN pair consists of two LUNs allocated from the primary storage subsystem and the secondary storage subsystem, and the LUN pairs are mirrored (PPRC source and target).

Important: The LUN that is used for the CAA repository disk must be allocated in an LSS that is separate from the other LSS that is used for the application data disk.

Tip: You can configure many LUNs from one or more LSS for application data disks. It is better to dedicate one LSS to one application. Do not let an LSS cross two or more applications.

Storage configuration details

We suggest that you create a diagram (or any other form of configuration documentation) that reflects the details of your environment. Figure 3-1 on page 27 shows an example of a storage configuration diagram. The actual data in Figure 3-1 on page 27 is for a different set of LPARs. For the detailed information about the storage configuration process, see 3.2.4, “Configuring the storage” on page 36.

We configure five LUN pairs from the two DS8800s. We run the `lss` (DSCLI) command on both the primary DS8800 and the secondary DS8800 to identify the systems, as shown in Example 5-1 on page 173.

Example 5-1 Identifying the storage subsystems

Primary storage (DS8800-05)

```
dscli> lssi
Name ID                Storage Unit      Model WNN          State  ESSNet
=====
-   IBM.2107-75XP411  IBM.2107-75XP410  951   500507630BFFC4C8 Online Enabled
```

Secondary storage (DS8800-03)

```
dscli> lssi
Name ID                Storage Unit      Model WNN          State  ESSNet
=====
DS8803 IBM.2107-75WT971  IBM.2107-75WT970  951   500507630BFFC1E2 Online Enabled
```

Example 5-2 shows the LUN configuration for our scenario.

Example 5-2 LUN configuration

Primary storage (DS8800-05)

```
dscli> lsfbvol -volgrp v3
Name ID    acstate  datastate  configstate  device  MTM  datatype  extpool  cap (2^30B)  cap (10^9B)  cap (blocks)
=====
OpenSwap_G1 A201 Online   Normal     Normal      2107-900  FB 512  P4        50.0        -            104857600
OpenSwap_G1 A202 Online   Normal     Normal      2107-900  FB 512  P4        50.0        -            104857600
OpenSwap_G1 A300 Online   Normal     Normal      2107-900  FB 512  P3        50.0        -            104857600
OpenSwap_G1 A301 Online   Normal     Normal      2107-900  FB 512  P3        50.0        -            104857600
OpenSwap_G1 C200 Online   Normal     Normal      2107-900  FB 512  P4        50.0        -            104857600
```

Secondary storage (DS8800-03)

```
dscli> lsfbvol -volgrp v13
Name ID    acstate  datastate  configstate  device  MTM  datatype  extpool  cap (2^30B)  cap (10^9B)  cap (blocks)
=====
OpenSwap_G1 A201 Online   Normal     Normal      2107-900  FB 512  P2        50.0        -            104857600
OpenSwap_G1 A202 Online   Normal     Normal      2107-900  FB 512  P2        50.0        -            104857600
OpenSwap_G1 A300 Online   Normal     Normal      2107-900  FB 512  P3        50.0        -            104857600
OpenSwap_G1 A301 Online   Normal     Normal      2107-900  FB 512  P3        50.0        -            104857600
OpenSwap_G1 C200 Online   Normal     Normal      2107-900  FB 512  P2        50.0        -            104857600
```

Example 5-3 shows information about all PPRC pairs (1spprc command issued from the DSCLI).

Example 5-3 PPRC pairs

```
dscli> lspprc -fullid a201-c200
```

```
ID                State          Reason Type          SourceLSS
Timeout (secs) Critical Mode First Pass Status
=====
IBM.2107-75XP411/A201:IBM.2107-75WT971/A201 Full Duplex      -      Metro Mirror
IBM.2107-75XP411/A2 unknown          Disabled      Invalid
IBM.2107-75XP411/A202:IBM.2107-75WT971/A202 Full Duplex      -      Metro Mirror
IBM.2107-75XP411/A2 unknown          Disabled      Invalid
IBM.2107-75XP411/A300:IBM.2107-75WT971/A300 Full Duplex      -      Metro Mirror
IBM.2107-75XP411/A3 unknown          Disabled      Invalid
IBM.2107-75XP411/A301:IBM.2107-75WT971/A301 Full Duplex      -      Metro Mirror
IBM.2107-75XP411/A3 unknown          Disabled      Invalid
.....<< Snippet >>.....
IBM.2107-75XP411/C200:IBM.2107-75WT971/C200 Full Duplex      -      Metro Mirror
IBM.2107-75XP411/C2 unknown          Disabled      Invalid
```

Example 5-4 shows information about the PPRC paths (`lspprcpath` command in the DSCLI).

Example 5-4 PPRC path information

```

dscli> lspprcpath -fullid a2-c2
Src          Tgt          State  SS  Port          Attached Port
Tgt WNNN
=====
=====
IBM.2107-75XP411/A2 IBM.2107-75WT971/A2 Success FFA2 IBM.2107-75XP411/I0102
IBM.2107-75WT971/I0102 500507630BFFC1E2
IBM.2107-75XP411/A2 IBM.2107-75WT971/A2 Success FFA2 IBM.2107-75XP411/I0202
IBM.2107-75WT971/I0132 500507630BFFC1E2
IBM.2107-75XP411/A3 IBM.2107-75WT971/A3 Success FFA3 IBM.2107-75XP411/I0102
IBM.2107-75WT971/I0102 500507630BFFC1E2
IBM.2107-75XP411/A3 IBM.2107-75WT971/A3 Success FFA3 IBM.2107-75XP411/I0202
IBM.2107-75WT971/I0132 500507630BFFC1E2
.....<< Snippet >>.....
IBM.2107-75XP411/C2 IBM.2107-75WT971/C2 Success FFC2 IBM.2107-75XP411/I0102
IBM.2107-75WT971/I0102 500507630BFFC1E2
IBM.2107-75XP411/C2 IBM.2107-75WT971/C2 Success FFC2 IBM.2107-75XP411/I0202
IBM.2107-75WT971/I0132 500507630BFFC1E2

```

5.2.2 Storage area network configuration

We suggest that you create a storage area network diagram (or any other form of configuration documentation) that reflects the details of your environment. An example of a diagram is shown in Figure 3-1 on page 27.

A detailed example of a zoning configuration is shown in 3.2.3, “Zoning configuration” on page 31.

5.2.3 LUN configuration in AIX and enabling HyperSwap

We configured the storage and SAN. In preparation for enabling HyperSwap, we need to identify the LUN data. We show only the main tasks that help to identify the disks used for this particular scenario.

You can find the step-by-step details for the AIX HyperSwap configuration in 3.2.6, “AIX configuration” on page 43.

Example 5-5 list all AIX disks used in this scenario and the associated LUN information.

Example 5-5 LUN information

```

root@PS1n01base:/> for i in `lspsv|awk '{print $1}'`; do lscfg -vpl $i|egrep "hdisk|Serial|Z7"; done

hdisk1          U78C0.001.DBJG630-P2-C2-T1-W500507630B1884C8-L40A240010000000  MPIO IBM 2107 FC Disk
Serial Number.....75XP411A
Device Specific.(Z7).....A201
hdisk2          U78C0.001.DBJG630-P2-C2-T1-W500507630B1884C8-L40A340000000000  MPIO IBM 2107 FC Disk
Serial Number.....75XP411A
Device Specific.(Z7).....A300
hdisk3          U78C0.001.DBJG630-P2-C2-T1-W500507630B1884C8-L40A340010000000  MPIO IBM 2107 FC Disk
Serial Number.....75XP411A
Device Specific.(Z7).....A301

```

```

hdisk5      U78C0.001.DBJG630-P2-C5-T1-W500507630B1001E2-L40A2400100000000  MPIO IBM 2107 FC Disk
Serial Number.....75WT971A
Device Specific.(Z7).....A201
hdisk6      U78C0.001.DBJG630-P2-C5-T1-W500507630B1001E2-L40A3400000000000  MPIO IBM 2107 FC Disk
Serial Number.....75WT971A
Device Specific.(Z7).....A300
hdisk7      U78C0.001.DBJG630-P2-C5-T1-W500507630B1001E2-L40A3400100000000  MPIO IBM 2107 FC Disk
Serial Number.....75WT971A
Device Specific.(Z7).....A301
.....<< Snippet >>.....
hdisk9      U78C0.001.DBJG630-P2-C2-T1-W500507630B1884C8-L40A2400200000000  MPIO IBM 2107 FC Disk
Serial Number.....75XP411A
Device Specific.(Z7).....A202
.....<< Snippet >>.....
hdisk13     U78C0.001.DBJG630-P2-C5-T1-W500507630B1001E2-L40A2400200000000  MPIO IBM 2107 FC Disk
Serial Number.....75WT971A
Device Specific.(Z7).....A202
.....<< Snippet >>.....
hdisk17     U78C0.001.DBJG630-P2-C5-T1-W500507630B1001E2-L40C2400000000000  MPIO IBM 2107 FC Disk
Serial Number.....75XP411C
Device Specific.(Z7).....C200
.....<< Snippet >>.....
hdisk19     U78C0.001.DBJG630-P2-C5-T1-W500507630B1001E2-L40C2400000000000  MPIO IBM 2107 FC Disk
Serial Number.....75WT971C
Device Specific.(Z7).....C200

```

Example 5-6 shows the output of the DSCLI command `lspprc`. We identify the PPRC pairs of the disks in AIX from Example 5-5 on page 174.

For example, we know that the LUN with Volume ID A201 in storage **IBM.2107-75XP411** has a Metro Mirror relationship with the LUN with Volume ID A201 of storage **IBM.2107-75WT971**.

As shown in Example 5-6 (DSCLI) and Example 5-5 on page 174 (AIX), LUN A201 of IBM.2107-75XP411 is hdisk1. LUN A201 of **IBM.2107-75WT971** is hdisk5. So, hdisk1 and hdisk5 will be paired when HyperSwap is enabled.

Example 5-6 Replicated LUNs identification

```

dscli> lspprc -fullid -remotedev IBM.2107-75XP411 A201-A301
ID                               State      Reason Type      SourceLSS
Timeout (secs) Critical Mode First Pass Status
=====
IBM.2107-75XP411/A201:IBM.2107-75WT971/A201 Full Duplex -      Metro Mirror IBM.2107-75XP411/A2
5                               Disabled   Invalid
IBM.2107-75XP411/A202:IBM.2107-75WT971/A202 Full Duplex -      Metro Mirror IBM.2107-75XP411/A2
5                               Disabled   Invalid
IBM.2107-75XP411/A203:IBM.2107-75WT971/A203 Full Duplex -      Metro Mirror IBM.2107-75XP411/A2
5                               Disabled   Invalid
IBM.2107-75XP411/A300:IBM.2107-75WT971/A300 Full Duplex -      Metro Mirror IBM.2107-75XP411/A3
5                               Disabled   Invalid
IBM.2107-75XP411/A301:IBM.2107-75WT971/A301 Full Duplex -      Metro Mirror IBM.2107-75XP411/A3
5                               Disabled   Invalid

```

From the DSCLI output, we also see that LUN A201 of **IBM.2107-75XP411** is the primary copy of the mirroring pair (State of PPRC is “Full Duplex”). In AIX, hdisk1 is the device that we will enable for HyperSwap.

After we identify the LUNs that will be used for the cluster, we need to perform the following steps to enable HyperSwap in AIX:

- ▶ Remove the Subsystem Device Driver Path Control Module (SDDPCM) if it is installed.
- ▶ Change the Path Control Module (PCM) to AIX_AAPCM for the 2107DS8K device driver.
- ▶ Reboot all nodes.
- ▶ After the disks are detected as replicated devices, change the **san_rep_cfg** attribute for the disks used in the configuration and verify, as shown in Example 5-7.

Example 5-7 HyperSwap enabled devices (AIX)

```

root@PS1n01base: /> manage_disk_drivers -d 2107DS8K -o AIX_AAPCM
***** ATTENTION *****

For the change to take effect the system must be rebooted

root@PS1n01base: /> chdev -l hdisk1 -a san_rep_cfg=migrate_disk -U
hdisk1 changed

root@ps1n01base: /> lsprrc -Ao
hdisk#   PPRC      Primary      Secondary    Primary Storage  Secondary Storage
         state    path group   path group    WWNN           WWNN
         ID      ID
hdisk3  Active  0(s)       1           500507630bffc4c8 500507630bffc1e2
hdisk1  Active  0(s)       1           500507630bffc4c8 500507630bffc1e2
hdisk2  Active  0(s)       1           500507630bffc4c8 500507630bffc1e2
<<snippet>>
hdisk9  Active  0(s)       1           500507630bffc4c8 500507630bffc1e2
hdisk17 Active  0(s)       1           500507630bffc4c8 500507630bffc1e2

root@ps1n01base: /> lsprrc -p hdisk1
path      WWNN          LSS  VOL   path
group id  group status
=====
1          500507630bffc1e2 0xa2 0x01  SECONDARY
0(s)      500507630bffc4c8 0xa2 0x01  PRIMARY

path      path path      parent connection
group id  id  status
=====
1         2   Enabled  fscsi4  500507630b1001e2,40a2400100000000
1         3   Enabled  fscsi6  500507630b1301e2,40a2400100000000
0         0   Enabled  fscsi0  500507630b5304c8,40a2400100000000
0         1   Enabled  fscsi2  500507630b1884c8,40a2400100000000

```

Also, ensure that the **reserve_policy** attribute is set to no_reserve. Example 5-8 shows the disk attributes after we enable HyperSwap.

Example 5-8 Disk attributes after we enable HyperSwap

```

root@ps1n01base: /> lsattr -El hdisk1
PCM                PCM/friend/aixmpiods8k                Path Control Module                False
PR_key_value       none                                    Persistent Reserve Key Value       True
algorithm          fail_over                               Algorithm                           True
clr_q              no                                       Device CLEARS its Queue on error   True
dist_err_pcmt     0                                       Distributed Error Percentage       True
dist_tw_width      50                                       Distributed Error Sample Time      True
hcheck_cmd         test_unit_rdy                           Health Check Command               True

```

hcheck_interval	60	Health Check Interval	True
hcheck_mode	nonactive	Health Check Mode	True
location		Location Label	True
lun_id	0x40a2400100000000	Logical Unit Number ID	False
lun_reset_spt	yes	LUN Reset Supported	True
max_coalesce	0x40000	Maximum Coalesce Size	True
max_retry_delay	60	Maximum Quiesce Time	True
max_transfer	0x80000	Maximum TRANSFER Size	True
node_name	0x500507630bffc4c8	FC Node Name	False
pvid	00cf8de6fcd21bbc00000000000000000	Physical volume identifier	False
q_err	yes	Use QERR bit	True
q_type	simple	Queuing TYPE	True
queue_depth	20	Queue DEPTH	True
reassign_to	120	REASSIGN time out value	True
reserve_policy	no_reserve	Reserve Policy	True
rw_timeout	30	READ/WRITE time out value	True
san_rep_cfg	migrate_disk	SAN Replication Device Configuration Policy	True+
san_rep_device	yes	SAN Replication Device	False
scsi_id	0x61a00	SCSI ID	False
start_timeout	60	START unit time out value	True
timeout_policy	fail_path	Timeout Policy	True
unique_id	352037355850343131413230310050b534ba072107900031BMfcp	Unique device identifier	False
ww_name	0x500507630b1884c8	FC World Wide Name	False

Table 5-1 Presents the disks that we use in our scenario and their designation.

Table 5-1 Disk used for test configuration

AIX disk	Replica	Designation
hdisk1	hdisk5	GRIDDG
hdisk2	hdisk6	GRIDDG
hdisk3	hdisk7	DATADG
hdisk9	hdisk13	DATADG
hdisk17	hdisk19	CAA_REPO

5.2.4 Oracle RAC cluster installation and configuration

We provide the specific information that is required to configure the Oracle RAC in our test environment.

Preparing the disks for the ASM disk groups

Resiliency to failures: The scenario that we present is only a guide for the actions that you must perform to create a running environment. We suggest that you follow the availability guidelines that are provided for the software and hardware that you use.

It is beyond the purpose of this document to present a configuration that eliminates all single points of failure (SPOFs).

Prepare the disks for the configuration of the ASM disk groups. You need to change the SCSI3 reservation policy for all disks used for ASM. This requirement applies equally to all disks managed by PowerHA HyperSwap, for example, the disk used for the CAA repository. Use the `chdev` command for all disks in all nodes used for Oracle ASM. Example 5-9 on page 178 shows the command that is executed on node ps1n01.

Example 5-9 Changing SCSI3 reservation policy

```
root@ps1n01base: /> for i in 1 2 3 9 17
>do chdev -l hdisk$i -a reserve_policy=no_reserve
done
hdisk1 changed
hdisk2 changed
hdisk3 changed
hdisk9 changed
hdisk17 changed
```

Preparing the remote command execution between cluster nodes

Prepare Secure Shell (SSH) for remote command execution. The Oracle installation requires remote command execution from any node to any node in the cluster without prompting for a password (no user interaction) or any kind of banner.

Configure the appropriate files by exchanging the host (SSH daemon) and user public keys (for the root, grid, and oracle users).

Grid configuration

Configuration data: In this document, we do not provide step-by-step Oracle Grid and Database installation steps. We only provide the relevant configuration data that we used in our environment.

For the detailed Oracle installation procedure, always check the latest Oracle documentation:

http://www.oracle.com/pls/db112/portal.portal_db?selected=11&frame=#aix_installation_guides

Prepare the grid user and oinstall group that are required for the software installation and management of the Oracle cluster software GRID and automatic storage management (ASM). Change the owner and group attributes of the raw hdisk devices /dev/rhdisk# that will be managed by Oracle to the grid user and oinstall group by using the **chown** command:

```
chown grid:oinstall /dev/rhdisk#
```

Also, prepare the oracle user and the dba group for the Oracle Database software installation and management.

We prepared the data in Table 5-2 for the four nodes. The IP labels must resolve to the IP addresses. (We use static name resolution in /etc/hosts).

Table 5-2 IP labels used in our environment

IP label types	Node1	Node2	Node3	Node4
Base IP label	ps1n01base	ps1n02base	ss1n03base	ss1n04base
Virtual IP (VIP)	ss1n01vip	ss1n01vip	ss1n01vip	ss1n01vip
RAC IP labels (private network)	ps1n01rac	ps1n02rac	ss1n03rac	ss1n04rac

ASM configuration

Table 5-3 on page 179 contains the data that we used to configure ASM and the disk groups.

Table 5-3 Disk information for ASM

Disk group	GRIDdg	datadg
Disk group members	hdisk1 and hdisk2	hdisk3 and hdisk9
Redundancy	External	External
Purpose	Data store for OCR/voting file	Data store for database data

Checking the ASM configuration

Example 5-10 shows the ASM configuration for our scenario.

Example 5-10 ASM configuration

```

grid@ps1n01base:/home/grid> asmcmd lsdk
Path
/dev/rhdisk1
/dev/rhdisk2
/dev/rhdisk3
/dev/rhdisk9

grid@ps1n01base:/home/grid> asmcmd lsdg
State   Type   Rebal Sector Block      AU   Total_MB Free_MB Req_mir_free_MB Usable_file_MB
Offline_disks Voting_files Name
MOUNTED EXTERN N      512  4096 1048576 102400 93470          0          93470
0              N DATADG/
MOUNTED EXTERN N      512  4096 1048576 102400 101493          0          101493
0              Y GRIDDG/

```

Example 5-11 shows the Oracle cluster status and the resources after the basic configuration (no database yet).

Example 5-11 CRS status of the resources

```

grid@ps1n01base:/home/grid> crs_stat -t
Name          Type          Target         State         Host
-----
ora.DATADG.dg ora....up.type ONLINE        ONLINE        ps1n01base
ora.GRIDDG.dg  ora....up.type ONLINE        ONLINE        ps1n01base
ora....ER.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora....RP.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora....P1.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora....N1.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora.asm        ora.asm.type  ONLINE        ONLINE        ps1n01base
ora.cvu        ora.cvu.type  ONLINE        ONLINE        ss1n03base
ora.gsd        ora.gsd.type  OFFLINE       OFFLINE
ora....network ora....rk.type ONLINE        ONLINE        ps1n01base
ora.oc4j       ora.oc4j.type ONLINE        ONLINE        ss1n04base
.....<< Snippet >>.....

```

Installing Oracle Database Real Application Cluster

The installation and creation of the Oracle database are not shown in detail here. We only show the result of the installation. Figure 5-2 on page 180 shows the Database Configuration Assistant: Summary window.

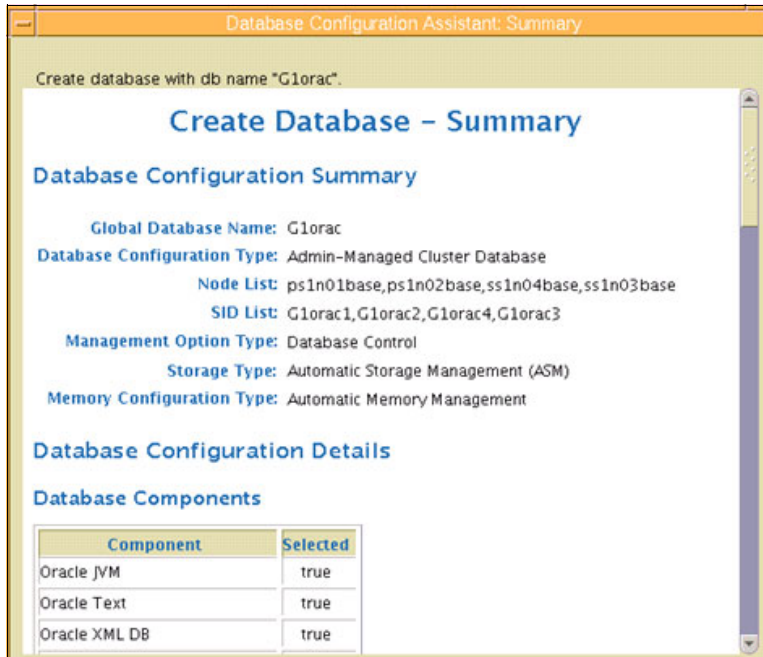


Figure 5-2 Database Configuration Assistant: Summary

Example 5-12 shows the resources managed by the Oracle Grid clustering infrastructure or Cluster Ready Services (CRS).

Example 5-12 CRS status with registered database

```
grid@ps1n01base:/home/grid> crs_stat -t
Name          Type          Target        State        Host
-----
ora.DATADG.dg ora....up.type ONLINE        ONLINE        ps1n01base
ora.GRIDDG.dg  ora....up.type ONLINE        ONLINE        ps1n01base
ora....ER.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora....N1.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora....ER.lsnr ora....er.type ONLINE        ONLINE        ps1n01base
ora.asm        ora.asm.type  ONLINE        ONLINE        ps1n01base
ora.cvu        ora.cvu.type  ONLINE        ONLINE        ps1n02base
ora.g1orac.db  ora....se.type ONLINE        ONLINE        ps1n01base
ora.gsd        ora.gsd.type  OFFLINE       OFFLINE
ora....network ora....rk.type ONLINE        ONLINE        ps1n01base
ora.oc4j       ora.oc4j.type ONLINE        ONLINE        ps1n02base
ora.ons        ora.ons.type  ONLINE        ONLINE        ps1n01base
.....<< Snippet >>.....
```

Database client configuration

We configure the Oracle Database client listener as shown in Example 5-13 on page 181 to enable load balancing and transparent application failover (TAF).

Example 5-13 Oracle client configuration stanza example

```
rp=(DESCRIPTION=
  (ENABLE = BROKEN)
  (ADDRESS_LIST=
    (LOAD_BALANCE=ON)
    (FAILOVER=ON)
    (ADDRESS=(PROTOCOL=TCP) (HOST=172.16.29.84) (PORT=1521))
    (ADDRESS=(PROTOCOL=TCP) (HOST=172.16.29.85) (PORT=1521))
    (ADDRESS=(PROTOCOL=TCP) (HOST=172.16.29.241) (PORT=1521))
    (ADDRESS=(PROTOCOL=TCP) (HOST=172.16.29.242) (PORT=1521))
  )
  (CONNECT_DATA=
    (SERVICE_NAME=rp.ibm.com)
    (FAILOVER_MODE =
      (TYPE = SELECT)
      (METHOD = BASIC)
      (RETRIES = 5)
      (DELAY = 20)
    )
  )
)
```

5.2.5 PowerHA cluster installation and configuration

The cluster configuration is based on a two-site (two nodes in each site), stretched cluster. The cluster configuration is a standard Extended Distance configuration. The differences consist in defining the following replicated storage resources:

- ▶ Storage systems:
 - DS8800-05 (Site_A)
 - DS8800-03 (Site_B)
- ▶ Mirror groups:
 - Repository mirror group
 - User mirror group

Checking the prerequisites

Check the installed cluster packages:

```
1s1pp -1 |grep cluster
```

Note: We installed PowerHA SystemMirror Enterprise Edition 7.1.2 Service Pack 2.

Important: In this scenario, the **PLANNED_HYPERSWAP_TIMEOUT** parameter is set to 20 seconds in the `ds8k_inband_mm.cfg` configuration file. The file is in the `/usr/es/sbin/cluster/xd_generic/xd_ds8k_mm` directory. The change becomes active when you restart the system.

Configure the cluster topology and resources

Because we use static IP name resolution, we add all IP addresses and labels to the `/etc/hosts` file. We create the `/etc/cluster/rhosts` file and add into it the IP label of each node's `en0`, as shown in Example 5-14.

Example 5-14 Cluster rhosts file

```
root@ps1n02base: /> cat /etc/cluster/rhosts
ps1n01base
ps1n02base
ss1n03base
ss1n04base
```

For the detailed steps for cluster configuration, see 3.2.7, “PowerHA cluster configuration” on page 50.

Example 5-15 shows the cluster topology. You can use the `cltopinfo` command or the System Management Interface Tool (SMIT) menu.

Example 5-15 PowerHA cluster topology

```
Cluster Name: G1cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: hdisk17
Cluster IP Address: 228.16.29.84
There are 4 node(s) and 1 network(s) defined
```

NODE ps1n01base:

```
Network net_ether_01
ps1n01rac      172.16.14.61
ps1n01base    172.16.29.84
```

NODE ps1n02base:

```
Network net_ether_01
ps1n02base    172.16.29.85
ps1n02rac     172.16.14.62
```

NODE ss1n03base:

```
Network net_ether_01
ss1n03base    172.16.29.241
ss1n03rac     172.16.14.71
```

NODE ss1n04base:

```
Network net_ether_01
ss1n04base    172.16.29.242
ss1n04rac     172.16.14.72
```

Defining the CAA repository

Because this is a new cluster deployment, we use a HyperSwap enabled disk for the CAA repository. See Example 5-16. We use the following command and menu selections to define the CAA repository:

smitty hacmp → Cluster Nodes and Networks → Multi Site Cluster Deployment → Define Repository Disk and Cluster IP Address

Example 5-16 Defining the CAA repository

```
                Define Repository Disk and Cluster IP Address
Type or select values in entry fields.
Press Enter AFTER making all desired changes.

* Cluster Name                [Entry Fields]
*RepositoryDisk?             G1cluster
Cluster IP Address           [(00f681f32cfe19a0)]
                             []
```

Verify and synchronize the cluster configuration by using the following SMIT command and menu selections:

smitty hacmp → Cluster Nodes and Networks → Verify and Synchronize Cluster Configuration

Cluster storage resources

Cluster storage resources are required for PowerHA to enable the kernel extension to send in-band commands to reconfigure the replicated storage resources (DS8800) under the control of PowerHA.

Use the following SMIT command and menu selections (see Example 5-17) to configure the primary and secondary storage devices:

smitty hacmp → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Storage Systems → Add a Storage System

Example 5-17 Primary storage device configuration

```
                Add a Storage System

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

* Storage System Name        [Entry Fields]
*SiteAssociation             PSDS8K
* Vendor Specific Identifier  siteA
* WNNN                       IBM.2107-00000XP411
                             500507630BFFC4C8      +
                             +
```

Repeat the task for the secondary storage device (Example 5-18 on page 184).

Example 5-18 Secondary storage device configuration

Add a Storage System

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

	[Entry Fields]	
* Storage System Name	SSDS8K	
*SiteAssociation	siteB	+
*VendorSpecificIdentifier	IBM.2107-00000WT971	+
* WNN	500507630BFFC1E2	+

Use the following SMIT command and menu selections to configure the cluster repository mirror group named repmg. This MG consists of the CAA repository disk as shown in Example 5-19.

smitty hacmp → **Cluster Applications and Resources** → **Resources** → **Configure DS8000 Metro Mirror (In-Band) Resources** → **Configure Mirror Groups** → **Add a Mirror Group**. Select **Cluster Repository**.

Example 5-19 Cluster repository mirror group definition

Add cluster Repository Mirror Group

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

	[Entry Fields]	
* Mirror Group Name	[repmg]	
* Site Name	siteA siteB	+
* Non Hyperswap Disk	hdisk17:9e9fc47a-30f5->	+
* Hyperswap Disk	hdisk17:9e9fc47a-30f5->	+
Hyperswap	Enabled	+
Consistency Group	Enabled	+
Unplanned HyperSwap Timeout (in sec)	[20]	#
Hyperswap Priority	High	+

Important: You need to set the value of “Unplanned HyperSwap Timeout (in sec)” to 20 seconds to ensure that the HyperSwap action finishes within the timeout period of Oracle RAC, which is 27 seconds, by default.

Use the following SMIT command and menu selections (Example 5-20 on page 185) to configure the user mirror group named datamg. This MG includes *all raw disks* used by Oracle RAC, including the OCR/VOTING disk group and the Data disk group (hdisk1, hdisk2, hdisk3, and hdisk9 in our scenario).

smitty hacmp → **Cluster Applications and Resources** → **Resources** → **Configure DS8000 Metro Mirror (In-Band) Resources** → **Configure Mirror Groups** → **Add a Mirror Group**. Select **User**.

Example 5-20 User mirror group definition

Add a User Mirror Group

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

	[Entry Fields]	
* Mirror Group Name	[datamg]	
Volume Group(s)		+
Raw Disk(s)	hdisk1:6841eddd-744f->	+
Hyperswap	Enabled	+
Consistency Group	Enabled	+
Unplanned HyperSwap Timeout (in sec)	[20]	#
Hyperswap Priority	Medium	+
Recovery Action	Automatic	+

Important: You need to set the value of “Unplanned HyperSwap Timeout (in sec) to 20 seconds to ensure that the HyperSwap action finishes within the timeout period of Oracle RAC, which is 27 seconds, by default.

Cluster resource groups

We can now configure the oracrg resource group and select the appropriate RG management policy. We use the following SMIT command and menu selections (Example 5-21):

smitty hacmp → **Cluster Applications and Resources** → **Resource Groups** → **Add a Resource Group**

Example 5-21 RG definition

Add a Resource Group (extended)

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

	[Entry Fields]	
* Resource Group Name	[oracrg]	
Inter-Site Management Policy	[Online On Both Sites]	+
* Participating Nodes from Primary Site	[ps1n01base ps1n02base]	+
Participating Nodes from Secondary Site	[ss1n03base ss1n04base]	+
Startup Policy	Online On All Availab>	+
Fallover Policy	Bring Offline (On Err>	+
Fallback Policy	Never Fallback	+

Resource group configuration: Two cluster management frameworks coexist in this configuration, Oracle Grid (Clusterware) and PowerHA:

- ▶ Oracle Grid provides resource management for Oracle Database Real Application Cluster (RAC).
- ▶ PowerHA provides handling for the replicated storage resources.

We configure a PowerHA resource group that manages only the HyperSwap enabled disks as RAW devices but no other resources (service IP addresses, file systems, or application controllers).

The resource group management policy resembles an intersite management policy to a concurrent resource group: Online On All Available Nodes (OOAN), Bring Offline On Error Node (BOOEN), or Never FallBack (NFB).

After we define the RG, we configure the resources and attributes that will be managed with this group (oracrg). We define the Raw Disk Universally Unique Identifiers (UUIDs) that correspond to the Oracle GRIDDG and DATADG ASM disk groups. We use the following SMIT command and menu selections (see Example 5-22 on page 187):

smitty hacmp → Cluster Applications and Resources → Resource Groups → Change/Show All Resources and Attributes for a Resource Group

Change/Show All Resources and Attributes for a Resource Group

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

[TOP]	[Entry Fields]
Resource Group Name	oracrg
Inter-site Management Policy	Online On Both Sites
Participating Nodes from Primary Site	ps1n01base ps1n02base
Participating Nodes from Secondary Site	ss1n03base ss1n04base
Startup Policy	Online On All Available Nodes
Fallover Policy	Bring Offline (On Error Node Only)
Fallback Policy	Never Fallback
Concurrent Volume Groups	<input type="checkbox"/> +
Use forced varyon of volume groups, if necessary	false+
Automatically Import Volume Groups	false +
Application Controller Name	<input type="checkbox"/> +
Tape Resources	<input type="checkbox"/> +
Raw Disk PVIDs	<input type="checkbox"/> +
Raw Disk UUIDs/hdisks	[07e37466-34b9-fe07-ea2d-04c749f206f6403]> +
PPRC Replicated Resources	<input type="checkbox"/> +
Workload Manager Class	<input type="checkbox"/> +
Disk Error Management?	no +
Miscellaneous Dataent?	<input type="checkbox"/>
SVC PPRC Replicated Resources	<input type="checkbox"/> +
EMC SRDF(R) Replicated Resources	<input type="checkbox"/> +
DS8000 Global Mirror Replicated Resources	<input type="checkbox"/> +
XIV Replicated Resources	<input type="checkbox"/> +
TRUECOPY Replicated Resources	<input type="checkbox"/> +
DS8000-Metro Mirror (In-band) Resources	datamg +

Finally, we verify and synchronize the PowerHA cluster configuration. After starting the cluster processes on all nodes, we can check the cluster process status and resource group status, as displayed in Example 5-23 on page 188.

Example 5-23 Cluster status

```
root@ss1n03base: /> /usr/sbin/clcmd lsrc -ls clstrmgrES|grep "Current state"
Current state: ST_STABLE
Current state: ST_STABLE
Current state: ST_STABLE
Current state: ST_STABLE
```

```
root@ss1n03base: /> /usr/sbin/cluster/utilities/clRGinfo
```

Group Name	State	Node
orarg	ONLINE	ps1n01base@sit
	ONLINE	ps1n02base@sit
	ONLINE	ss1n03base@sit
	ONLINE	ss1n04base@sit

5.3 Test scenarios

Important: The test results (reconfiguration times and service interruption times) are specific to our test environment. You must always test your environment to qualify the service level agreement (SLA) before committing the SLA.

We describe the tests that we performed to validate our configuration. We use various Oracle tools to interact with the test database. The configuration of the Oracle database (or the actions that we performed to create the configuration) is not described in detail. Consult with your database administrators about how to implement the testing scenario in your environment. We tested the following scenarios:

- ▶ Primary storage maintenance (planned HyperSwap)
- ▶ Node failure (unplanned)
- ▶ Primary storage failure (unplanned)
- ▶ Site failure (unplanned)

5.3.1 Test method description

Test conditions: The test methods that we used are basic programs. There is no guarantee that the results obtained can be used in real-life scenarios. You must always test your environment by using a test sequence that replicates (as closely as possible) the application that you intend to protect by using the proposed clustering infrastructure.

Follow these steps:

1. Create a test database named `rp.ibm.com`, and a test table named `testtable` of schema `testa` for verifying the database service and logging timestamps of every insertion record during the tests. Example 5-24 shows the test table description.

Example 5-24 Test table information

```
SQL> desc testa.testtable;
```

Name	Null?	Type
------	-------	------

```

PROCINST                                VARCHAR2(10) -- the instance
processing the record
RECSEQ                                  NUMBER -- record sequence
RECTIME                                TIMESTAMP(6) -- insertion time
of record, default value of the field is 'sysdate'
set the display format of the field "rectime"
alter session set nls_date_format='yyyy/mm/dd:hh24:mi:ss:ff';

```

We concurrently launch one `sqlplus` session on each node to keep inserting sequential records to the test table. The interval between every two consecutive insert operations is 1 second.

The SQL statement for every insertion record is shown in Example 5-25.

Example 5-25 Test program

```

insert into testa.testtable(procinst,recseq) values((select instance_name from
v$instance where rownum<2),1);
commit;
!sleep 1;
insert into testa.testtable(procinst,recseq) values((select instance_name from
v$instance where rownum<2),2);
commit;
!sleep 1;
.....<< Snippet >>.....

```

After finishing each test scenario, we search records indicating the timestamp interruption from the test table to calculate how long (the duration) the Oracle instance was frozen.

2. At the same time, during the SQL test, we also launch a read-only `dd` operation on each node to trace the disk I/O status from the system's perspective:

```

dd if=/dev/hdisk1 of=/dev/null bs=128&
iostat -T hdisk1 1|grep hdisk1

```

3. We configure the Network Time Protocol (NTP) service to synchronize the system clock of all of the nodes.

5.3.2 Primary storage maintenance (planned)

In this scenario, we perform a manual HyperSwap using the SMIT menus that are provided with PowerHA.

Expected behavior

The expectation is that the Oracle services are not affected during this test.

Display the status before the test

The following steps show the status:

1. The PowerHA resource group status at the beginning of the test is shown in Example 5-26.

Example 5-26 PowerHA cluster status

```

root@psln02base: /> /usr/sbin/cluster/utilities/clRGinfo
-----
Group Name      State           Node
-----

```

```

orarg          ONLINE          ps1n01base@sit
               ONLINE          ps1n02base@sit
               ONLINE          ss1n03base@sit
               ONLINE          ss1n04base@sit

```

2. List the PPRC status for the repository MG and user MG disks as shown in Example 5-27.

Example 5-27 PPRC status

```

root@ps1n02base: /> lspprc -Ao|egrep -w "hdisk1|hdisk3|hdisk2|hdisk9|hdisk17"
hdisk17 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk3 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk2 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk1 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk9 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2

```

3. List the Oracle services status by using the `crs_stat -t` command.

Perform the HyperSwap of the user mirror group manually

We performed these steps:

1. We emptied the test table:

```

Delete from testa.testable;
commit;

```

2. We concurrently launched the `sqlplus` sessions and the `dd` operations on every node (see 5.3.1, "Test method description" on page 188).

3. We performed the planned swap of the user MG by using the following SMIT command and menu selections:

```

smit hacmp → System Management (C-SPOC) → Storage → Manage Mirror
Groups → Manage User Mirror Group(s)

```

4. We waited for the system to perform the planned HyperSwap of the user MG.

5. We checked the PowerHA resource group status as shown in Example 5-28.

Example 5-28 Cluster RG status

```

root@ps1n02base: /> /usr/sbin/cluster/utilities/clRGinfo

```

```

-----
Group Name      State          Node
-----
orarg           ONLINE        ps1n01base@sit
                ONLINE        ps1n02base@sit
                ONLINE        ss1n03base@sit
                ONLINE        ss1n04base@sit

```

6. We listed the PPRC status for the repository MG and user MG disks as shown in Example 5-29.

Example 5-29 PPRC status

```

root@ps1n02base: /> lspprc -Ao|egrep -w "hdisk1|hdisk3|hdisk2|hdisk9|hdisk17"
hdisk17 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk3 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk2 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8

```

hdisk1	Active	1(s)	0	500507630bffc1e2	500507630bffc4c8
hdisk9	Active	1(s)	0	500507630bffc1e2	500507630bffc4c8

- We checked the Oracle services status. As expected, all services remained online.
- We calculated the duration of the Oracle instance that was being frozen by checking the timestamp interruption from the `testtable` and the interruption of the read operation from the `dd` output. The results are shown in Example 5-30.

Example 5-30 Checking the timestamp interruption and the interruption of the read operation

```
root@ps1n01base: /> sh /home/oracle/rundd.sh
hdisk1      28.0    10084.0   2521.0    10084      0 15:24:40
hdisk1      38.0    9360.0    2340.0    9360       0 15:24:41
hdisk1      38.0    9628.0    2404.0    9612      16 15:24:42
hdisk1      8.0    1428.0    357.0    1428       0 15:24:43
hdisk1      11.0    4012.0    1003.0    4012       0 15:24:44
hdisk1      26.0    10280.0   2567.0    10264     16 15:24:45
hdisk1      38.0    10228.0   2557.0    10228      0 15:24:46
```

```
root@ps1n02base: /> sh /home/oracle/rundd.sh
hdisk1      37.0    11812.0   2953.0    11812      0 15:24:40
hdisk1      45.0    11740.0   2935.0    11740      0 15:24:41
hdisk1      34.0    11812.0   2953.0    11812      0 15:24:42
hdisk1      24.0    5968.0    1492.0    5968       0 15:24:43
hdisk1      36.0    11804.0   2951.0    11804      0 15:24:44
hdisk1      35.0    11748.0   2937.0    11748      0 15:24:45
hdisk1      38.0    12040.0   3010.0    12040      0 15:24:46
```

```
root@ss1n03base: /> sh /home/oracle/rundd.sh
hdisk1      29.0    11772.0   2943.0    11772      0 15:24:40
hdisk1      38.0    11520.0   2880.0    11520      0 15:24:41
hdisk1      40.0    11992.0   2998.0    11992      0 15:24:42
hdisk1      23.0    8676.0    2169.0    8676       0 15:24:43
hdisk1      32.0    9660.0    2415.0    9660       0 15:24:44
hdisk1      48.0    11672.0   2918.0    11672      0 15:24:45
hdisk1      45.0    11536.0   2884.0    11536      0 15:24:46
```

```
root@ss1n04base: /> sh /home/oracle/rundd.sh
hdisk1      36.0    11424.0   2856.0    11424      0 15:24:40
hdisk1      33.0    11192.0   2798.0    11192      0 15:24:41
hdisk1      30.0    11352.0   2838.0    11352      0 15:24:42
hdisk1      25.0    6048.0    1512.0    6048       0 15:24:43
hdisk1      37.0    11196.0   2799.0    11196      0 15:24:44
hdisk1      42.0    11304.0   2826.0    11304      0 15:24:45
hdisk1      40.0    11400.0   2850.0    11400      0 15:24:46
```

```
SQL> select * from testa.testtable where procinst='rp1' and recseq < 21;
.....<<snippet>> .....
rp1      14 09-JAN-13 03.24.39.999261 PM
rp1      15 09-JAN-13 03.24.41.012581 PM
rp1      16 09-JAN-13 03.24.42.027474 PM
rp1      17 09-JAN-13 03.24.43.041724 PM
rp1      18 09-JAN-13 03.24.45.606533 PM
rp1      19 09-JAN-13 03.24.46.620032 PM
rp1      20 09-JAN-13 03.24.47.634644 PM
```

```
SQL> select * from testa.testtable where procinst='rp2' and recseq < 21;
.....<<snippet>> .....
rp2      14 09-JAN-13 03.24.40.055179 PM
rp2      15 09-JAN-13 03.24.41.064964 PM
rp2      16 09-JAN-13 03.24.42.074473 PM
rp2      17 09-JAN-13 03.24.43.084059 PM
rp2      18 09-JAN-13 03.24.44.097185 PM
rp2      19 09-JAN-13 03.24.45.106896 PM
rp2      20 09-JAN-13 03.24.46.116420 PM
```

```
SQL> select * from testa.testtable where procinst='rp3' and recseq < 21;
.....<<snippet>> .....
rp3      14 09-JAN-13 03.24.41.108267 PM
rp3      15 09-JAN-13 03.24.42.119223 PM
rp3      16 09-JAN-13 03.24.43.130317 PM
rp3      17 09-JAN-13 03.24.45.463860 PM
rp3      18 09-JAN-13 03.24.46.475668 PM
rp3      19 09-JAN-13 03.24.47.488081 PM
rp3      20 09-JAN-13 03.24.48.500076 PM
```

```
SQL> select * from testa.testtable where procinst='rp4' and recseq < 21;
.....<<snippet>> .....
rp4      14 09-JAN-13 03.24.41.029393 PM
rp4      15 09-JAN-13 03.24.42.040454 PM
rp4      16 09-JAN-13 03.24.43.051226 PM
rp4      17 09-JAN-13 03.24.44.065426 PM
rp4      18 09-JAN-13 03.24.45.076116 PM
rp4      19 09-JAN-13 03.24.46.087003 PM
rp4      20 09-JAN-13 03.24.47.097915 PM
```

9. We also checked the Oracle ASM alert log and database instance alert log on every node.
No event was generated.

Test result: *The interruption in the Oracle service was no more than 1 second in this case.*

Perform the HyperSwap of the repository mirror group manually

We performed this swap by using the following SMIT command and menu selections:

smi t hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage Cluster Repository Mirror Group(s)

We observed that the planned HyperSwap of the repository MG did not affect the Oracle services.

Path swap timing: After finishing the swap operation through the SMIT menu, we observed that the paths for hdisk17 (CAA repository) on the first node of the cluster were swapped immediately. *However, the swap required a few seconds to complete on the remaining three nodes.*

Revert to the initial configuration

At the end of this test, we performed a reverse swap of the user MG and the repository MG.

5.3.3 Node failure (unplanned)

In this test, we failed one node (by using the `halt -q` command) and observed the cluster reaction.

Expected behavior

It is expected that the Oracle services on the failing node will become unavailable, but the remaining nodes will continue to operate unaffected. New database connections will be redirected to the surviving nodes.

Display the status before the test

We performed these steps to show the status:

1. We displayed the PowerHA resource group status as shown in Example 5-31.

Example 5-31 RG status

```
root@ps1n02base: /> /usr/sbin/cluster/u*/clRGinfo
```

Group Name	State	Node
orarg	ONLINE	ps1n01base@sit
	ONLINE	ps1n02base@sit
	ONLINE	ss1n03base@sit
	ONLINE	ss1n04base@sit

2. We also checked the `clstrmgrES` service status.
3. We checked the Oracle services status by using the `crs_stat -t` command.

Perform node failure simulation

We performed these steps to simulate node failure:

1. We emptied the test table:

```
Delete from testa.testable;  
commit;
```

2. We concurrently launched the `sqlplus` sessions on every node.
3. We executed the `halt -q` command on node 2 (ps1n02base).
4. After node 2 was stopped, we checked the Oracle cluster services status by using the `crs_stat -t` command. The Oracle services on the other nodes were up and running (no disruption).
5. The PowerHA resource group status is shown in Example 5-32.

Example 5-32 RG status

```
root@ps1n02base: /> /usr/sbin/cluster/utilities/clRGinfo
```

Group Name	State	Node
orarg	ONLINE	ps1n01base@sit
	OFFLINE	ps1n02base@sit
	ONLINE	ss1n03base@sit
	ONLINE	ss1n04base@sit

6. Because there was no storage failover, there was no need to check the interruption of the read operation from the system's perspective.
7. We checked the test table and calculated the duration of time that the Oracle instance was frozen as shown in Example 5-33.

Example 5-33 Test results

```
SQL> select * from testa.testtable where recseq < 30 and procinst='rp1' order by
recseq;
```

```
.....<<snippet>> .....
rp1      18 11-JAN-13 01.12.18.999441 AM
rp1      19 11-JAN-13 01.12.20.013476 AM
rp1      20 11-JAN-13 01.12.21.021372 AM
rp1      21 11-JAN-13 01.12.22.031773 AM (starting point of time of freezing)
rp1      22 11-JAN-13 01.12.53.236481 AM (ending point of time of freezing)
rp1      23 11-JAN-13 01.12.54.248644 AM
rp1      24 11-JAN-13 01.12.55.258028 AM
rp1      25 11-JAN-13 01.12.56.266290 AM
rp1      26 11-JAN-13 01.12.57.273929 AM
rp1      27 11-JAN-13 01.12.58.284111 AM
rp1      28 11-JAN-13 01.12.59.294405 AM
rp1      29 11-JAN-13 01.13.00.303345 AM
```

29 rows selected.

```
SQL> select * from testa.testtable where recseq < 30 and procinst='rp2' order by
recseq;
```

```
rp2       1 11-JAN-13 01.12.02.144406 AM
rp2       2 11-JAN-13 01.12.03.156679 AM
rp2       3 11-JAN-13 01.12.04.163723 AM
rp2       4 11-JAN-13 01.12.05.171252 AM
rp2       5 11-JAN-13 01.12.06.179006 AM
rp2       6 11-JAN-13 01.12.07.188110 AM
rp2       7 11-JAN-13 01.12.08.197014 AM
rp2       8 11-JAN-13 01.12.09.207831 AM
rp2       9 11-JAN-13 01.12.10.215671 AM
rp2      10 11-JAN-13 01.12.11.223118 AM
rp2      11 11-JAN-13 01.12.12.230564 AM
rp2      12 11-JAN-13 01.12.13.238747 AM
rp2      13 11-JAN-13 01.12.14.247185 AM
rp2      14 11-JAN-13 01.12.15.254782 AM
rp2      15 11-JAN-13 01.12.16.261900 AM
rp2      16 11-JAN-13 01.12.17.271549 AM
rp2      17 11-JAN-13 01.12.18.281497 AM
rp2      18 11-JAN-13 01.12.19.290151 AM
rp2      19 11-JAN-13 01.12.20.298115 AM
rp2      20 11-JAN-13 01.12.21.306178 AM
(no more records since node was gone down)
20 rows selected.
```

```
SQL> select * from testa.testtable where recseq < 30 and procinst='rp3' order by
recseq;
```

```
.....<<snippet>> .....
rp3      18 11-JAN-13 01.12.19.679264 AM
rp3      19 11-JAN-13 01.12.20.694537 AM
rp3      20 11-JAN-13 01.12.21.706324 AM
```



```

rp3      21 11-JAN-13 01.12.22.991288 AM
rp3      22 11-JAN-13 01.12.53.227098 AM
rp3      23 11-JAN-13 01.12.54.241929 AM
rp3      24 11-JAN-13 01.12.55.249972 AM
rp3      25 11-JAN-13 01.12.56.261241 AM
rp3      26 11-JAN-13 01.12.57.277084 AM
rp3      27 11-JAN-13 01.12.58.291180 AM
rp3      28 11-JAN-13 01.12.59.301012 AM
rp3      29 11-JAN-13 01.13.00.312536 AM

```

29 rows selected.

```

SQL> select * from testa.testtable where recseq < 30 and procinst='rp4' order by
recseq;

```

```

.....<<snippet>> .....
rp4      18 11-JAN-13 01.12.20.265803 AM
rp4      19 11-JAN-13 01.12.21.276261 AM
rp4      20 11-JAN-13 01.12.22.289446 AM
rp4      21 11-JAN-13 01.12.53.226437 AM
rp4      22 11-JAN-13 01.12.54.342405 AM
rp4      23 11-JAN-13 01.12.55.360359 AM
rp4      24 11-JAN-13 01.12.56.383388 AM
rp4      25 11-JAN-13 01.12.57.397729 AM
rp4      26 11-JAN-13 01.12.58.416906 AM
rp4      27 11-JAN-13 01.12.59.431008 AM
rp4      28 11-JAN-13 01.13.00.441643 AM
rp4      29 11-JAN-13 01.13.01.452588 AM

```

29 rows selected.

Expected behavior: *The Oracle instance was frozen for 31 seconds in this case. This behavior is expected (as designed) and it is not influenced by the presence of the PowerHA framework or the HyperSwap enabled storage.*

8. We checked the Oracle ASM alert log and database alert log and noticed the information that is shown in Example 5-34.

Example 5-34 ASM alert log information and database alert log information about the node failure

Content segment of alert_+ASM1.log

```

Fri Jan 11 01:12:52 2013
Reconfiguration started (old inc 20, new inc 22)
List of instances:
 1 3 4 (myinst: 1)
Global Resource Directory frozen
* dead instance detected - domain 1 invalid = TRUE
* dead instance detected - domain 2 invalid = TRUE
Communication channels reestablished
Master broadcasted resource hash value bitmaps
Non-local Process blocks cleaned out

```

Content segment of alert_rp1.log

```

Fri Jan 11 01:12:52 2013
Reconfiguration started (old inc 20, new inc 22)
List of instances:

```

```

1 3 4 (myinst: 1)
Global Resource Directory frozen
* dead instance detected - domain 0 invalid = TRUE
Communication channels reestablished
Fri Jan 11 01:12:52 2013
* domain 0 not valid according to instance 3
* domain 0 not valid according to instance 4
Master broadcasted resource hash value bitmaps
Non-local Process blocks cleaned out

```

5.3.4 Primary storage failure (unplanned)

In this scenario, we simulate the primary storage failure by disabling the SAN communication between all nodes and the primary storage.

Expected behavior

PowerHA triggers an unplanned HyperSwap. All nodes will switch FC paths to access the secondary storage, which will become the PPRC source. There is no RG change as a result of this event. A short delay in the database service will be observed (the Oracle database froze for a short period during the storage reconfiguration), but there is no reconfiguration of the Oracle Grid-managed resources.

Display the status before the test

We performed these steps to display the status before the test:

1. We check the PPRC status before the test as shown in Example 5-35.

Example 5-35 PPRC status

```

root@psln02base: /> lsprrc -Ao|egrep -w "hdisk1|hdisk3|hdisk2|hdisk9|hdisk17"
hdisk17 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk3 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk2 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk1 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk9 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2

```

2. We check the system status, ensuring that the nodes are currently accessing the primary storage (Storage_A, DS8800-05).
3. We check the Oracle services status by using the `crs_stat -t` command.

Perform storage failure simulation

We performed these steps:

1. We emptied the test table:


```

Delete from test.testable;
commit;

```
2. We concurrently launched the `sqlplus` sessions and the `dd` operations on every node.
3. We disabled the zones (SAN switch) that allow all nodes to access the primary storage as shown in Example 5-36.

Example 5-36 Disabling zoning (IBM SAN switch)

```

switch#> cfgremove "powerswap","P7805LP1_fcs0_DS8805_I0234;
P7805LP1_fcs2_DS8805_I0302;P7805LP2_fcs0_DS8805_I0234;

```

```
P7805LP2_fcs2_DS8805_I0302;P7703LP1_fcs0_DS8805_I0302;
P7703LP1_fcs12_DS8805_I0234;P7703LP2_fcs0_DS8805_I0234;
P7703LP2_fcs2_DS8805_I0302"
switch#> cfgenable "powerswap"
```

4. After we disabled the zones, the unplanned HyperSwap was triggered.
5. When the unplanned HyperSwap is complete, we checked the PPRC status on all nodes. Example 5-37 shows the ps1n01base node.

Example 5-37 Checking the PPRC status on the ps1n01base node

```
root@ps1n01base: /> lspprc -Ao | egrep -w "hdisk1|hdisk2|hdisk3|hdisk9|hdisk17"
hdisk1 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk3 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk9 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk2 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk17 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
```

```
root@ps1n01base: /> lspprc -p hdisk1
path WWNN LSS VOL path
group id group status
=====
0 500507630bffc4c8 0xa2 0x01 SECONDARY
1(s) 500507630bffc1e2 0xa2 0x01 PRIMARY
```

```
path path path parent connection
group id id status
=====
0 0 Failed fscsi0 500507630b1884c8,40a2400100000000
0 1 Failed fscsi2 500507630b5304c8,40a2400100000000
1 2 Enabled fscsi4 500507630b1001e2,40a2400100000000
1 3 Enabled fscsi6 500507630b1301e2,40a2400100000000
```

6. We checked the Oracle service status. The Oracle services on all nodes were up and running.
7. We calculated the duration of the Oracle instance being frozen by checking the timestamp from the testtable and the interruption of the read operation from the dd output as shown in Example 5-38.

Example 5-38 Test results

```
root@ps1n01base: /> sh /home/oracle/rundd.sh
hdisk1 42.0 11000.0 2750.0 11000 0 21:43:56
hdisk1 26.0 11040.0 2760.0 11040 0 21:43:57
hdisk1 34.0 11348.0 2837.0 11348 0 21:43:58
hdisk1 36.0 11488.0 2872.0 11488 0 21:43:59
hdisk1 38.0 11556.0 2889.0 11556 0 21:44:00
hdisk1 40.0 11512.0 2878.0 11512 0 21:44:01
hdisk1 51.0 11352.0 2838.0 11352 0 21:44:02
hdisk1 38.0 11628.0 2907.0 11628 0 21:44:03
hdisk1 45.0 11348.0 2837.0 11348 0 21:44:04
hdisk1 39.0 11544.0 2886.0 11544 0 21:44:05
hdisk1 39.0 11256.0 2814.0 11256 0 21:44:06
hdisk1 47.0 11236.0 2809.0 11236 0 21:44:07
hdisk1 73.0 5576.0 1394.0 5576 0 21:44:08
hdisk1 100.0 0.0 0.0 0 0 21:44:09
```

```

hdisk1      100.0      0.0      0.0      0      0  21:44:10
hdisk1      100.0      0.0      0.0      0      0  21:44:11
hdisk1      100.0      0.0      0.0      0      0  21:44:12
hdisk1      100.0      0.0      0.0      0      0  21:44:13
hdisk1      100.0      0.0      0.0      0      0  21:44:14
hdisk1      100.0      0.0      0.0      0      0  21:44:15
hdisk1      100.0      0.0      0.0      0      0  21:44:16
hdisk1      100.0      0.0      0.0      0      0  21:44:17
hdisk1      100.0      0.0      0.0      0      0  21:44:18
hdisk1      100.0      0.0      0.0      0      0  21:44:19
hdisk1      100.0      0.0      0.0      0      0  21:44:20
hdisk1      100.0      0.0      0.0      0      0  21:44:21
hdisk1      100.0      0.0      0.0      0      0  21:44:22
hdisk1      26.0      0.0      0.0      0      0  21:44:23
hdisk1      0.0      0.0      0.0      0      0  21:44:24
hdisk1      27.0      7848.0    1962.0    7848    0  21:44:25
hdisk1      34.0      11040.0   2760.0    11040   0  21:44:26
hdisk1      37.0      10604.0   2651.0    10600   4  21:44:27
hdisk1      40.0      10424.0   2606.0    10424   0  21:44:28
hdisk1      37.0      10516.0   2629.0    10516   0  21:44:29
hdisk1      28.0      10532.0   2633.0    10532   0  21:44:30

```

```

SQL> select * from testa.testtable where recseq < 35 and procinst='rp1' order by
recseq;

```

```

.....<<snippet>> .....
rp1      13 10-JAN-13 09.44.02.334247 PM
rp1      14 10-JAN-13 09.44.03.346327 PM
rp1      15 10-JAN-13 09.44.04.357824 PM
rp1      16 10-JAN-13 09.44.05.368043 PM
rp1      17 10-JAN-13 09.44.06.380477 PM
rp1      18 10-JAN-13 09.44.07.393163 PM
rp1      19 10-JAN-13 09.44.08.404761 PM ( I/O suspended )
rp1      20 10-JAN-13 09.44.25.307246 PM ( I/O resumed )
rp1      21 10-JAN-13 09.44.26.319532 PM
rp1      22 10-JAN-13 09.44.27.332238 PM
rp1      23 10-JAN-13 09.44.28.346017 PM rp1      22 10-JAN-13 09.44.29.332238 PM
rp1      22 10-JAN-13 09.44.30.332238 PM
rp1      24 10-JAN-13 09.44.31.484704 PM
rp1      25 10-JAN-13 09.44.32.512650 PM
rp1      26 10-JAN-13 09.44.33.525361 PM
rp1      27 10-JAN-13 09.44.34.539131 PM
rp1      28 10-JAN-13 09.44.35.556596 PM
rp1      29 10-JAN-13 09.44.36.571180 PM

```

8. We checked the Oracle ASM alert log and the database instance alert log on every node. No related events were generated.

Test result: *The Oracle instance was frozen for 17 seconds in this case.*

Zoning configuration restore

After re-enabling the zoning, all nodes recovered from the failed path automatically. We list the PPRC status as shown in Example 5-39 on page 199.

Example 5-39 PPRC status

```
root@ps1n01base: /> lsspprc -Ao | egrep -w "hdisk1|hdisk3|hdisk2|hdisk9|hdisk17"
hdisk1 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk3 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk9 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk2 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
hdisk17 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8
```

```
root@ps1n01base: /> lsspprc -p hdisk1
path WWNN LSS VOL path
group id group status
=====
0 500507630bffc4c8 0xa2 0x01 SECONDARY
1(s) 500507630bffc1e2 0xa2 0x01 PRIMARY
```

```
path path path parent connection
group id id status
=====
0 0 Enabled fscsi0 500507630b1884c8,40a2400100000000
0 1 Enabled fscsi2 500507630b5304c8,40a2400100000000
1 2 Enabled fscsi4 500507630b1001e2,40a2400100000000
1 3 Enabled fscsi6 500507630b1301e2,40a2400100000000
```

5.3.5 Site failure (unplanned)

In this scenario, we simulate a complete site failure by simultaneously halting the ps1n01 and ps2n02 nodes and disabling the zones to Storage_A (DS8800-05).

Expected behavior

After experiencing a short freeze, the database service continues to provide service from the nodes in Site_B (ss1n03 and ss1n04). Oracle Clusterware will reconfigure the cluster resource for operation on the two surviving nodes. The storage access will be swapped to Storage_B without any impact on the Oracle services (other than a short freeze).

Display the status before the test

We performed these steps:

1. We checked the PPRC status before the test as shown in Example 5-40.

Example 5-40 Checking the PPRC status

```
root@ps1n02base: /> lsspprc -Ao | egrep -w "hdisk1|hdisk3|hdisk2|hdisk9|hdisk17"
hdisk17 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk3 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk2 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk1 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
hdisk9 Active 0(s) 1 500507630bffc4c8 500507630bffc1e2
```

2. We checked the system status, ensuring that the cluster nodes are currently accessing the primary storage.
3. We checked the Oracle services status by using the `crs_stat -t` command.

Perform site failure simulation

We performed these steps:

1. We emptied the test table:

```
Delete from test.testable;  
commit;
```

2. We concurrently launched the `sqlplus` sessions and `dd` operations on each node.
3. We halted two nodes at Site_A and disabled the primary storage zoning (between all nodes and the primary storage) at the same time to simulate the site failure.
4. After the site failure simulation, we checked the PPRC Status on the nodes at Site_B as shown in Example 5-41.

Example 5-41 Checking the PPRC status on the nodes at Site_B

```
root@ss1n03base: /> lspprc -Ao | egrep -w "hdisk1|hdisk2|hdisk3|hdisk9|hdisk17"  
hdisk2 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk17 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk3 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk9 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk1 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
  
root@ss1n04base: /> lspprc -Ao | egrep -w "hdisk1|hdisk2|hdisk3|hdisk9|hdisk17"  
hdisk9 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk3 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk2 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk1 Active 1(s) 0 500507630bffc1e2 500507630bffc4c8  
hdisk17 Active 1(s) 0 500507630bffc4c8,500507630bffc1e2
```

5. We observed that the Oracle services on the nodes at Site_B were up and running.
6. We checked the `iostat` command output and the test table and calculated the duration of the Oracle instance being frozen, as shown in Example 5-42.

Example 5-42 Test results

```
root@ss1n03base: /> sh /home/oracle/rundd.sh  
hdisk1 40.0 11140.0 2785.0 11140 0 14:18:52  
hdisk1 33.0 11584.0 2896.0 11584 0 14:18:53  
hdisk1 45.0 11320.0 2830.0 11320 0 14:18:54  
hdisk1 44.0 11012.0 2753.0 11012 0 14:18:55  
hdisk1 34.0 11052.0 2763.0 11052 0 14:18:56  
hdisk1 41.0 11300.0 2825.0 11300 0 14:18:57  
hdisk1 28.0 11168.0 2792.0 11168 0 14:18:58  
hdisk1 61.0 6764.0 1691.0 6764 0 14:18:59  
hdisk1 100.0 0.0 0.0 0 0 14:19:00  
hdisk1 100.0 0.0 0.0 0 0 14:19:01  
hdisk1 100.0 0.0 0.0 0 0 14:19:02  
hdisk1 100.0 0.0 0.0 0 0 14:19:03  
hdisk1 100.0 0.0 0.0 0 0 14:19:04  
hdisk1 100.0 0.0 0.0 0 0 14:19:05  
hdisk1 100.0 0.0 0.0 0 0 14:19:06  
hdisk1 100.0 0.0 0.0 0 0 14:19:07  
hdisk1 100.0 0.0 0.0 0 0 14:19:08  
hdisk1 100.0 0.0 0.0 0 0 14:19:09  
hdisk1 100.0 0.0 0.0 0 0 14:19:10  
hdisk1 100.0 0.0 0.0 0 0 14:19:11
```

hdisk1	100.0	0.0	0.0	0	0	14:19:12
hdisk1	100.0	0.0	0.0	0	0	14:19:13
hdisk1	41.0	0.0	0.0	0	0	14:19:14
hdisk1	0.0	0.0	0.0	0	0	14:19:15
hdisk1	26.0	6568.0	1642.0	6568	0	14:19:16
hdisk1	40.0	11468.0	2867.0	11468	0	14:19:17
hdisk1	43.0	10536.0	2634.0	10536	0	14:19:18
hdisk1	36.0	11136.0	2784.0	11136	0	14:19:19
hdisk1	37.0	10448.0	2612.0	10448	0	14:19:20
hdisk1	34.0	10612.0	2653.0	10612	0	14:19:21
hdisk1	38.0	10456.0	2614.0	10456	0	14:19:22
hdisk1	44.0	11052.0	2763.0	11052	0	14:19:23
hdisk1	46.0	10896.0	2724.0	10896	0	14:19:24
hdisk1	34.0	9920.0	2480.0	9920	0	14:19:25
hdisk1	35.0	11000.0	2750.0	11000	0	14:19:26
hdisk1	37.0	10868.0	2717.0	10868	0	14:19:27
hdisk1	34.0	10880.0	2720.0	10880	0	14:19:28
hdisk1	42.0	11164.0	2791.0	11164	0	14:19:29

```
SQL> select * from testa.testtable where recseq < 80 and procinst='rp3' order by
recseq;
```

```
.....<<snippet>> .....
```

```
rp3      60 11-JAN-13 02.18.53.111353 PM
rp3      61 11-JAN-13 02.18.54.122757 PM
rp3      62 11-JAN-13 02.18.55.136902 PM
rp3      63 11-JAN-13 02.18.56.147761 PM
rp3      64 11-JAN-13 02.18.57.160247 PM
rp3      65 11-JAN-13 02.18.58.177715 PM
rp3      66 11-JAN-13 02.18.59.192885 PM
rp3      67 11-JAN-13 02.19.00.203211 PM (I/O suspended)
rp3      68 11-JAN-13 02.19.16.320609 PM (I/O resumed, then RAC found nodes down
and started freeze)
rp3      69 11-JAN-13 02.19.58.448807 PM (RAC finished resource reconfiguration and
was unfreezed.)
rp3      70 11-JAN-13 02.19.59.472150 PM
rp3      71 11-JAN-13 02.20.00.496058 PM
rp3      72 11-JAN-13 02.20.01.522878 PM
rp3      73 11-JAN-13 02.20.02.537024 PM
rp3      74 11-JAN-13 02.20.03.550383 PM
rp3      75 11-JAN-13 02.20.04.568054 PM
rp3      76 11-JAN-13 02.20.05.585756 PM
rp3      77 11-JAN-13 02.20.06.603937 PM
rp3      78 11-JAN-13 02.20.07.617660 PM
rp3      79 11-JAN-13 02.20.08.637082 PM
```

```
79 rows selected.
```

Test result: *The Oracle instance was frozen for 57 seconds in this case.*

7. We checked the Oracle ASM alert log and the database alert log and noticed the messages shown in Example 5-43 on page 202.

Example 5-43 Oracle alert logs

Content segment of alert_+ASM3.log

Fri Jan 11 14:19:57 2013
Reconfiguration started (old inc 28, new inc 30)
List of instances:
3 4 (myinst: 3)
Global Resource Directory frozen
* dead instance detected - domain 1 invalid = TRUE
* dead instance detected - domain 1 invalid = TRUE
* dead instance detected - domain 1 invalid = TRUE
* dead instance detected - domain 2 invalid = TRUE
Communication channels reestablished
Master broadcasted resource hash value bitmaps
Non-local Process blocks cleaned out

Content segment of alert_+rp3.log

Fri Jan 11 14:19:57 2013
Reconfiguration started (old inc 28, new inc 30)
List of instances:
3 4 (myinst: 3)
Global Resource Directory frozen
* dead instance detected - domain 0 invalid = TRUE
Communication channels reestablished
Master broadcasted resource hash value bitmaps
Non-local Process blocks cleaned out

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following book provides additional information about the topic in this document. Publications referenced in this list might be available in softcopy only.

- ▶ *IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX*, SG24-8106-00

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

Online resources

This website is also relevant as a further information sources:

- ▶ IBM PowerHA SystemMirror Enterprise Edition V7 announcement letter
http://www-01.ibm.com/common/ssi/ShowDoc.wss?docURL=/common/ssi/rep_ca/4/760/ENUSJP12-0364/index.html&lang=en&request_locale=en
- ▶ Virtual I/O Server documentation
<http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=%2Fiphb1%2Fiphb1kickoff.htm>
- ▶ IBM PowerHA SystemMirror for AIX web page
<http://www-03.ibm.com/systems/power/software/availability/aix/>
- ▶ IBM DS8000 Copy Services documentation
<http://publib.boulder.ibm.com/infocenter/dsichelp/ds8000ic/index.jsp>
- ▶ *PowerHA Storage-based high availability and disaster recovery manual*
http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/hacmp_pprc_pdf.pdf
- ▶ PowerHA SystemMirror recommendations
http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/hacmp_pprc_pdf.pdf
- ▶ Installation guidelines provided by the Oracle documentation
http://www.oracle.com/pls/db112/portal.portal_db?selected=11
- ▶ Oracle support for environments using virtualized hardware resources
<http://www.oracle.com/technetwork/database/virtualizationmatrix-172995.html>
- ▶ Oracle Metalink
<http://bit.ly/YEgCBq>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



Deploying PowerHA Solution with AIX HyperSwap



Uses in-band storage replication management with PowerHA

This IBM Redpaper publication will help you plan, install, tailor, and configure the new IBM PowerHA with IBM HyperSwap clustering solution.

Describes AIX and storage configuration for HyperSwap

PowerHA with HyperSwap adds transparent storage protection for replicated storage, improving overall system availability by masking storage failures.

Shows Active-Active and Active-Standby clusters

The PowerHA cluster is an Extended Distance cluster with two sites. It manages, in principle, the replicated storage infrastructure through HyperSwap functionality.

The storage is provided by two DS8800s configured to replicate each other using Metro Mirror Peer-to-Peer Remote Copy (PPRC) synchronous replication. DS8800 supports in-band (SCSI commands) communication, which is used to manage (and automate) the replication using IBM AIX HyperSwap framework and PowerHA automation and management capabilities.

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**