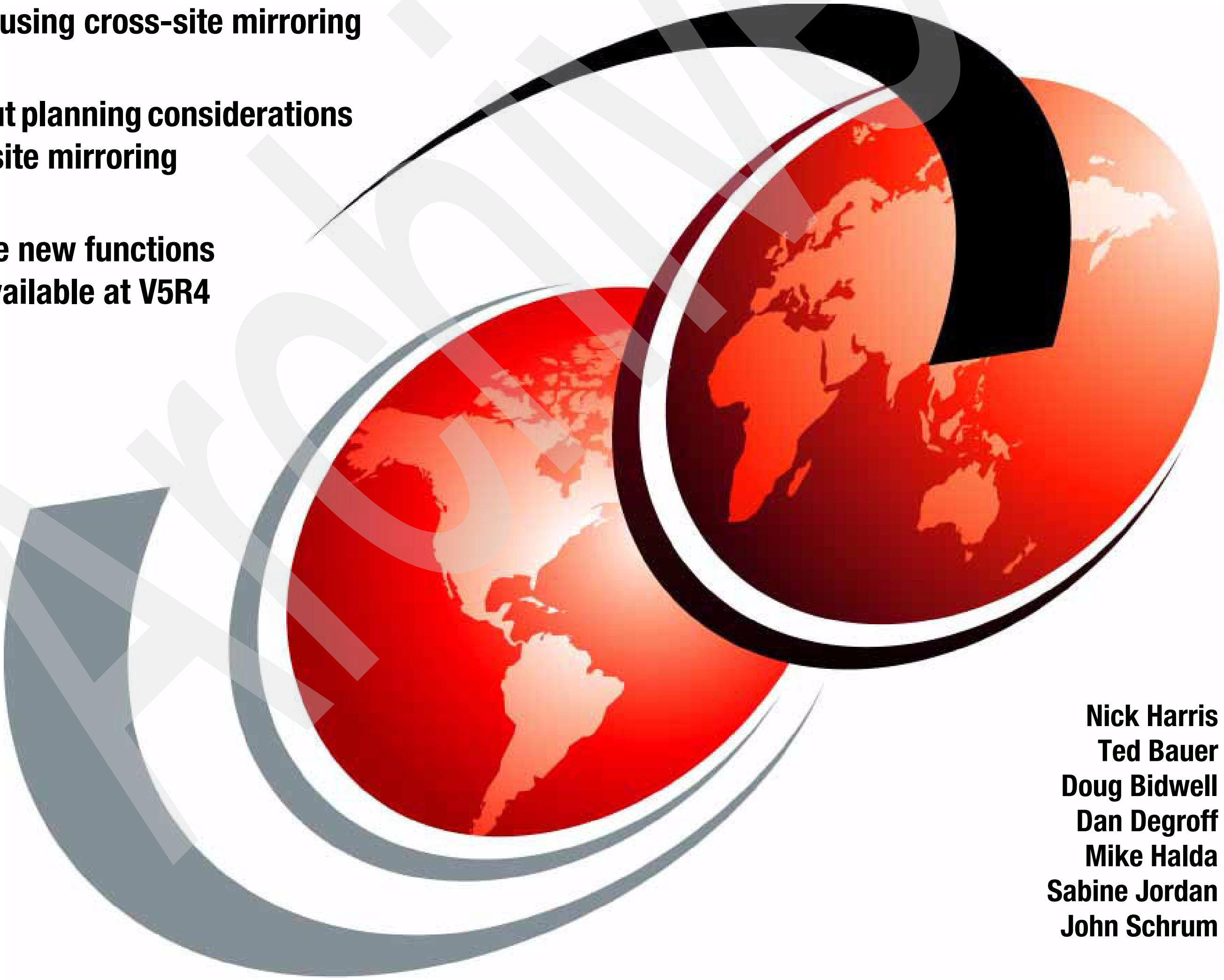
Availability Management

Planning and Implementing Cross-Site Mirroring on IBM System i5

Understand availability technology for IBM i5/OS using cross-site mirroring

Learn about planning considerations for cross-site mirroring

Explore the new functions that are available at V5R4



Nick Harris
Ted Bauer
Doug Bidwell
Dan Degroff
Mike Halda
Sabine Jordan
John Schrum

ibm.com/redbooks

International Technical Support Organization

Availability Management: Planning and Implementing Cross-Site Mirroring on IBM System i5

November 2007

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

Second Edition (November 2007)

This edition applies to V5R3 and V5R4 of IBM i5/OS.

Contents

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

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

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

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

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

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

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

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

Trademarks

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

AS/400®
ClusterProven®
Cross-Site®
DS6000™
DS8000™
Enterprise Storage Server®
eServer™
FlashCopy®
i5/OS®
IBM®
iSeries®
OS/400®
Redbooks®
Redbooks (logo) ®
System i™
System i5™
System/38™
TotalStorage®
xSeries®

The following terms are trademarks of other companies:

SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

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

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

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

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

Preface

In this IBM® Redbooks® publication, we introduce the concept of cross-site mirroring (XSM) as implemented in IBM OS/400® or i5/OS® at V5R3M0 and later. XSM describes the replication of data at multiple sites. It involves the use of clustering, cluster resource groups (CRGs), independent auxiliary storage pools (IASPs), and other components.

In this updated version, we include the new i5/OS V5R4 functions of Administrative Domain and Source Site Tracking. We also include preview information of the Target Site Tracking function.

An additional component of this highly available technology is that XSM keeps two identical copies of an independent disk pool at two sites to provide high availability and disaster recovery. These sites can be geographically close to one another or far apart, depending on the needs of the business.

This book is written for IBM technical professionals, Business Partners, and customers who are considering, planning, and implementing a highly available solution on the IBM System i5™ platform.

The team that wrote this book

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

Nick Harris is a Consulting IT Specialist for the IBM System i™ platform. He specializes in System i and IBM eServer™ iSeries® hardware, i5/OS and OS/400 software, logical partitioning, high availability (HA), external disk, Microsoft® Windows® integration, and Linux®. He writes IBM Redbooks publications and teaches IBM classes at ITSO technical forums worldwide in his areas of specialty. In his work, he explains how these areas are related to system design and server consolidation. He spent 13 years in the United Kingdom (U.K.) AS/400® Business, where he worked with S/36, S/38, AS/400, and iSeries servers. You can contact Nick by sending e-mail to: niharris@us.ibm.com

Ted Bauer is a Software Engineer working at the IBM Development Lab in Rochester, Minnesota. Ted specializes in System i5 and i5/OS HA. Ted coordinates the Rochester cluster lab and builds HA configurations for customer proofs of concept, performance testing, and application testing. You can contact Ted by sending e-mail to: twbauer@us.ibm.com

Doug Bidwell is a consultant for DB Associates, an IBM Business Partner. He specializes in iSeries and System i5 HA. Doug has implemented most aspects of availability for System i5 including clusters, cross-site mirroring, geographic mirroring, and logical replication. You can contact Doug by sending e-mail to: tuner@dlbassoc.com

Dan Degroff is an Availability Specialist working in the High Availability competency center in the IBM Development Lab in Rochester. Dan specializes in System i5 and i5/OS HA. He works with the Rochester cluster lab and builds HA configurations for customer proofs of concept, performance testing, and application testing. You can contact Dan by sending e-mail to: degroff@us.ibm.com

Mike Halda is a a consultant for Vision Solutions, an IBM High Availability Business Partner. Mike specializes in the planning, implementation, and performance of the OMS solution from Vision Solutions. Mike works out of the IBM Rochester Lab where he maintains a close relationship between Vision Solutions and System i5 Development. You can contact Mike by sending e-mail to: mhalda@visionsolutions.com

Sabine Jordan is a consulting IT Specialist working in IBM Germany. She has worked on several XSM implementations, for both SAP® and non-SAP environments. Among these implementations, she has created concepts for the XSM design and implemented the whole project (cluster setup, application changes), as well as performed customer education and testing. In addition, Sabine presents and delivers workshops (internal and external) on XSM as well as high availability and disaster recovery. You can contact her by sending e-mail to: sabine_jordan@de.ibm.com

John Schrum is a consultant for Vision Solutions, an IBM High Availability Business Partner. John specializes in planning, implementation and maintenance of the OMS solution from Vision Solutions with its customers. You can contact John by sending e-mail to: jschrum@visionsolutions.com

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

Michael Cain
Patrick L. Cawley
Tom Dylla
Scott Helt
Henry May
Bob Miller
Vicki Morey
Ron Peterson
James Ranweiler
Curt Schemmel
Mike Snyder
Kiswanto Thayib
John Unterholzner
Geoffrey Warren
IBM Development Lab, Rochester, MN

Chris Felner-Goesh
David Granum
Steve Hershey
James Johnson
Christopher Labonte
Rob Larson
Andrei Matetic
Nguyen Nguyen
Pietro Pastori
Shirley Pintos
Marilyn Oliver
Shawn Reding
Bob Smoot
IBM Support, Rochester, MN

Become a published author

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

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

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

`ibm.com/redbooks/residencies.html`

Comments welcome

Your comments are important to us!

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

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

  `ibm.com/redbooks`

- Send your comments in an e-mail to:

  `redbook@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

1

Introduction to cross-site mirroring

In this chapter, we introduce the concept of cross-site mirroring (XSM) as implemented in OS/400 or i5/OS at V5R3M0. *Cross-site mirroring* is the replication of data at multiple sites. It involves the use of clustering, cluster resource groups (CRGs), independent auxiliary storage pools (IASPs), and other components. A subfunction of XSM is *geographic mirroring*, which describes the storage management replication function of XSM.

XSM keeps two identical copies of an independent disk pool at two sites to provide high availability (HA) and disaster recovery. These sites can be geographically close to one another or far apart, depending on the needs of the business. The copy owned by the primary node is the *production copy*, and the copy owned by a backup system at the other site is the *mirror copy*. User operations and applications access the independent disk pool on the primary node, which is the node that owns the production copy. By having a second copy of critical data at a remote location, you ensure greater protection and availability for protection against such events as fire or a natural disaster.

XSM provides synchronous and asynchronous send modes. Changes that are made to the source system are guaranteed to be made in the same order on the target system.XSM is intended for clustered systems environments.

In this chapter, we explain these concepts of XSM further.

1.1 What is new in V5R4

In this section, we provide a brief description of what is new with i5/OS V5R4 in regard to XSM. We explain the concepts of Source Site Tracking and Target Site Tracking as well as the new Administrative Domain and its functions. In addition to positioning these new features in later chapters, we explain how to set them up.

1.1.1 Source Site Tracking

Before the availability of Source Site Tracking with V5R4, you had to fully re-synchronize your backup IASP to your production IASP by sending all the data in the IASP over your communication lines. You did this whenever the production system was unable to send changes to the backup system, for the following reasons, for example:

- You needed to take down the backup system, for example to install program temporary fixes (PTFs).
- The communication link used by XSM to send data to the backup system was not available.

Source Site Tracking helps to eliminate the need to fully re-synchronize from the production copy to the backup copy in cases where XSM is *suspended*.

1.1.2 Administrative Domain

A Cluster Administrative Domain monitors and synchronizes changes to selected resources from the system auxiliary storage pool (ASP) within a cluster. The Cluster Administrative Domain provides easier management and synchronization of attributes for resources that are shared within a cluster, such as environment variables or user profiles. It can be used to help maintain a consistent environment on the systems included in the Administrative Domain, thereby ensuring that applications behave the same on all systems included in the cluster.

A Cluster Administrative Domain is represented by a peer CRG. When a Cluster Administrative Domain is created, the peer CRG is automatically created by the system. The name of the Cluster Administrative Domain becomes the name of the peer CRG. Membership in the Cluster Administrative Domain can be changed by adding and removing nodes to the recovery domain of the peer CRG. The nodes that make up the Cluster Administrative Domain are defined by the recovery domain of the peer CRG.

All of the nodes are peer nodes. This means that changes that are made to objects controlled by the peer CRG on *any* of the nodes in the peer CRG are synchronized to *all other* nodes within the peer CRG. Replicate nodes are not allowed in a Cluster Administrative Domain. A cluster node can be defined only in one Cluster Administrative Domain within the cluster.

1.1.3 Target Site Tracking

Target Site Tracking has been introduced with the PTF MF40053 for i5/OS V5R4. Target Site Tracking, together with Source Site Tracking, allows you to perform the following actions:

- Detach XSM
- Vary on the backup copy on the backup system
- Use the copy on the backup system for any activity that you want, such as doing a backup or performing tests

- Re-attach XSM
- Have the systems resynchronize to each other without sending all your IASP data over your communication lines

Be aware that changes made to the backup copy of your data while in detached mode will be overwritten from the production site.

1.2 Clustering

In this section, we briefly describe clustering, its basic components, and concepts. We provide the basic elements that are required before you can configure XSM.

For a complete discussion about clustering and how to set up a cluster, refer to *Clustering and IASPs for Higher Availability on the IBM eServer iSeries Server*, SG24-5194.

1.2.1 Concepts of clustering

A *cluster* can be defined as a collection of interconnected complete computers, or *nodes*, that appear on a network as a single machine. The cluster is managed as a single system or operating entity. It is designed specifically to tolerate component failures and to support the addition or subtraction of components in a way that is transparent to users.

Clustering becomes an important concept for both high availability and disaster recovery discussions. The main purpose of clustering is to achieve *high availability*. HA allows important production data and applications to be available during periods of planned system outages.

Clustering is also used for *disaster recovery* plans. Disaster recovery typically refers to ensuring that the same important production data and applications are available in the event of an unplanned system outage, caused many times by natural disasters.

Cluster Resource Services, a component of OS/400 or i5/OS, provides the following functions:

- Tools to create and manage clusters, the ability to detect a failure within a cluster, and switchover and failover mechanisms to move work between cluster nodes for planned or unplanned outages
- A common method for setting up object replication for nodes within a cluster

  This includes the data objects and program objects that are necessary to run applications that are cluster-enabled.
- Mechanisms to automatically switch applications and users from a primary to a backup node within a cluster for planned or unplanned outages
- Heartbeat monitoring, which uses a low-level message function to constantly ascertain that every node can communicate with other nodes in the cluster

  If a node fails or a break occurs in the network, heartbeat monitoring tries to re-establish communications. If communications cannot be re-established within a designated time, the heartbeat monitoring reports the failure to the rest of the nodes within the cluster.

1.2.2 Cluster components

A cluster is made up of the following components:

- Cluster node
  - Primary node
  - Backup node
  - Replicate node
- Cluster resource group
  - Data resilient CRG (type-1)
  - Application resilient CRG (type-2)
  - Device resilient CRG (type-3)
  - Peer CRG
- Cluster resource services
- Cluster version
- Device domain
- Resilient resources
- Cluster management support and clients

Figure 1-1 shows the components of a cluster.

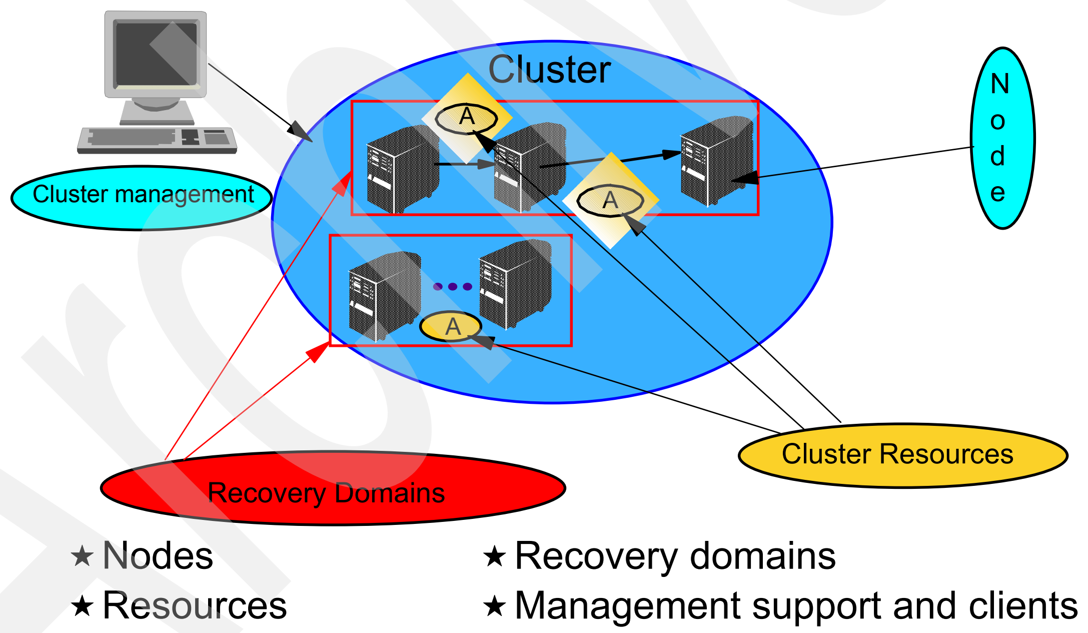


Figure 1-1 Cluster components

A *cluster node* is any System i environment or partition that is a member of a cluster. Cluster communications that run over IP connections provide the communications path between cluster services on each node in the cluster. A cluster node can operate in one or more of the three following nodes:

- A *primary node* is the cluster node that is the primary point of access for cluster resources.
- A *backup node* is a cluster node that can assume the primary role if the primary node fails or a manual switchover is initiated.

- A *replicate node* is a cluster node that maintains copies of the cluster resources but is unable to assume the role of primary or backup.

A *cluster resource group* (CRG) is an OS/400 or i5/OS external system object that is a set or group of cluster resources. The CRG describes a list of recovery domain nodes and supplies the name of the CRG exit program that manages cluster-related events for that group. One such event is moving the users from one node to another node in case of a failure.

CRG objects are defined either as data resilient, application resilient, or device resilient:

- A *data resilient CRG (type-1)* allows multiple copies of data to be maintained on more than one node in a cluster.
- An *application resilient CRG (type-2)* allows an application (program) to run on any of the nodes in a cluster.
- A *device resilient CRG (type-3)* allows a hardware resource to be switched between systems. The device CRG contains a list of device configuration objects that are used for clustering. Each object represents an IASP.

A *peer CRG* defines nodes in the recovery domain with peer roles. It is used to represent the Cluster Administrative Domain. It contains Monitored Resource Entries, for example user profiles, that can be synchronized between the nodes in the CRG.

Cluster resource services are the set of OS/400 or i5/OS system service functions that support System i cluster implementations.

The *cluster version* identifies the communication level of the nodes in the cluster.

A *device domain* is a subset of cluster nodes, across which a set of resilient devices, such as an IASP, can be shared. The sharing is not concurrent for each node, which means that only one node can use the resilient resource at one time. Through the configuration of the primary node, the secondary node is made aware of the individual hardware within the CRG and is ready to receive the CRG should the resilient resource be switched. A function of a device domain is to prevent conflicts that would cause the failure of an attempt to switch a resilient device between systems.

Figure 1-2 shows a device domain with a primary and a secondary node and a switchable device (an IASP) that can be switched from Node 1 to Node 2.

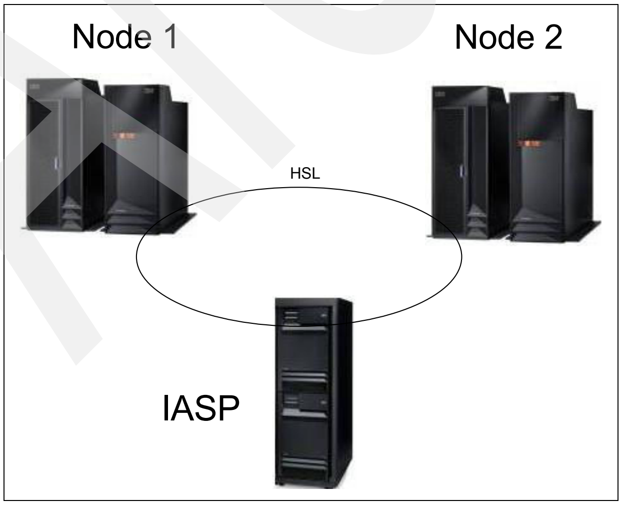


Figure 1-2 An example of a device domain

A *resilient resource* is a device, data, or an application that can be recovered if a node in the cluster fails:

- *Resilient data* is data that is replicated, or copied, on more than one node in a cluster.
- *Resilient applications* can be restarted on a different cluster node without requiring the clients to be reconfigured.
- *Resilient devices* are physical resources, represented by a configuration object such as a device description, that are accessible from more than one node in a cluster through the use of switched disk technology and independent disk pools.

Cluster management support and clients

IBM provides a cluster management graphical user interface (GUI) that is accessible through iSeries Navigator and is available through Option 41 of OS/400 or i5/OS. The utility allows you to create and manage a cluster that uses switchable IASPs and to ensure data availability. The cluster management GUI features a wizard that steps you through the creation of the cluster and all of its components.

1.3 Auxiliary storage pools

In this section, we provide an introduction to types of auxiliary storage pools and explain how they relate to clustering and XSM.

1.3.1 Concepts of auxiliary storage pools

Since the announcement of the AS/400 in 1988, *auxiliary storage pools* have existed. These ASPs allow you to divide the total disk storage on the system into logical groups, or *disk pools*, in order to limit the impact of storage-device failures and to reduce recovery time. You can then isolate one or more applications or data in one or more ASPs, for various reasons related to backup and recovery, performance, or other purposes.

ASPs include the system ASP and user ASPs:

- The *system ASP* contains SLIC and OS/400 or i5/OS code. There is only one system ASP per system or partition, and it is always numbered 1.
- *User ASPs* are any other ASPs that are defined on the system, other than the system ASP.
  - *Basic user ASPs* are numbered 2 through 32. Data in a basic user ASP is always accessible whenever the server is up and running.
  - *Independent user ASPs* are numbered 33 through 255.

1.3.2 Concepts of independent auxiliary storage pools

IASPs are a type of user ASP, numbered 33 through 255. The system assigns the IASP number, where, for a basic ASP, the user can choose the number. IASPs are different from basic ASPs.

IASPs can be used on a single system or *switched* between multiple systems or logical partitions (LPARs) when the IASP is associated with a switchable hardware group, otherwise known as a *device CRG*. When used on a single system, the IASP can be dynamically varied on or off, without restarting the system. In iSeries Navigator, the IASP and its contents can be dynamically made available or unavailable to the system.

When used across multiple systems, clustering support is required between the systems, and the cluster management GUI in iSeries Navigator is used to switch the IASP across systems in the cluster. This is referred to as a *switchable IASP*. At any given time, the IASP can be used by only one of those systems. That is multiple systems cannot simultaneously use the IASP.

IASPs also differ from basic ASPs because they are identified by a device name on the System i5 machine. This device can be varied on or off to make it available or unavailable, respectively. This can be done without a system initial program load (IPL), which saves a lot of time and increases the flexibility offered by ASPs.

1.4 Cross-site mirroring

In this section, we describe the general relationship between the XSM functions of clustering and ASPs.

1.4.1 Concepts of cross-site mirroring

XSM is part of OS/400 or i5/OS Option 41 - High Availability Switchable Resources. It provides the following functions:

- Data resilience
  - Mirroring an ASP group from one location to a second location
  - Switchover or automatic failover to the secondary copy in the event of an outage at the primary location
- Extended capabilities for basic switchable IASPs
  - Addresses single point of failure
  - Provides the possibility of multiple data copies
  - Alleviates switchable tower connectivity restrictions
- Site data resiliency protection in addition to high availability
  - Keeps the second copy of the IASP at another “site”
  - Can make the other site geographically remote
- Additional backup nodes for resilient data
  - Can stores both copies of IASP in switchable devices
  - Can switch each copy between nodes locally

XSM provides the ability to replicate changes made to the production copy of an IASP to a mirror copy of that IASP. As data is written to the production copy of an IASP, the operating system mirrors that data to a second copy of the IASP through another system. This process keeps multiple identical copies of the data.

Changes written to the production copy on the source system are guaranteed to be made in the same order to the mirror copy on the target system. If the production copy of the IASP fails or is shut down, you have a hot backup, in which case the mirror copy becomes the production copy.

The IASP used in XSM has the benefits of any other IASP, with its ability to be made available or unavailable (varied on or off), and you have greater flexibility for the following reasons:

- You can protect the production IASP and mirror IASP with the protection that you prefer, either disk unit mirroring or device parity protection (RAID-5 or RAID-6). Moreover, the production IASP and the mirror IASP are not required to have the same type of protection. While no protection is required for either IASP, we highly recommend that you use some type of protection for most scenarios.

- You can set the threshold of the IASP to warn you when storage space is running low. The server sends a message, allowing you the time to add more storage space or to delete unnecessary objects. Be aware that, if the user ignores the warning and the production IASP becomes full, the application stops and objects cannot be created. That is, with IASPs, there is no overflowing of data into the system disk pool the way there is with basic user ASPs.
- The mirror copy can be detached and then separately be made available to perform save operations, create reports, or perform data mining. However, when the mirror copy is reattached, it is synchronized with the production copy, and all modifications made to the detached copy are lost. Be aware that synchronization can be a lengthy process unless you use the V5R4 Source Side Tracking function.
- If you configure the IASPs to be switchable, you increase your options to have more backup nodes that allow for failover and switchover methods.

1.4.2 Geographic mirror

Geographic mirroring was available in i5/OS V5R3M0. It is currently the only subfunction of XSM. Do not use the two terms interchangeably, however. Geographic mirroring specifically refers to storage-based management replication. XSM is a concept that describes replication of data at multiple sites.

Geographic mirroring is intended for use by clustered system environments and uses data port services. *Data port services* are Licensed Internal Code that supports the transfer of large volumes of data between a source system and one of any specified target systems. This is a transport mechanism that communicates over TCP/IP. It provides both synchronous and asynchronous send modes. Be aware of the fact that, even in asynchronous mode, a local write waits for the data to reach the main storage of the backup node before the write operation is considered to be finished.

While geographic mirroring is performed, users cannot access the mirror copy of the data. Figure 1-3 and Figure 1-4 on page 9 show a simple geographically mirrored IASP and an environment that also incorporates switchable IASPs at both sites.

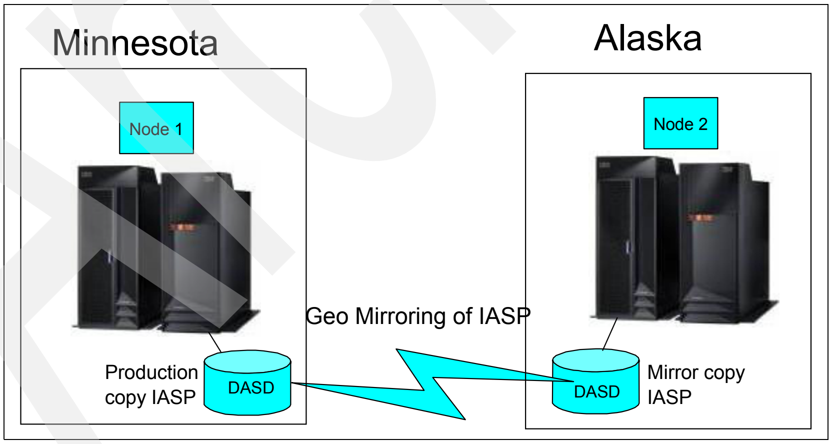


Figure 1-3 A simple example of geographic mirroring

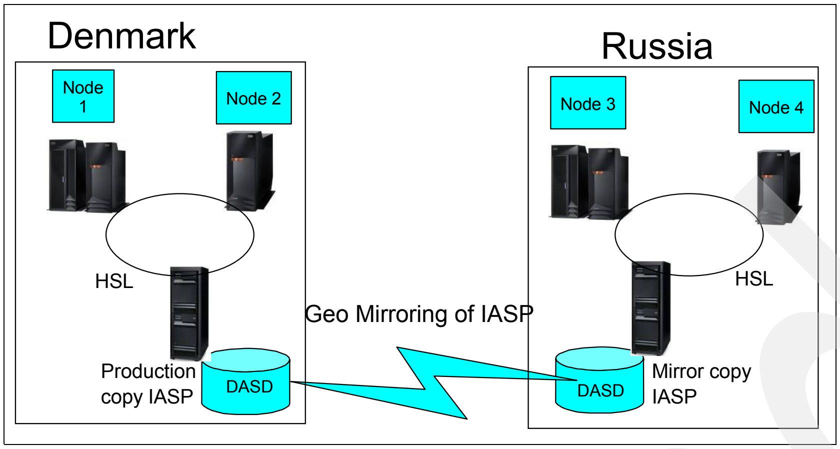


Figure 1-4 An example of geographic mirroring and switched IASPs

1.4.3 Failover and switchover

Two important concepts that are related to clustering and XSM are failover and switchover capabilities from the source system to the target system:

- A *failover* means that the source or primary system has failed and that the target or secondary system takes over. This term is used in reference to unplanned outages.
- A *switchover* is user-initiated. The user can perform a switchover if the primary system has to be shut down for maintenance, for example. In this case, production work is switched over to the target system (backup node), which takes over the role as the primary node.

1.4.4 Rolling upgrades

Upgrades of OS/400 or i5/OS releases made to any nodes that are involved in XSM require a rolling upgrade. A *rolling upgrade* is simply a strategic plan to upgrade all nodes at once, with minimal impact to your application that is writing data to the IASP.

XSM can occur from a production copy IASP to a mirror copy IASP, if the node that owns the mirror copy is at a later release of i5/OS than the production copy. However, if the mirror copy is at an earlier release than the production copy, XSM is suspended. This forces the nodes to be upgraded in an order that is dictated by the recovery domain order, starting with the node that is the last backup. During the rolling upgrade, the production copy and mirroring copy are moved to their recovery nodes.

1.4.5 Supported and unsupported OS/400 object types

Before you decide to base your HA setup on XSM, you must consider the object types that OS/400 or i5/OS allows you to put into an IASP.

Table 1-1 lists objects that are *not* supported at V5R3M0 for use in IASPs.

Table 1-1 Unsupported OS/400 objects

*AUTHLR	*DEVD	*JOBQ	*PRDDFN
*AUTL	*DOC	*JOBSCD	*PRDLOD
*CFGL	*DSTMF[a]	*LIND	*RCT
*CNNL	*EDTD	*MODD	*SOCKET
*COSD	*EXITRG	*M36	*SSND
*CRG	*FLR	*M36CFG	*S36
*CSPMAP	*IGCSRT	*NTBD	*USRPRF
*CSPTBL	*IGCTBL	*NWID	[b]
*CTLD	*IMGCLG	*NWSD	
*DDIR	*IPXD	*PRDAVL	

[a] *DSTMF is the object type returned for stream files that are accessed through the QNTC file system from a remote server. Therefore, you should never see *DSTMF when accessing the IASP directories from the local system.

[b] IBM Q-libraries cannot reside in an IASP.

Table 1-2 lists the objects that are *supported* at V5R3M0 for use in IASPs.

Table 1-2 Supported object types

*ALRTBL[a]	*FILE[b]	*MSGF	*SCHIDX
*BLKSF	*FNTRSC	*MSGQ[c]	*SPADCT
*BNDDIR	*FNTTBL	*NODGRP	*SPLF
*CHRSF	*FORMDF	*NODL	*SQLPKG
*CHTFMT	*FTR	*OUTQ	*SQLUDT
*CLD	*GSS	*OVL	*SRVPGM
*CLS[d]	*IGCDCT	*PAGDFN	*STMF
*CMD	*JOBD[e]	*PAGSEG	*SVRSTG
*CRQD	*JRN	*PDG	*SYMLNK
*CSI	*JRNRCV	*PGM[f]	*TBL
*DIR	*LIB[g]	*PNLGRP	*USRIDX
*DTAARA	*LOCALE	*PSFCFG	*USRQ
*DTADCT	*MEDDFN	*QMFORM	*USRSPC
*DTAQ	*MENU	*QMQRY	*VLDL
*FCT	*MGTCOL	*QRYDFN	*WSCST
*FIFO	*MODULE	*SBSD[h]	

[a] *ALRTBL: If network attributes reference the alert table, this object needs to exist in the system ASP.

b. *FILE: Database files that are either multi-system database files or that have DataLink fields that are created as Link Control cannot be located in an IASP. If an active subsystem references the file object, *FILE must exist in the system ASP, for example, the sign-on display file.
c. *MSGQ: If network attributes reference the message queue, *MSGQ needs to exist in the system ASP.
d. *CLS: If an active subsystem references the class object, *CLS must exist in the system ASP.
e. *JOBD: If an active subsystem references the job description object, *JOBD must exist in the system ASP, for example, autostart job entry, communication entry, remote location name entry, or workstation entry.
f. *PGM: If an active subsystem references the program object, *PGM must exist in the system ASP; for example, routing entries and prestart job entries. The program that is associated with the attention key must reside in the system ASP.
g. *LIB: The library that is specified by CRTSBSD SYSLIBLE() must exist in the system ASP.
h. *SBSD: You cannot start a subsystem whose description is located in an IASP.

1.5 Working with cross-site mirroring

In this section, we provide a general overview for using XSM in a production environment. We also provide the benefits and limitations based on the design of the XSM environment.

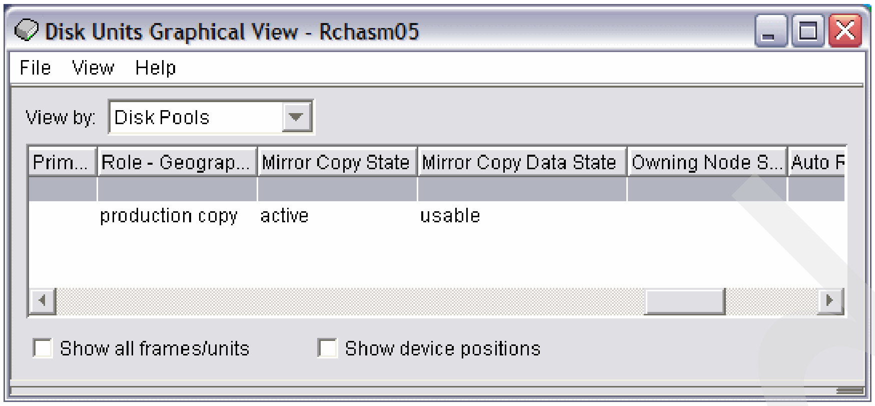
1.5.1 Ensuring that cross-site mirroring is working

You can think of XSM as another type of hardware mirroring, similar to disk mirroring that has been available for years. As with disk mirroring, there is no way that the user can determine if the IASP mirror copy is *synched* with the production copy. Also, you cannot check whether the mirroring process is current in the same way that you can with software mirroring (see Chapter 2, “Cross-site mirroring versus other high availability solutions” on page 13).

Although mirroring occurs at the hardware level with XSM, there are still several questions that you must ask to keep XSM up and running:

- Is the cluster active?
- Are the production copy IASP and mirror copy IASP close to being full?
- Is TCP/IP up and running?
- Are the communication lines between the systems active?
- Are there messages in the QSYSOPR message queue about problems with clustering?
- Does the device CRG show that the node of the mirror copy IASP is available for switchover or failover?

iSeries Navigator offers information about the current status of your XSM environment. You can find this information under **Configuration and Service** → **Hardware** → **Disk Unit** → **Disk Pool**. Then right-click the IASP for which you want further information and select **Graphical View**. Figure 1-5 on page 12 shows an example of part of the information that you can find in the Graphical View window, especially that geographic mirroring is working (mirror copy state is active).



View by: Disk Pools

Prim...	Role - Geograp...	Mirror Copy State	Mirror Copy Data State	Owning Node S...	Auto F
	production copy	active	usable		

☐ Show all frames/units ☐ Show device positions

Figure 1-5 iSeries Navigator view of disk pool status

1.5.2 Benefits of cross-site mirroring

Now that we have discussed the considerations regarding failover, switchover, rolling upgrades, and supported object types, here is a summary of the benefits provided by XSM:

- XSM provides site disaster protection by keeping a copy of the IASP at another site, which can be geographically distant, by using the geographic mirror function. Having an additional copy at another remote site improves availability.
- XSM can provide several backup nodes. In addition to having a production copy and a mirrored copy, backup node possibilities are expanded when the IASP is configured as switchable in an expansion unit, on an input/output processor (IOP) on a shared bus, or on an IOP that is assigned to an input/output (I/O) pool.

1.5.3 Limitations of cross-site mirroring

XSM has the following limitations:

- While XSM is performed, you cannot access the mirror copy. This ensures that the data integrity of the mirror copy is maintained.
- If you detach the mirror copy to perform a save operation, to perform data mining, or to create reports, you must reattach the mirror copy to resume XSM. With V5R3, this method requires a full synchronization with the production copy after it is reattached. This can be a lengthy process, possibly several hours, during which time your production system is unprotected. See Chapter 9, “Performance considerations” on page 153.

  Starting with V5R4 and a special PTF, you can also use Target Site Tracking, which can shorten synchronization times. For a detailed discussion about this new function, see 7.2.1, “How Target Site Tracking works” on page 140.
- As mentioned previously, not all object types can be mirrored via XSM. This means that you have to maintain important objects, such as user profiles and authorization lists, on both systems yourself.

  V5R4 introduced the Administrative Domain to support you with this task. See Chapter 6, “Cluster Administrative Domain” on page 123, for a detailed discussion about how the Administrative Domain works and how to set it up.
- XSM can only be performed on objects in an IASP, and not on objects in the system ASP or basic user ASPs.

2

Cross-site mirroring versus other high availability solutions

In this chapter, we provide a brief overview of the other solutions for high availability (HA) that you can use on the System i5 platform. The different characteristics offered by each solution are summarized to give a perspective on how cross-site mirroring (XSM) compares to these alternative solutions. We discuss the following alternative solutions in this chapter:

- Hardware-based solutions using storage area networks, including IBM TotalStorage® solutions

  For this solution, we briefly discuss the Copy Services feature and the following functions:

  - Metro Mirror
  - Global Mirror
  - FlashCopy®

- Software-based mirroring solutions using remote journaling

  Some of these solutions are offered through various IBM Business Partners. We draw comparisons between these options and XSM.

2.1 Storage area networks

In this section, we briefly introduce the concept of a storage area network (SAN) and its relationship to storage solutions.

2.1.1 Definition of a storage area network

Traditionally, System i users have used internal disks to meet their storage needs. In many of today's businesses, the System i5 platform is not the only one that has to be maintained. External SANs are gaining popularity because of their ability to be shared among multiple, heterogeneous platforms.

A *storage area network* is a dedicated network that is separate from local area networks (LANs) and wide area networks (WANs). A SAN generally refers to interconnected storage-related resources that are connected to one or more servers. The term SAN usually refers to hardware, but often includes specialized software for management, monitoring, and configuration.

SANs provide many benefits in your System i network, including the following benefits:

- Scalability

  Storage is independent of the server itself, so you are not limited by the number of disks that you can attach directly to the server.

- Improved availability of applications

  Storage is independent of applications and is accessible through alternative data paths.

- Centralized and consolidated storage

  Storage capacity can be connected to servers at a greater distance, and storage resources can be disconnected from individual hosts. The results can be lower overall costs through better use of the storage, lower management costs, increased flexibility, and increased control.

- Data transfer for storage at remote sites

  You can keep a remote copy of data for disaster recovery.

- Simplified centralized management

  A single image of storage media simplifies management.

SANs are a newer development in the disk and tape attachment business. They consolidate the storage of multiple, heterogeneous platforms into a single set of centrally managed resources. To do so, they employ a combination of technologies, including hardware, software, and networking components. Figure 2-1 shows an example of a SAN.

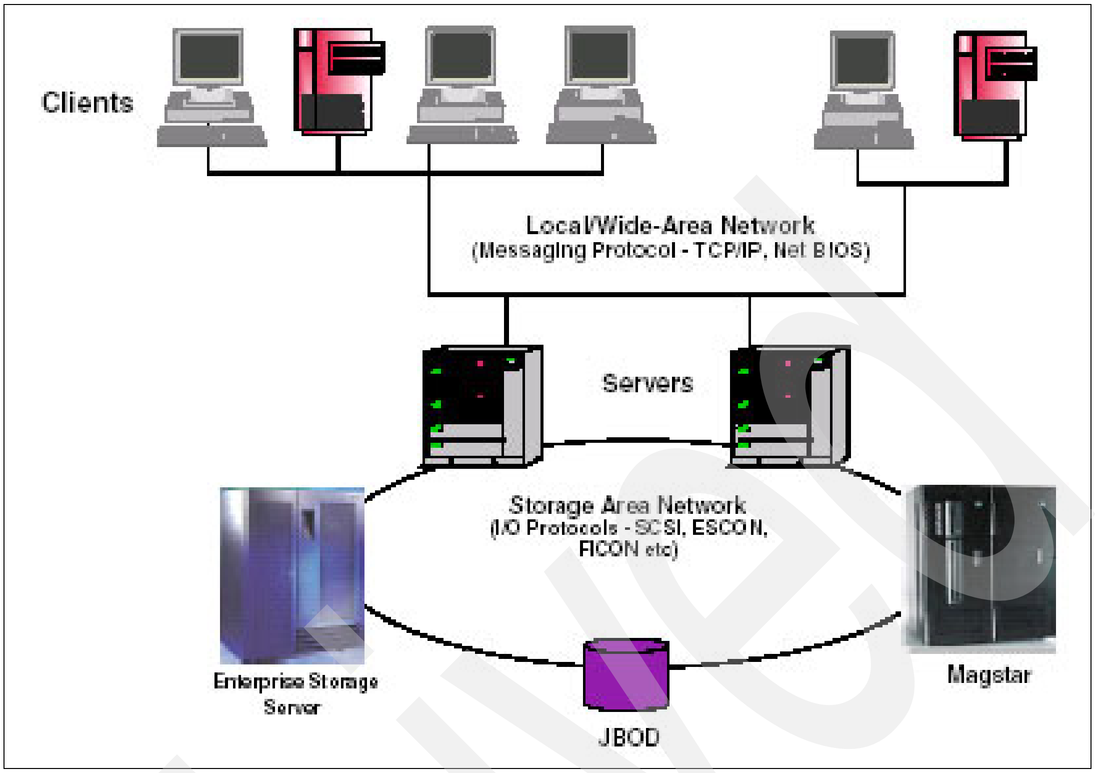

Clients
Local/Wide-Area Network
(Messaging Protocol - TCP/IP, Net BIOS)
Servers
Storage Area Network
(I/O Protocols - SCSI, ESCON, FICON etc)
Enterprise Storage Server
JBOD
Magstar

Figure 2-1   A storage area network

There are limitations to SANs because of hardware-based replication technology:

- The storage subsystem is unaware of the line of business (LOB) applications that are running on the system.
- The second copy (disaster recovery copy) of the data is not available for offline use unless a point-in-time copy (a third copy) is made of it.
- The second copy (or third copy) is a clone of the source data. You must use extreme care when using this data in an online mode or restoring partial data from either the second or third copy to the source system.

2.1.2 External storage to meet business needs

The System i platform already has its own SAN-like integrated storage subsystem, so why do you need external storage?

Many organizations are suffering from an enormous and almost uncontrolled explosion of open systems. Storage consolidation is an attractive proposal for IT departments under extreme business pressure to maintain continuity. When the number of servers grows, the complexity of managing them grows even faster. Such simple tasks as monitoring and increasing storage can be a major business problem. External disk storage subsystems address these issues.

What should you do? There is no easy answer, but here are some ideas:

- If you are already using the System i platform and are under no pressure to consolidate, do nothing. Your System i environment and OS/400 or i5/OS will manage your disks for you.

- Customers who have large System i storage capacity will undoubtedly run important applications. They might already be looking at a higher level of availability than consolidated storage can offer at this time. Therefore, the best option for these customers is to go for a System i5 cluster with OS/400 or i5/OS replication services.

2.2 IBM TotalStorage solutions

In this section, we introduce IBM TotalStorage solutions as a SAN-based storage solution. We also discuss the Copy Services feature and the services that it provides, as well as the usage of the product in general.

2.2.1 IBM TotalStorage solutions as a SAN-based storage solution

IBM TotalStorage solutions bring new advanced storage capabilities to the System i5 platform by allowing more storage consolidation and flexibility in an enterprise environment. These capabilities include multi-server connectivity, fully redundant hardware including non-volatile storage cache, RAID-5 or RAID-10 protection, Copy Services solutions, and advanced storage allocation capabilities.

In many customer environments, through strategic direction, all servers use SAN-based storage, and IBM TotalStorage solutions are designed to meet those needs.

The System i5 platform benefits from SAN storage in the following ways:

- Consolidated storage
- Redeployment of disks among systems and platforms
- Movement of disks among systems
- Copy Services solutions
- Scalable caches for reads and writes

A SAN works well with the System i5 platform in the following conditions:

- When it is sized for performance and not just for terabytes
- When it is well managed

SAN storage is about sharing disk units, or *DASD*, with other platforms, among other objects. However, it has the following HA advantages:

- Since V5R1M0 of OS/400, support exists for internal DASD, such as the load source unit, to be mirrored to external DASD. This enables disaster recovery across the SAN.
- At V5R3M5 of OS/400 or i5/OS (as of October 2005), the System i platform has the ability to have an external load source. In previous releases, the System i platform required an internal disk controller for a load source, which can reside either in the base unit or in a slot of an I/O tower, in cases where customers use tower-level partitioning. With the October 2005 announcement, the load source can reside on the external storage network.

2.2.2 IBM offerings through IBM TotalStorage solutions

IBM has a wide range of product offerings that are based on open standards and that share a common set of tools, interfaces, and innovative features. The IBM TotalStorage DS family, the Enterprise Storage Server® (ESS), the DS8000™, and the DS6000™, give you the freedom to choose the right combination of solutions for your current needs and the flexibility to help your infrastructure evolve as your needs change.

With V5R3M0 and later of i5/OS, IBM delivers significant enhancements in the areas that improve availability with the introduction of multipath I/O for storage subsystems attached via

SANs. Multipath support takes advantage of multiple paths between a host system and a storage subsystem logical unit number (LUN) or set of LUNs. In a case where an adapter fails, the system automatically reroutes I/O operations to another available path.

With the announcement of V5R3M5 of i5/OS in October 2005, IBM supports the capability to install a load source disk unit inside an IBM TotalStorage subsystem, enabling direct boot support from a SAN-attached IBM TotalStorage subsystem for the ESS, DS6000, and DS8000 families.

Figure 2-2 shows a SAN with multiple platforms.

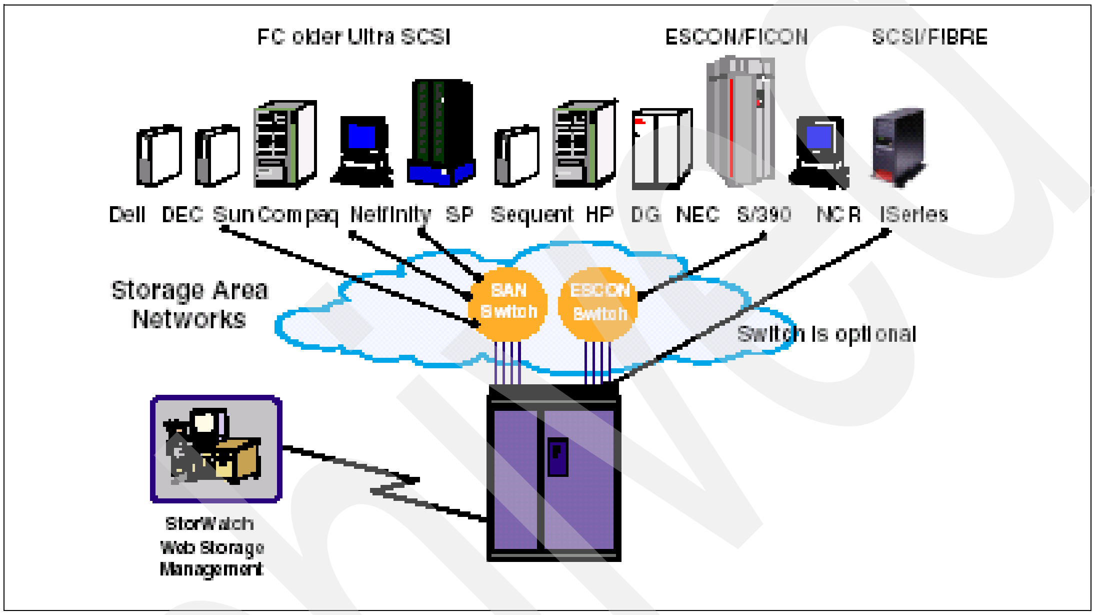


Figure 2-2 A SAN with multiple platforms

2.2.3 Disaster recovery and high availability with SAN

The ESS as well as the DS6000 and DS8000 are fault-tolerant subsystems, which, in general, can be managed and upgraded concurrently with customer operation. With advanced functions, such as Copy Services, you can copy data from one set of disks within a single ESS or to an ESS at another site.

As with any mixed server environment, there are special considerations. For any server that has data in main storage, which has not been written to disk, availability is limited to data that resides on disk. The storage pool on the IBM System Storage series that is provided does not take into account data that is still in System i5 memory, which can be up to 1 TB with L2/L3 cache of 1.6/36 MB. The memory stored in the pool is not copied to disk unless the system is shut down. Therefore, although ESS as a server has high reliability and redundancy features for the disks that it controls, the total solution (both application and storage) cannot be considered as a true HA solution. This applies to the System i platform as well as others.

2.3 Copy Services

In this section, we provide a brief overview of the functions that are supported with Copy Services, as well as the usage of this feature.

2.3.1 Definition of Copy Services

Copy Services is an optional feature of the IBM TotalStorage ESS, DS6000, and DS8000. It brings powerful data copying and mirroring technologies to open systems environments that were previously available only for mainframe storage.

With Copy Services Version 2, the following functions are supported on Licensed Internal Code level 2.4.0 and later for ESS and all DS6000 and DS8000 code levels:

- Metro Mirror
- Global Mirror
- FlashCopy

This is only a brief overview of all the possible copy functions with IBM TSS. For more information, refer to *iSeries and IBM TotalStorage: A Guide to Implementing External Disk on eServer i5*, SG24-7120.

2.3.2 Metro Mirror

Metro Mirror, previously known as Synchronous Peer-to-Peer Remote Copy (PPRC), is a function of a storage server that constantly updates a secondary copy of a volume to match changes that are made to a primary volume. The primary and secondary volumes can be on the same storage server or on separate storage servers. In the case of two DASD subsystems, the secondary subsystem can be located at another site up to 300 km away. However, there can be performance implications when using synchronous communications over this distance, and it can be more practical to consider shorter distances to minimize the performance impact.

Metro Mirror is application independent. Because the copying function occurs at the disk subsystem level, the application has no knowledge of its existence. Figure 2-3 shows the architecture of Metro Mirror.

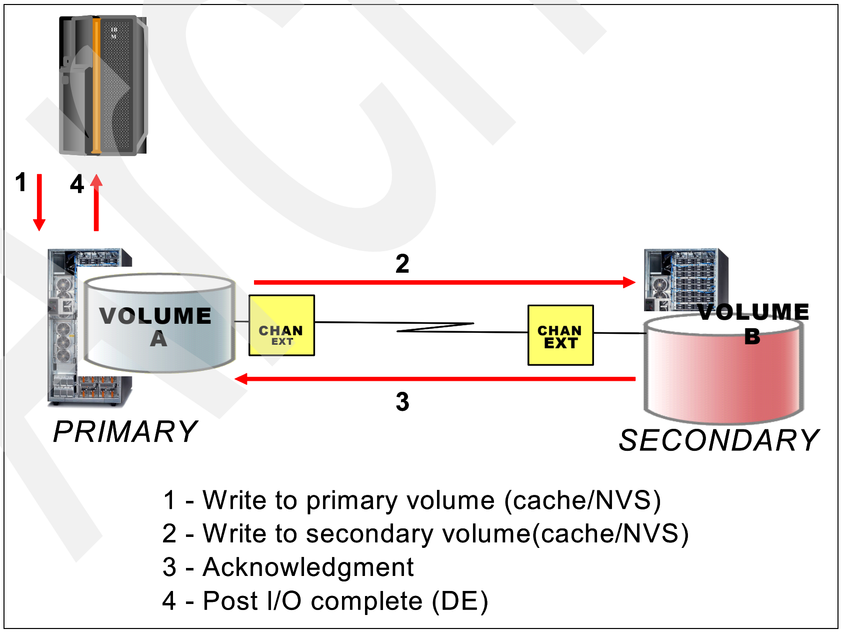

1
4
2
VOLUME A
CHAN EXT
CHAN EXT
VOLUME B
3
PRIMARY
SECONDARY
1 - Write to primary volume (cache/NVS)
2 - Write to secondary volume(cache/NVS)
3 - Acknowledgment
4 - Post I/O complete (DE)

Figure 2-3 Metro Mirror architecture

The synchronous protocol guarantees that the secondary copy is up-to-date and consistent by ensuring that the primary copy will be committed only if the primary copy receives acknowledgment that the secondary copy has been written.

> **Important:** Metro Mirror provides a remote copy that is synchronized with the source copy. For this remote copy to be made, the disaster recovery system is at the point of failure when it is recovered. However, this does not take into account any further recovery actions that are necessary to bring the applications to a clean recovery point, such as applying or removing journal entries. These actions will happen at the database recovery stage, and you must expect that your recovery time objective should take this into account. However, using Metro Mirror probably take less time than if you were to do recovery from tape.

2.3.3 Global Mirror

Global Mirror, previously known as Asynchronous PPRC, or PPRC Extended Distance (PPRC-XD), is designed to provide a long-distance remote copy solution across two sites using asynchronous technology. It operates over high-speed, Fibre Channel communication links. It is designed to maintain a complete and consistent remote mirror of data asynchronously at virtually unlimited distances with next to no application response time degradation.

The asynchronous technique provides better performance at unlimited distances, by allowing the secondary site to trail in currency a few seconds behind the primary site. This two-site data mirroring function is designed to provide a high-performance, cost-effective global distance data replication and disaster recovery solution. Figure 2-4 shows the architecture of Global Mirror.

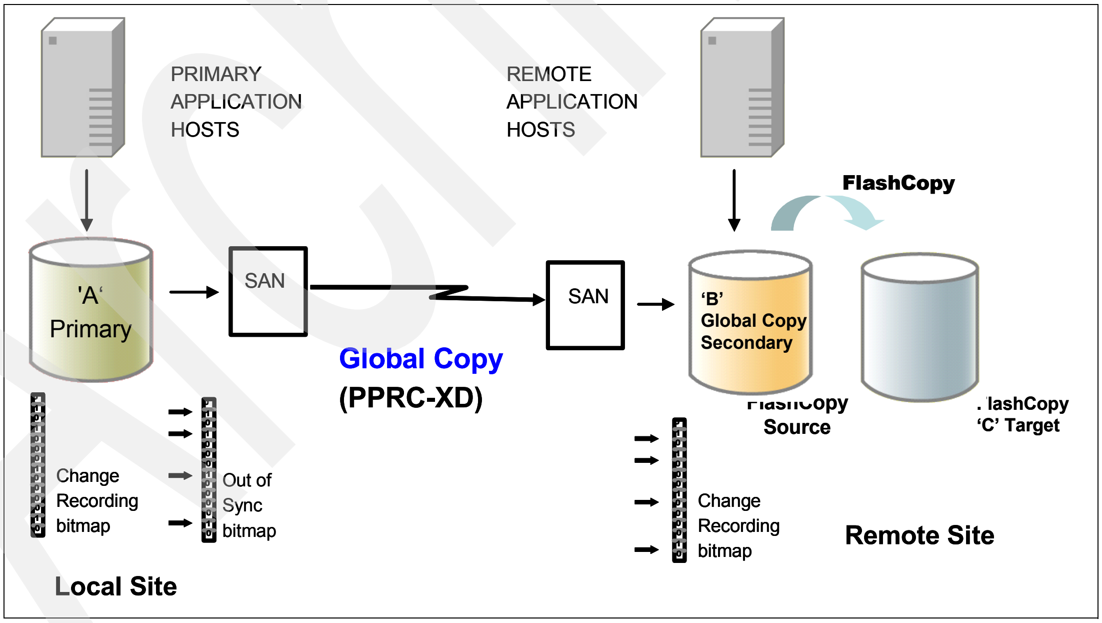


Figure 2-4 Global Mirror architecture

Important: Global Mirror provides a remote copy that is behind the source copy by some time. It gives a Recovery Point Objective (RPO) of between a few seconds and a few minutes, depending on write I/O activity and bandwidth. This means that the Disaster Recovery system will be a few seconds or minutes behind the production system point of failure when it is recovered. As with Metro Mirror, this happens at the database recovery state, and again, it can be expected to be a lot faster than recovery from tape.

2.3.4 FlashCopy

In a storage system, LUNs are virtual representations of a disk as seen from the host system. In reality, a LUN can span multiple disks, and the size of the LUN is set when it is defined to the storage system. FlashCopy makes a single point-in-time copy of a LUN, which is also known as a *time-zero copy*. The target copy is independent of the source LUNs and is available for both read and write access after the FlashCopy command is processed.

FlashCopy provides an instant or point-in-time copy of a logical volume. Point-in-time copy functions give you an instantaneous copy, or view, of the original data as it looked at a specific point-in-time. Figure 2-5 shows the relationship between the source volume on the left and the target volume to the right. When FlashCopy runs, a bitmap that represents the relationship between the source and target volume is created. This is represented by the grid in the upper half of Figure 2-5.

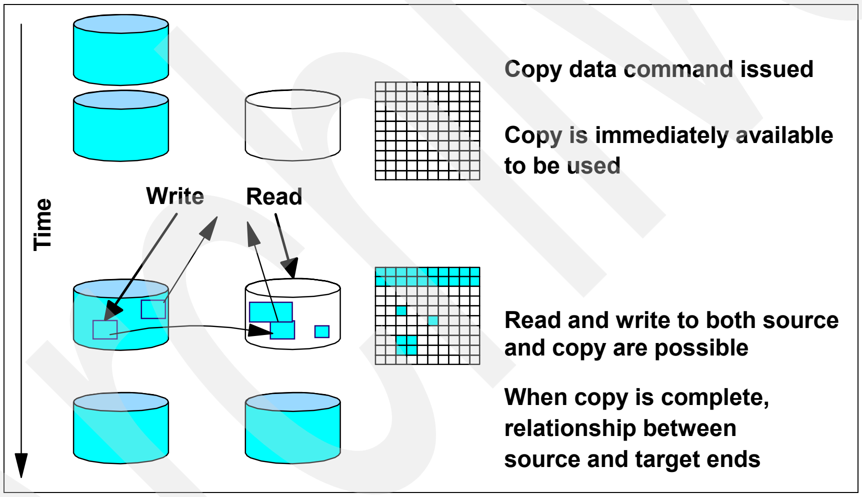

Copy data command issued
Copy is immediately available to be used
Time
Write
Read
Read and write to both source and copy are possible
When copy is complete, relationship between source and target ends

Figure 2-5 FlashCopy

FlashCopy can be used with either COPY or NOCOPY options. In both cases, the target is available as soon as the FlashCopy command is processed. When the COPY option is used, a background task runs to copy the data, sequentially track-by-track, from source to target. With NOCOPY, the data is only copied “on write”, meaning that only those tracks that have been written to on either the source or target are written to the target volume. This is shown by the shaded segments in the second grid in Figure 2-5, indicating those tracks that have been updated on the target volume.

When the source volume is accessed for read, data is read from the source volume. If a write operation is done to the source, it is done immediately if the track has already been copied to the target, either because a COPY option was already copied it or because a “copy on write” occurred. If the source track to be written has not already been copied, the unchanged track is copied to the target volume, and then the source is updated. When the target volume is accessed for read, the bitmap shows which tracks are already copied. If the track is already copied, either by a background COPY or by a “copy on write” that has been performed, the read is done from the target volume. If the track is not copied, either because the background copy has not reached that track or because it has not been updated on either the source or target volume, the data is read from the source volume.

In this way, both source and target volumes are independent. If the COPY option is used, the relationship between source and target is ended when all target tracks have been written. With NOCOPY, the relationship is maintained until explicitly ended. Although only one source and one target volume are shown in Figure 2-5, the same applies for all volumes in a relationship. Because OS/400 or i5/OS stripes its data across all available volumes in the system or user auxiliary storage pool (ASP) or independent auxiliary storage pool (IASP), any FlashCopy must treat all source volumes as one group.

The point-in-time copy that is created by FlashCopy is typically used where you need a copy of production data to be made with minimal application downtime. The copy can be used for online backup, testing of new applications, or copying a database for data mining purposes. The copy looks exactly like the original source volume and is instantly available.

For copies that are usually accessed only once, such as for creating a point-in-time backup, the NOCOPY option is used. When the target is required for regular access, such as populating a data warehouse or creating a cloned environment, we recommend that you use the COPY option.

Incremental FlashCopy (V2)

Incremental FlashCopy is a new feature of FlashCopy that is available with Copy Services Version 2. This function is to be used in conjunction with the background copy option to track the changes on the source volume since the last FlashCopy relationship was invoked.

When this option is selected, only the tracks that have been changed on the source are copied to the target. The direction of the *refresh* can also be reversed, so that the changes made to the new source (originally the target volume) are copied to the new target volume (originally the source volume).

Inband FlashCopy (V2)

The new inband management capability feature that comes with Copy Services Version 2 allows you to invoke FlashCopy on a remote site storage subsystem. If you have two sites, one local and one remote, which are in a PPRC relationship, a FlashCopy task on the remote site storage subsystem can be invoked from the primary site storage subsystem with a PPRC inband connection.

Multiple relationship FlashCopy (V2)

With Copy Services Version 2, one FlashCopy source volume can have up to 12 FlashCopy target volumes. This gives you more flexibility, because you can initiate the multiple relationships using the same source volume without waiting for other relationships to end.

FlashCopy consistency groups

New options are available to facilitate the creation of FlashCopy consistency groups. With FlashCopy consistency groups, I/O activity to a volume is held off until the Consistency Created task with the FlashCopy Consistency Group option is used.

Using FlashCopy and BRMS

Backup Recovery and Media Services (BRMS) is the strategic solution by IBM for performing backups and recovering System i5 environments. BRMS has a wealth of features, including the ability to work in a network with other systems to maintain a common inventory of tape volumes.

FlashCopy creates a clone of the source system onto a second set of disk drives that are then attached and used by another system (or a logical partition (LPAR)). The BRMS implementation of ESS FlashCopy provides a way to perform a backup on a system that is copied by using FlashCopy, and the BRMS history appears as though the backup was performed on the production system.

For a complete description about using BRMS and FlashCopy, refer to *iSeries and IBM TotalStorage: A Guide to Implementing External Disk on eServer i5*, SG24-7120*.*

2.3.5 Using Copy Services with OS/400 or i5/OS

Prior to the announcement of the #2487 PCI-X IOP for SAN Load Source, it was not possible to boot from an external storage server. However, the introduction of #2847 IOP for Fibre Channel load source for SAN further expands OS/400 or i5/OS availability options. With the new IOP, you no longer have to mirror your internal load source to a SAN-attached load source. Instead, you can now place the load source directly inside a SAN-attached storage subsystem and then enable mirroring to provide protection. This capability enables easier startup of a system environment that has been copied using Copy Services functions such as Metro Mirror, Global Mirror, or FlashCopy. During the restart of a copied environment, you no longer have to perform the “Recover Remote Load Source Disk Unit” through Dedicated Service Tools (DST). By not performing this function, you reduce the time and overall steps that are required to start a point-in-time system image after Metro Mirror, Global Mirror, and FlashCopy functions have completed.

Now, with the announcement of the #2487 PCI-IX IOP for the SAN Load Source, it is possible to have the entire disk space, including the load source unit, contained in the external storage subsystem. This means that is it much easier to replicate the entire storage space.

For simple environments or for object types that are not supported in an IASP, having everything contained in *SYSBAS (the system ASP plus user ASPs 2 through 32) is the easiest environment on which to implement Copy Services. However, you must keep in mind that *SYSBAS includes all work areas and swap space. Replicating these areas requires greater bandwidth and can introduce other operational complexities that must be managed. For example, the target is an exact replica of the source system and has exactly the same network attributes as the source system. If used for disaster recovery purposes, this does not cause a problem. However, if you want to test the disaster recovery environment or if you want to use FlashCopy for daily backups, you need to change the network attributes so that both systems can be in the network at the same time.

For more advanced availability and disaster recovery environments, we recommend that you use IASPs for the applications and data and use the iSeries Copy Services Toolkit. This approach follows the architected solutions for availability that are also used with switchable IASPs and XSM. IASPs allow for more automated failover techniques that are available with clustering for OS/400 or i5/OS.

2.3.6 The iSeries Copy Services Toolkit

The iSeries Copy Services Toolkit is a Services Offering developed by IBM Rochester. It blends two technologies that use the iSeries Availability architecture provided by IASPs, along with the advanced functions provided by IBM TotalStorage Copy Services. The toolkit is a combination of management software to control the IASP environment and services to implement it.

2.4 Copying the entire DASD space

With the new support for the SAN Load Source in System i5 platforms, it is now possible to have the entire disk space in an IBM TotalStorage server. This provides new opportunities that were previously impossible for System i5 customers.

Now you can create a complete copy of your entire system in minimal time. You can then use this copy for a variety of purposes such as to minimize your backup windows, protect yourself from a failure during an upgrade, or use it as a fast way to provide yourself with a backup or test system. All of these actions can be done by copying the entire DASD space with minimal impact to your production operations.

These facilities fall into two main areas:

- Creating a clone of the system or partition
- Performing daily operational tasks such as backups and disaster recovery

2.4.1 Creating a clone

You should consider cloning a system as an infrequent task and not part of daily operations. Typically, the clone is used as a whole to avoid doing a lengthy restore from tape. You use the clone for complete backups whenever major hardware or software maintenance is to be performed on your system.

With the ability to create a complete copy of the whole environment, you have a copy on disk that can be attached to the system or partition and for which an initial program load (IPL) can be done normally. For example, if you have planned a release upgrade over a weekend, you can now create a clone of the entire environment on the same disk subsystem. You do this by using FlashCopy immediately after doing the system shutdown and performing the upgrade on the original copy. If problems or delays occur, you can continue with the upgrade until just prior to the time that the service needs to be available for the users. If the maintenance is not completed, you can abort it and reattach the target copy, or do a “fast reverse/restore” to the original source LUNs and do a normal IPL, rather than doing a full system restore.

Cloning a system can save you a lot of time, not only for total system backups in connection with hardware or software upgrades, but also for such tasks as creating a new test environment.

2.4.2 System backups using FlashCopy

Creating a clone by creating regular copies of the entire DASD space can also be a part of the day-to-day tasks in order to minimize the downtime associated with taking backups.

When you want to make a full backup of your system, you can use standard OS/400 or i5/OS commands with or without *Save While Active (SWA)*. Both require the applications to be quiesced to some extent. Sometimes it is faster to go to restricted state than to wait for the SWA checkpoint to be reached, which can take a considerable amount of time in a system with a complex library structure. When the backup is finished, the user subsystems must be started again. The downtime in connection with this type of backup can take several hours.

With FlashCopy, you can take a copy of the entire DASD space, but it requires that you shut down your system to ensure that all of the data in main memory is flushed to disk. After you shut down your system, the actual copy for the clone is done in minutes. After the clone is made, you can perform an IPL on your production system and return it to service while you perform your backup on a second system or partition.

2.4.3 Using a copy for disaster recovery

As described in the previous section, FlashCopy can be done only within the same external disk subsystem, so it is not suitable for disaster recovery. In order to provide an off-site copy for disaster recovery purposes, you should use either Metro Mirror or Global Mirror for metro and global distances between two external disk subsystems.

Important: As with all iSeries Availability techniques, use OS/400 or i5/OS journaling. The use of journaling ensures that, even if objects remain in main memory, the journal receiver will be written to disk. Consequently, they will be copied to the disaster recovery site using Metro Mirror or Global Mirror and will be available on the disaster recovery server to apply to the database when the system is started.

Unlike cloning or making copies using FlashCopy, Metro Mirror and Global Mirror are constantly updating the target copy, so you cannot be assured of having a clean starting point. With both Metro Mirror and Global Mirror, you have a restartable copy, but the restart is the same as though an IPL was done on the original system after repair. With Metro Mirror, the recovery point is the same as the point at which the production system failed. With Global Mirror, it is at the point when the last consistency group was formed. This depends on the bandwidth and write rate.

Consistency group: The consistency group function can help create a consistent point-in-time copy across multiple LUNs or volumes and across multiple DS8000s.

2.4.4 Considerations for copying the entire DASD space

Because copying the entire DASD space creates a copy of the whole source system, you must take into consideration the following items:

- The copy is an exact copy of the original source system in every respect.
- The system name and network attributes are identical.
- The TCP/IP settings are identical.
- The BRMS network information is identical.
- The user profiles and passwords are identical.
- The Job Schedule entries are identical.
- Relational database entries are identical.

You should be extremely careful when you activate a partition that was built from a total copy of the DASD space. In particular, you have to ensure that it does not automatically connect to the network because this causes substantial problems within both the copy and its parent system. You must ensure that your copy environment is correctly customized before you attach it to a network.

Remember that booting from a SAN and copying the entire DASD space is not a high-availability solution. It involves a large amount of subsequent work to make sure that the copy works in the environment where it is used.

2.5 Software-based high availability solutions

Software-based solutions offer an extremely different set of possibilities in regard to high availability and business continuity. In the following sections, the generalizations of these software-based high availability solutions are based on the functionality of IBM *High Availability Business Partners (HABPs),* such as Data Mirror and Vision Solutions. In this section, we emphasize the features and functions that each of the individual business partners have in common when compared to hardware-based solutions such as IBM TotalStorage solutions and XSM with IASPs.

2.5.1 Journaling

Software-based high availability solutions are mostly based on OS/400 or i5/OS journaling. With journaling, you set up a journal and a journal receiver and then define the physical files, data queues, data areas, or integrated file system objects that are to be journaled to this particular journal. Whenever a record is changed, a journal entry is written into the *journal receiver*. It contains information about the record that was changed, the file to which it belonged, the job that changed it, the actual changes, and so forth. Journaling has been around since the days of the System/38™. In fact, a lot of user applications are journaled for various purposes such as keeping track of user activity against the file or for being able to roll back in case of a user error or a program error.

In this discussion, journaling is classified as *local journaling* or *remote journaling*. The difference is simply the manner in which the data is transferred between systems.

Local journaling

High availability solutions based on local journaling have a reader job on the source system that reads the journal entries for the files that are defined to be mirrored and sends the changes across to the receiver job on the target system. Here, an apply process (job) applies the changes to the target database. The job that is used to transmit the changes from the source system to the target system is not a built-in journaling job, but part of the software-based high availability program or package.

Figure 2-6 shows a high availability solution with local journaling.

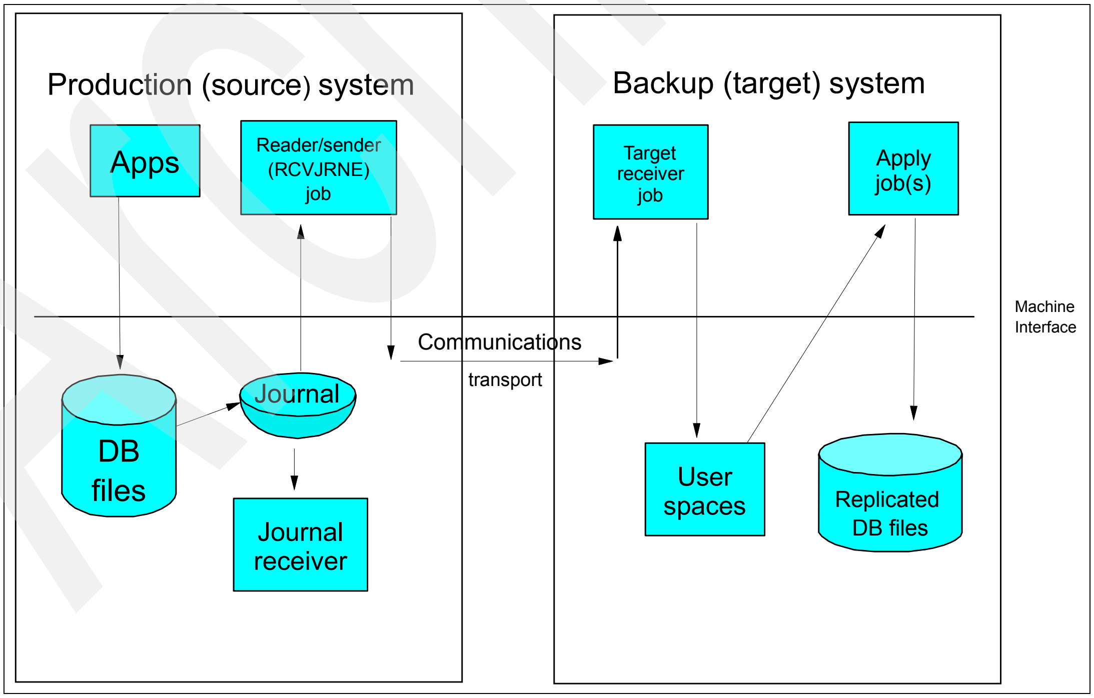


Figure 2-6 A general example of a high availability solution with local journaling

Remote journaling

With V4R2M0 of OS/400, the concept of remote journaling enhanced communication between source and target systems. The changes can be sent more quickly from the source system to the target system than what is possible with local journaling and the use of a reader/sender job. Remote journaling is implemented at the Licensed Internal Code layer, providing for faster processing between systems.

With remote journaling, you set up local journaling on your source system as you normally do. You then use the ADDRMTJRN command to associate your local journal to a remote journal, through the use of a relational database. When a transaction is put into the local journal receiver, it is then immediately sent to the remote journal and its receiver via the communications path that is designated in the relational database directory entry.

Remote journaling allows you to establish journals and journal receivers on the target system that are associated with specific journals and journal receivers on the source system. After the remote journaling function is activated, the source system continuously replicates journal entries to the target system as described previously.

The remote journaling function is a part of the base OS/400 or i5/OS system and is not a separate product or feature.

The use of remote journaling offers the following advantages:

- It lowers processor consumption on the source machine by shifting the processing required to read the journal entries from the source system to the target system. Most of the workload is moved to the target system because the reader job that normally reads from the journal on the source system is moved to the target system.
- It eliminates the need to buffer journal entries to a temporary area before transmitting them from the source machine to the target machine. This translates into fewer disk writes and greater DASD efficiency on the source system.
- Since it is implemented at the Licensed Internal Code level, it significantly improves the replication performance of journal entries and allows database images to be sent to the target system in real time. This real-time operation is called *synchronous delivery mode*. If synchronous delivery mode is used, the journal entries are guaranteed to be in main storage on the target system prior to control being returned to the application on the source machine.
- It allows the journal receiver save and restore operations to be moved to the target system. This way, the resource utilization on the source machine can be reduced.

For more information about remote journaling, refer to *AS/400 Remote Journal Function for High Availability and Data Replication*, SG24-5189.

Figure 2-7 shows an example of a high availability solution that uses remote journaling with a reader job on the target side. The remote journal function provides a much more efficient transport of journal entries than the traditional approach. In this scenario, when a user application makes changes to a database file, there is no need to buffer the resulting journal entries to a staging area on the production (source) system. Efficient system microcode is used instead to capture and transmit journal entries directly from the source system to associated journals and journal receivers on a target system. Much of the processing is done below the Machine Interface (MI). Therefore, more processor cycles are available on the production machine for other important tasks.

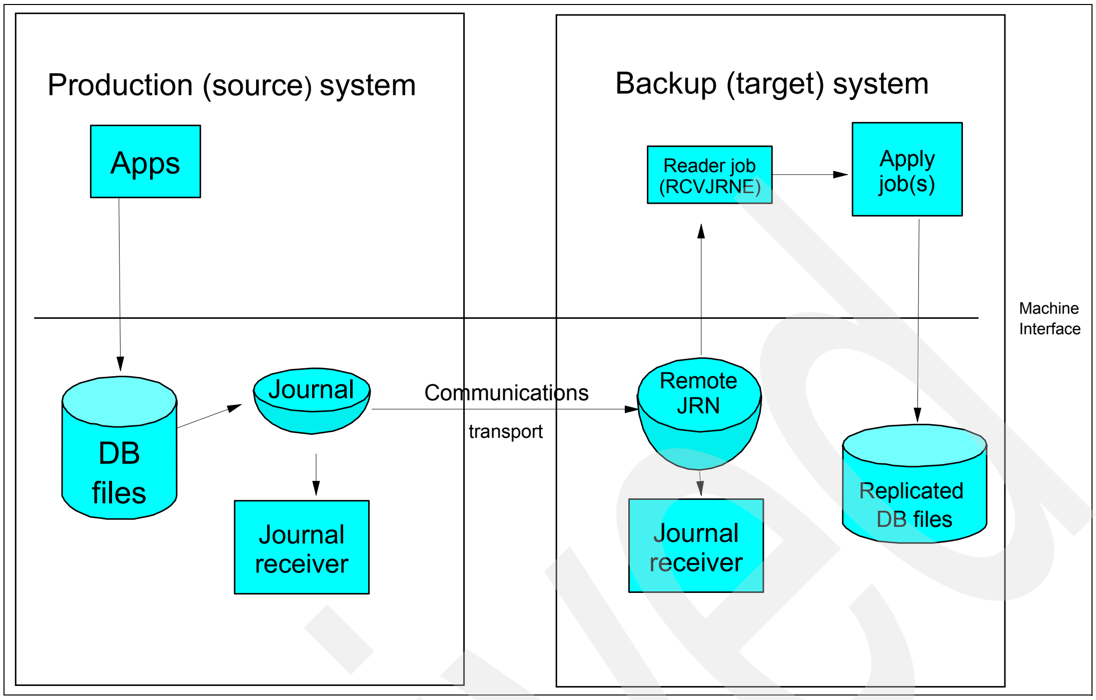


Figure 2-7 A general example of a high availability solution with remote journaling

Object types not journaled

With journaling, whether local or remote, you can replicate changes from the source system to the target system for the following object types:

- Physical files
- Data areas
- Data queues
- Integrated file system objects

At V5R3M0 of OS/400 or i5/OS, only these object types are allowed to be journaled.

However, a usable backup system usually requires more than just database files. The backup system must have all of the applications and objects that are required to continue critical business tasks and operations.

Users also need access to the backup system. This means that they need to have a user profile on the target system with the same attributes as that profile on the source system, and their devices must be able to connect to the target system.

The applications that a business requires for its daily operations dictate the other objects that are required on the backup system. Not all of the applications that are used during normal operations are required on the backup system. In the event of an unplanned outage, the business might choose to run with a subset of those applications. This allows the business to use a smaller system as the backup system or to reduce the impact of the additional users when the backup system is already being used for other purposes.

The exact objects that comprise an application vary widely. The following list shows some object types that are commonly part of an application:

- Authorization lists (*AUTL)
- Job descriptions (*JOBD)
- Libraries (*LIB)
- Programs (*PGM)
- User spaces (*USRSPC)

For many of the objects in the list, the content, attributes, and security of the object affect how the application operates. The objects must be continuously synchronized between the production and backup systems. For some objects, replicating the object content in near real time can be as important as replicating the database entries.

Replicating non-journaled object types

Most of the HABPs have solutions for mirroring non-journaled objects of the types listed in the previous section. This replication is typically done by using the system audit journal (QAUDJRN), which is configured to monitor for creations, deletions, and other object-related events. The HABP solution reads from the QAUDJRN. Based on its list of objects that are defined to be mirrored, it sends the whole object to the target system via temporary save files, or with ObjectConnect/400 (included in the operating system as option 22), if configured on the systems.

Figure 2-8 shows a general view of replication of non-journaled objects.

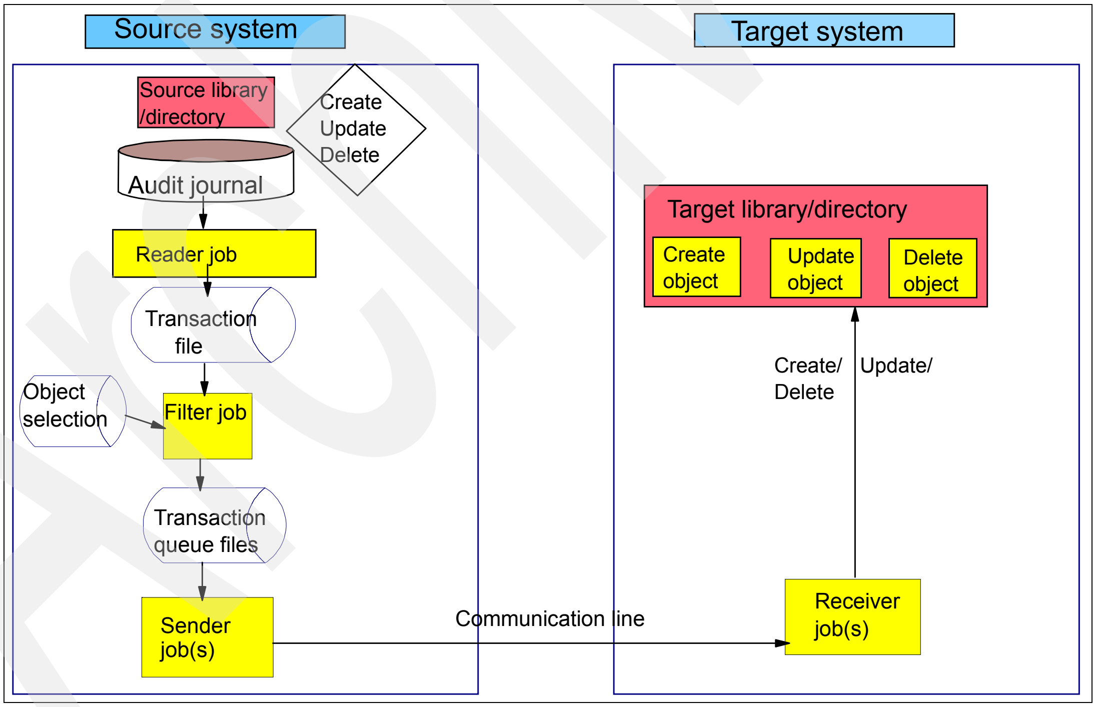


Figure 2-8 A general view of mirroring non-database objects from source to target

2.5.2 Planning for mirroring with an HABP software solution

Planning for and configuration of a HABP software solution is no simple task. In this section, we present the areas that you must consider carefully.

Hardware considerations

Make sure that the source system can handle the additional workload of journaling and the mirroring software. The HABP can provide information about the extra workload that you can expect in your particular environment.

The target system must have sufficient disk capacity to contain all the mirrored objects, but what about processor capacity? Often a less powerful processor on the target side is sufficient in case of a failover or switchover, because non-critical business applications such as queries or statistics are not allowed to run until the users are switched back to the source system.

Another important consideration is the number of disk arms on the target system. If there are many transactions on the source system, you run the risk of the target system falling behind, if the number of disk arms is considerably smaller on the target system than on the source. Keep in mind that, in normal daily circumstances, the main responsibility of the target system is to apply changes to the mirrored objects and databases.

Communications between the systems

The usual recommendation is to set up a dedicated connection between the systems for performance and higher security. This can mean additional network hardware and software configuration objects.

The speed of the communication line is another important consideration. Some HABPs have tools that can help you estimate the line capacity that is necessary to ensure that the target system does not fall behind the source system.

What is to be mirrored

In some environments, customers simply want to mirror everything on the system, which is perfectly possible. Keep in mind, however, that it is not possible to mirror the operating system (OS/400 or i5/OS), other licensed programs, and PTFs. That is, the two systems still have to be maintained individually.

Setting up a software mirrored environment can be a complex and time-consuming affair, especially in environments where objects are journaled to many different journals.

Some consideration also needs to be given to temporary objects, such as user spaces, that exist only on the system while an application is running, or while an individual user is active on the system. If the mirroring process cannot get to these objects because an application has an exclusive lock on them, various errors and extra processor workload can result. In this case, it is necessary to exclude these objects from the mirroring process.

Be aware that files that contain license codes or keys for various software products are often based on the system serial number. If you mirror these files to the target system, the application will fail on the target when you try to start it, because the target serial number is different from that of the source.

Setting up a mirroring solution might provide a good opportunity to get rid of old, unused libraries and objects, so that only the libraries and objects necessary for your business are replicated.

2.5.3 Working with an HABP software solution

Although working with a HABP software solution should not be a time-consuming daily task for the system administrator, you must still check a few things on a daily basis to make sure that the data has been mirrored correctly to the target system so that it is usable when it is needed:

- Is the HABP application running properly?

  The various software solutions provide status panels that let you determine if everything is running as expected. More importantly, they indicate if the target system is current, which means that all of the transactions that were sent across have been applied.

- Are any objects out of synch or in an error status?

  The status panels also provide this information. If necessary, objects that cannot be mirrored successfully need to be resent, or further investigation is needed to determine why they cannot be sent.

- Is there a communications problem?

  The status panels might indicate that there is a problem with communications. If this is the case, the system administrator or network administrator has to analyze the status of the interfaces or if the communications jobs have ended.

- How are the journal receivers managed?

  Most of the HABP solutions offer tools for journal management, which allows for automatic deletion of journal receivers after the mirroring process is finished with them. If your journaling is based on existing journals that are being used for purposes other than mirroring, the system administrator must manage the change and deletion of receivers manually.

2.5.4 Benefits of using HABP software solutions

So far, we have outlined some additional efforts and costs that are associated with HABP software mirroring solutions. These solutions provide the following benefits:

- Target system is available during switchovers and failovers.

  In case of a failure on the source system or a manual switchover to the target system, the users can continue working from the target system while the source system is being repaired. If you perform planned maintenance on the source system, such as upgrades or model changes, users can switch to the target system and let production continue there.

  The source and target systems can be at different OS/400 releases. This means that you can upgrade one system at a time, with or without switching users to the other system while the upgrade is performed.

- Data on the target system remains available to users.

  The data on the target system is available for all sorts of read-only jobs. You are not allowed to update the data on the target side, but you can run jobs such as queries and statistics efficiently on the target system. This can be an advantage for using a target system in a testing environment as well.

- Data on the target system is available for backups.

  This is an important feature of this type of solution. When a backup is to be performed, you only stop the apply process in order to have a consistent backup of your data. Meanwhile, the mirroring process from source to target is still running, and data is stored in temporary spaces on the target system. After the backup is finished, the apply jobs are started again, and the data that was sent from the source system to the target system while the backup was running is applied. You do not have to shut down applications on the source system to perform backups, and these applications are not unprotected during backups.

2.6 Comparing mirroring solutions

Table 2-1 and Table 2-2 provide an overview of the differences between the various implementations for HA and disaster recovery as described in this chapter. Table 2-1 compares the system functions to save methods.

Table 2-1 Comparison of system functions to save methods

	Save while active	**Savlib *NONSYS**	**FlashCopy backup**	**Backup of software mirrored data**
System power down required to save data	No	No	Yes (see 2.4.2, “System backups using FlashCopy” on page 23)	No
Restricted state required to save data	No, but time to establish sync point	Hours	N/A; system powered down	No
Duration of outage	Minutes	Hours	Minutes, time to power down and to IPL	None
Performance impact while saving	Some	N/A, dedicated	Maybe, separate partition or system required	None
Time to recover	Hours	Hours	Hours if restoring from backup created with FlashCopy. Minutes if using live FlashCopy volumes	Minutes
Incremental save/restore	Yes	Yes	No	Yes
Ease of save	Easy	Easy	Moderate	Easy
Protection while backup is being performed	None	None	Some	Full

Table 2-2 compares the availability features.

Table 2-2 Availability comparison

	SAN solutions	**XSM**	**HABP software solutions**
Switchover time	Hours	Minutes	Less than 30 minutes
Failover time	Hours	Minutes	Less than 30 minutes
Based on journaling	No	No, but journaling is highly recommended to ensure data integrity	Yes
Data on target system usable in normal mode	No	No	Yes, read-only
Resynch time after use of target data	None (if read-only)	Hours (less time required with V5R4 and Target Site Tracking)	None (if read-only)
Supported by OS/400 cluster services	No	Yes	Yes

	SAN solutions	**XSM**	**HABP software solutions**
Object types supported	All	Limited	Most
Save/restore required for initial setup	No	No	Yes

These tables compare some characteristics among the various methods of mirroring solutions. All of these solutions have their own benefits and limitations, based on their design. Any of these solutions can be better than another, based on your environment. You need to understand the features and the capabilities of each solution. Then you must select the method that best fits your environment and provides the functions and possibilities that your business requires.

For a detailed comparison of these solutions, refer to *Data Resilience Solutions for IBM i5/OS High Availability Clusters*, REDP-0888.

3

Conceptual view of cross-site mirroring

In this chapter, we provide a more detailed description of cross-site mirroring (XSM). We compare the use of switchable independent auxiliary storage pools (IASPs) to that of cross-site or geographically mirrored IASP copies. We describe an environment that uses both switched IASP and mirror technology simultaneously. We also discusses DASD considerations related to XSM, as well as the implication of using a replicate node in a cross-site mirrored environment.

3.1 Reviewing switchable IASPs and device CRGs

Before we go into detail about XSM, which became available at V5R3M0 of i5/OS, it is helpful to review the high availability options that existed prior to XSM. In this section, we discuss switchable IASP support. Switchable IASPs can be combined with an XSM environment to provide even more high availability solutions, but either can operate independently of the other.

3.1.1 Simple view of a switchable IASP between two cluster nodes

A switchable IASP is usually configured as a tower of DASD that is set up between two nodes in a cluster. It can be as simple as an input/output processor (IOP) that is switchable between two logical partitions (LPARs) acting as nodes. A device cluster resource group (CRG) is defined in order to associate the switchable IASP with the nodes in the cluster. The connection between the tower of DASD and the cluster nodes is a *high-speed link (HSL) loop*, as shown in Figure 3-1.

HSL loop: The term *HSL loop* is often used to describe the physical connection between the two nodes and the switchable tower. Do not confuse this with the chargeable licensed product, OptiConnect (Option 23), which only requires installation for system-to-system communications.

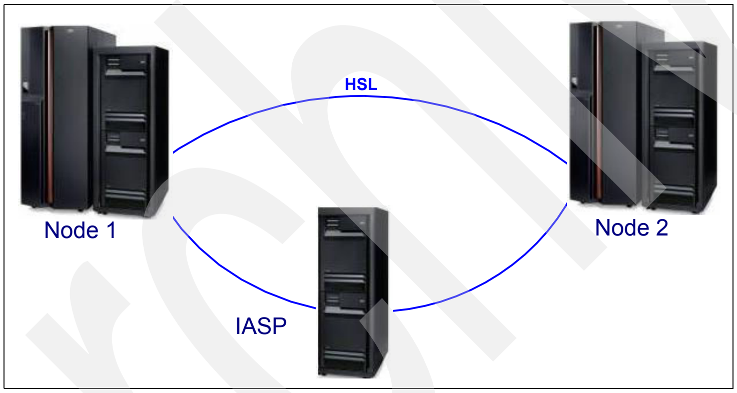


Figure 3-1 Example of a switchable IASP

Clustering requires some type of TCP/IP communications link to exist between the nodes. This connection supports the low level cluster communications. The HSL loop can be shared for both the data transport and the clustering communications, but we do not recommend this option. The cluster communications should ideally have multiple paths between nodes.

As mentioned in Chapter 1, “Introduction to cross-site mirroring” on page 1, a switchable IASP can be connected only to one system for access at any one time, as shown in Figure 3-2. While the IASP is attached to Node 1, only users on Node 1 can use the IASP.

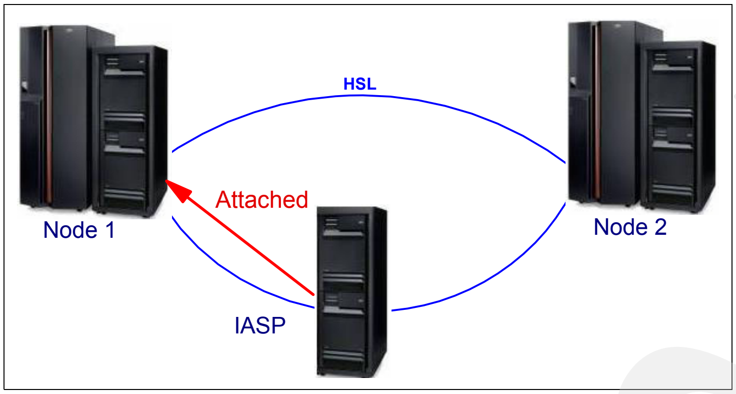


Figure 3-2 Example of an IASP attached to Node 1 in a cluster

If a switchover or failover occurs, clustering moves the IASP, which is part of a device CRG, to Node 2 as shown in Figure 3-3.

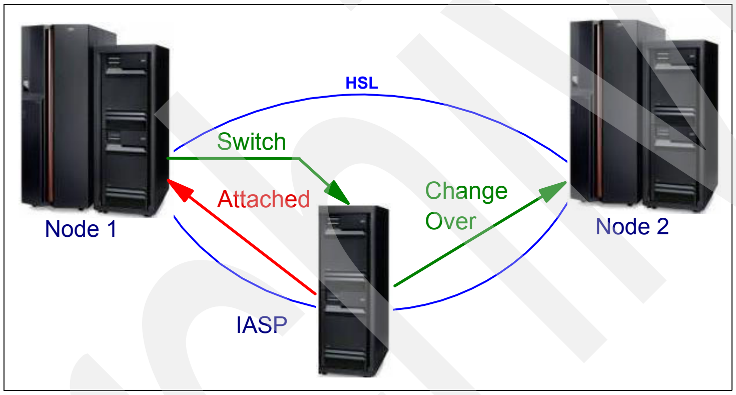


Figure 3-3 Example of a device CRG switchover

During the switchover, the tower is removed from Node 1 and reports into Node 2. When the resources for the IASP report from the tower, then the IASP is varied on to Node 2 as shown in Figure 3-4.

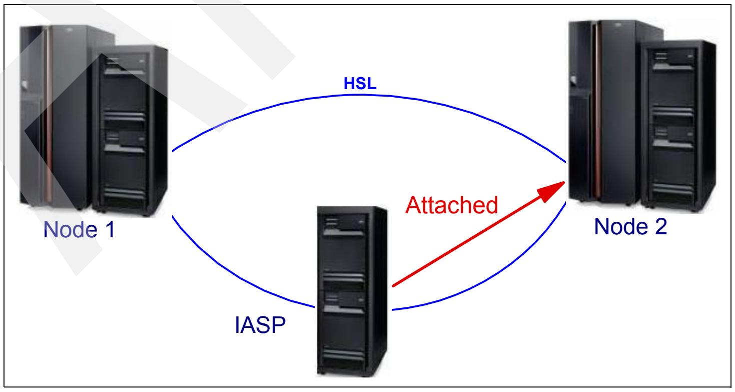


Figure 3-4 Result of an IASP attached to Node 2 following a switchover

All the hardware in the tower switches. If there are additional resources in that tower, for example, communications cards, they switch with the tower and IASP to the other system. In some cases, this can be desirable, for example, for the Integrated xSeries® Server (IXS) and tape adapters. However, if you are considering a planned switchover, and some devices in the switchable tower are required or needed on the source system, it is not desirable for them to be located in the switchable tower.

Even though the entire tower switches over, all the resources in the tower should report into the other system. When these additional resources report in, they might adopt different resource names. Some customization might need to occur if you want to use them with an existing application. For example, a communication device might not get the same communications name.

3.1.2 Relationship between clustering components

Sometimes it is easier to take a more abstract perspective of a cluster to understand the relationship between the clustering components. Figure 3-5 shows the relationship between the cluster, device CRG, and IASP. The IASP is defined in the CRG. The CRG belongs to a particular cluster.

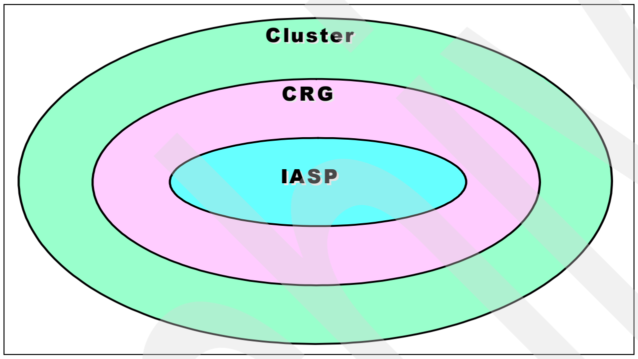


Figure 3-5 Relationship between a cluster, device CRG, and an IASP

It is possible to define an IASP that spans more than one tower. Defining more than one tower might be required when you need more DASD allocated to the IASP, but that amount of DASD will not fit into one tower. In this case, both towers switch together. These towers can be on the same HSL loop or on two separate loops. Keeping them on separate loops is a method of achieving ring-level mirroring. Figure 3-6 illustrates this concept.

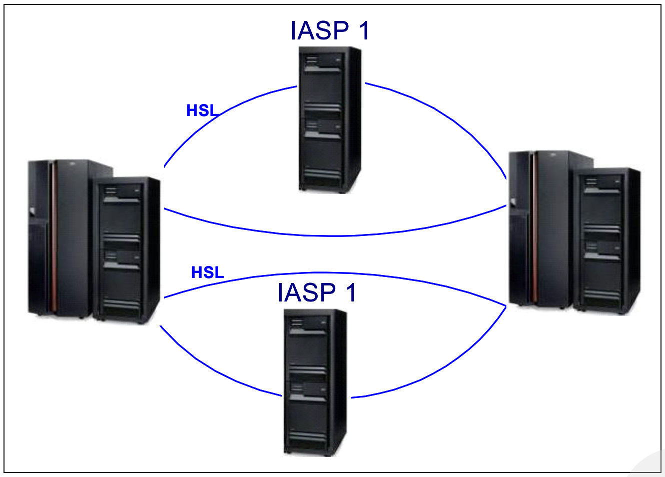


Figure 3-6 Example of one IASP spanning two HSL rings

Figure 3-7 shows an example of an IASP (IASP B) that spans two towers, but those towers are on the same loop. It also shows an example of two IASPs (IASP C and IASP D) that are contained in the same tower. Clustering enforces the requirement that they are both added to the same device CRG. In this example, clustering ensures that both IASPs are switched together from node to node in the cluster.

3.1.3 Device domains

Clustering uses device domains to define how system resources are allocated to device CRGs. Figure 3-7 shows an example of using device domains. Device domains are not required by data CRGs or application CRGs.

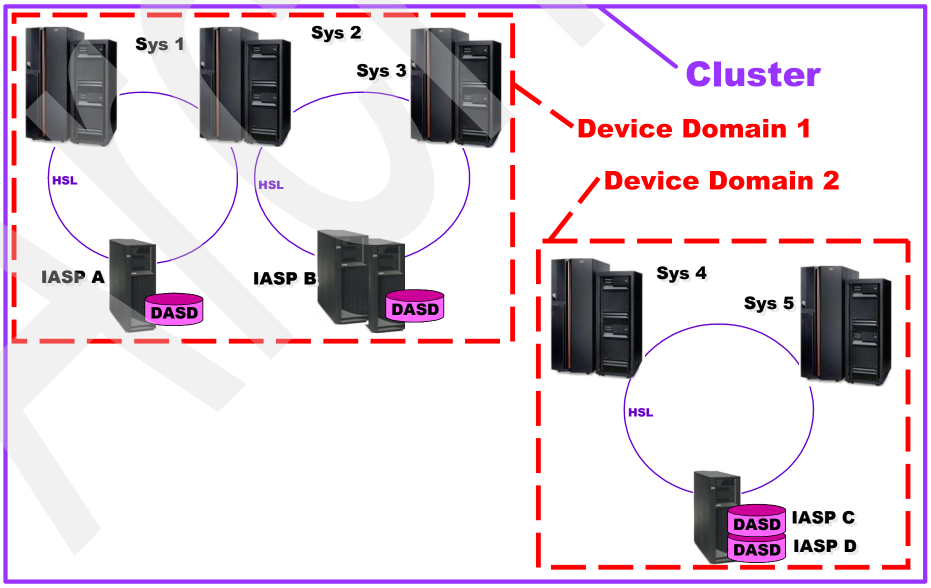


Figure 3-7 Multiple CRG setup with multiple device domains

A device CRG defines how a cluster give out information such as IASP numbers, DASD numbers, virtual address, and so forth. Unlike basic ASPs, in which the user defines the ASP number to use, the system assigns the ASP number to an IASP. The system does not allow more than one IASP in a device domain to have the same IASP number.

Using the example in Figure 3-7, IASP A and IASP C can both use ASP number 33. Since IASP A and IASP B are in the same device domain, they are assigned different ASP numbers. Since it is possible for IASP A and IASP B to both be attached to Node Sys 2 at the same time, the use of a device domain prevents problems that can otherwise arise if IASP A and IASP B were allowed to be assigned the same ASP number.

In Figure 3-7, a switchable IASP on another HSL loop cannot be added between Node Sys 3 and Node Sys 4 because they are a part of two different device domains. The only way they can be added is to remove Sys 4 and Sys 5 from Device Domain 2 and then to add them to Device Domain 1. However, a system with an IASP cannot be added to a device domain that already exists because there is a potential problem that can occur for resources that were already assigned. In this example, IASP C and IASP D have to be deleted and then recreated after Node Sys 4 and Node Sys 5 are added to Device Domain 1.

On a new setup, you can create a device domain on a node that has an associated IASP. If you want to add two nodes, each with their own IASP, to the device domain, the first node to be added to the device domain can keep its existing IASP. The other node must have its IASP deleted and recreated after it is added to the device domain.

Remember to plan for the existing nodes as well as for any future changes that you might have for the cluster.

Figure 3-8 shows the relationship of the device domain in the abstract view that we created previously. A device domain is part of a cluster and can be part of only one cluster. A device CRG can only be part of one device domain.

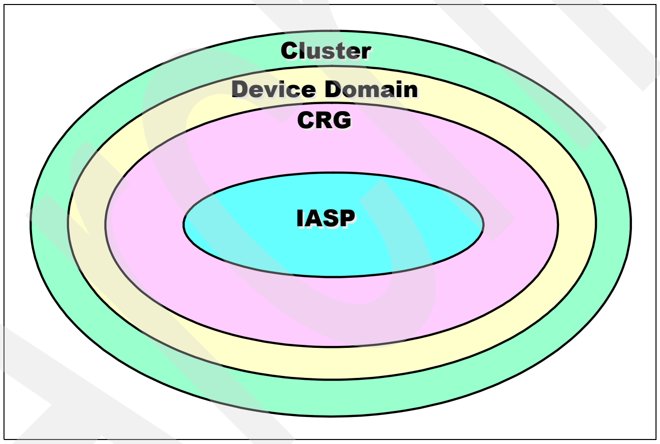


Figure 3-8 Relationship of a device domain to other clustering components

3.1.4 ASP types

To understand ASP groups, let us review the types of ASPs that we discussed briefly in Chapter 1, “Introduction to cross-site mirroring” on page 1. The two main divisions of ASPs are the system ASP and user ASPs. The *system ASP* is ASP number 1. *User ASPs* have existed since 1988 and allow data to be divided logically for performance and data resiliency.

The first type of user ASP is called a *basic ASP*. A basic ASP can contain both library data and integrated file system data. Basic ASP numbers are 2 through 32 and can be chosen and assigned by the user. With the introduction of IASPs, a new term was needed to identify all ASPs that were not IASPs. This term is *SYSBAS*. SYSBAS refers to basic user ASPs as well as the system ASP, so ASP numbers 1 through 32. This is where all of the operating system resides.

With the introduction of IASPs in V5R1M0 of OS/400, the first type of IASP includes a *user-defined file system (UDFS)* type of IASP. This type of IASP can contain only integrated file system data and has no ability to journal that integrated file system data. UDFS IASPs are a great choice if you do not have to journal your data and you need only integrated file system data.

Primary and secondary IASPs were introduced in V5R2M0 of OS/400. These IASPs allow both library data and integrated file system data. The integrated file system data in a primary or secondary IASP can be journaled.

UDFS IASPs have a lot less overhead than primary or secondary IASPs that allow library type objects. This is because of the extra tracking that is required for database objects with the system and IASP cross-reference files, among other reasons. Because of this, a UDFS IASP can vary on much faster than a primary or secondary IASP since it does not have to go through the other recovery steps. Figure 3-9 shows the relationship between the different ASP types.

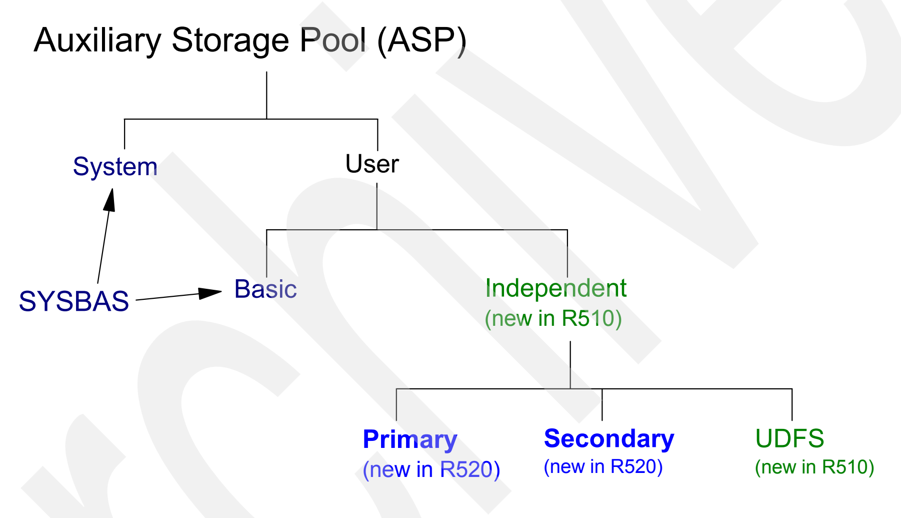


Figure 3-9 Types of ASPs

IASPs can be switchable or non-switchable. Clustering uses switchable IASPs and requires the i5/OS optional product, HA Switchable Resources (option 41). Non-switchable IASPs can be used without clustering. The support for use of a non-switchable IASP is provided in the base components of i5/OS.

IASPs have advantages even if you do not use the switchable feature. As shown in Figure 3-10 on page 40, you can use IASPs to divide your work into different logical units. You can isolate payroll data from order entry data, for example. This can even provide greater security, because you can restrict certain users from the IASP at the IASP level. It also provides hardware isolation. If something happens to the IASP that contains the order entry data, and its hardware is isolated, the IASP that contains the payroll information is not affected.

Figure 3-10 also shows an example of using IASPs to isolate data that is used by different companies. Libraries and library objects of the same name can reside in separate IASPs. For example, Company 1 can have the same library names as Company 2. However, you cannot duplicate library names between an IASP and the system ASP. Each primary IASP defines its own database system and can have its own relational database directory entry.

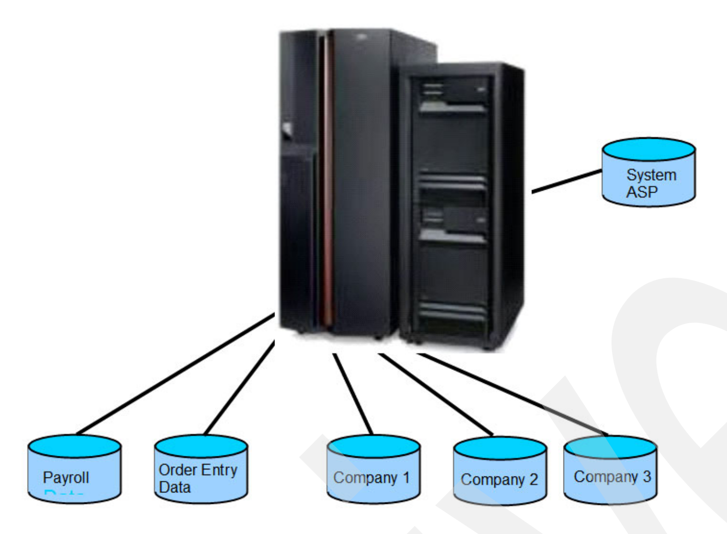


Figure 3-10 Non-switchable IASPs

Figure 3-9 on page 39 shows secondary IASPs as well as primary IASPs. A secondary IASP is logically separate from a primary IASP, but it is associated with the primary IASP. A secondary IASP always varies on with its primary IASP. If something happens to the primary IASP, a secondary IASP is not able to vary on. The secondary IASP is most often used when you need a way to isolate data from the primary IASP and have it always available, usually for performance metrics.

A common use for secondary IASPs is journal receiver isolation from the file and journal in the primary IASP, where the journal receiver exists in the secondary IASP. A secondary IASP does not create its own database environment, but it uses the database environment from the primary IASP, including its relational database directory entry.

3.1.5 Definition of ASP groups

The last component in our clustering model is an *ASP group*. An ASP group is always a primary IASP and any number of secondary IASPs. Figure 3-11 shows an example of several different ASP groups. The simplest form of an ASP group is the ASP group named “Green”. Any primary IASP has its own ASP group. The ASP group has the same name as the IASP. Primary IASP Green defines an ASP group named Green.

ASP groups “Blue” and “Red” contain a primary IASP and one or more secondary IASPs. The secondary IASPs are in the same ASP group as the primary IASP. There is no limit to the number of secondary IASPs that can be associated with a primary IASP. Note that the system ASP and basic user ASPs are all part of SYSBAS. This is not exactly an ASP group, such as those used with IASPs, but they are all grouped together in SYSBAS as non-IASPs.

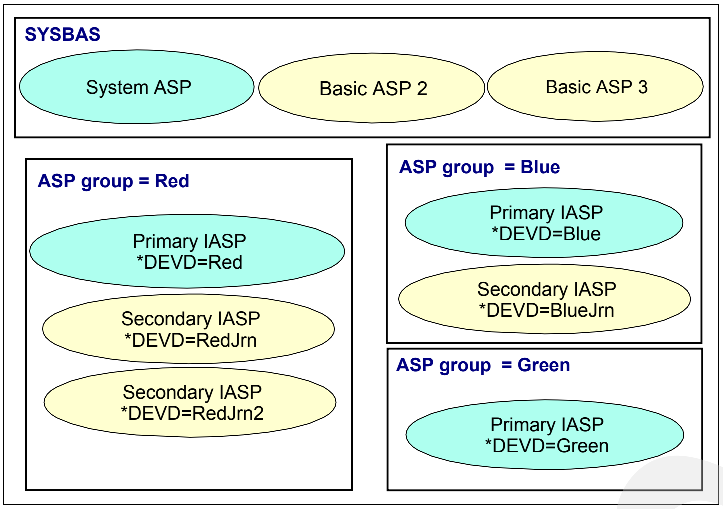


Figure 3-11 ASP groups

If we add ASP groups to our previous conceptual diagram, we notice that an IASP is part of an ASP group, which is part of a device CRG. Figure 3-12 shows this relationship.

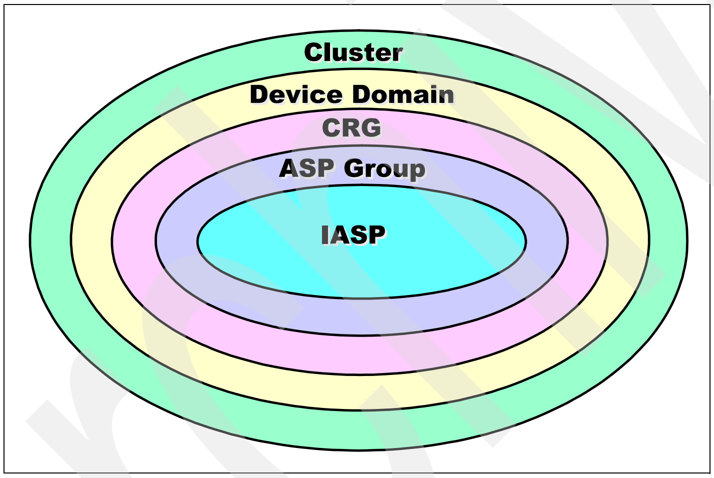


Figure 3-12 Relationship between an ASP group and other clustering components

3.1.6 Switchable IASPs using LPARs

Switchable IASPs can be used between different physical systems. Figure 3-1 on page 34 illustrates this example. You can also use switchable IASPs between LPARs within one physical system unit.

The switchable unit between two physical systems is a tower. The switchable unit between two LPARs on the same physical system is an IOP.

For LPARs on systems that are not managed by a Hardware Management Console (HMC), the bus on which the IOP is on is defined (through System Service Tools) as “own bus shared” or “use bus shared” for each partition. The DASD under the IOP can then be used in an IASP as part of a device CRG. See Figure 3-13.

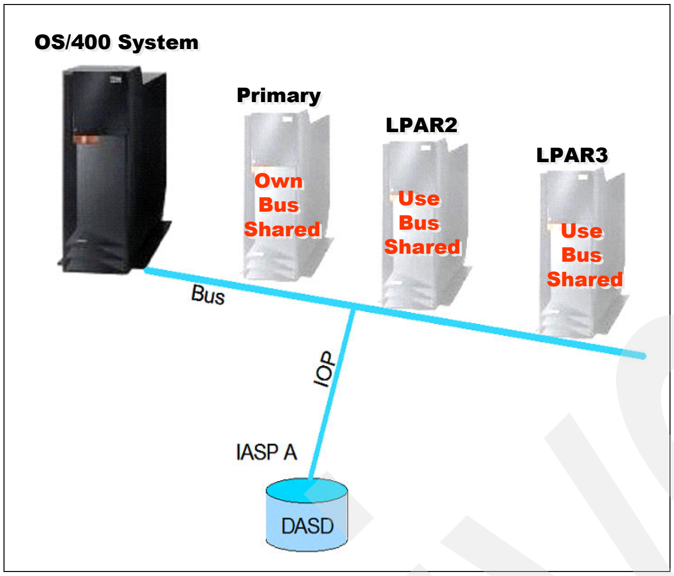


Figure 3-13 A switchable IASP on LPARs without an HMC

On partitions that are managed by an HMC, the concept of bus sharing does not exist. To allow the IOP to be switchable between the partitions, you have to create an I/O pool to which the IOP and input/output adapter (IOA) belong, as shown in Figure 3-14. In this case, each of the partitions is defined according to whether it needs the hardware. We discuss this in detail in Chapter 5, “Configuring cross-site mirroring” on page 79.

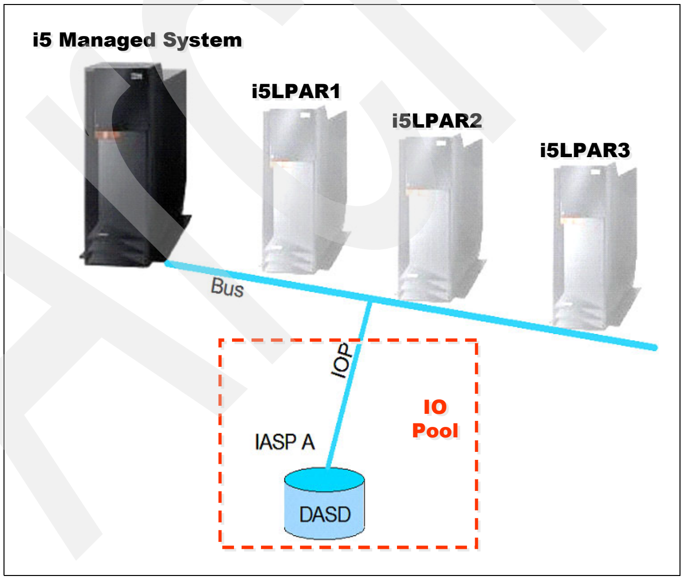


Figure 3-14 A switchable IASP on LPARs with an HMC

To provide even greater flexibility, a switchable IASP can be switched between LPARs on two different physical systems, as shown in Figure 3-15. An HSL OptiConnect Loop allows only two physical systems on any one loop with a switchable tower, but it does not limit the number of partitions on each physical system that can use the switchable IASP.

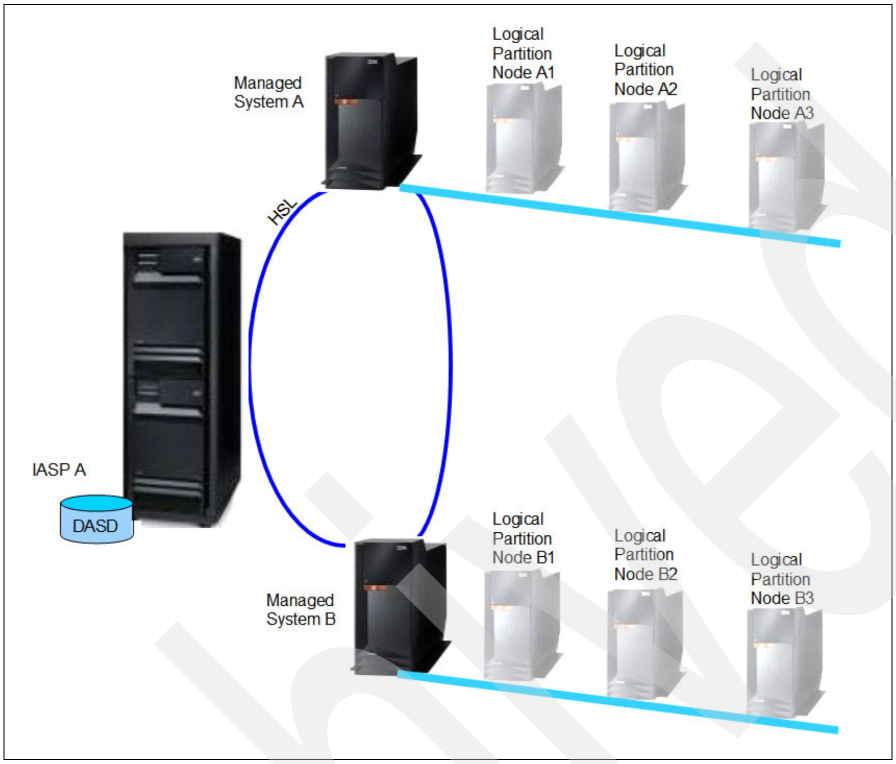

Managed System A
Logical Partition Node A1
Logical Partition Node A2
Logical Partition Node A3
HSL
IASP A
DASD
Managed System B
Logical Partition Node B1
Logical Partition Node B2
Logical Partition Node B3

Figure 3-15 A switchable IASP on an HSL OptiConnect Loop with LPARs

3.2 Basics of cross-site mirroring

XSM extends the advantage of using switchable IASPs by giving the data further resilience through distance. It requires the installation of the licensed program, HA Switchable Resources (option 41 of 5722-SS1), the use of a device CRG, and a device domain. This section provides a conceptual view of XSM.

3.2.1 Switchable IASPs versus cross-site mirroring

XSM provides site disaster protection, because another copy of the IASP is kept at a different location, as shown in the example in Figure 3-16 on page 44. The example shows an XSM environment between a system in Russia and a system in Denmark. When using switchable IASPs, you only have one copy of the IASP; with XSM, there is a copy of the IASP at each location. However, only one copy of the IASP is accessed at any given time. This copy is called the *production copy*. In our example, Sys 1 in Russia owns the production copy. When changes are made on the production copy IASP, those changes are transferred over TCP/IP to Sys 2 in Denmark and written to the *mirror copy* IASP. This movement of data over TCP/IP uses a service called *SLIC Data Port Services*. Figure 3-16 shows a two-node XSM environment.

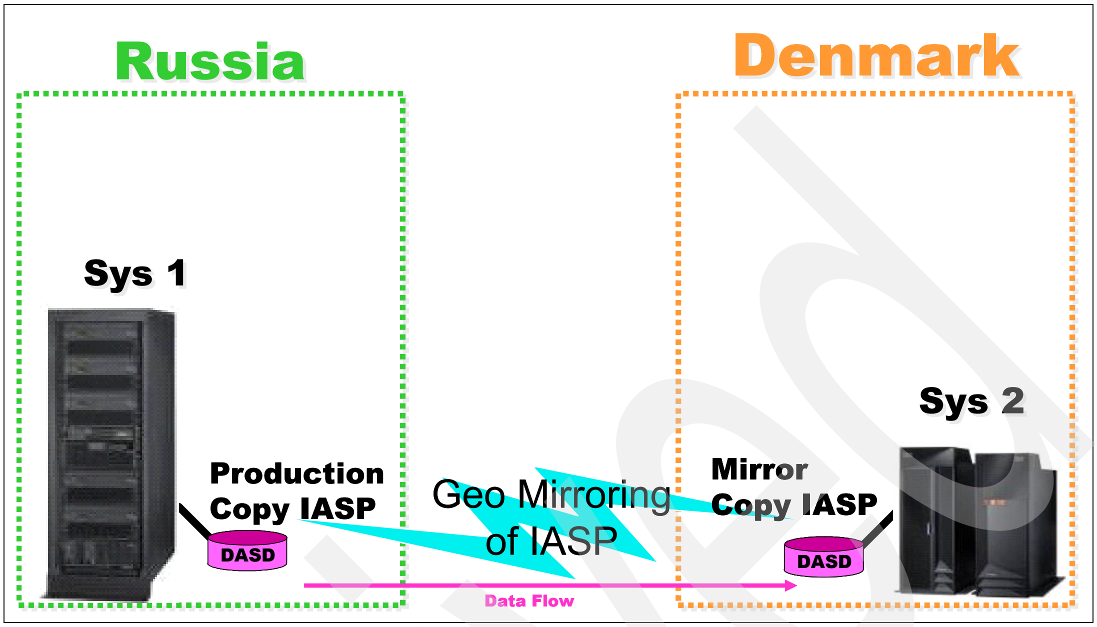


Figure 3-16 Two-node XSM environment

With switchable IASPs, the nodes in the cluster can change during a switchover or failover. Similarly, the roles of the IASP copies used in XSM can be reversed. Using our example, we can perform a switchover so that Sys 2 in Denmark becomes the production copy and users can read and write to that IASP copy. The data is then transferred to Sys 1 in Russia, which now has the mirror copy of the IASP, but cannot access the IASP.

Switchable IASPs and XSM are not mutually exclusive options. As Figure 3-17 shows, they can be used simultaneously to provide more flexibility and higher availability. XSM can only mirror between two different sites. When switchable IASPs are combined with cross-site mirrors, the IASP can switch between multiple systems (cluster nodes) at two different geographic sites.

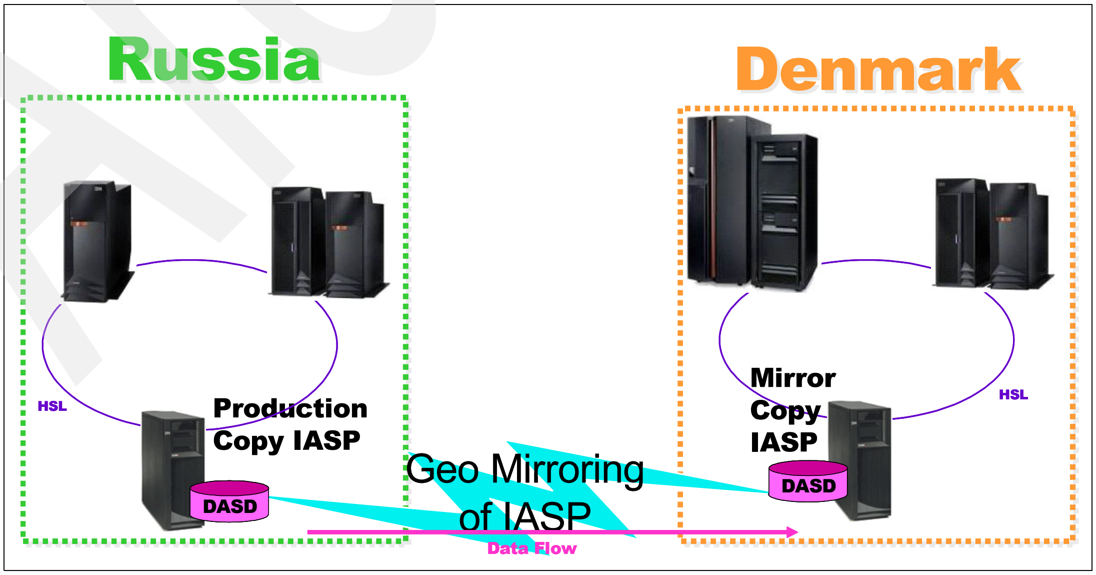


Figure 3-17 Example of switchable IASPs on four nodes

3.2.2 Cross-site mirroring communications

XSM can communicate over one communication line. However, for redundancy and, in some situations, better performance, the transfer of the data can be divided among as many as four communication lines, as shown in Figure 3-18.

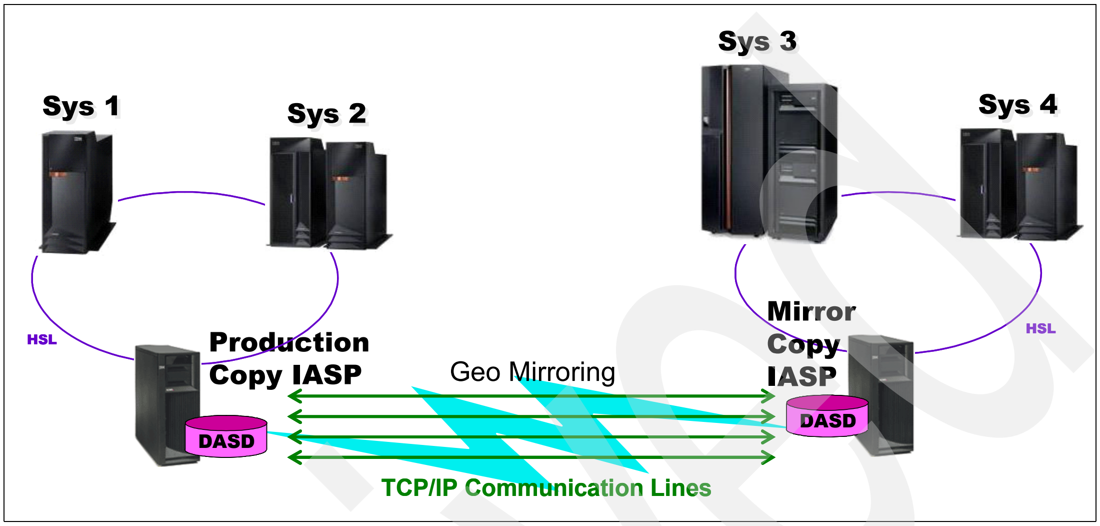


Figure 3-18 Example of data transfer over four communication lines

In addition to performance, using multiple lines also gives you more protection against failures. If there is a failure on one communication line at a point as shown in Figure 3-19, and other lines are unaffected, the cross-site mirror does not terminate.

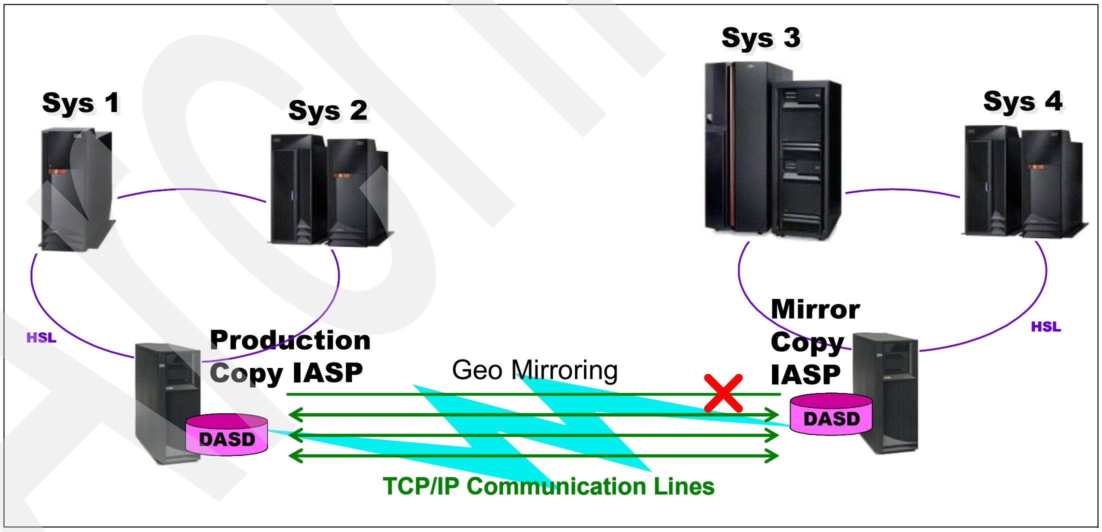


Figure 3-19 Communications failure

XSM uses a round-robin method of spreading the workload over multiple communication lines. It divides the data transfer evenly and uses the lines equally, as shown in Figure 3-20 on page 46. This method has the potential to create a problem if one communication line is slower than the others. We describe this mixing communication line bandwidth in Chapter 9, “Performance considerations” on page 153.

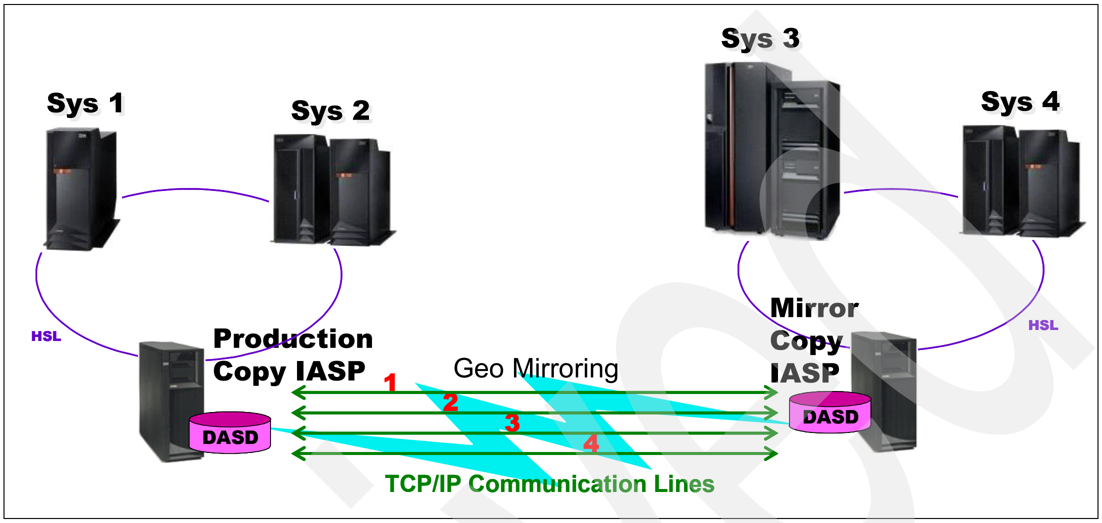


Figure 3-20 Round-robin approach to load balancing over multiple communication lines

It is important to note that XSM does not require the systems to be geographically distant from one another. The systems in a cross-site mirror environment can sit next to each other in the same room or be on two LPARs in the same physical system. Communication for XSM is done using a synchronous or asynchronous send mode.

3.2.3 Synchronous send mode

Figure 3-21 shows a graphical view of how the synchronous send mode works. The process entails the following tasks.

1. In *synchronous mode*, an application or user writes to the production copy of the IASP.
2. While that write is occurring, the synchronous send mode task transfers the information to the system that owns the mirror copy of the IASP.
3. The information is written to disk.
4. When the data is written to disk on the mirror copy of the IASP, an acknowledgement is sent to the system owning the production copy of the IASP.
5. Assuming that the write is also complete on the production copy of the IASP, control is returned to the user or application of the original write request.

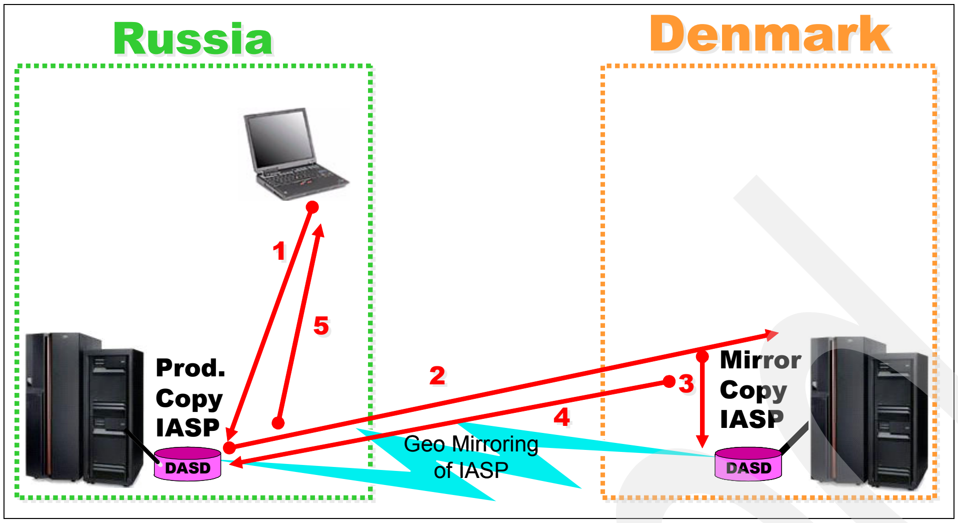


Figure 3-21 Synchronous send mode

3.2.4 Asynchronous send mode

Figure 3-22 shows how the *asynchronous send mode* works. It entails the following actions:

1. An application or user writes to the production copy of the IASP.
2. While that write is occurring, the asynchronous send mode task transfers the information to the main memory of the system that owns the mirror copy of the IASP.
3. That system then sends an acknowledgement back to the system owning the production copy of the IASP. In asynchronous send mode, the data on the system owning the mirror copy of the IASP is not written to disk before the acknowledgement is sent back.
4. Assuming the write is also complete on the production copy of the IASP, control is returned to the user or application of the original write request.

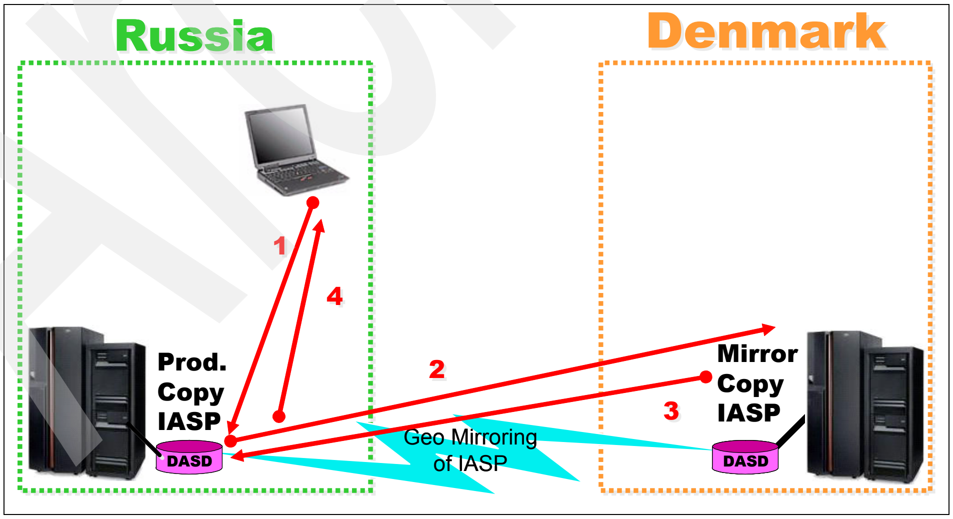


Figure 3-22 Asynchronous send mode

3.2.5 Failure to receive acknowledgements

A user-configurable threshold can be set for each ASP group. The threshold allows you to specify how long the system that owns the production copy of the IASP should wait for an acknowledgement from the system that owns the mirror copy of the IASP. If the threshold is exceeded, XSM to the mirror copy of the IASP is suspended. Control is returned to the user or application of the original write request.

For example, if the threshold is set to three minutes, the system that owns the production copy of the IASP waits up to three minutes for an acknowledgement from the other system during a possible failure. At this time, the system that owns the production copy of the IASP suspends its attempt to mirror data to the backup copy of the IASP. Because of this, the mirror copy of the IASP is no longer in a proper state to allow a switchover. The mirror copy of the IASP is no longer in synch with the production copy of the IASP.

After the communication issue is resolved, the data on the mirror copy of the IASP needs to be synchronized with the production copy. Because this process can take some time, we recommend that you take measures to prevent these situations. One way to help prevent this situation is to have at least two separate communication lines to decrease the single points of failure in communications between the two systems.

It is important to set the threshold to the optimum value, based on your particular environment. If the value is set too high, it takes longer for the system that owns the production copy of the IASP to detect the communications failure. The user jobs or application jobs that use the IASP seem to hang until the threshold is reached. If the value is set too low, XSM for that IASP might be suspended prematurely, when there is no communication issue. If there is a delay in the acknowledgements, but the delay is not as long as the threshold value, XSM for that IASP continues as though there is no problem.

3.2.6 Suspend and resume functions

XSM for a particular IASP can be suspended automatically, as described previously, or a user can manually suspend XSM to the backup copy of the IASP. A *suspend operation* is done per IASP, so XSM on a secondary IASP can be suspended, while the primary IASP in the same ASP group is not affected. When you are ready to restart XSM for an IASP, you choose the option to *resume*.

If XSM for the IASP is suspended in V5R3, a full synchronization is required after XSM for that IASP resumes. Starting with V5R4, Source Site Tracking has been implemented, which when used, enables partial synchronization. For more information, refer to 7.1.1, “How Source Site Tracking works” on page 136. In addition, refer to 7.2, “Target Site Tracking” on page 140.

Be aware that while in a suspended state, you still cannot access data on the mirrored copy.

3.2.7 Detach and reattach functions

While XSM is occurring, neither users nor applications can use the backup copy of the IASP. An additional function of XSM is *detach*. When a detach is performed, the data between the production copy and mirror copy of the IASP is no longer synchronized. However, the mirror copy can be varied on and used by users and applications at the same time that users and applications are using the production copy.

A detach operation is done at the ASP group level, so all of the IASPs in the ASP group are detached together. A *reattach* is required to restart XSM and begin resynchronization of the IASP copies.

A detach operation can be used to perform backups of the IASP, to run queries for data mining that do not affect the production copy, or even to set up a test environment for an application prior to moving it into a production environment. However, during detach, without the use of a switchable IASP, the production copy has no system to which it can fail over or switch over. Also, when the backup copy is reattached, the data in the backup copy is lost due to resynchronization.

If the data on the production copy is not important and you need to save the data on the backup copy, you can redefine the backup copy as the production copy and the production copy as the backup copy using the CHGCRG command. On the next switchover, however, you need to recover from a signature violation. We discuss this in detail in Chapter 10, “Troubleshooting” on page 175.

3.2.8 Source Site Tracking and Target Site Tracking

In V5R4, new site object tracking options were introduced. *Source Site Tracking* has been implemented, which when used, enables partial synchronization. T*arget Site Tracking* has also been introduced for i5/OS V5R4 with PTF MF40053. To learn more about these options, refer to Chapter 7, “Site object tracking” on page 135.

3.3 Switchover and failover recovery examples

In this section, we provide examples of switchover or failover scenarios with XSM.

3.3.1 Two-node example of switchover and failover

A two-node scenario is the easiest example to explain. Figure 3-23 shows a basic environment between system Sys1 in Denmark and system Sys2 in Russia. Sys1 owns the production copy of the IASP, while Sys2 owns the backup copy. Through XSM, the data in the IASP is flowing from Sys1 to Sys2.

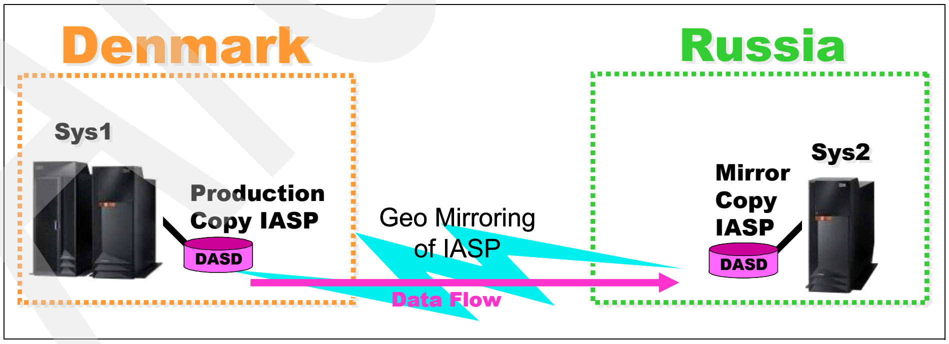


Figure 3-23 Example of a two-node environment

If we perform a switchover using the CHGCRGPRI command, for example, Sys2 then owns the production copy of the IASP, and Sys1 owns the mirror copy. The data then flows in the opposite direction, as shown in Figure 3-24.

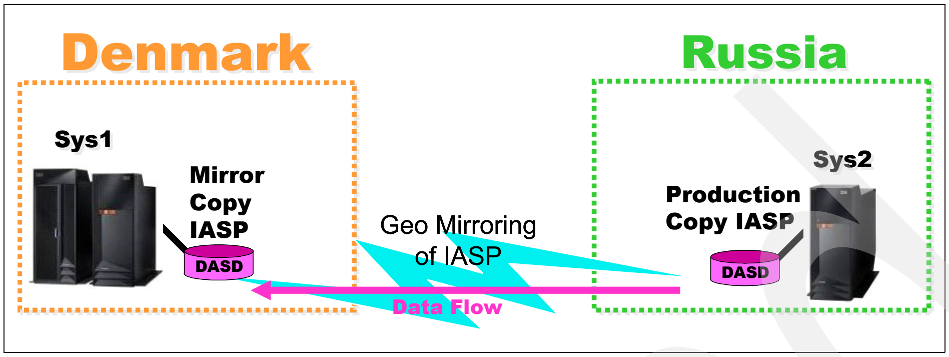


Figure 3-24 Data flow reversed after a switchover is performed

Now instead of a switchover, let us imagine a failover scenario. We begin with our original environment, which is shown in Figure 3-23. If Sys1 in Denmark goes down unexpectedly, we have a failover situation. In this case, Sys2 in Russia now owns the production copy of the IASP, and Sys1 owns the mirror copy. However, since Sys1 is down, the mirror copy is a suspended copy, and a resume operation is necessary from Sys2 when Sys1 comes back into service, as shown in Figure 3-25.

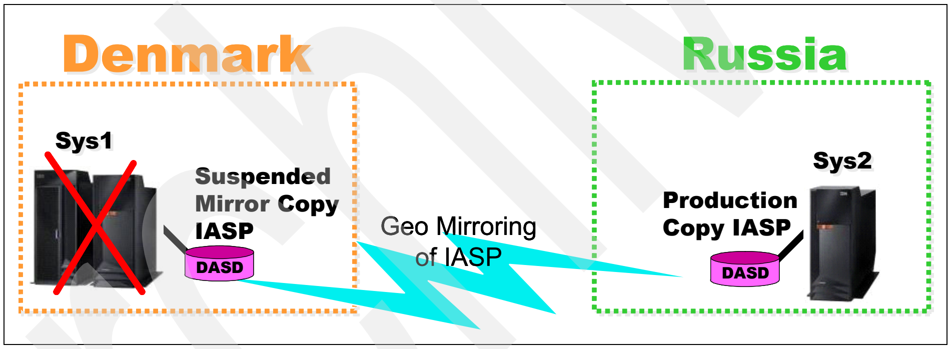


Figure 3-25 Flow of data after a failover

3.3.2 First scenario of four-node recovery domain

In Figure 3-17 on page 44, we show that you can combine cross-site mirrors with switchable IASP functionality at both sites. In this example, the two systems in Russia are Sys1 and Sys2. The two systems in Denmark are Sys3 and Sys4. This environment provides far more functional possibilities and protection. However, it gets more complex to manage and understand the expectations of how a switchover or failover will affect the recovery.

When we have an environment that combines both XSM and switchable IASPs, the CRG recovery domain and the site names determine which node owns which copy of the IASP. The first node in the recovery domain owns the production copy. The highest node in the recovery domain that is at a different site than the first node owns the mirror copy.

Table 3-1 shows the recovery domain for the cluster in this scenario.

Table 3-1 First scenario of a recovery domain in a four-node cluster

Node name	Site name
Sys1	Russia
Sys2	Russia
Sys3	Denmark
Sys4	Denmark

Sys1 is currently the first node in the recovery domain, so it owns the production copy of the IASP. Sys3 is the first node in our recovery domain with a different site name than the first node, so Sys3 is the node that owns the mirror copy of the IASP. This is illustrated in Figure 3-26. The site primary node is labeled, and an arrow from the IASP copy to the system is shown. The data flows from the production copy of the IASP to the mirror copy of the IASP.

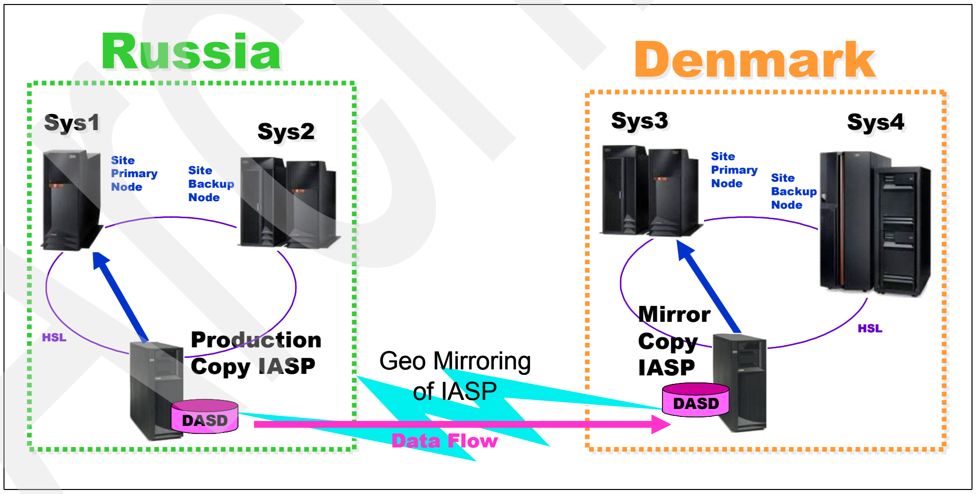


Figure 3-26 First scenario of a four-node recovery domain

If we need to do maintenance on Sys1 and need to bring down the system, we use the CHGCRGPRI command to move the CRG with the production copy of the IASP to the site backup node. The same is true in a failover scenario if Sys1 goes down unexpectedly, as shown in Figure 3-27. In this case, Sys2 takes ownership of the production copy of the IASP. Because Sys3 is still the highest node with a different site name, it keeps ownership of the mirror copy of the IASP.

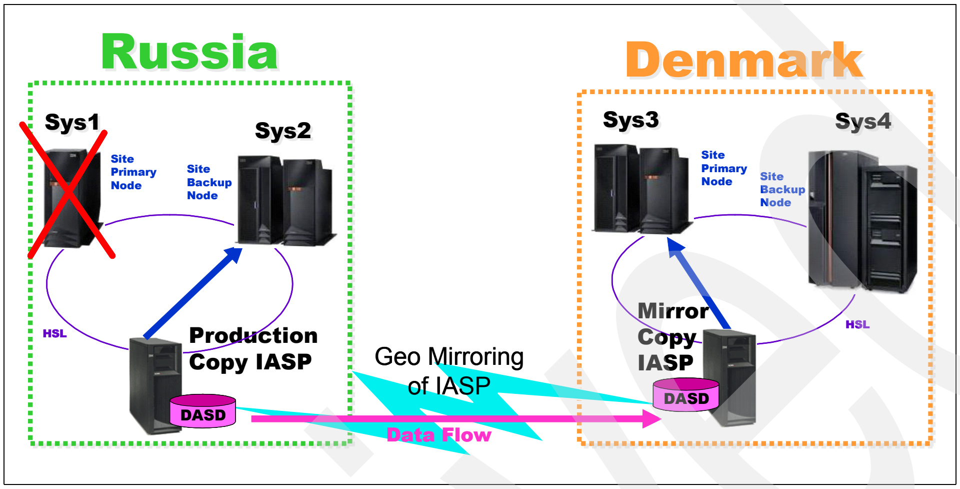


Figure 3-27 Four-node recovery domain after a switchover or failover of the primary node

When Sys1 is back up, Sys2 still retains ownership of the production copy of the IASP. Sys1 needs to be re-added to the cluster using the STRCLUNOD command. The current recovery domain is changed, similar to that shown in Table 3-2. In order to move Sys1 to the top of the recovery domain, a CHGCRG command is needed to change the current recovery domain.

Table 3-2 Recovery domain change after a switchover

Node name	Site name
Sys2	Russia
Sys3	Denmark
Sys4	Denmark
Sys1	Russia

If no change is made to the recovery domain, Sys3 becomes the highest node in the recovery domain after Sys2. If Sys2 is brought down at this point, due to a switchover or failover, Sys3 now gains ownership of the production copy of the IASP. Since the production copy of the IASP is at a different site (Denmark), data flow reverses, and Sys1 owns the mirror copy of the IASP. These changes occur because Sys1 is now the highest node in the recovery domain at a different site than the production copy of the IASP. It now becomes the site primary node for the mirror copy of the IASP, as shown in Figure 3-28.

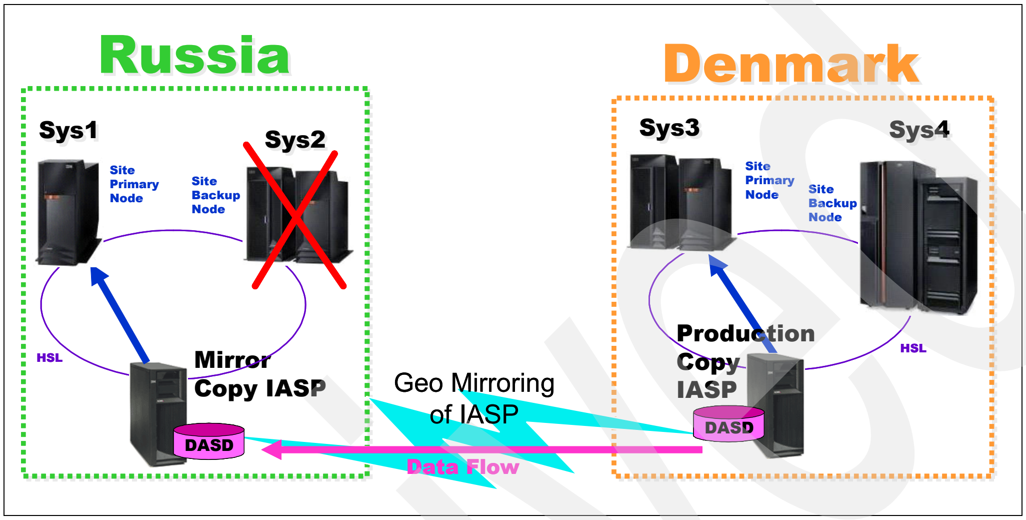


Figure 3-28 Four-node recovery domain after a second failover or switchover

At this point, Sys3 moves to the top of the recovery domain as shown in Table 3-3.

Table 3-3 Recovery domain change after a second failover or switchover

Node name	Site name
Sys3	Denmark
Sys4	Denmark
Sys1	Russia
Sys2	Russia

If Sys1 is not backed up before Sys2 goes down (or fails), neither Sys1 nor Sys2 is eligible to own the mirror copy of the IASP. Therefore, mirroring is suspended. However, the production copy is still owned by Sys3 and is available for use. When Sys1 and Sys2 are back up, the STRCLUNOD for each node allows them to rejoin the cluster. A resume function on Sys3 allows XSM to restart. Sys2 again owns the mirror copy of the IASP, since it last owned a copy prior to being suspended, allows the production copy and mirror copy of the IASP to resynch, and neither Sys1 nor Sys2 is eligible for switchover or failover until synchronization is complete.

After all systems are back up, Sys3 owns the production copy of the IASP, Sys1 owns the mirror copy of the IASP, and Sys4 becomes the next highest node in the recovery domain after Sys3. If Sys3 is brought down at this point, and a switchover or failover occurs, then Sys4 owns the production copy of the IASP, and Sys1 maintains ownership of the mirror copy, as shown in Figure 3-29.

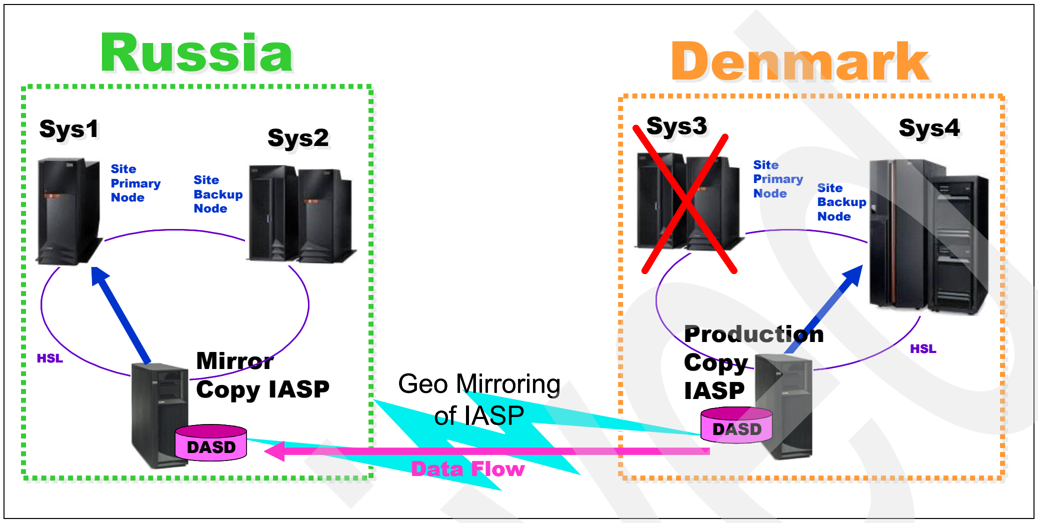


Figure 3-29 Four-node recovery domain after a third failover or switchover

At this point, the recovery domain changes again, moving Sys4 to the top, with Sys1 being the next highest node in the recovery domain, as shown in Table 3-4.

Table 3-4 Recovery domain change after a third failover or switchover

Node name	Site name
Sys4	Denmark
Sys1	Russia
Sys2	Russia
Sys3	Denmark

When Sys3 is back up and added back to the cluster, if Sys4 goes down, Sys1 gains ownership of the production copy of the IASP, and Sys3 now owns the mirror copy of the IASP, as shown in Figure 3-30. The recovery domain looks like it did in the beginning.

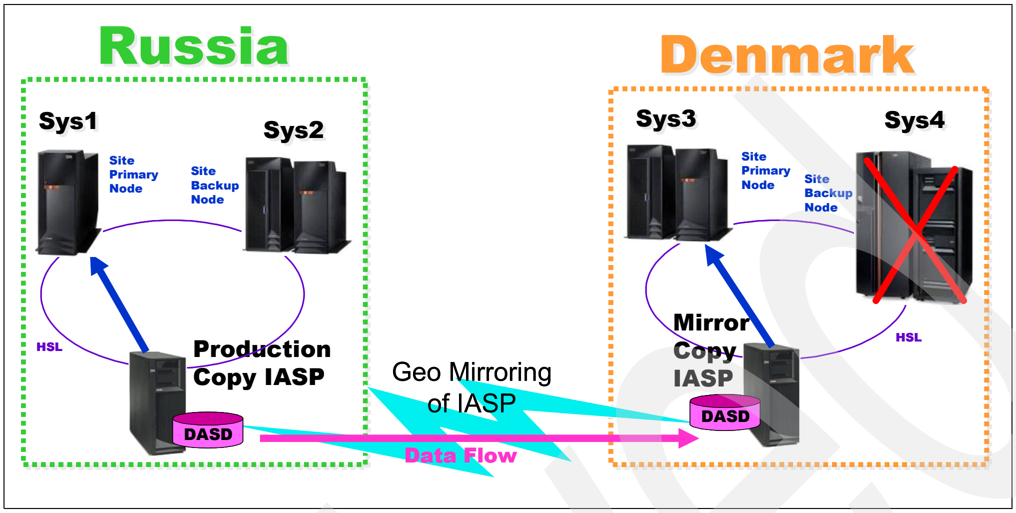


Figure 3-30 Four-node recovery domain after a final failover or switchover

3.3.3 Second scenario of four-node recovery domain

In this second scenario, the recovery domain is configured as shown in Table 3-5.

Table 3-5 Second scenario of a recovery domain in a four-node cluster

Node name	Site name
Sys1	Russia
Sys3	Denmark
Sys2	Russia
Sys4	Denmark

In this scenario, Sys3 moves ahead of Sys2 in the recovery domain. Sys1 is currently the highest node in the recovery domain, and therefore, Sys1 owns the production copy of the IASP. Sys3 is the first node in the recovery domain with a different site name than the first node, and therefore, Sys3 owns the mirror copy of the IASP, as shown in Figure 3-31.

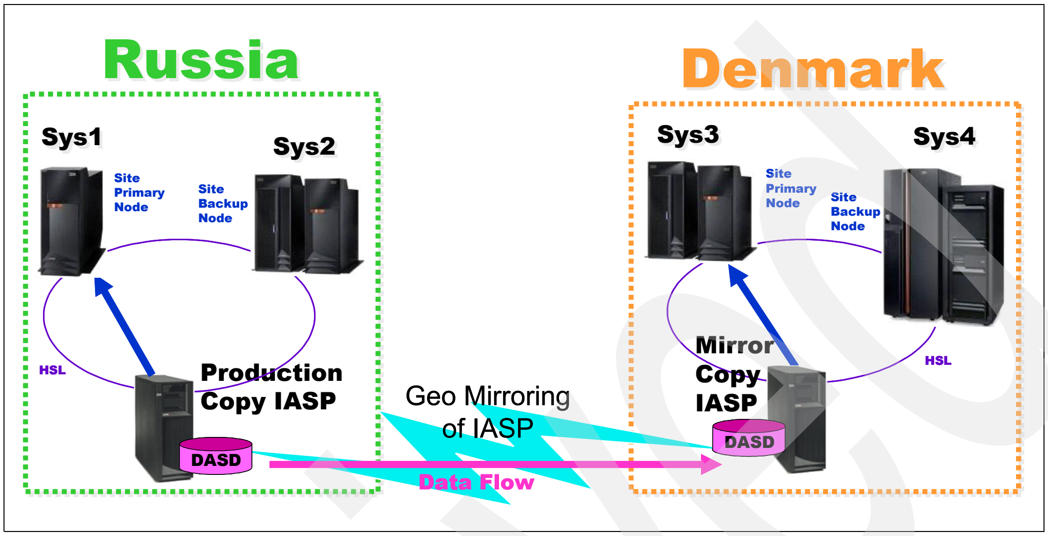


Figure 3-31 Second scenario of a four-node recovery domain

If Sys1 is brought down, and a switchover or failover occurs, Sys3 gains control of the production copy of the IASP because it is the next highest node in the recovery domain. Sys2 owns the mirror copy of the IASP, because it is the next highest node in the recovery domain that has a different site name than the node that owns the production copy. The data flow reverses, since the location of the production copy is changed, as shown in Figure 3-32.

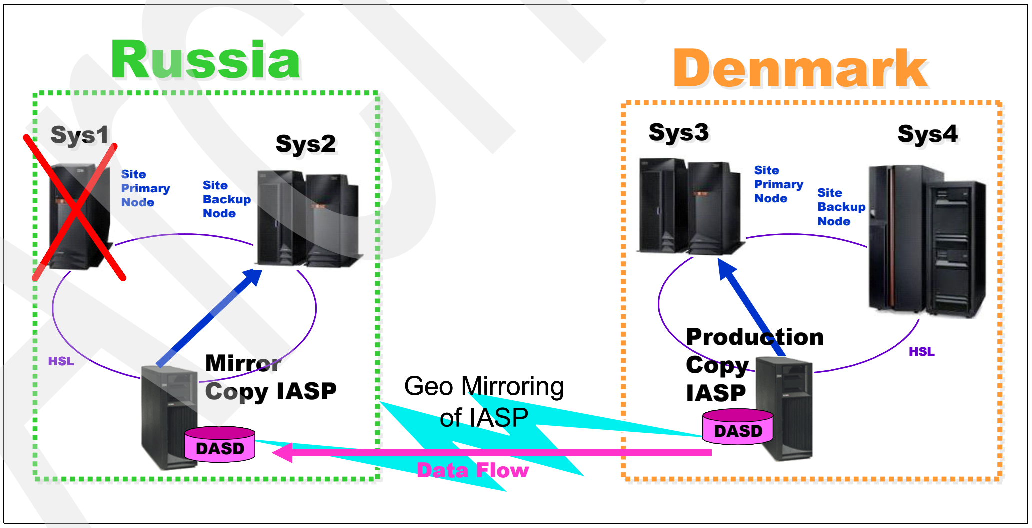


Figure 3-32 Four-node recovery domain after the first failover or switchover

At this point, the recovery domain is changed as well, as shown in Table 3-6.

Table 3-6 Recovery domain changed after the first failover or switchover

Node name	Site name
Sys3	Denmark
Sys2	Russia
Sys4	Denmark
Sys1	Russia

If Sys3 is the next system to be brought down, and a switchover or failover occurs, Sys2 is the next highest node in the recovery domain. Sys2 gains control of the production copy of the IASP. Sys4 owns the mirror copy of the IASP, because it is the next highest node in the recovery domain at a different site than the production copy. Again, the flow of data is reversed (Figure 3-33).

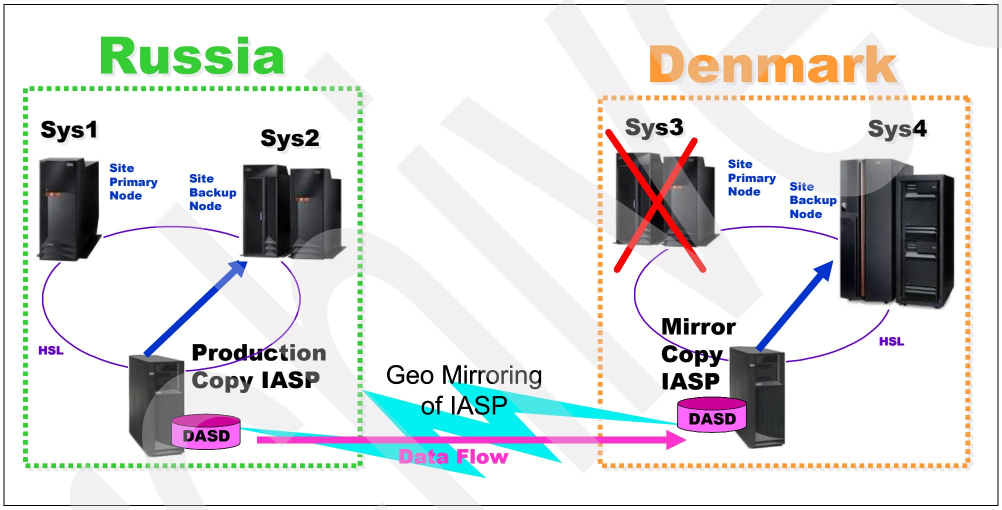


Figure 3-33 Four-node recovery domain after a second failover or switchover

The recovery domain is changed again, as shown in Table 3-7.

Table 3-7 Recovery domain changed after a second failover or switchover

Node name	Site name
Sys2	Russia
Sys4	Denmark
Sys1	Russia
Sys3	Denmark

If Sys2 is the next system to be brought down, and another switchover or failover occurs, Sys4 gains control of the production copy of the IASP, and Sys1 owns the mirror copy. Again, the data flow is reversed, as shown in Figure 3-34.

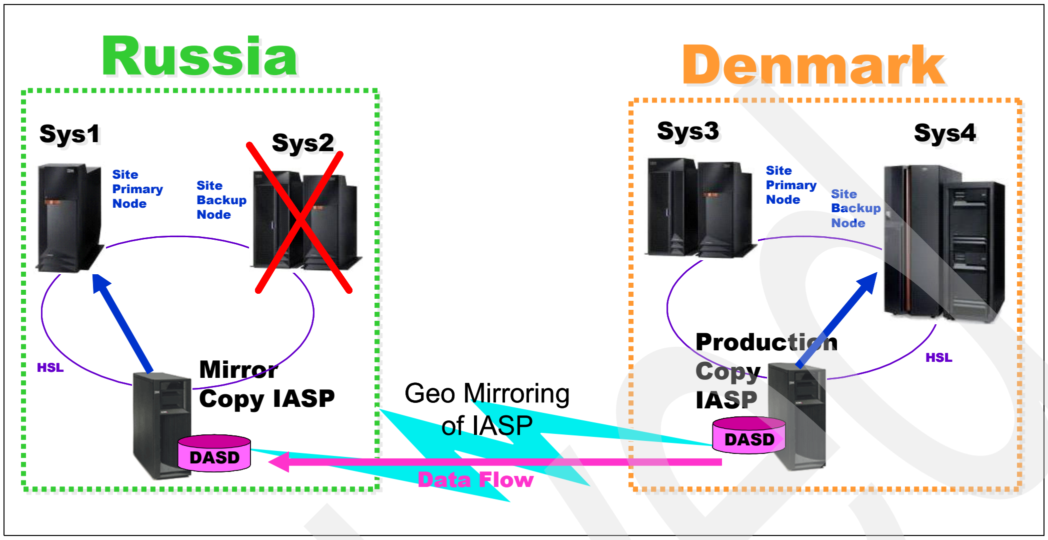


Figure 3-34 Four-node recovery domain after a third failover or switchover

The recovery domain changes as shown in Table 3-8.

Table 3-8 Recovery domain changed after a third failover or switchover

Node name	Site name
Sys4	Denmark
Sys1	Russia
Sys3	Denmark
Sys2	Russia

Finally, if Sys4 is brought down and a switchover or failover occurs, Sys1 regains ownership of the production copy of the IASP, Sys3 owns the mirror copy of the IASP, and data flow is reversed. Figure 3-35 reflects the same scenario with which we started.

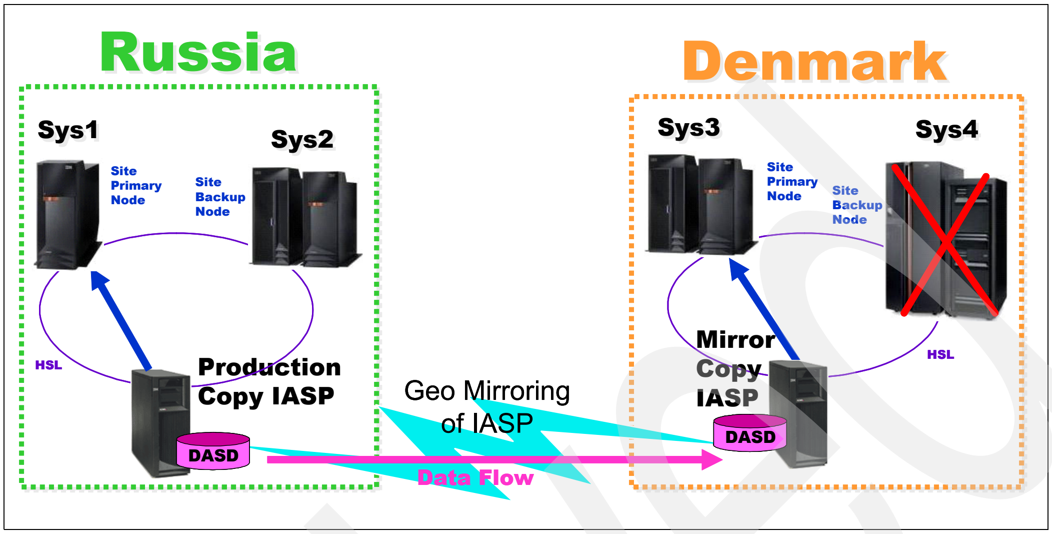


Figure 3-35   Four-node recovery domain after a final failover or switchover

3.3.4 Mirror copy failover for a four-node cluster

So far, our examples have shown what happens if the system that owns the production copy is brought down or fails. If the system that owns the mirror copy is brought down or fails, the scenario changes slightly. Let us look at an example, using the original recovery domain, as shown in Table 3-9.

Table 3-9   Recovery domain

Node name	Site name
Sys1	Russia
Sys2	Russia
Sys3	Denmark
Sys4	Denmark

Figure 3-31 on page 56 shows the original environment, which illustrates the flow. If Sys3 is brought down or fails, Sys4 is the next highest node at the same site as Sys3, so it now owns the mirror copy of the IASP. There is no change to the production copy of the IASP as shown in Figure 3-36.

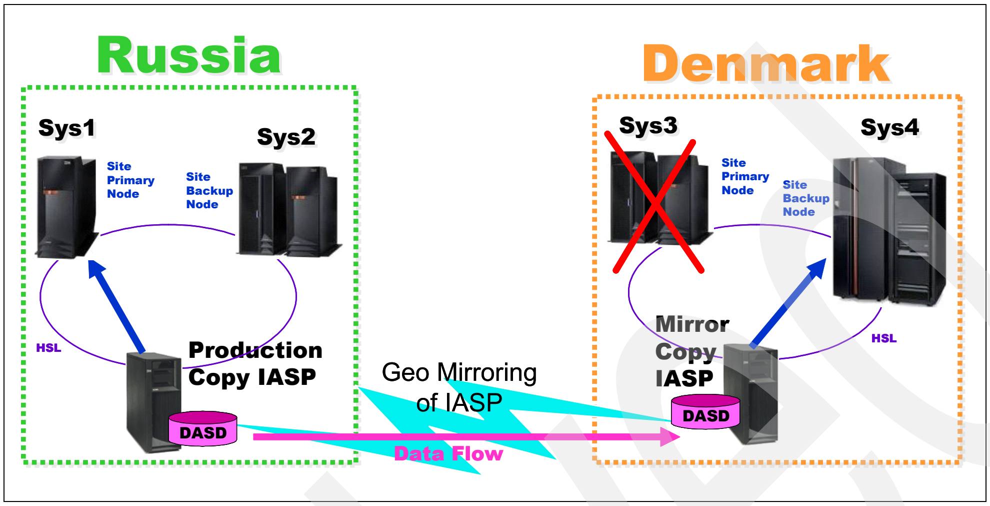


Figure 3-36 Four-node mirror copy failure

Unfortunately, since the production copy of the IASP did not change, the mirror copy is out of synch when Sys4 owns it, and a full resynch is required.

3.4 DASD considerations with cross-site mirroring

Unlike using switchable IASPs, you must consider the amounts of DASD capacity for both the production copy and mirror copy of the IASPs with XSM. DASD protection can vary between sites as well. For performance reasons, you must also consider the amount of disk arms on each system. We discuss this in detail in Chapter 9, “Performance considerations” on page 153.

3.4.1 Amount of DASD capacity

XSM requires that the mirror copy of the IASP contains at least 95% as much storage capacity as that of the production copy, but the copy can exceed much more than 100%. This amount is verified only during the original configuration. The amount becomes a problem if the mirror copy has a lot more storage capacity than the production copy and then becomes the production copy after a switchover or failover. At that point, the production copy has far more storage capacity than the mirror copy.

If the production copy of the IASP has a lot more storage capacity than the mirror copy, the copy can be suspended because the mirror copy runs out of storage long before the production copy, as shown in Figure 3-37.

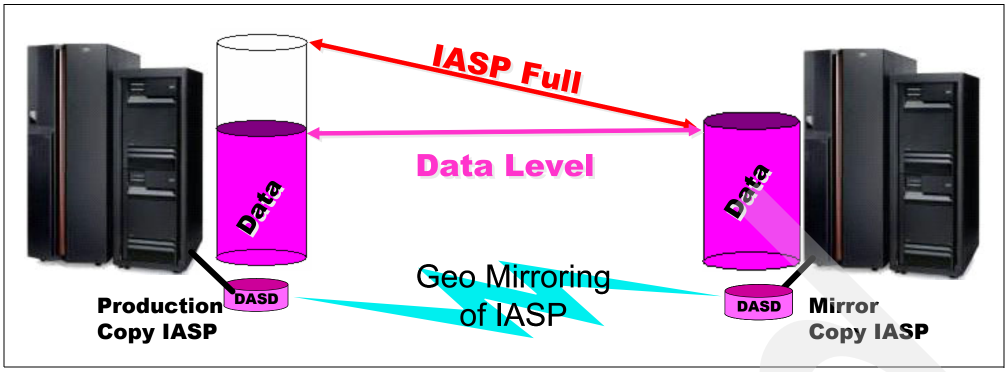
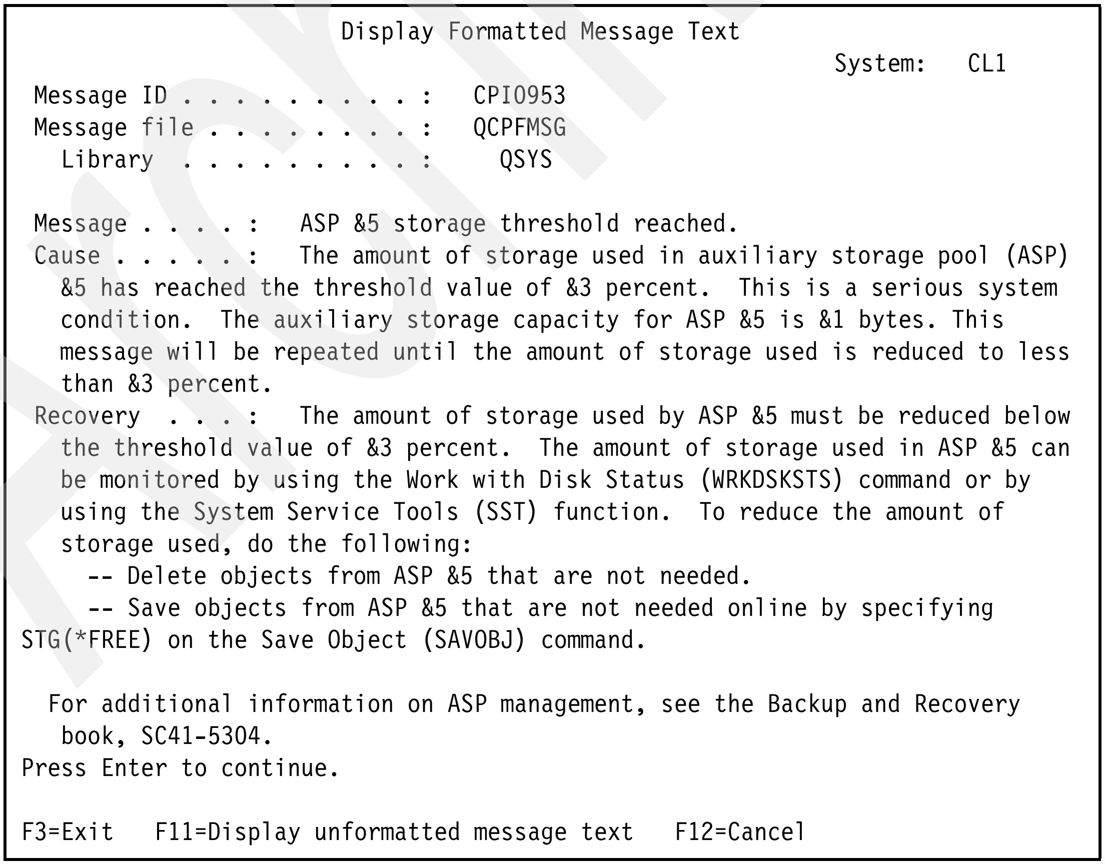


Figure 3-37 Uneven DASD capacity

It is important to set the ASP threshold to be about the same on each side so you know when your IASP is nearing its capacity. If you ignore the warning and the *mirror copy* IASP becomes full, XSM is suspended. If you ignore the warning and the *production copy* IASP becomes full, the application stops and objects cannot be created.

If the threshold is met in the mirror copy, error message CPI0953 is posted in the history log (QHST) of the system that owns the mirror copy. Error message CPI095A is posted in QHST of the system that owns the production copy.

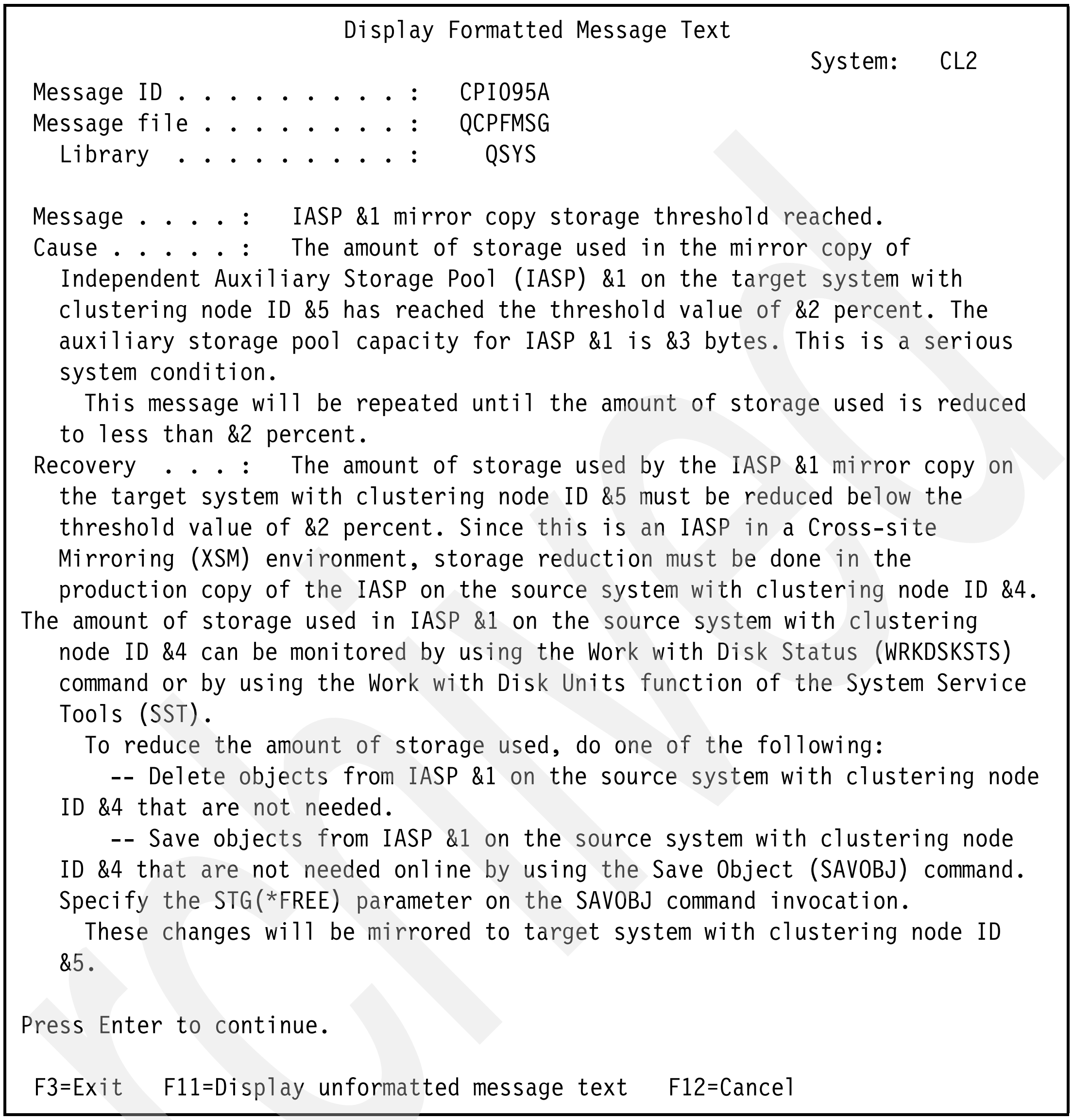
If the threshold is met in the production copy, error message CPI0953 is posted in QHST of the system that owns the production copy. Figure 3-38 shows an example of error message CPI0953.



```
                           Display Formatted Message Text
                                                                 System:   CL1
 Message ID . . . . . . . . . :   CPI0953
 Message file . . . . . . . . :   QCPFMSG
   Library  . . . . . . . . . :     QSYS

 Message . . . . :   ASP &5 storage threshold reached.
 Cause . . . . . :   The amount of storage used in auxiliary storage pool (ASP)
   &5 has reached the threshold value of &3 percent.  This is a serious system
   condition.  The auxiliary storage capacity for ASP &5 is &1 bytes. This
   message will be repeated until the amount of storage used is reduced to less
   than &3 percent.
 Recovery  . . . :   The amount of storage used by ASP &5 must be reduced below
   the threshold value of &3 percent.  The amount of storage used in ASP &5 can
   be monitored by using the Work with Disk Status (WRKDSKSTS) command or by
   using the System Service Tools (SST) function.  To reduce the amount of
   storage used, do the following:
     -- Delete objects from ASP &5 that are not needed.
     -- Save objects from ASP &5 that are not needed online by specifying
STG(*FREE) on the Save Object (SAVOBJ) command.

  For additional information on ASP management, see the Backup and Recovery
   book, SC41-5304.
Press Enter to continue.

F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 3-38 Example of message CPI0953

Figure 3-39 shows an example of error message CPI095A.



```
                          Display Formatted Message Text
                                                                    System:   CL2
 Message ID . . . . . . . . . :   CPI095A
 Message file . . . . . . . . :   QCPFMSG
   Library  . . . . . . . . . :     QSYS

 Message . . . . :   IASP &1 mirror copy storage threshold reached.
 Cause . . . . . :   The amount of storage used in the mirror copy of
   Independent Auxiliary Storage Pool (IASP) &1 on the target system with
   clustering node ID &5 has reached the threshold value of &2 percent. The
   auxiliary storage pool capacity for IASP &1 is &3 bytes. This is a serious
   system condition.
     This message will be repeated until the amount of storage used is reduced
   to less than &2 percent.
 Recovery  . . . :   The amount of storage used by the IASP &1 mirror copy on
   the target system with clustering node ID &5 must be reduced below the
   threshold value of &2 percent. Since this is an IASP in a Cross-site
   Mirroring (XSM) environment, storage reduction must be done in the
   production copy of the IASP on the source system with clustering node ID &4.
The amount of storage used in IASP &1 on the source system with clustering
   node ID &4 can be monitored by using the Work with Disk Status (WRKDSKSTS)
   command or by using the Work with Disk Units function of the System Service
   Tools (SST).
     To reduce the amount of storage used, do one of the following:
       -- Delete objects from IASP &1 on the source system with clustering node
   ID &4 that are not needed.
       -- Save objects from IASP &1 on the source system with clustering node
   ID &4 that are not needed online by using the Save Object (SAVOBJ) command.
   Specify the STG(*FREE) parameter on the SAVOBJ command invocation.
     These changes will be mirrored to target system with clustering node ID
   &5.

Press Enter to continue.

 F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 3-39 Example of message CPI095A

3.4.2 DASD protection

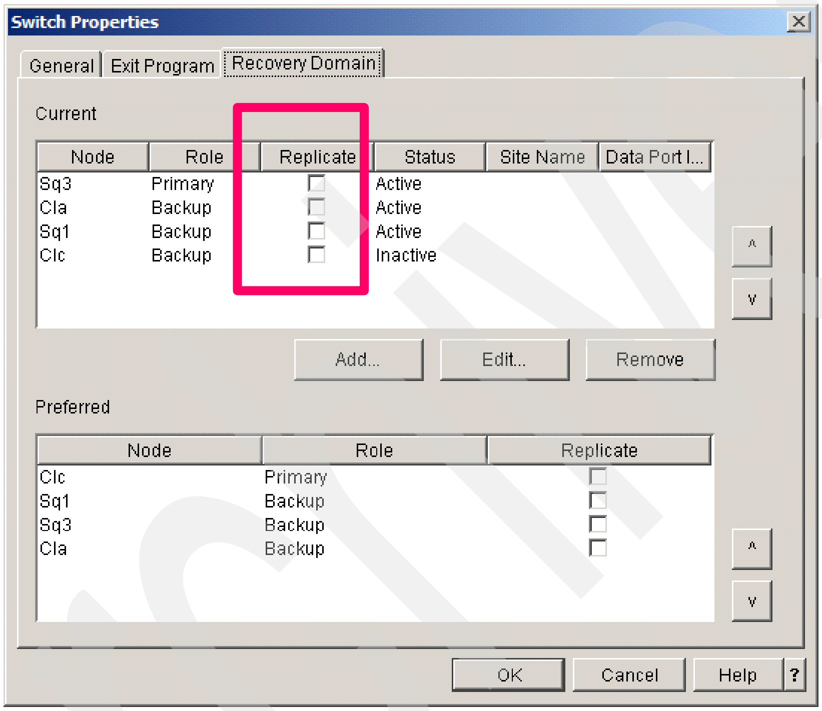
XSM does not change the type of DASD protection that you can use, such as DASD mirroring or parity protection. These same protection schemes can be used for each copy of an IASP in an environment. Also, each protection scheme can be unique for each copy of the IASP. For example, the production copy might use DASD mirroring, while the mirror copy uses parity protection. While neither copy is required to use any type of protection, one of these forms of protection is highly recommended.

3.5 Cluster replicate nodes and cross-site mirroring

Replicate nodes in a cluster have been mistaken for nodes that are involved in XSM. Replicate nodes should not be used in an XSM environment. If replicate nodes are configured in an XSM environment, the full functionality for XSM is lost.

A *replicate node* is a cluster node that contains copies of cluster resources but is unable to assume the role of primary or backup. Failover or switchover to a replicate node is not allowed. If you want a replicate node to become a primary, you must first change the role of the replicate node to that of a backup node. You accomplish this by changing the recovery domain for the device CRG.

By using iSeries Navigator, you can change the role of a cluster node (see Figure 3-40).



Current

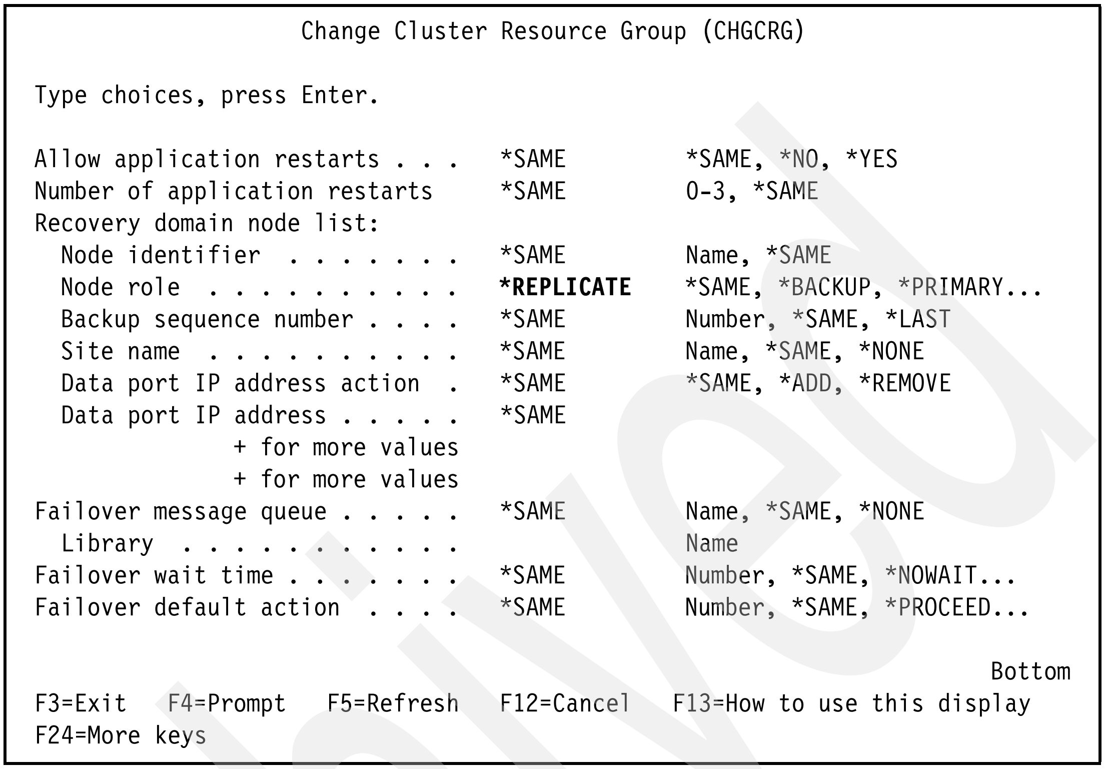
Node	Role	Replicate	Status	Site Name	Data Port I...
Sq3	Primary	□	Active		
Cla	Backup	□	Active		
Sq1	Backup	□	Active		
Clc	Backup	□	Inactive		

Preferred

Node	Role	Replicate
Clc	Primary	□
Sq1	Backup	□
Sq3	Backup	□
Cla	Backup	□

Figure 3-40 Changing a cluster node role using iSeries Navigator

You can also change the role of a cluster node by using the CHGCRG command, as shown in Figure 3-41.



```
                     Change Cluster Resource Group (CHGCRG)

 Type choices, press Enter.

 Allow application restarts . . .   *SAME         *SAME, *NO, *YES
 Number of application restarts     *SAME         0-3, *SAME
 Recovery domain node list:
   Node identifier  . . . . . . .   *SAME         Name, *SAME
   Node role  . . . . . . . . . .   *REPLICATE    *SAME, *BACKUP, *PRIMARY...
   Backup sequence number . . . .   *SAME         Number, *SAME, *LAST
   Site name  . . . . . . . . . .   *SAME         Name, *SAME, *NONE
   Data port IP address action  .   *SAME         *SAME, *ADD, *REMOVE
   Data port IP address . . . . .   *SAME
                + for more values
                + for more values
 Failover message queue . . . . .   *SAME         Name, *SAME, *NONE
   Library  . . . . . . . . . . .                 Name
 Failover wait time . . . . . . .   *SAME         Number, *SAME, *NOWAIT...
 Failover default action  . . . .   *SAME         Number, *SAME, *PROCEED...

                                                                         Bottom
 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 3-41 Changing a cluster node role using the CHGCRG command

One reason to use a replicate node is if you are using a data CRG. You can send data to multiple nodes in the cluster, as shown in Figure 3-42.

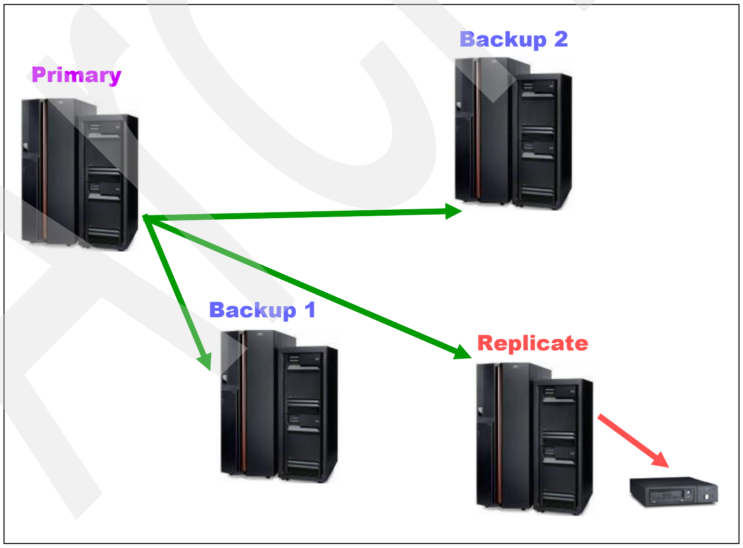


Figure 3-42 Using a replicate node in a data CRG

In this example, the replicate node can be smaller than the other nodes in the cluster. You do not want that node to have the ability to assume the role of primary after a switchover or failover. The node is therefore identified as a replicate node, so that it can still receive a copy of the data. This node can be used to perform backups of the data.

With a device CRG, however, this node cannot be used for a switchover or failover. The IASP needs to be detached to perform a backup of the data, and a full resynch is needed after the IASP is reattached.

4

Planning and installation

In this chapter, we provide a guideline for the initial preparation, planning, installation, and configuration changes that are needed prior to considering the implementation of geographic mirroring. We discuss the hardware requirements, software requirements, and necessary configurations to consider before creating your geographic mirroring environment.

Because a geographic mirroring environment requires a local area network (LAN), wide area network (WAN), or a combination of both, we also discuss the network requirements and security recommended for geographic mirroring.

4.1 Hardware requirements

In this section, we discuss the necessary hardware components and characteristics of the systems that are used in a geographic mirroring environment. You must consider the System i5 model, processor requirements, memory requirements, and disk requirements that are used in a geographic mirroring environment.

4.1.1 Server requirements

In order to use geographic mirroring, you must have at least two System i machines that support V5R3M0 of OS/400 or i5/OS. Geographic mirroring *can* be used only with the following AS/400, iSeries, and System i5 models:

- AS/400 models
  - 170
  - 250, 270
  - 720, 730, 740
  - 820, 830, 840
  - SB1
- iSeries and System i5 models
  - 250, 270
  - 520, 550, 570, 595
  - 800, 810, 820, 825, 830, 840, 870, 890
  - SB2, SB3

4.1.2 Processor requirements

Geographic mirroring creates an additional 15% to 20% workload to the system processors on both the system that owns the production copy of the independent auxiliary storage pool (IASP) and the system that owns the mirror copy of the IASP. There is no formula to calculate this exactly, because it depends on many factors in the environment and the configuration. The Workload Estimator applet does not support this type of workload either, so you cannot use it to estimate the additional processor required in order to add geographic mirroring to your environment. You can assume that approximately an additional 20% is required.

If you are an independent software vendor (ISV), consider using the IBM Innovation Center to help you evaluate and build sizing guides based on load testing of your solution on IBM products and technology. If you have never used the facilities of the IBM Innovation Centers before, you can start exploring center-based offerings by visiting this Web site at:

http://www-1.ibm.com/partnerworld/pwhome.nsf/weblook/mkt_innovation.html

If you are a customer, consider using the IBM System i Benchmark Center to help you evaluate and build sizing guides based on load testing of your solution on IBM products and technology. If you have never used the facilities of the IBM System i Benchmark Center before, you can start exploring center-based offerings by visiting this Web site at:

http://www-03.ibm.com/servers/eserver/iseries/benchmark/cbc/

Regarding the backup system, be especially careful in sizing that system's processor. It should not be a small percentage of your production system because this might slow down synchronization times considerably.

4.1.3 Memory requirements

Geographic mirroring requires extra memory of machine pool storage. To calculate the exact amount of this extra storage, use following formula:

```
Extra Machine Pool Size = 271.5 MB + (0.2 x Number of disks in the IASP)
```

This extra memory is needed particularly during the synchronization process on the system that owns the mirror copy of the IASP. However, you must add extra storage on every cluster node that is involved in geographic mirroring. Any node in the cluster can become the primary owner of the mirror copy of the IASP if a switchover or failover occurs. Use the WRKSHRPOOL command to set the machine pool size.

We recommend that you create a separate storage pool for the geographic mirroring jobs, especially if you have a large IASP.

The dynamic tuning support that is provided by the operating system can dynamically change the storage pool size to improve overall performance. The Performance Adjuster task moves storage from storage pools that have minimal use to pools that can benefit from more storage. This tuning also sets activity levels to balance the number of threads in the pool with the storage allocated for the pool. If your environment does not require automatic adjustment, you can set the system value QPFRADJ to zero to prohibit the Performance Adjuster task from changing the size of the storage pools. Setting this value to 2 or 3 (automatic adjustment) does not drop the machine pool size to less than what is set with the WRKSHRPOOL command, so this is also acceptable with a geographic mirroring environment.

4.1.4 Disk requirements

For geographic mirroring, you must have at least one non-configured disk drive on each of the nodes that is participating in geographic mirroring. Disks might not be the same capacity, but the Geographic Mirroring Configuration wizard requires that the total disk space on the system that owns the mirror copy of the IASP is no less than the size of the production copy of the IASP. You can use any disk supported by V5R3M0 of OS/400 or i5/OS. You do not have to use a dedicated adapter if you do not use a switchable IASP in your configuration. Stand-alone IASPs can be used for geographic mirroring.

If you plan to use a switchable IASP, your disk choices are limited to the disks and controllers that are supported in the switchable towers. The controllers must be dedicated, meaning that everything that is attached to the controllers, such as disks, CD drives, tape drives, and so forth, is switched to a backup node in a cluster switchover or failover scenario. If this behavior is unwanted, move these disks and other hardware to a different I/O tower.

The following recommendations are related to disk configuration for geographic mirroring:

- Use dedicated adapters for IASPs, even if you do not plan to switch the IASPs between cluster nodes.
- Use drives of equal capacity to build the IASP and its mirrored copy.
- Use an equal number of drives on the nodes that are involved in geographic mirroring.
- Use the *iSeries Disk Arm Requirements* white paper to determine if your system has adequate disk configuration. You can find this white paper on the Web at:

  http://www-03.ibm.com/servers/eserver/iseries/perfmgmt/diskarm.html

- Correct any disk performance issues *before* you configure geographic mirroring. Use Collection Services and iSeries Navigator or Performance Tools to help determine if more disk drives and controllers are needed to support your current workload.
- Remember that temporary storage that is needed by jobs that use data in the IASP require storage space from the disk pools in SYSBAS. A general rule is to use a ratio of one to three (1:3) for the relationship between the size of the SYSBAS and IASPs.

4.2 Software requirements

In this section, we discuss the necessary licensed products and software fix packages that you need to install prior to configuring geographic mirroring.

4.2.1 Operating system level (5722-SS1)

Geographic mirroring is supported with V5R3M0 or higher of OS/400 or i5/OS. If you are currently using V5R2M0 or an earlier release of OS/400 or i5/OS, you can obtain the latest version of OS/400 or i5/OS at no charge from IBM or an IBM Business Partner provided that your software maintenance agreement is in force. You can find information about OS/400 or i5/OS features and the latest release enhancements at the following Web site:

http://www-03.ibm.com/servers/eserver/iseries/software/os/i5os.html

4.2.2 High Availability Switchable Resources (5722-SS1, option 41)

High Availability Switchable Resources provide the capability to achieve a highly available environment using switchable resources. Option 41 includes support for the following features:

- Switchable IASPs

  Switchable IASPs allow you to move the data in an IASP to a backup system to keep the data constantly available. The data is contained in a collection of switchable disk units such as an I/O tower.

- iSeries Navigator Cluster Management GUI

  This GUI allows you to create and manage a switched disk cluster by using up to four cluster nodes. The utility includes wizards and help text that simplify the tasks that are involved in creating and managing the cluster.

To define switchable IASPs or to use the iSeries Navigator Cluster Management Utility, HA Switchable Resources (5722-SS1, Option 41) is required. This product is a keyed and chargeable feature, but it is non-chargeable for enterprise systems.

4.2.3 IBM eServer iSeries Access for Windows (5722-XE1)

IBM eServer iSeries Access for Windows, previously known as iSeries Client Access Express for iSeries, is a component of the IBM eServer iSeries Access Family (5722-XW1). It offers a powerful set of capabilities to connect PCs to System i models. It also enables users and application programmers to leverage business information, applications, and resources across an enterprise by extending System i resources to the PC desktop.

You must install V5R3M0 or a later version of the iSeries Access for Windows product if you want to use geographic mirroring. IBM eServer iSeries Access for Windows 5722-XE1 is a no-charge product.

4.2.4 Host Servers (5722-SS1, option 12)

The Host Servers product option handles requests from client PCs or devices, such as running applications, querying a database, printing a document, or performing a backup or recovery procedure. These servers are used by iSeries Access, and you must install this product option to configure clusters and geographic mirroring via iSeries Access.

This is a no-charge product.

4.2.5 IBM TCP/IP Connectivity Utilities for iSeries (5722-TC1)

OS/400 and i5/OS are equipped with a complete and robust suite of TCP/IP protocols, servers, and services. TCP/IP networking on the System i platform is administered and managed directly from iSeries Navigator running on a PC client or from the command line interface. The TCP/IP protocol stack on the System i platform is tuned for robust, secure, and scalable TCP/IP services and servers. This product is required to set up clustering and geographic mirroring.

IBM TCP/IP Connectivity Utilities for iSeries (5722-TC1) is a no-charge product.

4.2.6 IBM Developer Kit for Java (5722-JV1)

We recommend that you also install IBM Developer Kit for Java™ 1.4 or later and IBM Toolbox for Java (5722-JC1) as part of your preparation to configure geographic mirroring.

To define switchable IASPs or to use the IBM Cluster Management Utility, OS/400 Option 41, HA Switchable Resources is required. iSeries Navigator works optimally if the latest Java code and features are installed on the system.

This product is a keyed and chargeable feature.

4.2.7 OptiConnect for iSeries (5722-SS1, option 23)

OptiConnect for iSeries provides high-speed transparent access to data through system product division (SPD) fiber optic bus connections or high-speed link (HSL) fiber optic and copper bus connections. Using OptiConnect for iSeries among systems that share the same bus, connected with SPD fiber or HSL fiber or copper cables only, can achieve transport efficiencies that are not possible with more general purpose, wide-area communications protocols. OptiConnect for iSeries can also be an integral part of high availability (HA) configurations if you prefer to use this transport between systems.

This product is a chargeable feature.

Note: This product is *not* required for connecting switchable I/O towers to systems via an HSL OptiConnect loop. It is required only for system-to-system communications over the HSL OptiConnect loop. Clustering and geographic mirroring can use another communications transport, if desired.

4.2.8 Fixes and service packs

Prior to configuring geographic mirroring, we highly recommend that you install the latest cumulative program temporary fix (PTF) package, the HIPER group PTFs, and the database group PTFs. To find this information, go to the following Web site:

http://www-912.ibm.com/s_dir/slkbase.nsf/recommendedfixes

The first section, called “Recommended for all systems”, provides information about the latest cumulative PTF package, HIPER PTFs, and database group PTFs.

The second section is called “Recommended for specific products or functions.” Select **High Availability: Cluster, IASP, XSM, and Journal**. We recommend that you refer to all topics that pertain to your environment.

In the third section, called “Additional fix information”, you can find specific information about additional fixes that you need for your installation, data about required firmware levels, and useful links to the Software Knowledge Base documents. If needed, load individual Management Central, IASP, and cluster fixes.

Recommendation: Install the latest group PTF for IBM Developer Kit for Java.

Download and install the latest iSeries Access service pack from the IBM eServer iSeries Access Web page at:

http://www-03.ibm.com/servers/eserver/iseries/access/

4.3 Communication requirements

Because geographic mirroring involves two different systems or partitions, it also requires the capability for those systems to communicate to one another. This can be partition-to-partition communications, LAN communications, or even WAN communications over greater distances.

In this section, we describe the planning requirements regarding these various communication methods that are used in geographic mirroring.

4.3.1 Partition-to-partition geographic mirroring

The concept of geographic mirroring implies that two nodes in a cluster each own a copy of an IASP. The primary reasons to set up geographic mirroring are to provide an HA solution or to provide disaster protection and recovery for vital company data. It can, however, be set up in a partition-to-partition environment for such purposes as testing and metrics.

For geographic mirroring, at least one IP connection must exist between the two nodes. However, cluster communications need the use of a connection for low-level heartbeat messaging, and geographic mirroring requires the use of a connection for data transfer. Therefore, we recommend that, at a minimum, you set up two IP connections between the two nodes.

For partition-to-partition communications, the system bus transport can be used to set up Virtual Ethernet connections or OptiConnect connections. Remember that OptiConnect for iSeries (5722-SS1, Option 23) is a chargeable feature and is required for system-to-system communications using OptiConnect.

Two partitions can also communicate with one another using an external network, but they require that the appropriate hardware is installed for each partition.

4.3.2 HSL and system-powered control network requirements

High-speed link is the name of the system internal bus technology of the latest System i5 machines that connect system processors to industry-standard Peripheral Component Interconnect (PCI) buses. As faster processors, larger cache, faster memory, a super-fast cross-bar switch complex, faster DASD, and much faster IOPs and IOAs continue to emerge, the HSL bus infrastructure provides more speed, capacity, flexibility, and power for today's System i5 machines.

> **Tip:** For more information about HSL technology, refer to *High-speed Link Loop Architecture for the IBM eServer iSeries Server: OS/400 Version 5 Release 2*, REDP-3652.

You can use HSL technology in a geographic mirroring environment in two ways. First you can use switchable IASPs. HSL (Optical, HSL, or HSL-2) cables must be used to attach the expansion units to the servers (nodes) in the cluster.

In this case, the switchable expansion unit must be physically adjacent in the HSL loop to the alternate system or expansion unit that is owned by the alternative system. You can include a maximum of three servers (cluster nodes) on each HSL loop, although each server can be connected to multiple HSL loops. You can include a maximum of four expansion units on each HSL loop, although a maximum of three expansion units can be included on each loop segment. Each Integrated xSeries Adapter (IXA) card counts as an I/O tower.

On an HSL loop that contains two servers, two segments exist that are separated by the two servers. All expansion units on one loop segment must be contained in the same device cluster resource group (CRG).

The switchable expansion unit must be system powered control network (SPCN)-cabled to the system unit that will initially serve as the primary node for the switchable hardware group (device CRG). The primary node might be a logical partition (LPAR) within the system unit. If using LPARs, the system buses in the intended expansion unit must be owned and dedicated by the partition that is involved in the cluster. Figure 4-1 on page 74 shows the correct HSL setup for the two-node cluster with switchable towers. This setup includes the following components:

- Two systems (System A and System B) are connected via an HSL loop.
- Towers 1, 2, 3, and 4 belong to this HSL loop. Tower 5 is connected to System B on another HSL loop.
- Three towers (1, 2, and 3) are on one loop segment (one of the connections between two systems). Tower 4 belongs to another loop segment.
- All switchable external towers on one loop segment are in the same device CRG and switch together. You can switch towers 1, 2, and 3 together. If tower 1 is managed by system A, and tower 3 is managed by system B, and these towers are not switchable, you can only switch tower 2 in this segment. You can also make tower 4 switchable.

Figure 4-1 shows an example of an HSL loop.

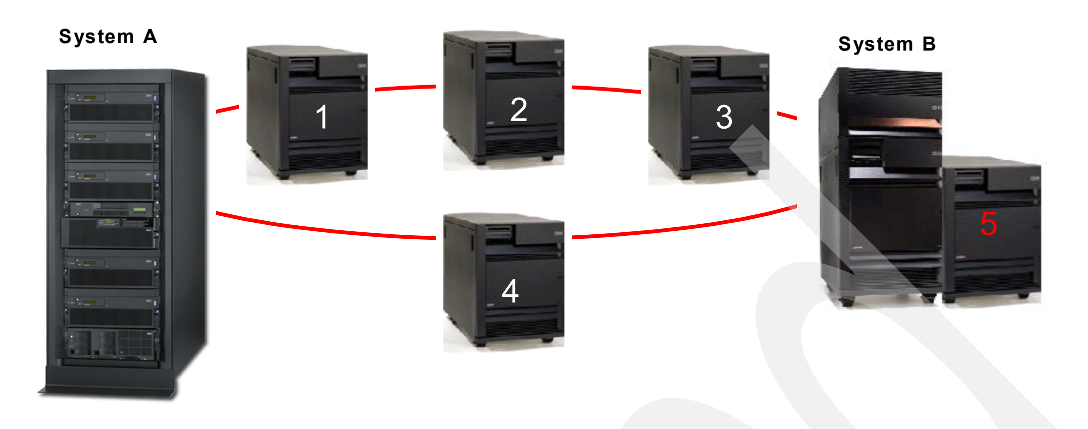


Figure 4-1 HSL loop

The second way in which you can use HSL technology in a geographic mirroring environment is to use the HSL OptiConnect transport to allow data transfer for geographic mirroring and cluster heartbeat monitoring between the cluster nodes.

Using the previous example, as shown in Figure 4-1, HSL is used to connect switchable towers to the cluster nodes as well as to provide a high-speed connection between nodes for system-to-system (node-to-node) communications, using OptiConnect for iSeries. OptiConnect for OS/400 or i5/OS supports point-to-point connectivity in a multiple loop network up to 64 nodes, including all HSL and virtual HSL connections between LPARs.

HSL OptiConnect provides fast, reliable, and secure communication between System i5 machines to the maximum distance of 250 meters for optical HSL, while the longest copper HSL cable is 15 meters. OS/400 or i5/OS allows up to three System i5 machines to be connected on one HSL OptiConnect loop. However, when three systems are connected on one loop, there must be no expansion tower between systems. The IASPs for geographic mirroring must be configured for the internal disks.

Therefore, implementing geographic mirroring is more difficult on modern System i5 machines that have fewer than four HSL ports (two HSL loops per server), because System i5 machines have a limited (six or eight) number of disk drives in a central electronic complex. For instance, a four-node cluster setup requires two HSL OptiConnect loops per server, as shown in Figure 4-2. In this example, you see three HSL OptiConnect loops and four iSeries 825 models with internal disks that are used to implement geographic mirroring.

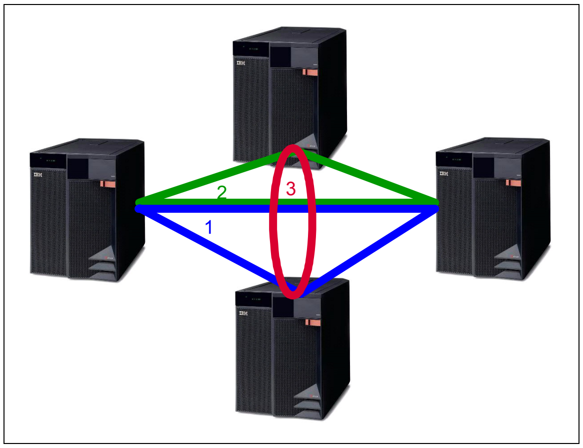


Figure 4-2 Four-node OptiConnect connections

4.3.3 Local area network requirements

Geographic mirroring over an external LAN requires a minimum of at least one network card per cluster node. Additional network cards are not required, but we highly recommend them for the following reasons:

- One IP connection is required for clustering communications. This low-level heartbeat messaging between cluster nodes is needed to ensure that other members of the cluster are still active and no failover is required. We recommend that you use two IP addresses and two network cards for cluster messaging to guarantee redundancy. You can use 10/100 Ethernet cards for this purpose, because clustering communications does not require high throughput.
- Up to four dedicated network cards can be used exclusively for the geographic mirroring data port service. It is preferable to use Gigabit Ethernet cards. You must adhere to PCI card placement rules. Do not attach two or more high-performance PCI cards (for example, Gigabit Ethernet, Fibre Channel, SCSI, and so forth) to the same IOP or bus. You can validate the card setup with the help of the LPAR Validation Tool even if you do not have LPARs on your system. Do not use these cards for purposes other than geographic mirroring of data port services in the production environment, especially for heartbeat monitoring. If data port services generates heavy network traffic during synchronization, the cluster can lose inter-node connectivity, and unwanted failover might occur.

  > **Tip:** Do not mix Gigabit and 10/100 Megabit network IOAs for data port services unless you can accept that total throughput will be limited to the slowest network card performance.

- If you plan to use IP takeover, you might want to provide an additional dedicated network card on each of the geographic mirroring cluster nodes.

4.3.4 Wide area network considerations

The greater the distance is between the cluster nodes that are used in geographic mirroring, the more factors you have to consider regarding the WAN configuration prior to implementing geographic mirroring. The two primary considerations include the topology of the network and the security used over the network.

Network topology

Data port services and heartbeat monitoring performance can be limited by your network topology. The type of networking equipment, the quality of service, and the distance between nodes can all affect the communications latency. As a result, these become additional factors that affect the speed at which geographic mirroring operates. In production environments, however, these factors have minimal effect on geographic mirroring performance when the distance between the nodes is fewer than 48 kilometers (30 miles). Geographic mirroring can be used for virtually any distance, but you must consider the reliability and performance of the network connection prior to configuring geographic mirroring.

Network security

Geographic mirroring does not use a specified TCP “server” port. The ports that are used for geographic mirroring are negotiated between the nodes through cluster messaging services. Any ephemeral port on the system can be chosen, depending on which ports are available at the time at which geographic mirroring starts.

Consider a case where the systems are connected via a campus area network or an intranet that is safe from other public networks. In this example, the network equipment, such as routers and switches, can be configured to allow all traffic between the IP addresses that are configured for geographic mirroring over any TCP port.

If the network connection between the systems involves an external public network, such as the Internet, greater consideration is required. Geographic mirroring does not have any built-in functionality for encrypting the data. It is typically unheard of to configure firewalls and routers over a WAN to allow all TCP ports to be left open. The best option in this type of environment is to configure a virtual private network (VPN) between the routers at either end of the external network to encrypt the data before it is sent over the public network. A VPN configuration can be used on the System i5 machine, but it adds far more overhead to the system and can affect geographic mirroring performance.

4.4 Additional configuration requirements

Before you configure any geographic mirroring scenario, you must complete the following additional tasks:

- Install the optional components of Logical Systems and Configuration and Service for iSeries Navigator. We recommend that you perform a full setup for iSeries Access for Windows.

  You can find instructions about how to add iSeries Navigator components in *IBM eServer iSeries Independent ASPs: A Guide to Moving Applications to IASPs*, SG24-6802.

  Make sure that you re-install the latest service pack after adding components to iSeries Access.

- Set up Management Central in iSeries Navigator.

  For information about Management Central, refer to the following Web address:

  http://publib.boulder.ibm.com/infocenter/iseries/v5r3/topic/rzaih/rzaih1.htm

- Ensure that you have configured TCP/IP on both systems, including valid IP addresses, subnets, and routes. Geographic mirroring does not use host names. Therefore, the host table and domain name server are not required.

- Make sure the Internet daemon (INETD) server is started on each of the System i5 machines in the cluster. This is required for a node to be added or started, as well as for merging partition processing. To start this server, the licensed product HA Switchable Resources (5722-SS1, Option 41) must be installed.

  To start this server from a command line, use the following command:

  ```
  STRTCPSVR SERVER(*INETD)
  ```

  You can also set the INETD server to start automatically when TCP/IP is started by using iSeries Navigator. Click **Network** → **Servers** → **TCP/IP**. Then right-click **INETD** and select **Properties**.

- Allow each node in your cluster to be added to a cluster by changing the network attributes on the System i5 machine. From a command line, enter *either* of the following commands:

  ```
  CHGNETA ALWADDCLU(*ANY)
  CHGNETA ALWADDCLU(*RQSAUT)
  ```

  The value of `*ANY` is recommended for test environments, since it does not require any authorization.

- Make sure you configure and can use the Service Tools Server (STS) on each of the systems, and you have a valid Dedicated Service Tools (DST) user ID. To configure STS, from a command line, enter the following command:

  ```
  ADDSRVTBLE  SERVICE('as-sts')  PORT(3000)  PROTOCOL('tcp')  TEXT('Service Tools Server')
  ALIAS('AS-STS')
  ```

  The single quotation marks are required for the SERVICE and PROTOCOL parameters. If the values are typed in uppercase characters, the server will not start.

  After you create the server table entry, you must end and restart TCP/IP.

  > **Switchover scenario:** If the system or partition is part of a clustered environment, ending TCP/IP can result in a switchover scenario.

- Verify that your user ID has at least *IOSYSCFG special authority on all systems in the cluster and geographically mirrored environment.

5

Configuring cross-site mirroring

In this chapter, we provide an example of configuring a simple, two-node geographically mirrored cluster. You can configure this type of cluster partly through the command line interface, and partly through iSeries Navigator. We provide step-by-step instructions for configuring this type of scenario from scratch, assuming that no cluster has been configured yet. This is not an ideal environment for cross-site mirroring (XSM), since suspending the mirror at any point provides no failover possibility without using a switchable independent auxiliary storage pool (IASP) and having additional nodes in the cluster. The examples in this chapter are merely to provide details about how to implement XSM.

5.1 Benefits of a two-node cluster

Although a cluster of three or more nodes provides much more resilience for your data and applications, this two-node scenario has the following advantages:

- It is more reliable than any stand-alone configuration.
- You need to acquire minimal additional hardware.
- It is the easiest scenario to configure.

These advantages make a two-node geographically mirrored cluster the best solution for the proof-of-concept deployment and for the test environment. If you are new to clustering and XSM, you should start with a two-node scenario.

5.2 Configuring a two-node cluster

To configure a two-node XSM cluster scenario:

1. Create a cluster.
2. Create an IASP (disk pool) on the primary node.
3. Create a device cluster resource group (CRG) for the IASP, and define roles for the nodes in a recovery domain.
4. Launch the Configure Wizard in iSeries Navigator and follow the instructions. The wizard enables you to specify the parameters for mirroring as well as creates the IASP on the backup node.
5. Start the CRG and make the mirror copy of the IASP available.

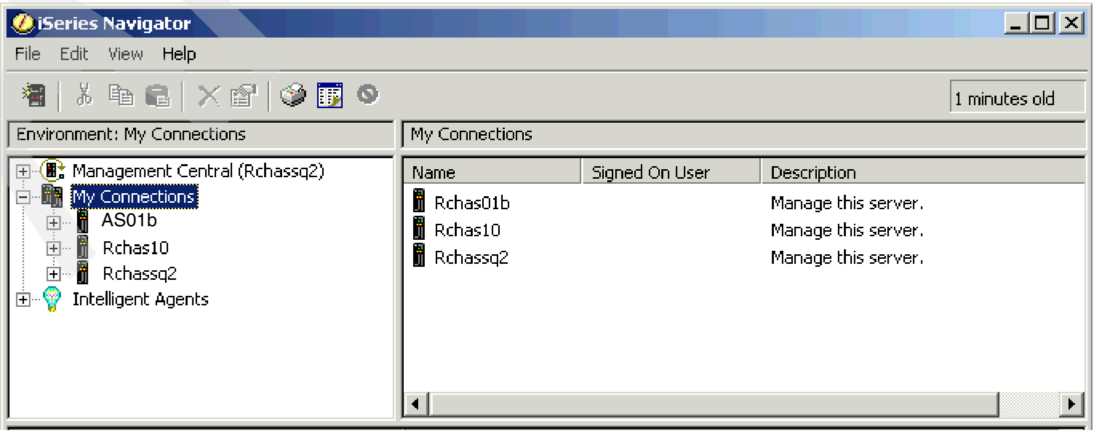
5.2.1 Creating a two-node cluster

To create a two-node cluster, you can use iSeries Navigator or a 5250 interface.

Creating a cluster with iSeries Navigator

You can use the Create New Cluster wizard in iSeries Navigator to create a cluster:

1. Launch iSeries Navigator and sign in with an appropriate user name and password.
2. In the left navigation bar of the iSeries Navigator window, expand **Management Central** and click **Cluster**. See Figure 5-1.

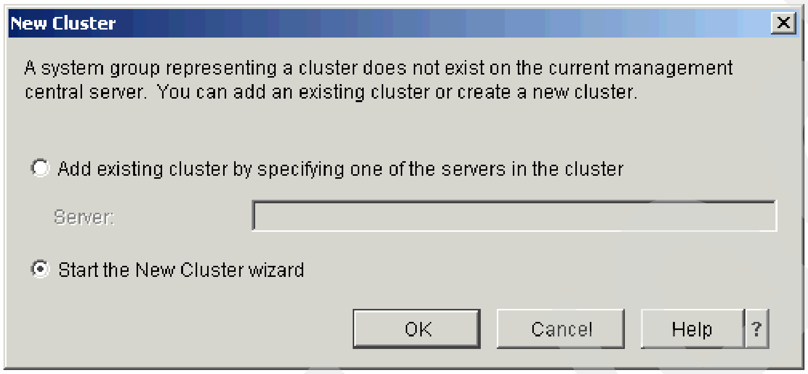


Name	Signed On User	Description
Rchas01b		Manage this server.
Rchas10		Manage this server.
Rchassq2		Manage this server.

Figure 5-1 iSeries Navigator window

3. In the New Cluster window (Figure 5-2), accept the default of **Start the New Cluster wizard** and click **OK**.

> **Tip:** If you do not see the New Cluster window (Figure 5-2), in the iSeries Navigator window (Figure 5-1), click **Management Central**, right-click **Clusters** and select **New Cluster**.



A system group representing a cluster does not exist on the current management central server. You can add an existing cluster or create a new cluster.

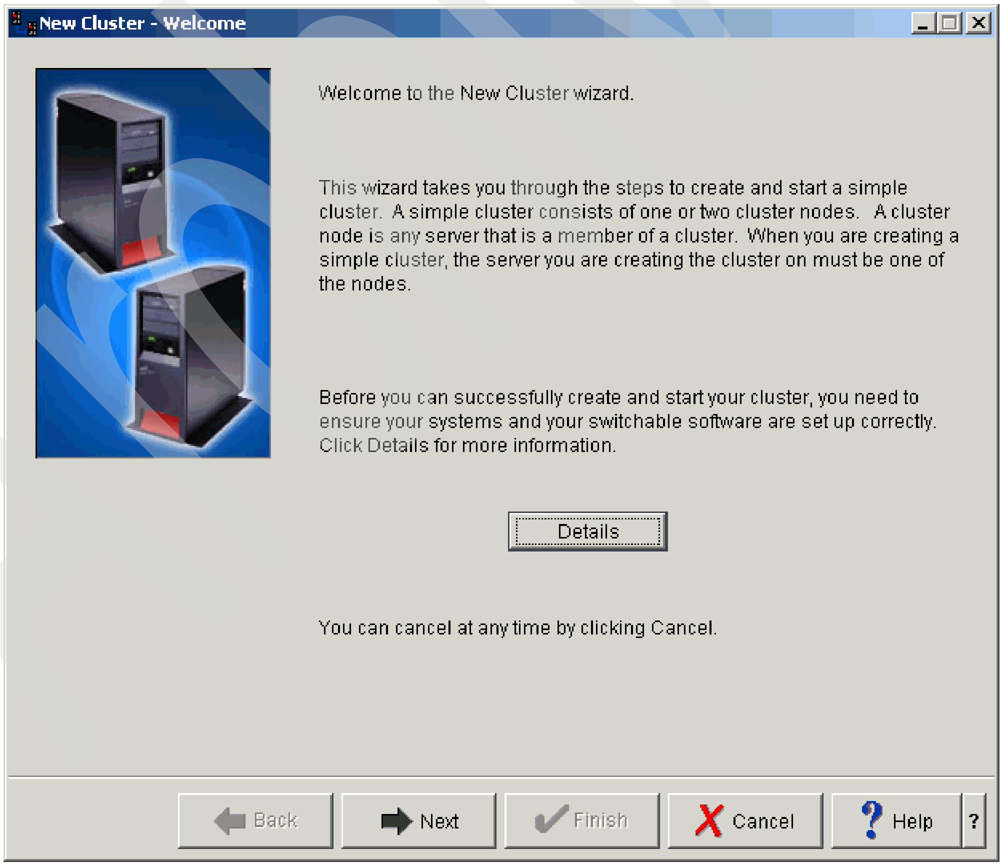
○ Add existing cluster by specifying one of the servers in the cluster

Server:

● Start the New Cluster wizard

Figure 5-2 Selecting the New Cluster wizard option

4. In the New Cluster - Welcome window (Figure 5-3), click **Next**.



Welcome to the New Cluster wizard.

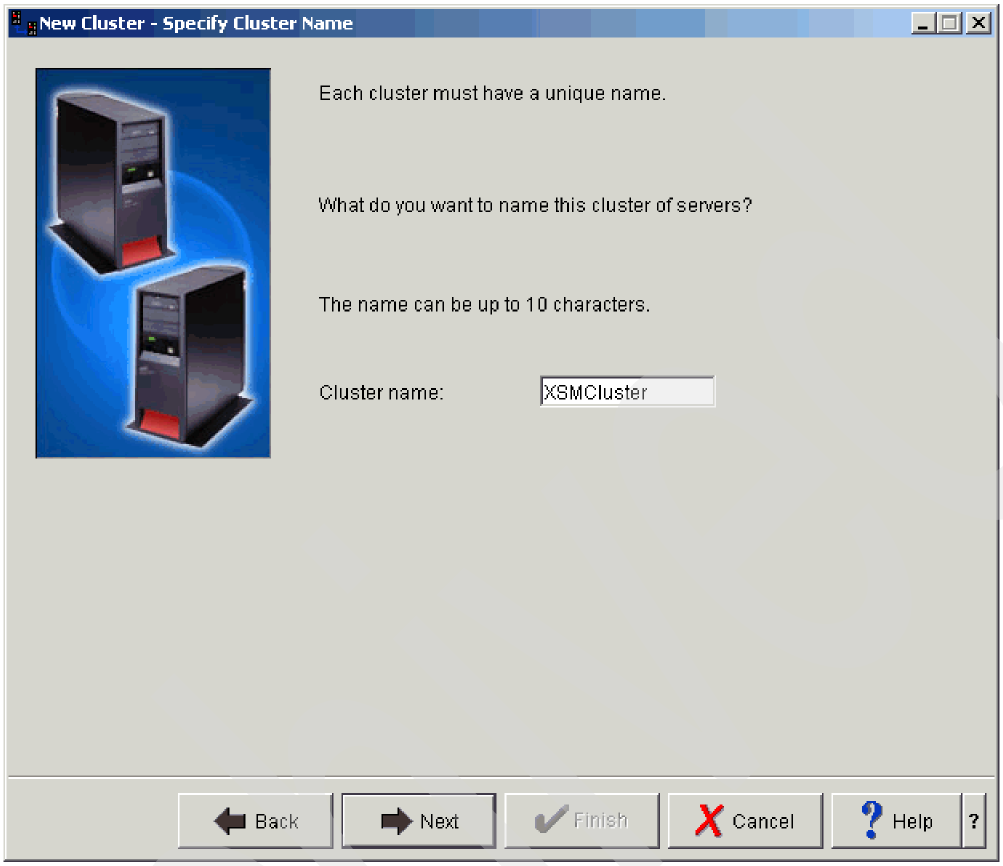
This wizard takes you through the steps to create and start a simple cluster. A simple cluster consists of one or two cluster nodes. A cluster node is any server that is a member of a cluster. When you are creating a simple cluster, the server you are creating the cluster on must be one of the nodes.

Before you can successfully create and start your cluster, you need to ensure your systems and your switchable software are set up correctly. Click Details for more information.

You can cancel at any time by clicking Cancel.

Figure 5-3 New Cluster - Welcome window

5. In the Specify Cluster Name window (Figure 5-4), type a name for the new cluster, and click **Next**.



Each cluster must have a unique name.

What do you want to name this cluster of servers?

The name can be up to 10 characters.

Cluster name: XSMCluster

Figure 5-4 Specify Cluster Name window

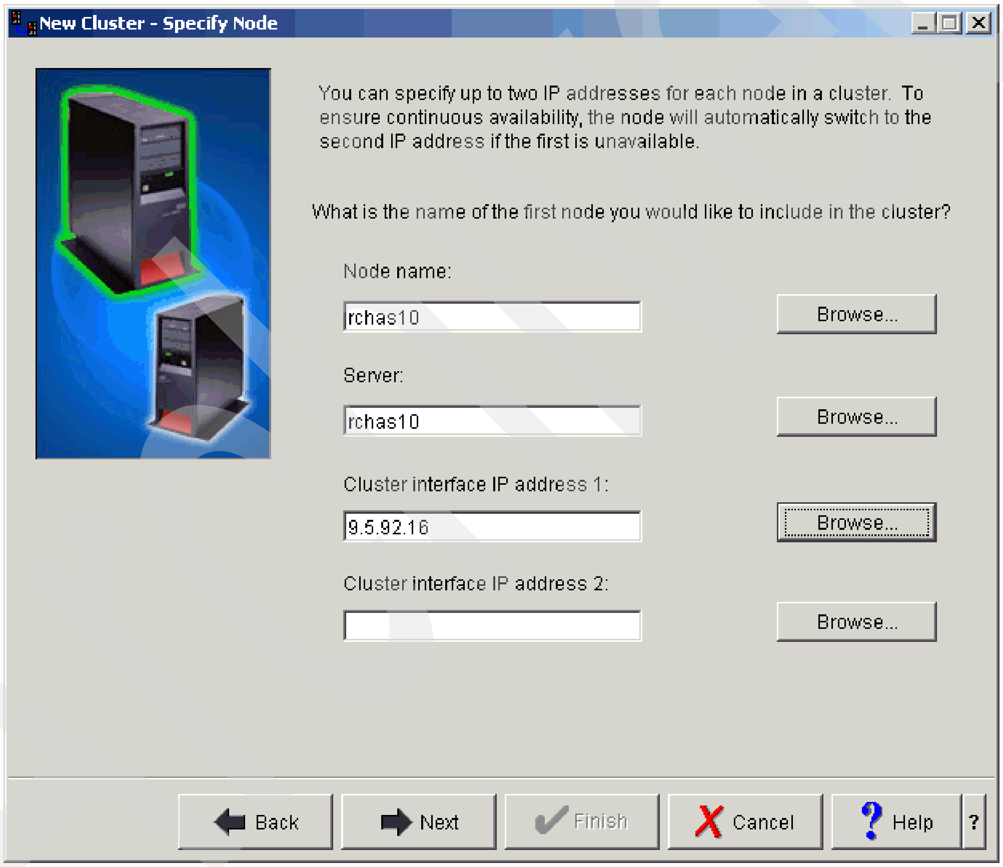
6. In the Specify Node window (Figure 5-5), complete the following tasks:

   a. Enter the information about the server that will become a primary node of the new cluster or click **Browse** to search within the endpoint nodes known by Management Central. Make sure that the server name is known by all systems, that is, the primary node, the backup node, and Management Central system in your cluster configuration via Domain Name System (DNS) or host tables.

      You might prefer to use IP addresses in this field instead. However, the node name is the alias that cluster software uses instead of IP addresses or server names in iSeries Navigator and 5250 panels. The name that you enter for Node name can be the same as your server name or different, but it cannot be an IP address.

   b. Type up to two IP addresses to be used for cluster heartbeat communications. The second interface IP address is optional and is used if the primary address becomes unavailable.

   c. Click **Next**.



You can specify up to two IP addresses for each node in a cluster. To ensure continuous availability, the node will automatically switch to the second IP address if the first is unavailable.

What is the name of the first node you would like to include in the cluster?

Node name: rchas10

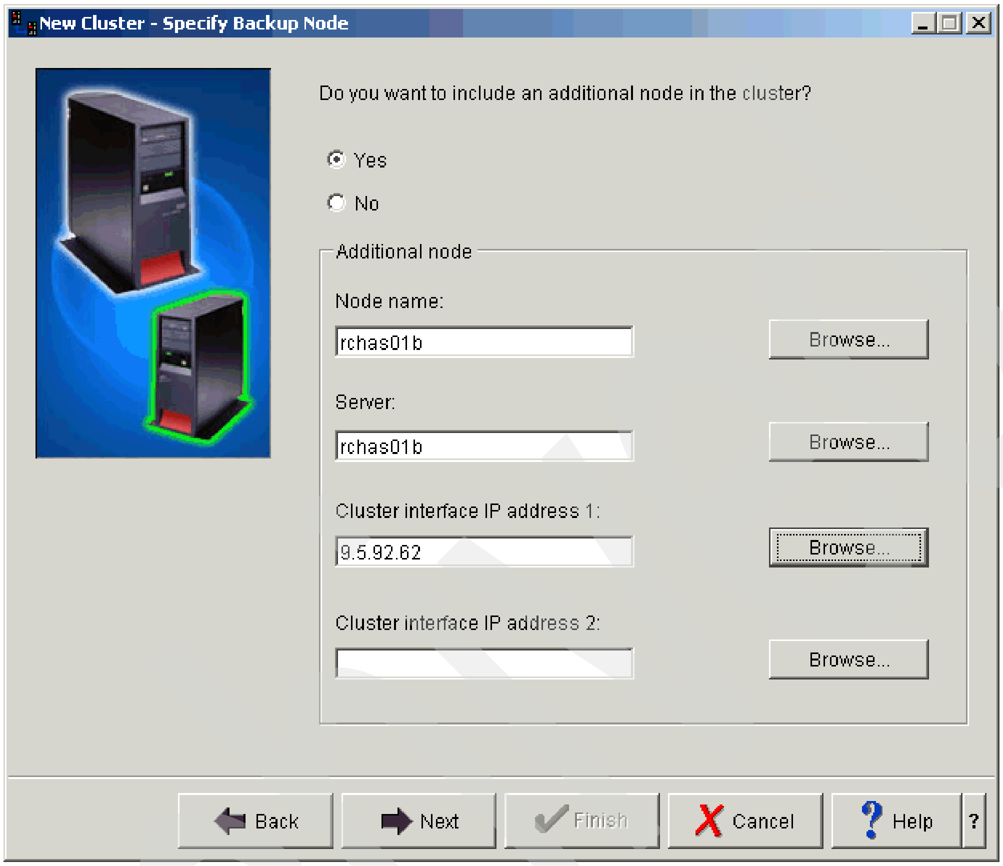
Server: rchas10

Cluster interface IP address 1: 9.5.92.16

Cluster interface IP address 2:

Figure 5-5 Specify Node window

7. In the Specify Backup Node window (Figure 5-6), type the name of the backup node, the server name, and one or two IP addresses for the internal cluster communications. Then click **Next**.



Do you want to include an additional node in the cluster?

● Yes

○ No

Additional node

Node name: rchas01b

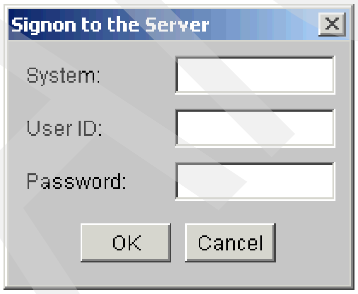
Server: rchas01b

Cluster interface IP address 1: 9.5.92.62

Cluster interface IP address 2:

Figure 5-6 Specify Backup Node window

8. A small window (Figure 5-7) opens in the top left corner, prompting you to sign on to the backup node. This window might open behind other windows if you have many other tasks running at the same time on your computer. Sign on to the backup node, and click **OK**.



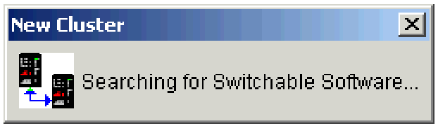
System:

User ID:

Password:

Figure 5-7 Signon to the Server window

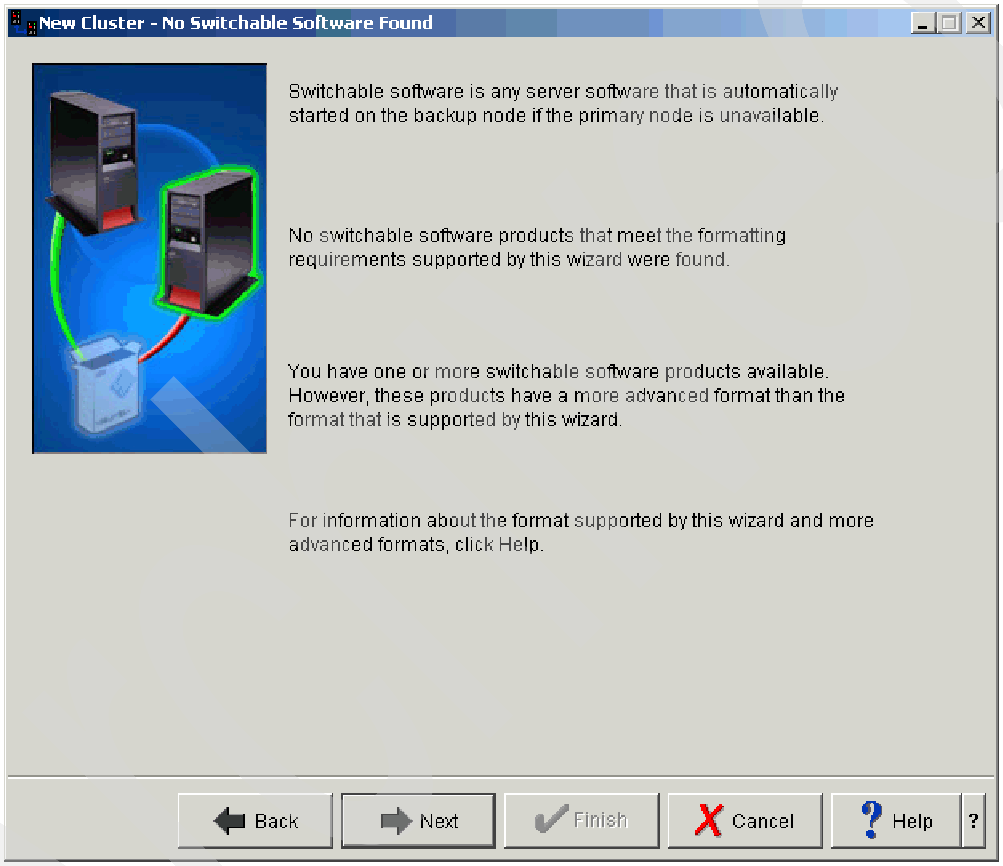
9. The New Cluster wizard searches for switchable software as shown in Figure 5-8.



Searching for Switchable Software...

Figure 5-8 Searching for switchable software

10. If you do not have any ClusterProven® software product installed, the No Switchable Software Found window (Figure 5-9) opens and indicates that no such software is found. Click **Next**.



Switchable software is any server software that is automatically started on the backup node if the primary node is unavailable.

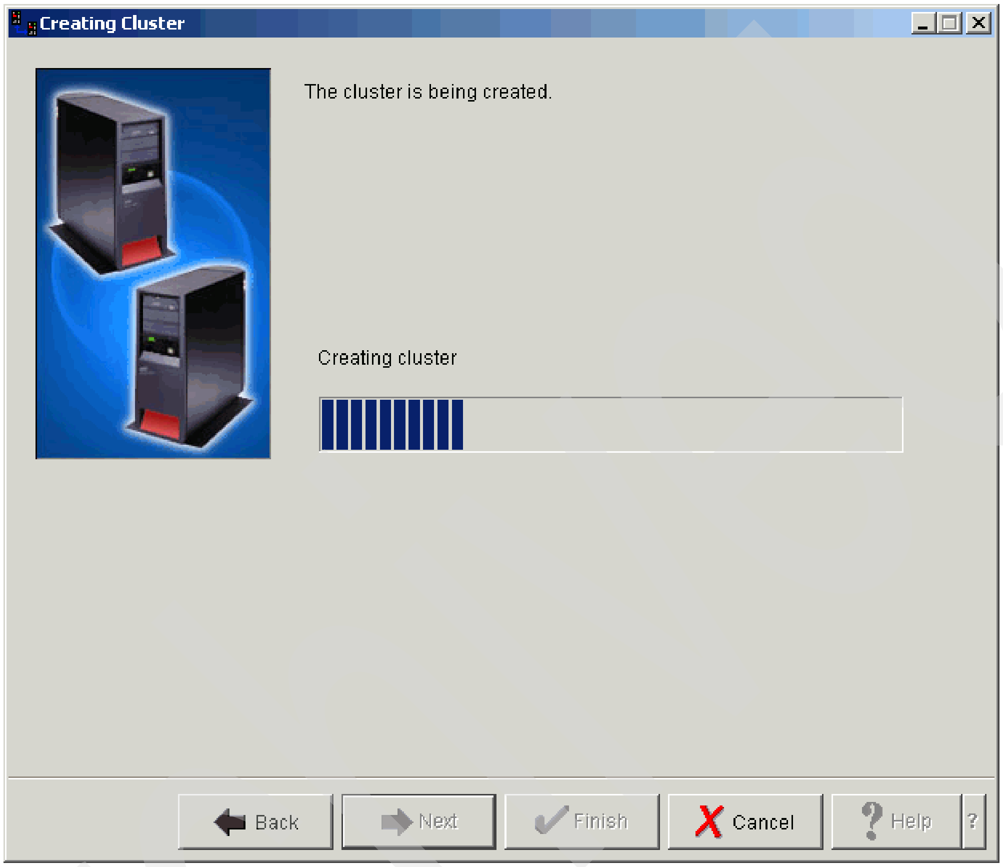
No switchable software products that meet the formatting requirements supported by this wizard were found.

You have one or more switchable software products available. However, these products have a more advanced format than the format that is supported by this wizard.

For information about the format supported by this wizard and more advanced formats, click Help.

Figure 5-9 No Switchable Software Found window

11. When the cluster is being created, you see the Creating Cluster window (Figure 5-10), which indicates the progress of this operation. Usually it takes a couple of minutes to create a cluster. If the wizard stops or errors occur, refer to Chapter 10, “Troubleshooting” on page 175.

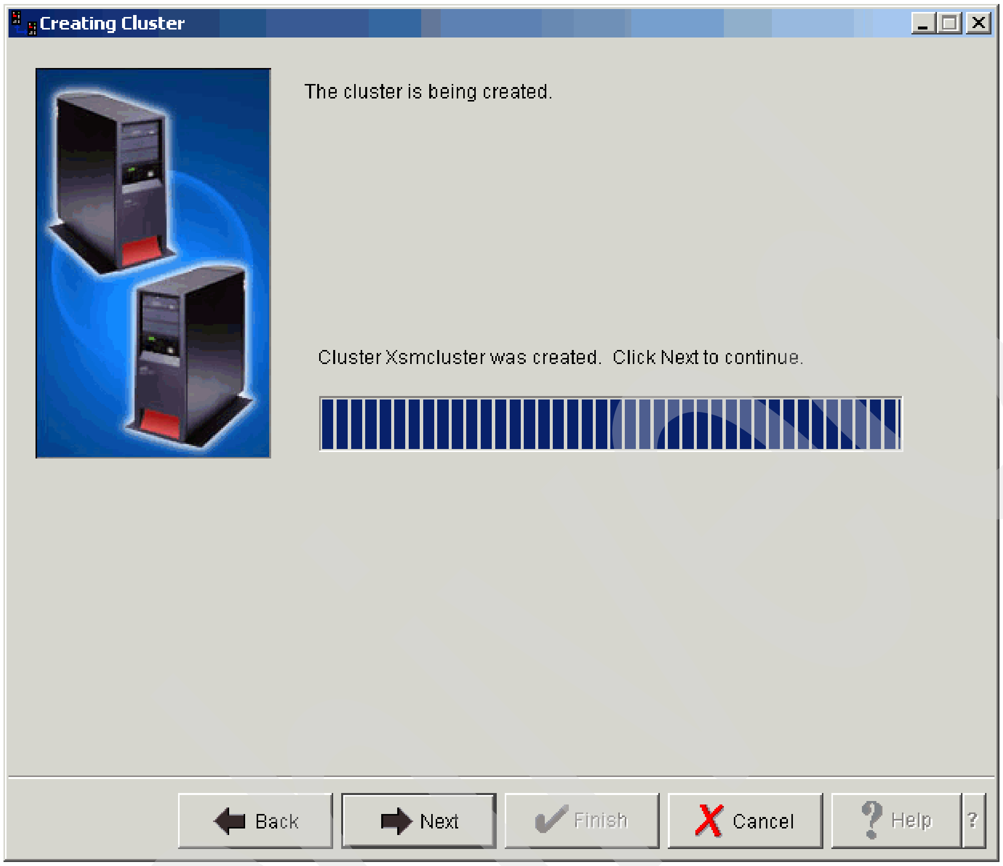


The cluster is being created.

Creating cluster

Figure 5-10 Creating Cluster window

When the cluster is created successfully, the window shows a message indicating the creation of the cluster (Figure 5-11). Click **Next**.

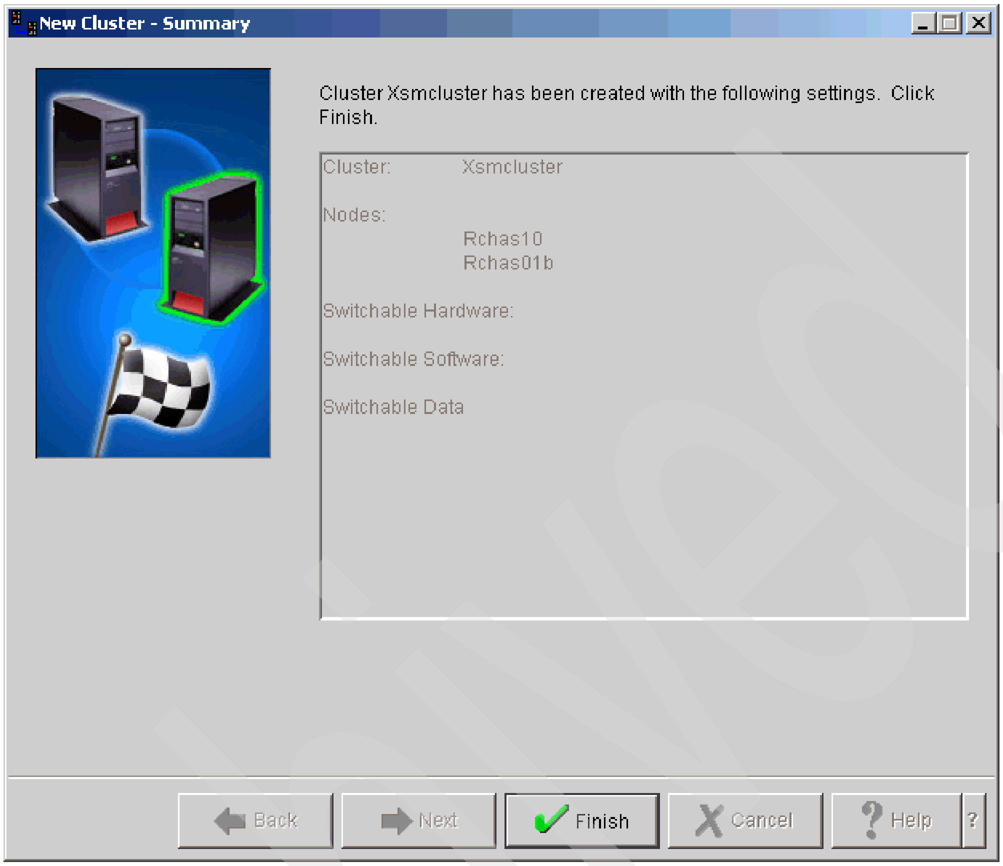


The cluster is being created.

Cluster Xsmcluster was created. Click Next to continue.

Figure 5-11   Cluster created successfully

12.In the Summary window (Figure 5-12), click **Finish**.

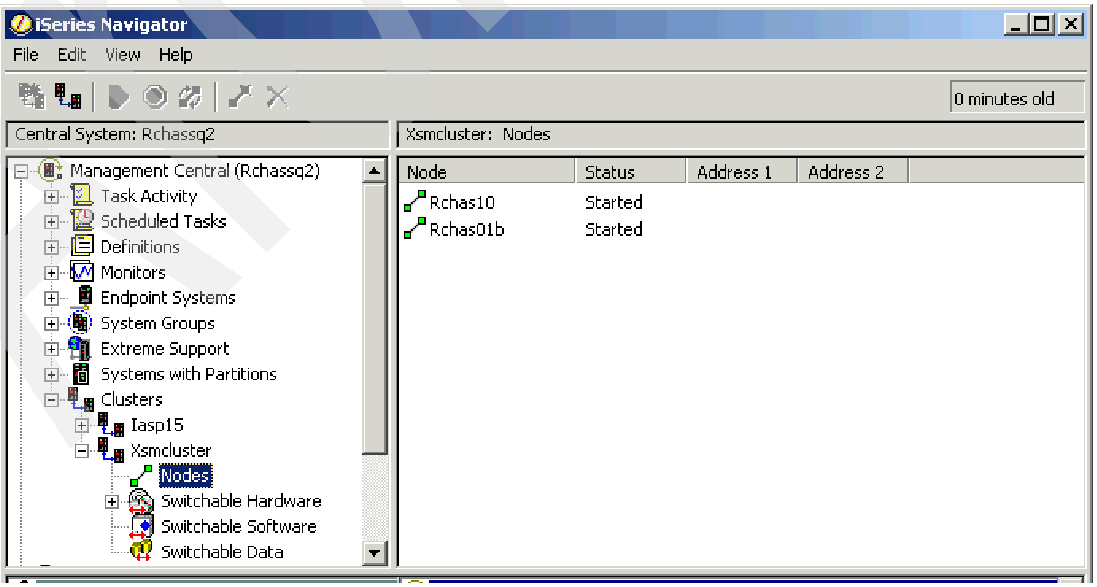


Cluster Xsmcluster has been created with the following settings. Click Finish.

```
Cluster:        Xsmcluster

Nodes:
                Rchas10
                Rchas01b

Switchable Hardware:

Switchable Software:

Switchable Data
```

Figure 5-12 Xsmcluster successfully created

13.In iSeries Navigator, select **Management Central** → **Clusters**. Then you see the name of the newly created cluster as shown in the example in Figure 5-13.



Xsmcluster: Nodes

Node	Status	Address 1	Address 2
Rchas10	Started		
Rchas01b	Started		

Figure 5-13 Cluster XSMCluster on Management Central

Creating a two-node cluster using the 5250 interface

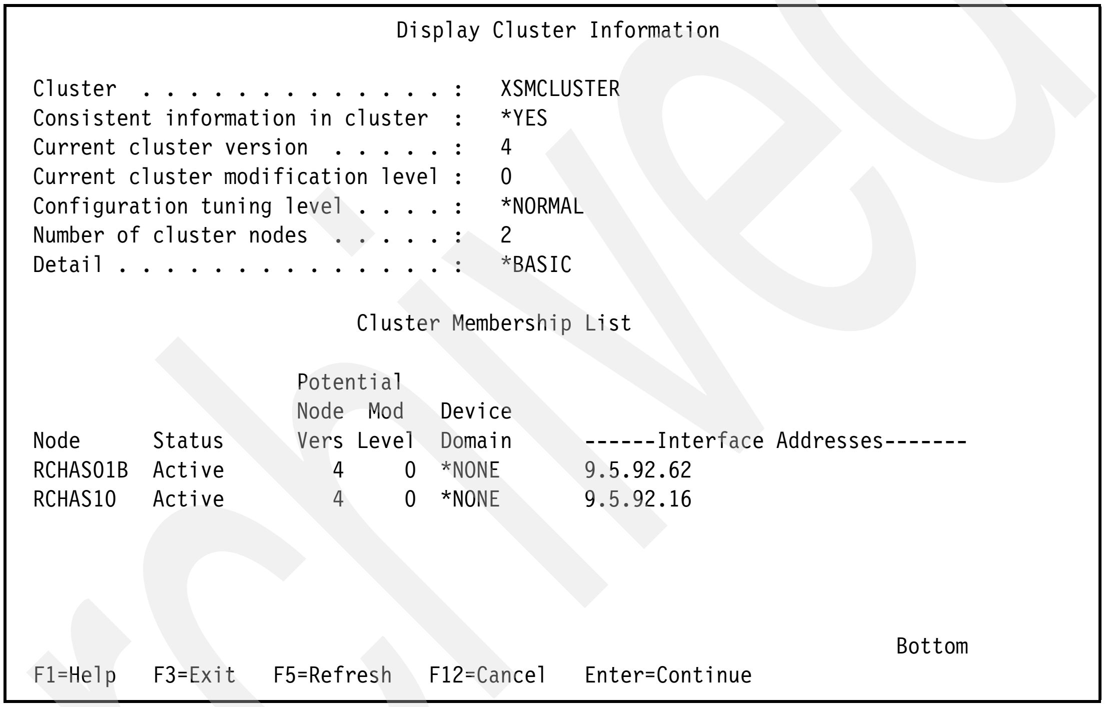
To create a two-node cluster using the 5250 interface:

1. Log in to the primary node and type the following commands:

```
CRTCLU CLUSTER(XSMCLUSTER) NODE((RCHAS01B ('9.5.92.62')) (RCHAS10 ('9.5.92.16')))
STRCLUNOD CLUSTER(XSMCLUSTER) NODE(RCHAS01B)
STRCLUNOD CLUSTER(XSMCLUSTER) NODE(RCHAS10)
ADDDEVDMNE CLUSTER(XSMCLUSTER) DEVDMN(XSMCLUSTER) NODE(RCHAS01B)
ADDDEVDMNE CLUSTER(XSMCLUSTER) DEVDMN(XSMCLUSTER) NODE(RCHAS10)
```

   After these commands are complete, your cluster is up and running.

2. Enter the DSPCLUINF command to check the status of your cluster as shown in Figure 5-14.



```
                              Display Cluster Information

Cluster  . . . . . . . . . . . . . :   XSMCLUSTER
Consistent information in cluster  :   *YES
Current cluster version  . . . . . :   4
Current cluster modification level :   0
Configuration tuning level . . . . :   *NORMAL
Number of cluster nodes  . . . . . :   2
Detail . . . . . . . . . . . . . . :   *BASIC

                           Cluster Membership List

                      Potential
                      Node  Mod   Device
Node      Status      Vers Level  Domain      ------Interface Addresses-------
RCHAS01B  Active         4     0  *NONE       9.5.92.62
RCHAS10   Active         4     0  *NONE       9.5.92.16




                                                                    Bottom
F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 5-14 Display Cluster Information panel

3. After the cluster is created through the command line interface, you are unable to see it in Management Central. To add an existing cluster to the Clusters applet in Management Central, right-click **Clusters** and select **Add Existing Cluster**, as shown in Figure 5-15.

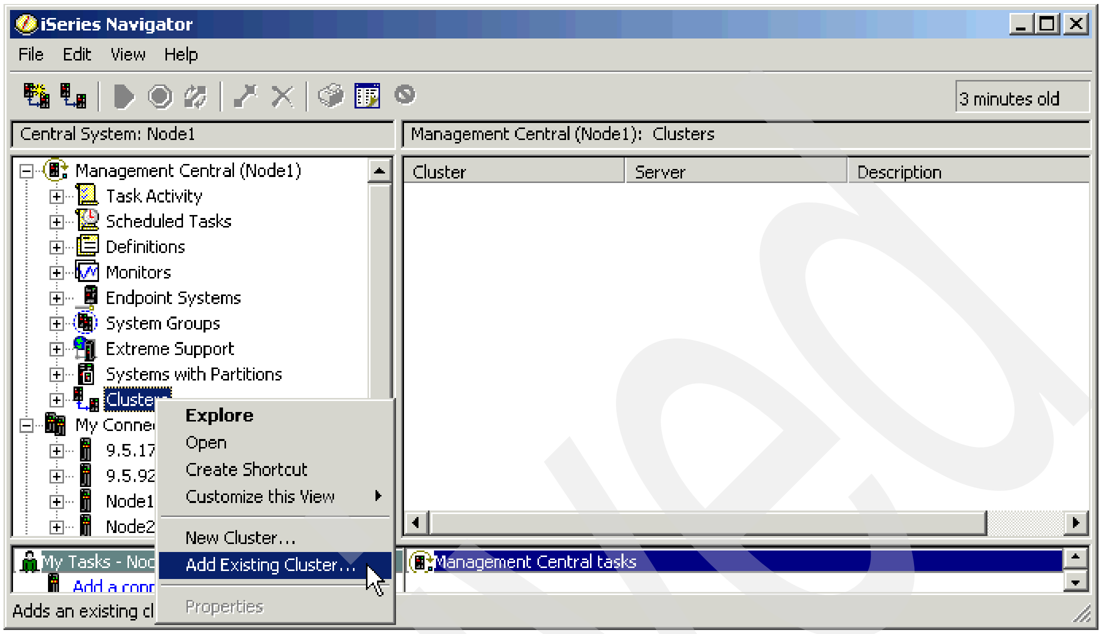
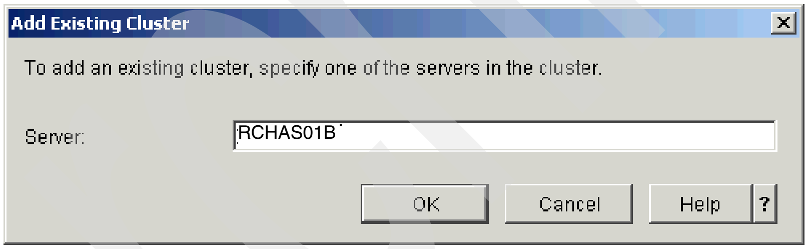


Figure 5-15 Selecting the Add Existing Cluster option

4. In the Add Existing Cluster window (Figure 5-16), enter the name of one of the nodes and click **OK**.


Add Existing Cluster

To add an existing cluster, specify one of the servers in the cluster.

Server: RCHAS01B

Figure 5-16 Add Existing Cluster window

Now you can manage your cluster via the graphical interface provided by Management Central functions.

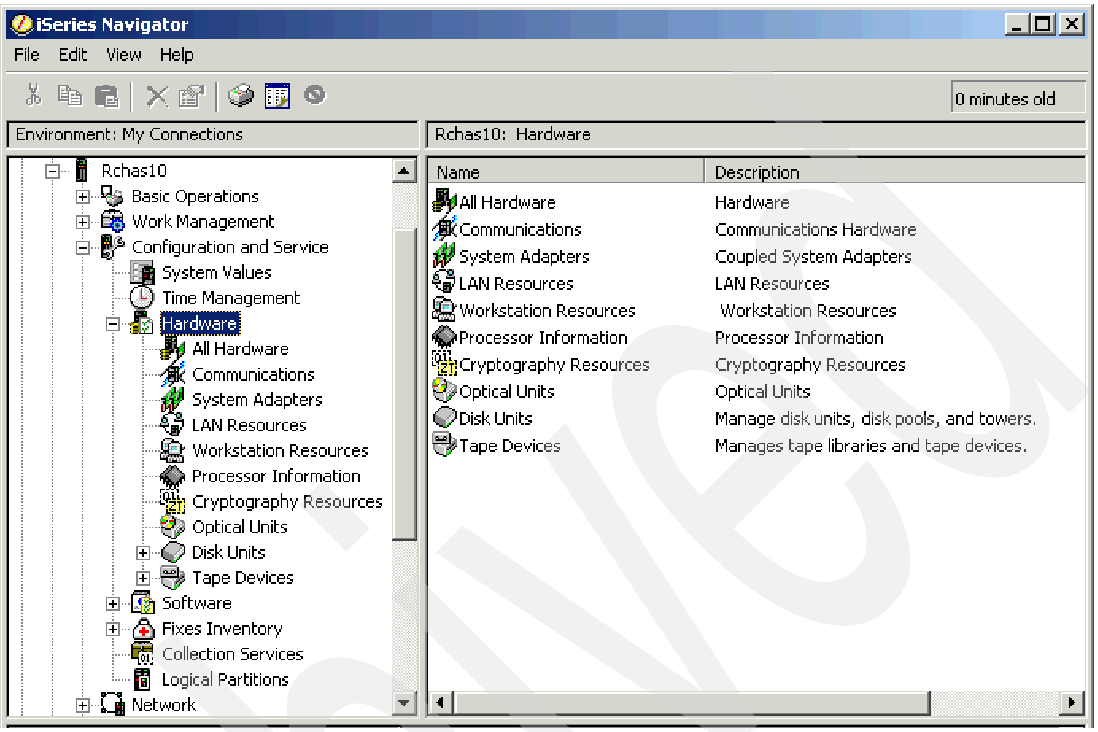
5.2.2 Creating a stand-alone IASP

In this section, we explain how to create a stand-alone non-switchable IASP. Before you create an IASP, make sure that you have enough disk space on both nodes. When you configure the IASP, the total size of the IASP on the second node *must not* be less than the size of the IASP that you configure in this chapter.

Important: You cannot create an IASP by using the command line interface. Use iSeries Navigator for this task instead.

To create a non-switchable IASP using iSeries Navigator:

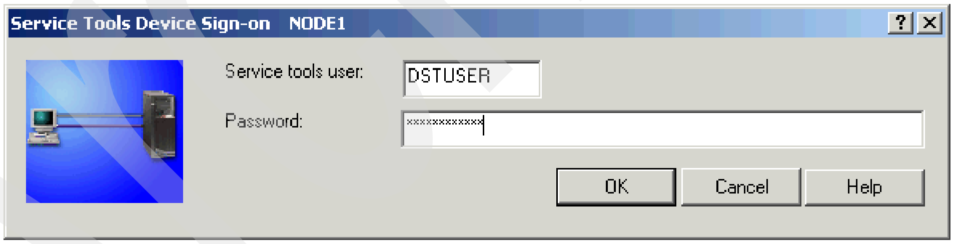
1. Expand **My Connections** → **Primary Resource** → **Configuration and Service** → **Hardware**, as shown in Figure 5-17. In the right pane, click **Disk Units**.



Name	Description
All Hardware	Hardware
Communications	Communications Hardware
System Adapters	Coupled System Adapters
LAN Resources	LAN Resources
Workstation Resources	Workstation Resources
Processor Information	Processor Information
Cryptography Resources	Cryptography Resources
Optical Units	Optical Units
Disk Units	Manage disk units, disk pools, and towers.
Tape Devices	Manages tape libraries and tape devices.

Figure 5-17 iSeries Navigator with Configuration and Service section expanded

2. Sign on to Service Tools (Figure 5-18) with a valid DST user name and password. Click **OK**.



Service tools user: DSTUSER

Password: ************

Figure 5-18 Service Tools Device Sign-on window

3. After you sign on to Service Tools, you can manage disk drives in iSeries Navigator. You *must* use *non-configured* drives to create the stand-alone IASP. If the cluster node has several non-configured drives and you do not want to use all of them for the geographically mirrored IASP, the easiest way to find desired disks is to select **Disk Units → By Location** in iSeries Navigator, as shown in Figure 5-19. Write down the disk unit names.

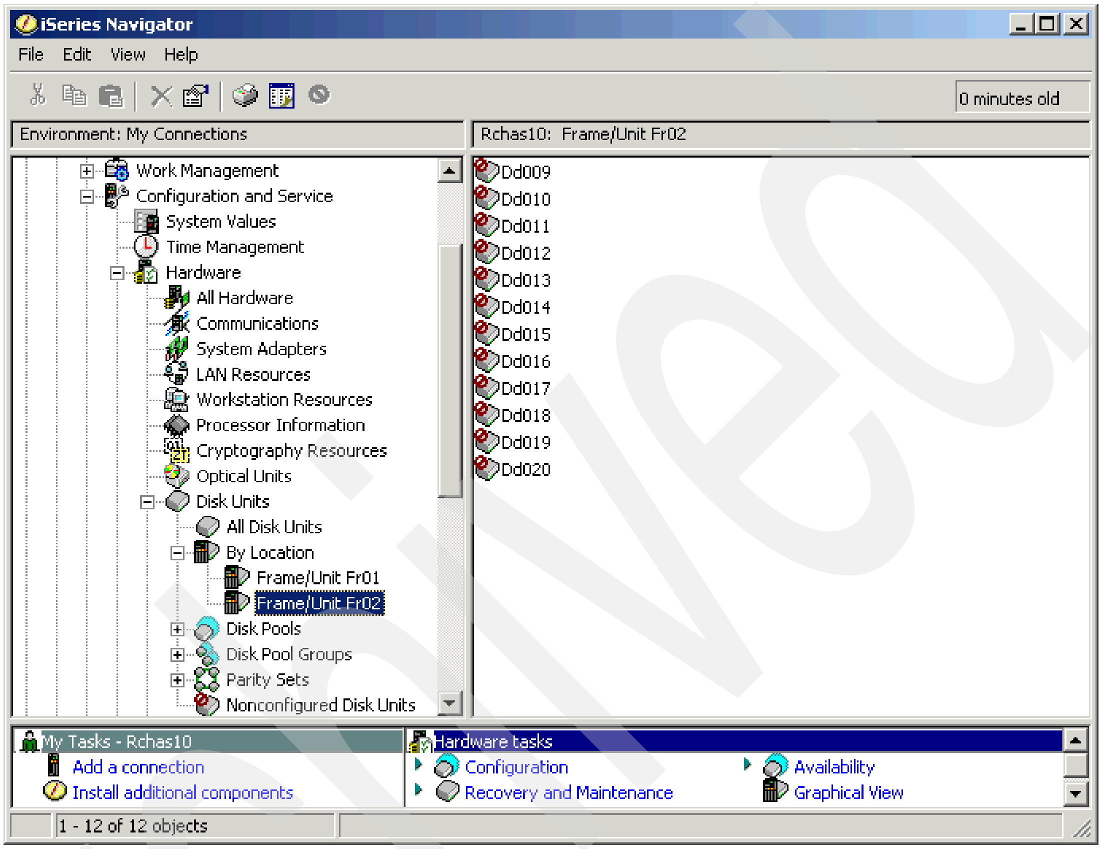
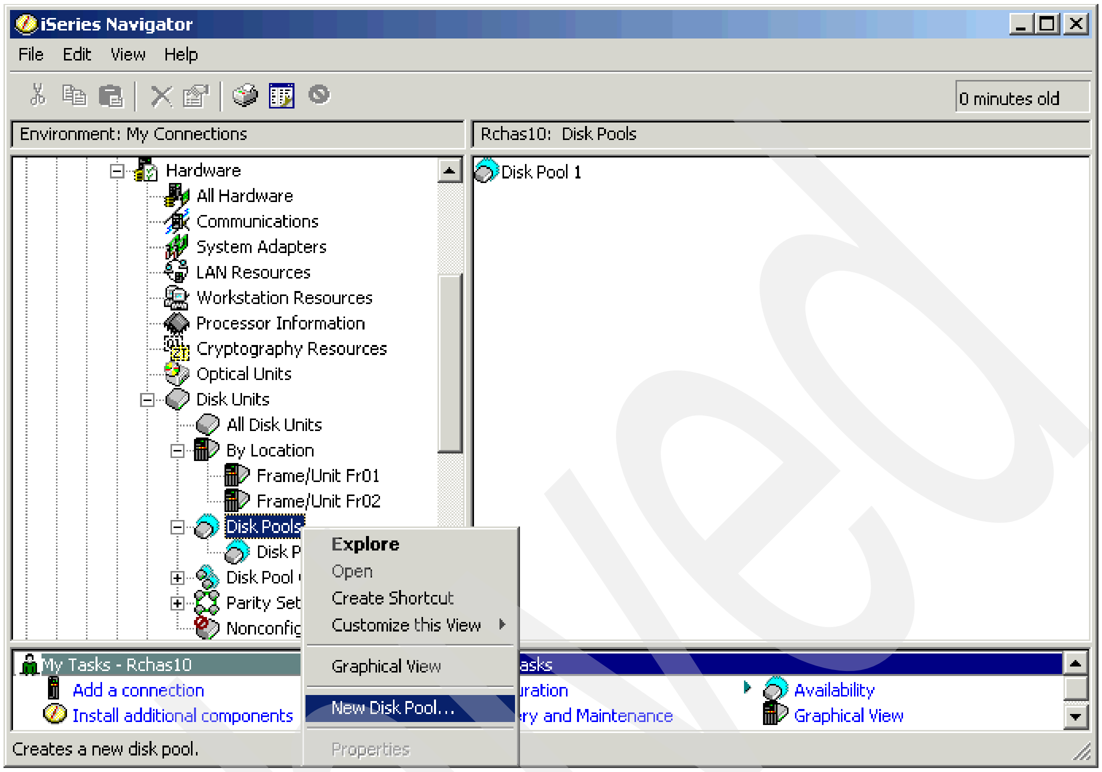


Figure 5-19 Disks Units → By Location

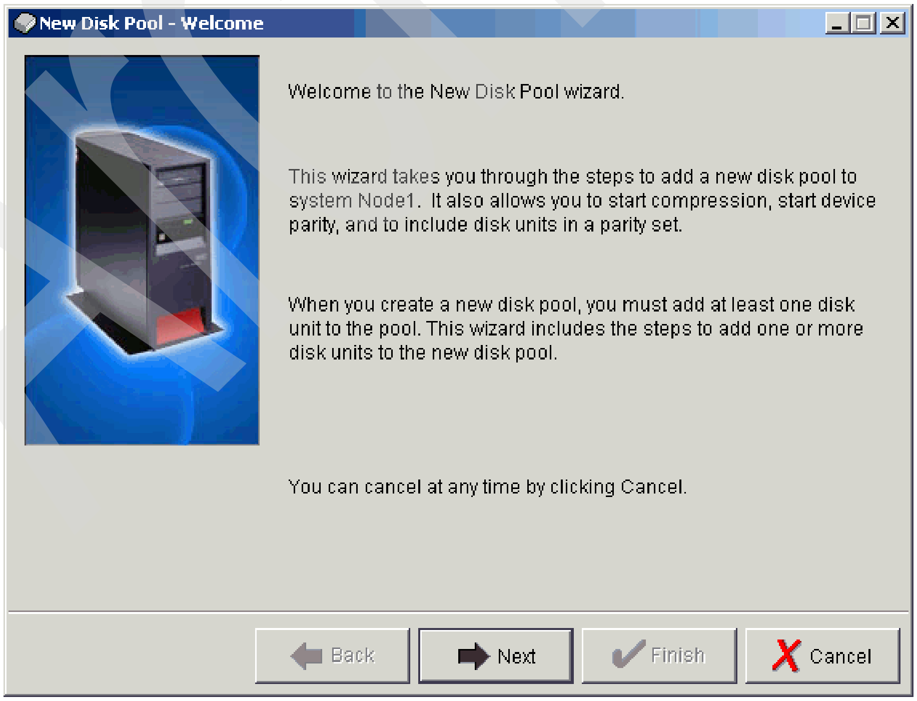
4. Select **Disk Units** → **Disk Pools,** right-click **Disk Pools,** and select **New Disk Pool**, as shown in Figure 5-20, to launch the New Disk Pool wizard.



Creates a new disk pool.

Figure 5-20 Launching the New Disk Pool wizard

5. In the New Disk Pool - Welcome window (Figure 5-21), click **Next**.



Welcome to the New Disk Pool wizard.

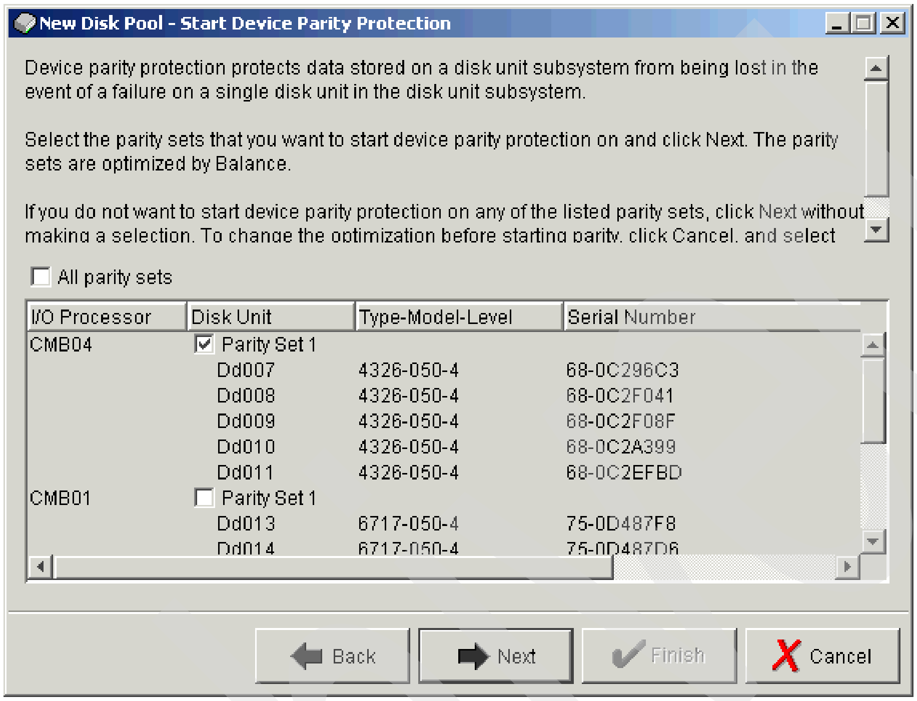
This wizard takes you through the steps to add a new disk pool to system Node1. It also allows you to start compression, start device parity, and to include disk units in a parity set.

When you create a new disk pool, you must add at least one disk unit to the pool. This wizard includes the steps to add one or more disk units to the new disk pool.

You can cancel at any time by clicking Cancel.

Figure 5-21 Welcome window for the New Disk Pool wizard

6. We recommend that you start device parity protection when creating a new disk pool. In the New Disk Pool - Start Device Parity Protection window (Figure 5-22), select the parity set that contains the disk drives that you selected in step 3 on page 92. The device parity protection task does not start until you complete all the steps in the wizard. Click **Next**.



Device parity protection protects data stored on a disk unit subsystem from being lost in the event of a failure on a single disk unit in the disk unit subsystem.

Select the parity sets that you want to start device parity protection on and click Next. The parity sets are optimized by Balance.

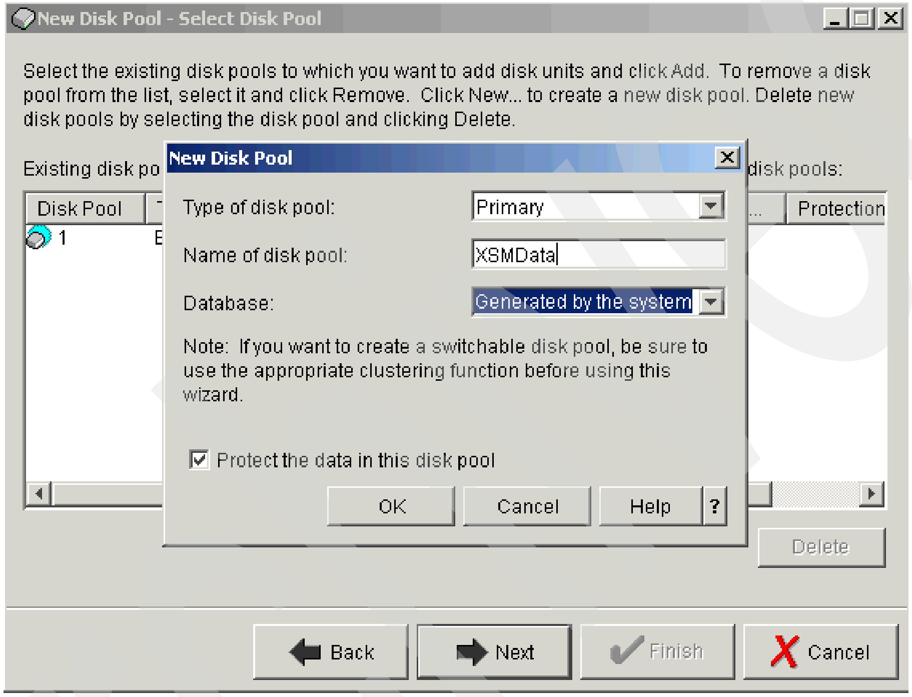
If you do not want to start device parity protection on any of the listed parity sets, click Next without making a selection. To change the optimization before starting parity, click Cancel, and select

□ All parity sets

I/O Processor	Disk Unit	Type-Model-Level	Serial Number
CMB04	☑ Parity Set 1		
	Dd007	4326-050-4	68-0C296C3
	Dd008	4326-050-4	68-0C2F041
	Dd009	4326-050-4	68-0C2F08F
	Dd010	4326-050-4	68-0C2A399
	Dd011	4326-050-4	68-0C2EFBD
CMB01	□ Parity Set 1		
	Dd013	6717-050-4	75-0D487F8
	Dd014	6717-050-4	75-0D487D6

Figure 5-22 Start Device Parity Protection window

7. Set the configuration parameters for the new disk pool:
   a. In the New Disk Pool window (inset in Figure 5-23), type the new disk pool name.
   b. Ensure that you create a **Primary** disk pool.
   c. You can use the default value for the database name, which is **Generated by the system**. In this case, the database name is the same as a disk pool name. Alternately, you can type over the default value to give a different name to the database that is associated with the new IASP.
   d. Select **Protect the data in this disk pool**.
   e. Click **OK**.



New Disk Pool - Select Disk Pool

Select the existing disk pools to which you want to add disk units and click Add. To remove a disk pool from the list, select it and click Remove. Click New... to create a new disk pool. Delete new disk pools by selecting the disk pool and clicking Delete.

New Disk Pool

Type of disk pool: Primary

Name of disk pool: XSMData

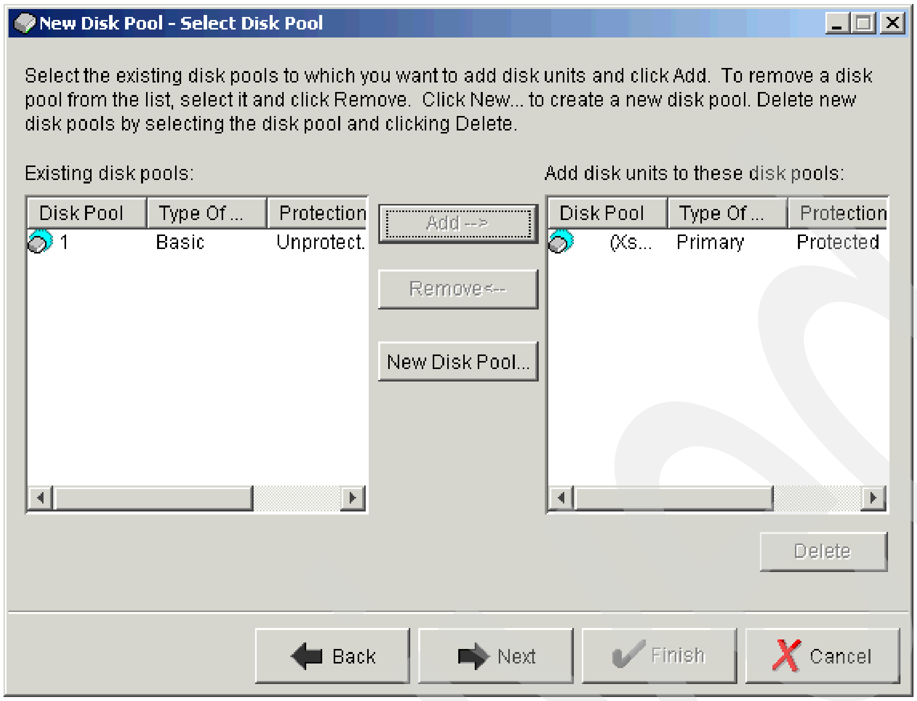
Database: Generated by the system

Note: If you want to create a switchable disk pool, be sure to use the appropriate clustering function before using this wizard.

☑ Protect the data in this disk pool

Figure 5-23 New Disk Pool settings window

f. Click **Next** to confirm your selection of the new disk pool (see Figure 5-24).



New Disk Pool - Select Disk Pool

Select the existing disk pools to which you want to add disk units and click Add. To remove a disk pool from the list, select it and click Remove. Click New... to create a new disk pool. Delete new disk pools by selecting the disk pool and clicking Delete.

Existing disk pools:

Disk Pool	Type Of ...	Protection
1	Basic	Unprotect.

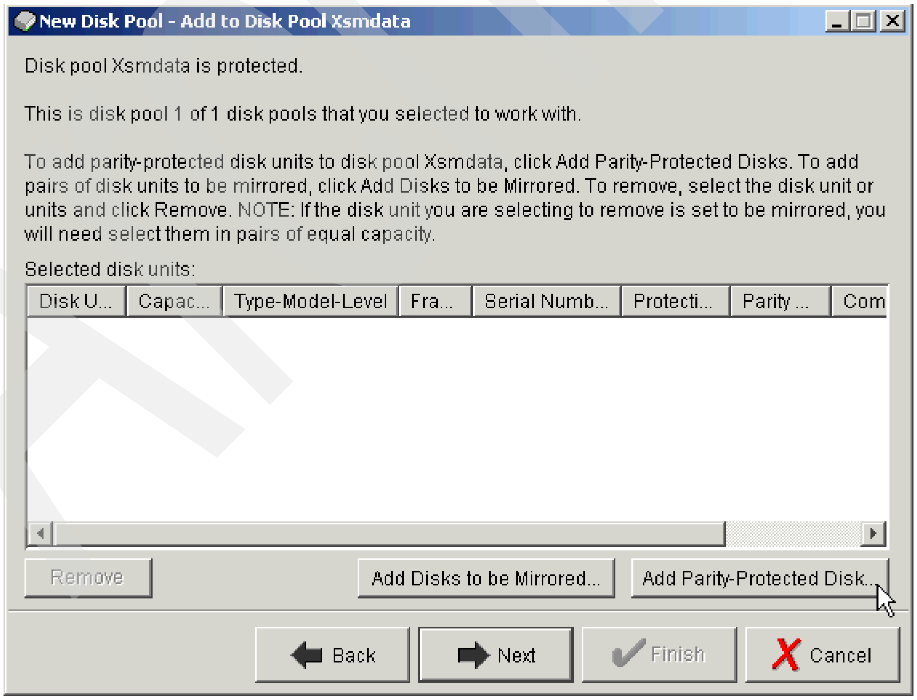
Add disk units to these disk pools:

Disk Pool	Type Of ...	Protection
(Xs...	Primary	Protected

Figure 5-24 Select Disk Pool window

8. In the Add to Disk Pool window (Figure 5-25), complete these steps:

   a. Click **Add Parity-Protected Disks**.



New Disk Pool - Add to Disk Pool Xsmdata

Disk pool Xsmdata is protected.

This is disk pool 1 of 1 disk pools that you selected to work with.

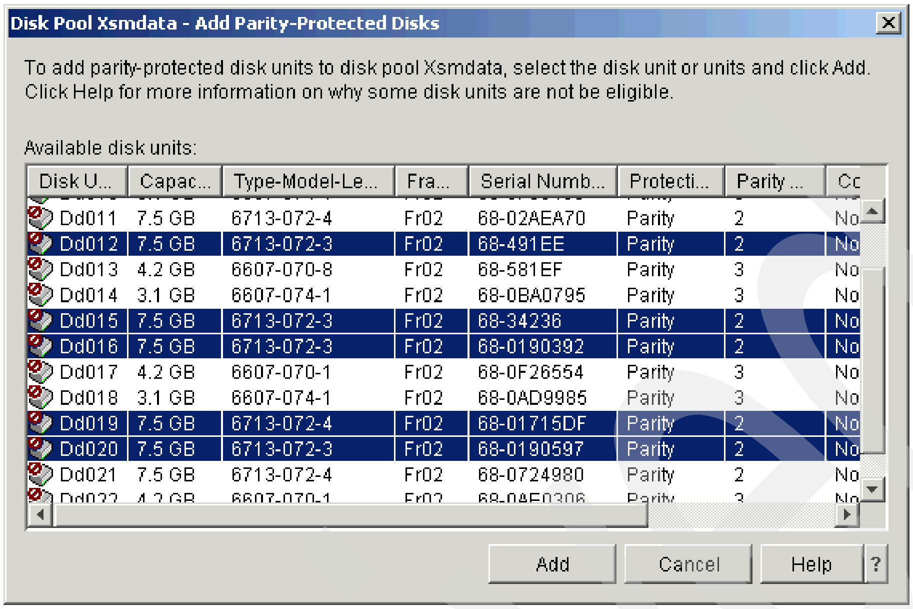
To add parity-protected disk units to disk pool Xsmdata, click Add Parity-Protected Disks. To add pairs of disk units to be mirrored, click Add Disks to be Mirrored. To remove, select the disk unit or units and click Remove. NOTE: If the disk unit you are selecting to remove is set to be mirrored, you will need select them in pairs of equal capacity.

Selected disk units:

Disk U...	Capac...	Type-Model-Level	Fra...	Serial Numb...	Protecti...	Parity ...	Com

Figure 5-25 Add to Disk Pool window

b. In the Add Parity-Protected Disks window (Figure 5-26), select the disks that you chose in step 3 on page 92 and click **Add**.



Disk Pool Xsmdata - Add Parity-Protected Disks

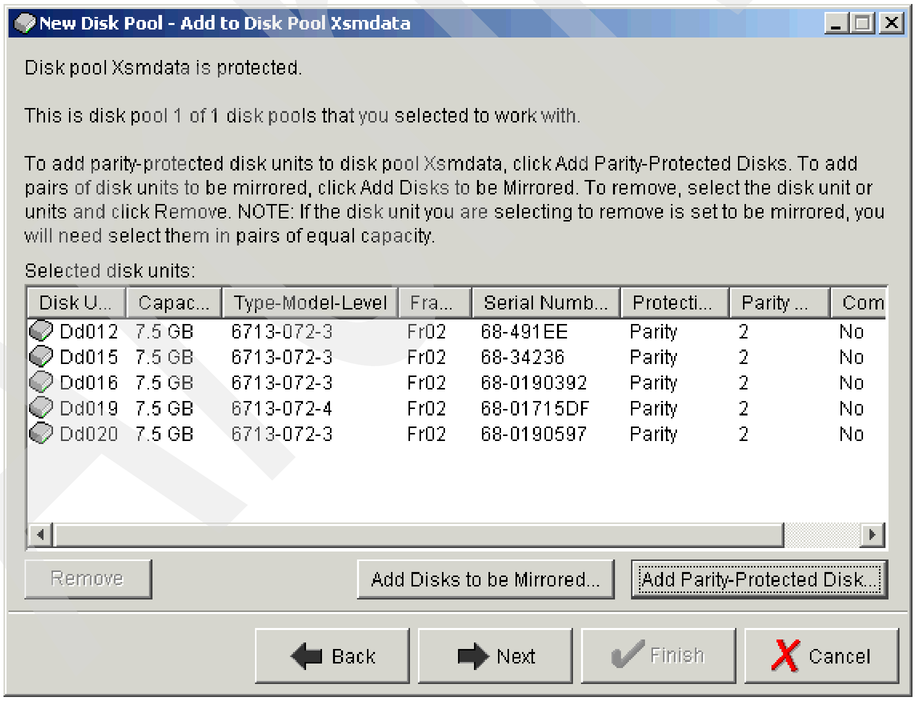
To add parity-protected disk units to disk pool Xsmdata, select the disk unit or units and click Add. Click Help for more information on why some disk units are not be eligible.

Available disk units:

Disk U...	Capac...	Type-Model-Le...	Fra...	Serial Numb...	Protecti...	Parity ...	Cc
Dd011	7.5 GB	6713-072-4	Fr02	68-02AEA70	Parity	2	No
Dd012	7.5 GB	6713-072-3	Fr02	68-491EE	Parity	2	No
Dd013	4.2 GB	6607-070-8	Fr02	68-581EF	Parity	3	No
Dd014	3.1 GB	6607-074-1	Fr02	68-0BA0795	Parity	3	No
Dd015	7.5 GB	6713-072-3	Fr02	68-34236	Parity	2	No
Dd016	7.5 GB	6713-072-3	Fr02	68-0190392	Parity	2	No
Dd017	4.2 GB	6607-070-1	Fr02	68-0F26554	Parity	3	No
Dd018	3.1 GB	6607-074-1	Fr02	68-0AD9985	Parity	3	No
Dd019	7.5 GB	6713-072-4	Fr02	68-01715DF	Parity	2	No
Dd020	7.5 GB	6713-072-3	Fr02	68-0190597	Parity	2	No
Dd021	7.5 GB	6713-072-4	Fr02	68-0724980	Parity	2	No
Dd022	4.2 GB	6607-070-1	Fr02	68-0AE0306	Parity	3	No

Figure 5-26 Add Parity-Protected Disks window

c. When you return to the Add to Disk Pool window (Figure 5-27), click **Next** to confirm your selection.



New Disk Pool - Add to Disk Pool Xsmdata

Disk pool Xsmdata is protected.

This is disk pool 1 of 1 disk pools that you selected to work with.

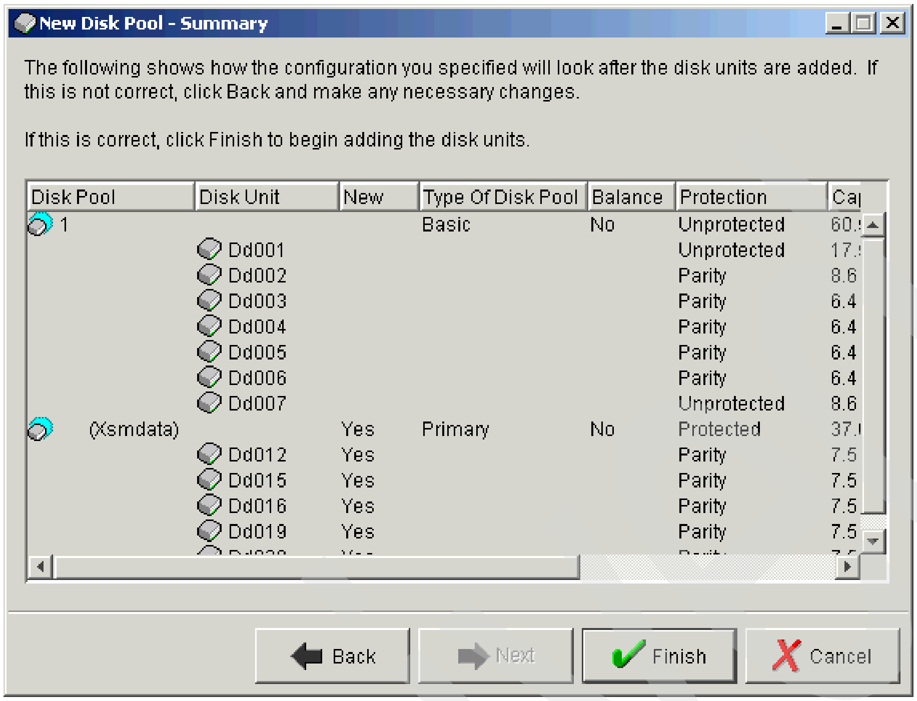
To add parity-protected disk units to disk pool Xsmdata, click Add Parity-Protected Disks. To add pairs of disk units to be mirrored, click Add Disks to be Mirrored. To remove, select the disk unit or units and click Remove. NOTE: If the disk unit you are selecting to remove is set to be mirrored, you will need select them in pairs of equal capacity.

Selected disk units:

Disk U...	Capac...	Type-Model-Level	Fra...	Serial Numb...	Protecti...	Parity ...	Com
Dd012	7.5 GB	6713-072-3	Fr02	68-491EE	Parity	2	No
Dd015	7.5 GB	6713-072-3	Fr02	68-34236	Parity	2	No
Dd016	7.5 GB	6713-072-3	Fr02	68-0190392	Parity	2	No
Dd019	7.5 GB	6713-072-4	Fr02	68-01715DF	Parity	2	No
Dd020	7.5 GB	6713-072-3	Fr02	68-0190597	Parity	2	No

Figure 5-27 Add to Disk Pool window with disks selected

9. In the Summary window (Figure 5-28), click **Finish**.



New Disk Pool - Summary

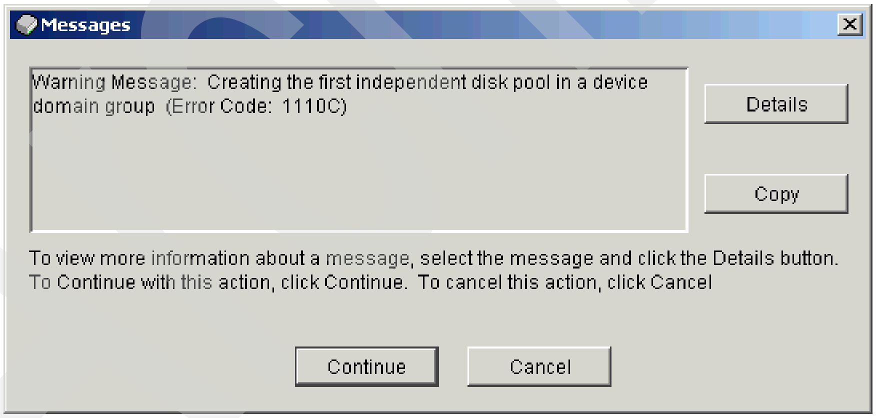
The following shows how the configuration you specified will look after the disk units are added. If this is not correct, click Back and make any necessary changes.

If this is correct, click Finish to begin adding the disk units.

Disk Pool	Disk Unit	New	Type Of Disk Pool	Balance	Protection	Ca
1			Basic	No	Unprotected	60.
	Dd001				Unprotected	17.
	Dd002				Parity	8.6
	Dd003				Parity	6.4
	Dd004				Parity	6.4
	Dd005				Parity	6.4
	Dd006				Parity	6.4
	Dd007				Unprotected	8.6
(Xsmdata)		Yes	Primary	No	Protected	37.
	Dd012	Yes			Parity	7.5
	Dd015	Yes			Parity	7.5
	Dd016	Yes			Parity	7.5
	Dd019	Yes			Parity	7.5

Figure 5-28 New Disk Pool - Summary window

10. The New Disk Pool wizard might issue warnings before it starts to create the IASP. Observe every warning thoroughly. Figure 5-29 shows an information message about creating the first independent disk pool in a device domain group. Click **Continue**.



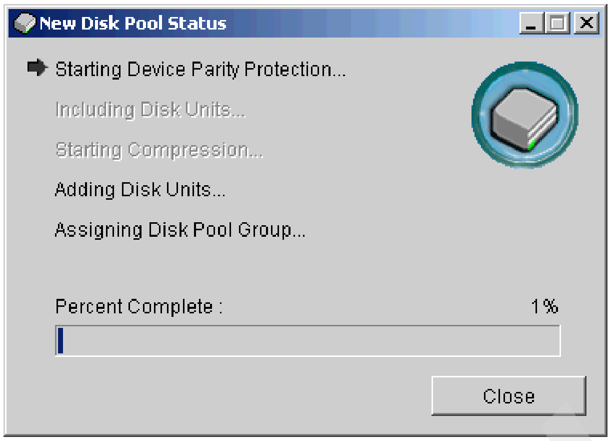
Messages

Warning Message: Creating the first independent disk pool in a device domain group (Error Code: 1110C)

To view more information about a message, select the message and click the Details button. To Continue with this action, click Continue. To cancel this action, click Cancel

Figure 5-29 Warning message

11. A status window (Figure 5-30) is displayed that indicates the progress of the operations that you have performed. If you click **Close** in this window, the status window closes, but the wizard continues its work. It prevents you from using the function Work with disk units in Service Tools or iSeries Navigator.


New Disk Pool Status

Starting Device Parity Protection...

Including Disk Units...

Starting Compression...

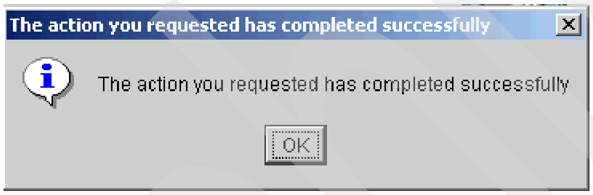
Adding Disk Units...

Assigning Disk Pool Group...

Percent Complete : 1%

Figure 5-30 New Disk Pool Status window

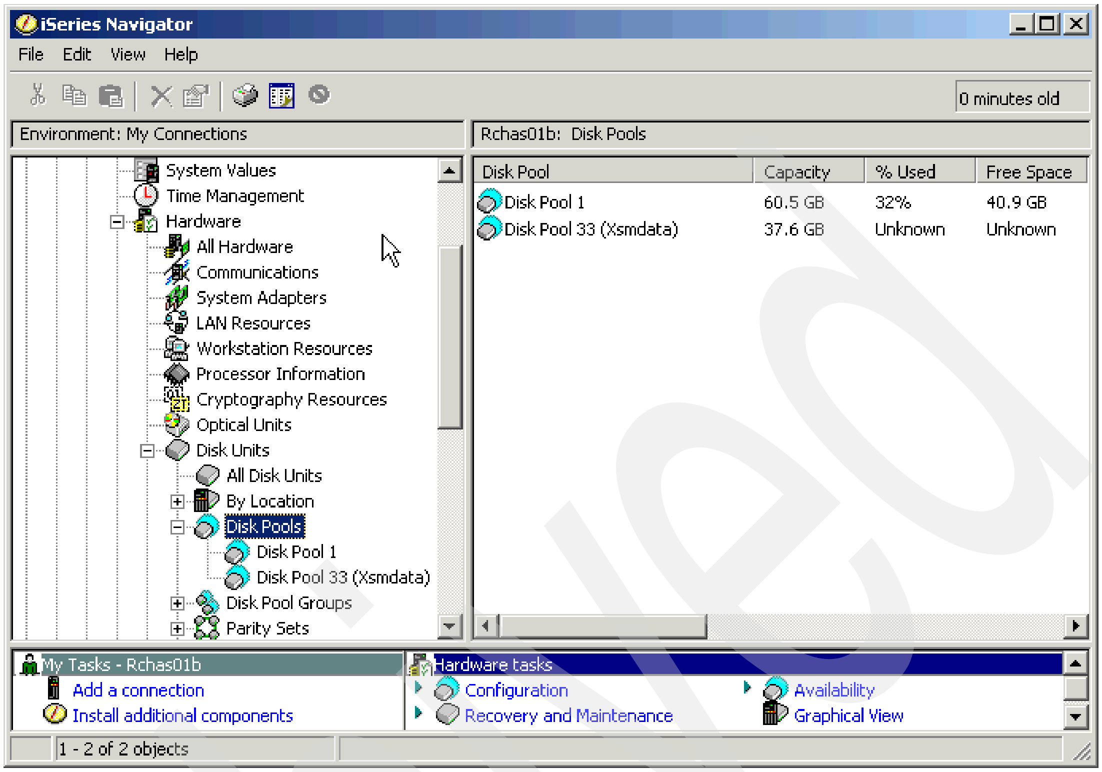
12. After all the wizard's tasks are completed, in the message window (Figure 5-31), click **OK** to confirm the successful creation of the IASP.


The action you requested has completed successfully

The action you requested has completed successfully

Figure 5-31 New disk pool created successfully

Now you can manage this new IASP in iSeries Navigator, as shown in Figure 5-32.



Rchas01b: Disk Pools

Disk Pool	Capacity	% Used	Free Space
Disk Pool 1	60.5 GB	32%	40.9 GB
Disk Pool 33 (Xsmdata)	37.6 GB	Unknown	Unknown

1 - 2 of 2 objects

Figure 5-32 New disk pool in iSeries Navigator

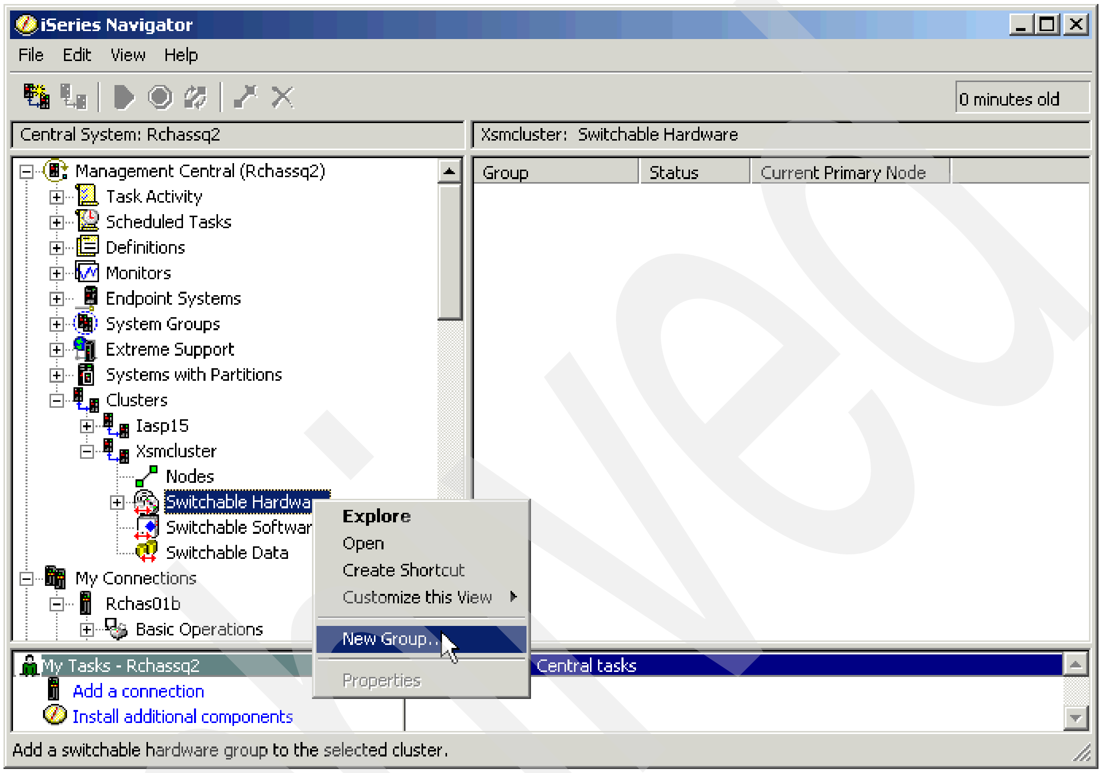
5.2.3 Creating a new device CRG for cross-site mirroring

You have to define a device CRG for the independent disk pool using iSeries Navigator or a 5250 terminal.

Creating a device CRG using iSeries Navigator

To create a device CRG using iSeries Navigator:

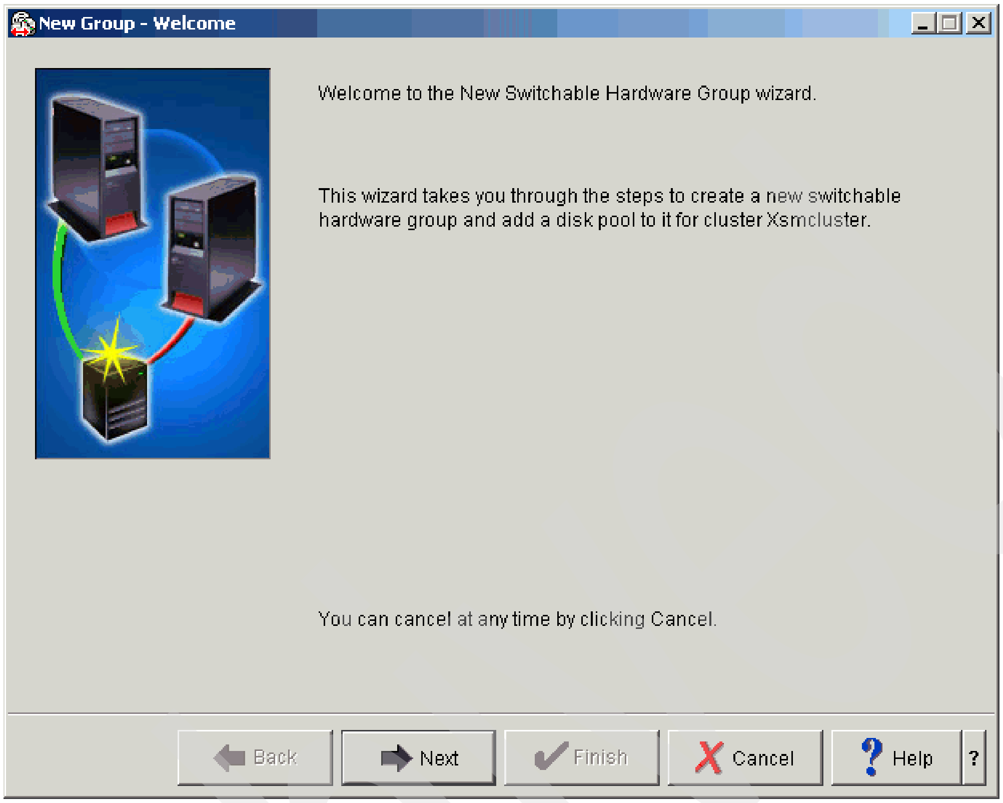
1. Expand your cluster environment. Select **Management Central** → **Clusters** → **Switchable Hardware**. Right-click **Switchable Hardware** and select **New Group** as shown in Figure 5-33. The New Switchable Hardware Group wizard is launched.



Add a switchable hardware group to the selected cluster.

Figure 5-33 New switchable hardware group

2. In the Welcome window (Figure 5-34), click **Next**.



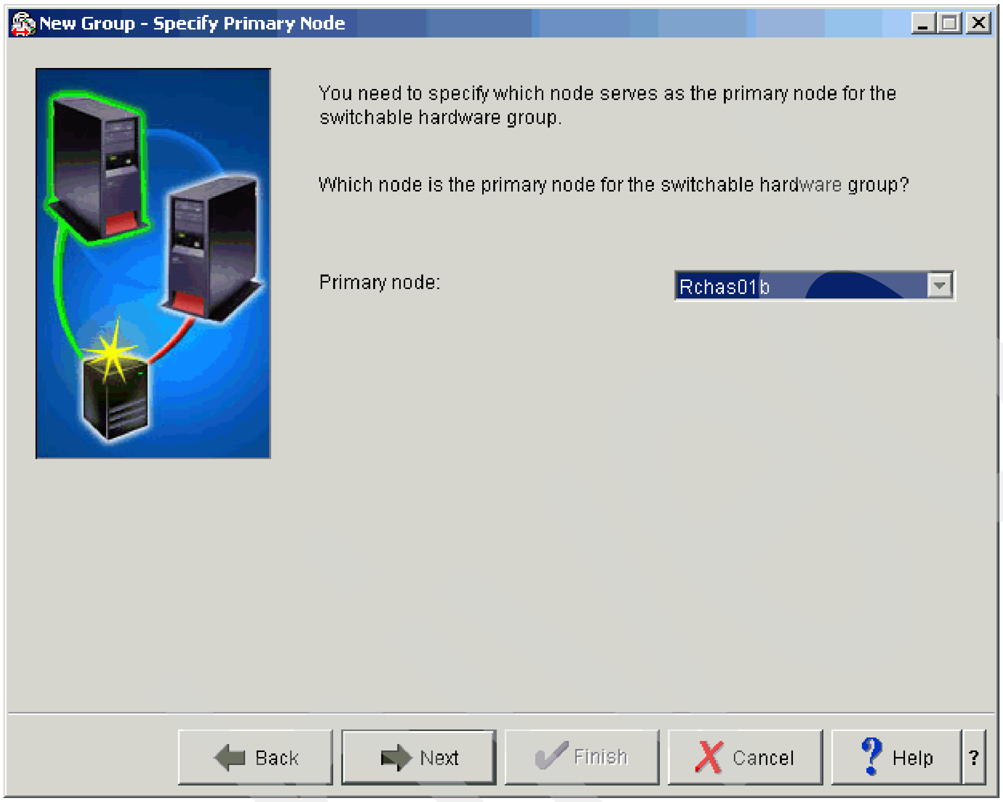
Welcome to the New Switchable Hardware Group wizard.

This wizard takes you through the steps to create a new switchable hardware group and add a disk pool to it for cluster Xsmcluster.

You can cancel at any time by clicking Cancel.

Figure 5-34 New Group - Welcome window

3. In the Specify Primary Node window (Figure 5-35), select the primary node of your cluster environment and click **Next**.



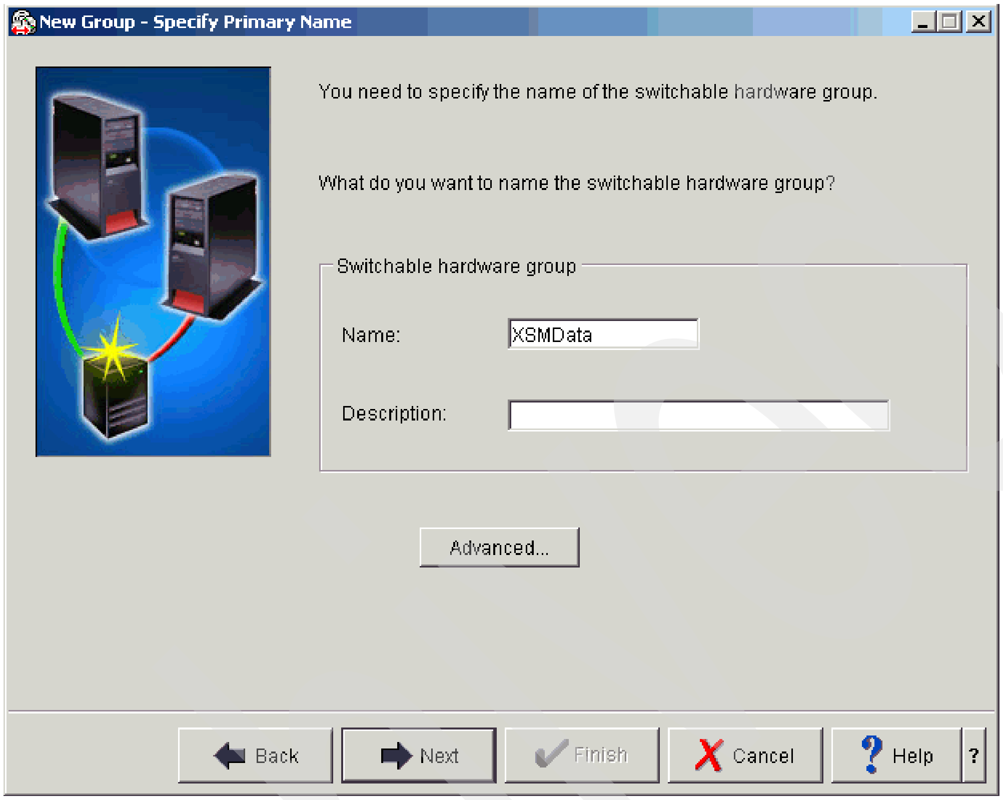
You need to specify which node serves as the primary node for the switchable hardware group.

Which node is the primary node for the switchable hardware group?

Primary node: Rchas01b

Figure 5-35 Specify Primary Node window

4. In the Specify Primary Name window (Figure 5-36), specify the name of the switchable hardware group. The name of the CRG can be the same as the name of your IASP. Click **Next**.



You need to specify the name of the switchable hardware group.

What do you want to name the switchable hardware group?

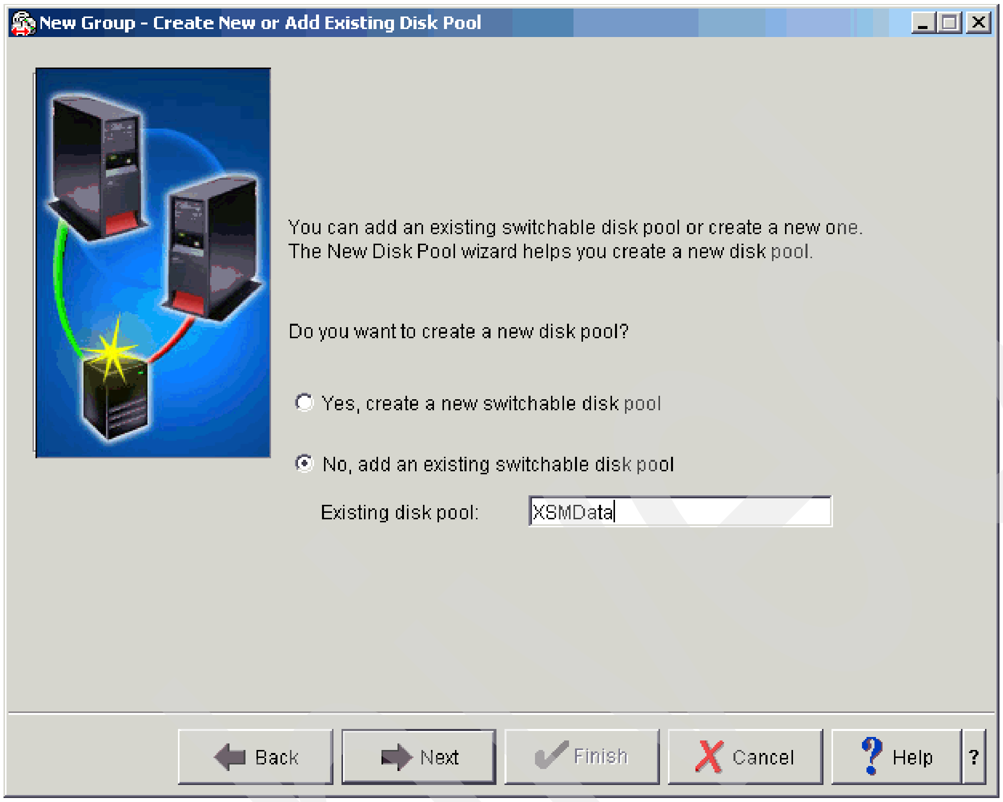
Switchable hardware group

Name: XSMData

Description:

Figure 5-36 Specify Primary Name window

5. In the Create New or Add Existing Disk Pool window (Figure 5-37), select **No, add an existing switchable disk pool**, type the name of your IASP, and then click **Next**.



You can add an existing switchable disk pool or create a new one. The New Disk Pool wizard helps you create a new disk pool.

Do you want to create a new disk pool?

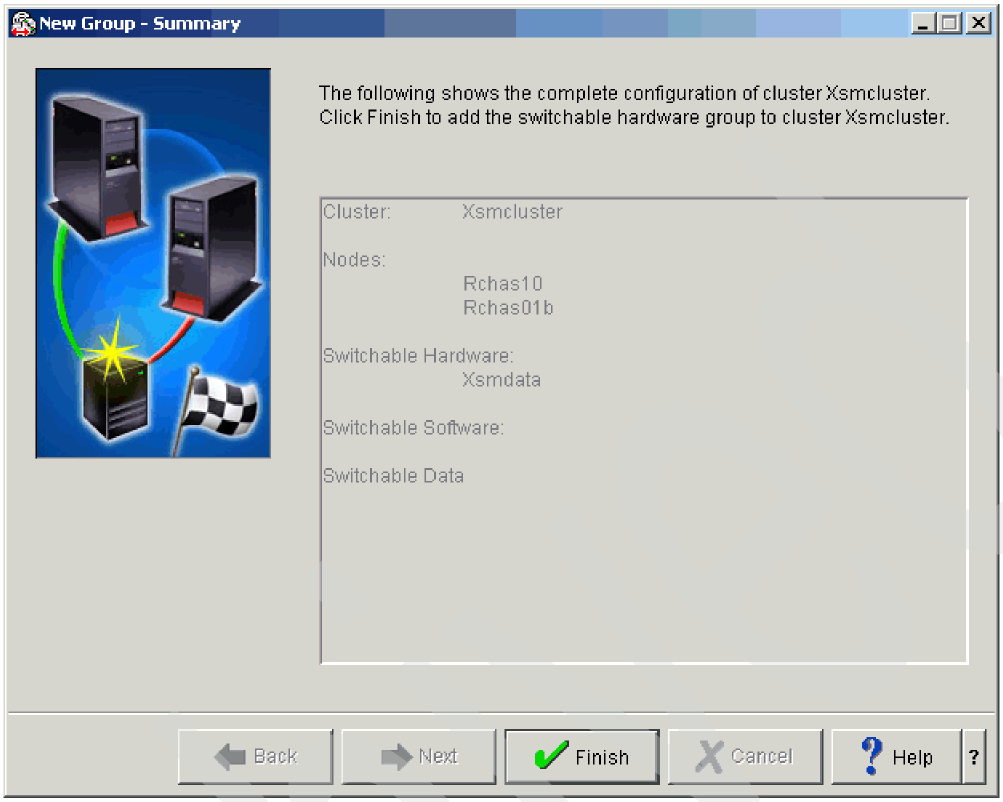
○ Yes, create a new switchable disk pool

● No, add an existing switchable disk pool

Existing disk pool: XSMData

Figure 5-37 Create New or Add Existing Disk Pool window

6. In the Summary window (Figure 5-38), click **Finish**.



The following shows the complete configuration of cluster Xsmcluster.
Click Finish to add the switchable hardware group to cluster Xsmcluster.

Cluster: Xsmcluster

Nodes:
Rchas10
Rchas01b

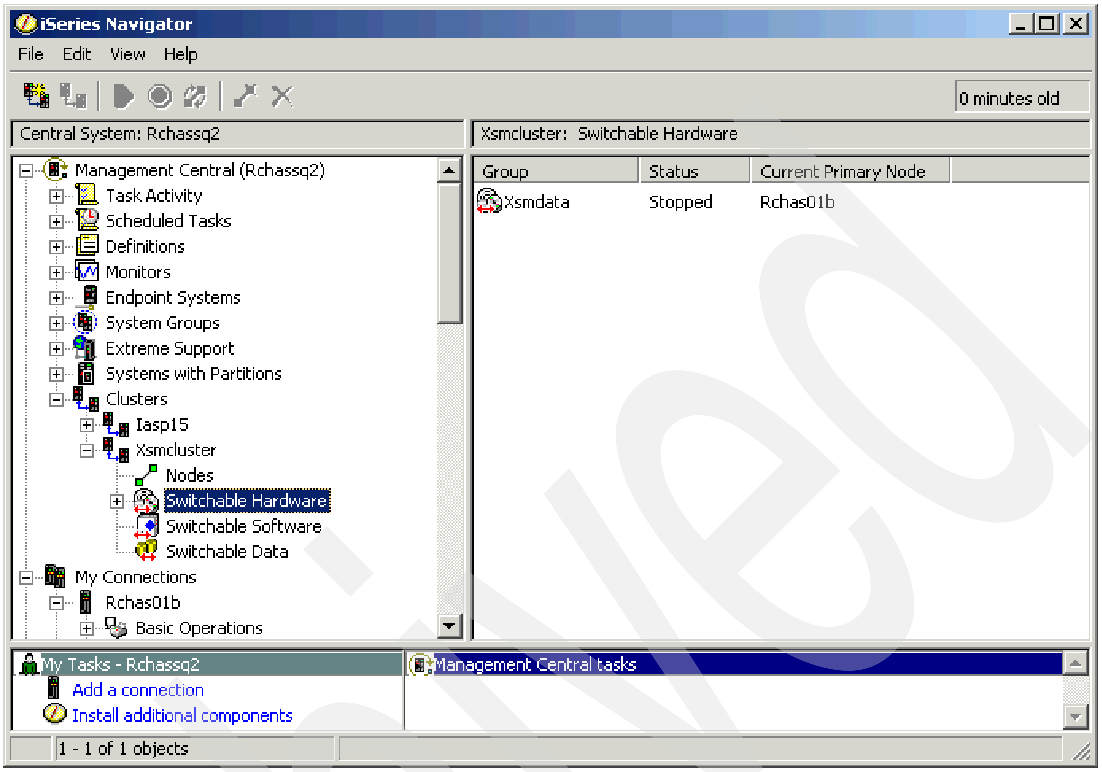
Switchable Hardware:
Xsmdata

Switchable Software:

Switchable Data

Figure 5-38 Summary window

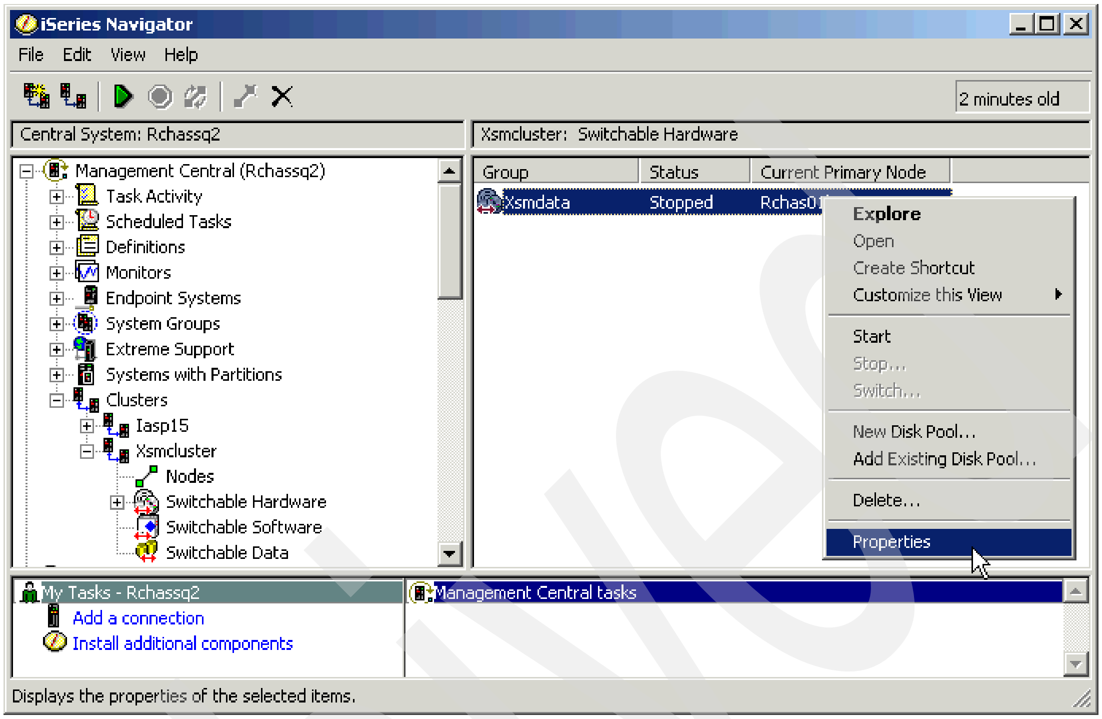
7. After the new device CRG is created, it is displayed in iSeries Navigator in your cluster environment, as shown in Figure 5-39.



Group	Status	Current Primary Node
Xsmdata	Stopped	Rchas01b

Figure 5-39 Switchable hardware view shows newly created cluster resource group

To modify the properties of the new CRG, right-click it and select **Properties**, as shown in Figure 5-40.



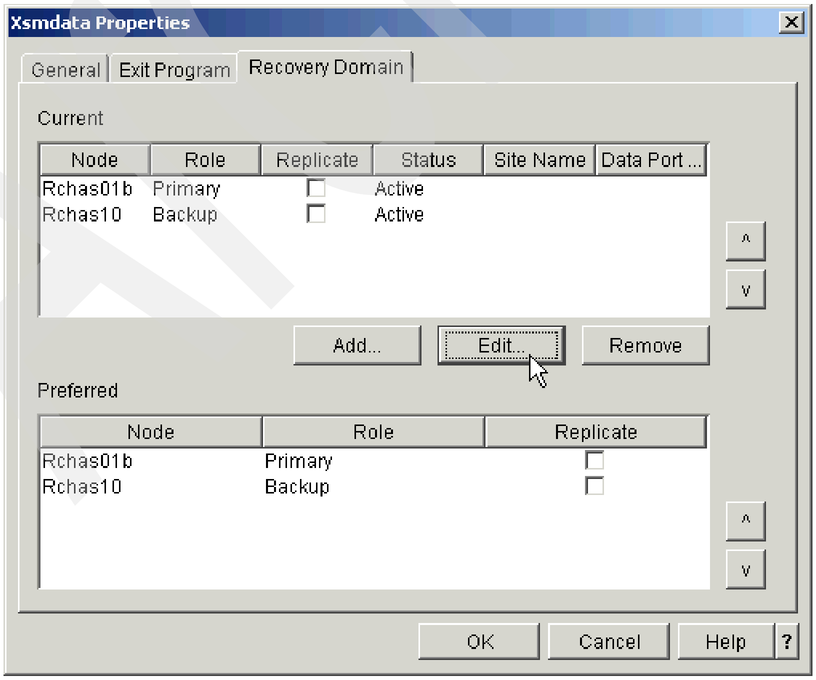
Group	Status	Current Primary Node
Xsmdata	Stopped	Rchas01

Displays the properties of the selected items.

Figure 5-40 Modifying the new CRG properties

8. In the CRG Properties window (Figure 5-41), complete the following actions:

   a. Click the **Recovery Domain** tab.
   b. Select the **primary node**.
   c. Click **Edit**.



Current

Node	Role	Replicate	Status	Site Name	Data Port ...
Rchas01b	Primary	☐	Active		
Rchas10	Backup	☐	Active		

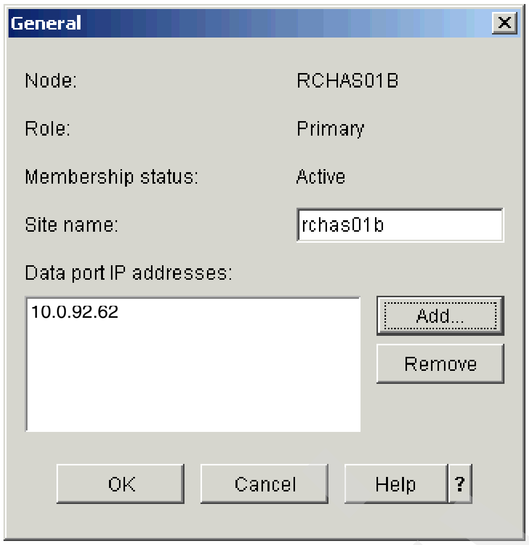
Preferred

Node	Role	Replicate
Rchas01b	Primary	☐
Rchas10	Backup	☐

Figure 5-41 Recovery Domain properties

d. In the General window (Figure 5-42), add one or more TCP/IP addresses that are defined on your primary system that you want to use for the function. Click **OK** and repeat this procedure for the backup node.

e. In the **Properties** window, click **OK**.



Node: RCHAS01B

Role: Primary

Membership status: Active

Site name: rchas01b

Data port IP addresses: 10.0.92.62

Figure 5-42 Adding a data port IP address

Now you have now created a device CRG for XSM using iSeries Navigator.

Creating a device CRG using the CRTCRG command

You can create a device CRG by using a 5250 terminal and performing the following steps:

1. On the backup system, manually create the device description for the IASP by entering the following command:

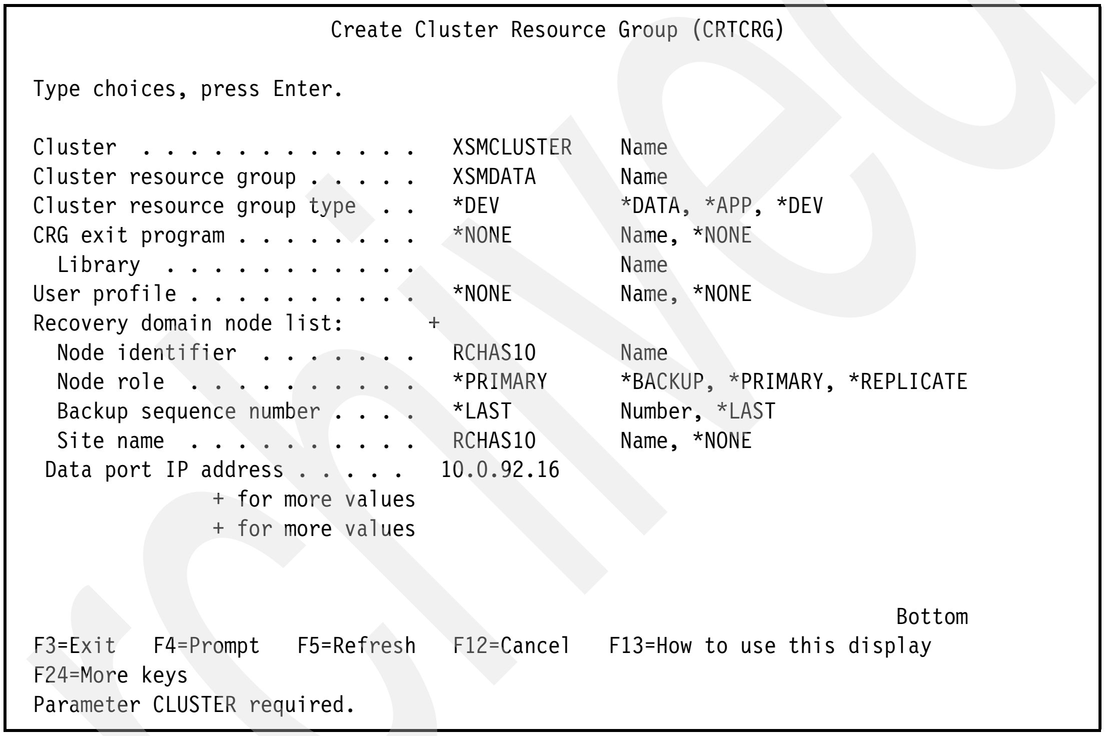
   ```
   CRTDEVASP DEVD(name) RSRCNAME(name)
   ```

   Here *name* must be the name of your ASP device on the production system.

2. Type the following command and then press F4:

   ```
   CRTCRG
   ```

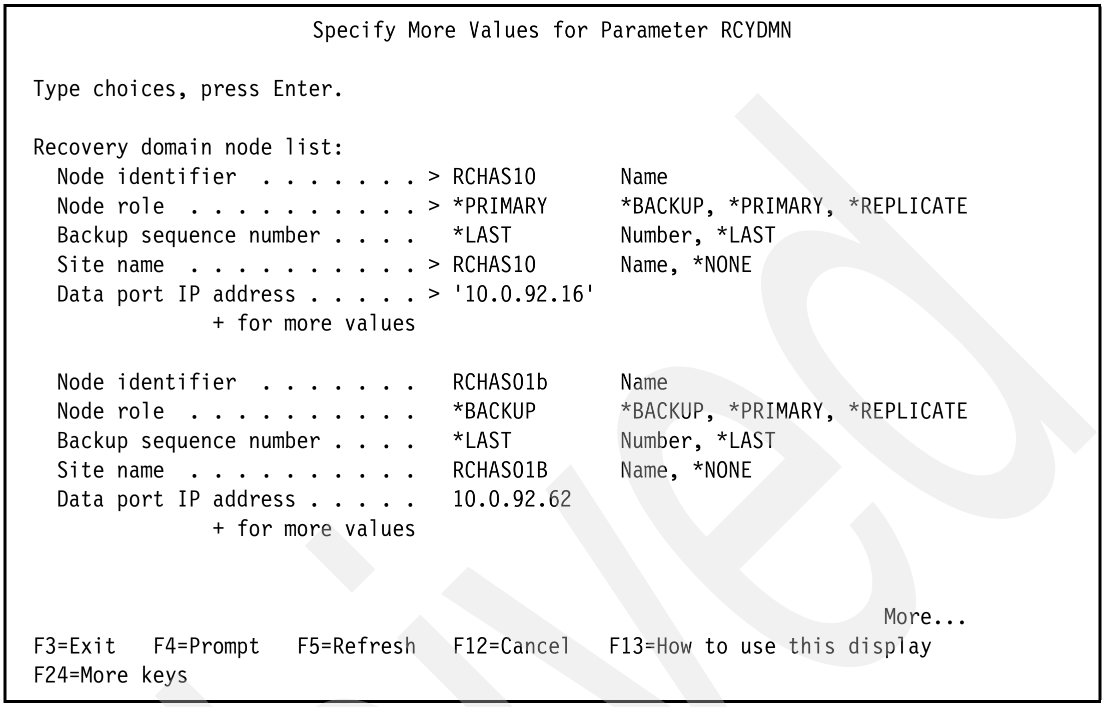
3. In the Create Cluster Resource Group (CRTCRG) panel (Figure 5-43), complete the following steps:
   a. Type the cluster name and the name of the CRG. The type of CRG must be `*DEV`.
   b. In the CRG exit program and User profile fields, type `*NONE`.
   c. Add nodes to the recovery domains. In the Node identifier and Site name fields, type the name of the first node.
   d. In the Node Role field, type `*PRIMARY`.
   e. Add one or more valid data port addresses.
   f. Next to Recovery domain node list, type the plus (+) sign.
   g. Press Enter.



```
                         Create Cluster Resource Group (CRTCRG)

Type choices, press Enter.

Cluster  . . . . . . . . . . . .   XSMCLUSTER    Name
Cluster resource group . . . . .   XSMDATA       Name
Cluster resource group type  . .   *DEV          *DATA, *APP, *DEV
CRG exit program . . . . . . . .   *NONE         Name, *NONE
  Library  . . . . . . . . . . .                 Name
User profile . . . . . . . . . .   *NONE         Name, *NONE
Recovery domain node list:       +
  Node identifier  . . . . . . .   RCHAS10       Name
  Node role  . . . . . . . . . .   *PRIMARY      *BACKUP, *PRIMARY, *REPLICATE
  Backup sequence number . . . .   *LAST         Number, *LAST
  Site name  . . . . . . . . . .   RCHAS10       Name, *NONE
 Data port IP address . . . . .   10.0.92.16
               + for more values
               + for more values



                                                                       Bottom
F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
F24=More keys
Parameter CLUSTER required.
```

Figure 5-43 Create Cluster Resource Group (CRTCRG) panel

4. In the Specify More Values for Parameter RCYDMN panel (Figure 5-44), add a backup node to your CRG. Type the information about the backup node and press Enter.

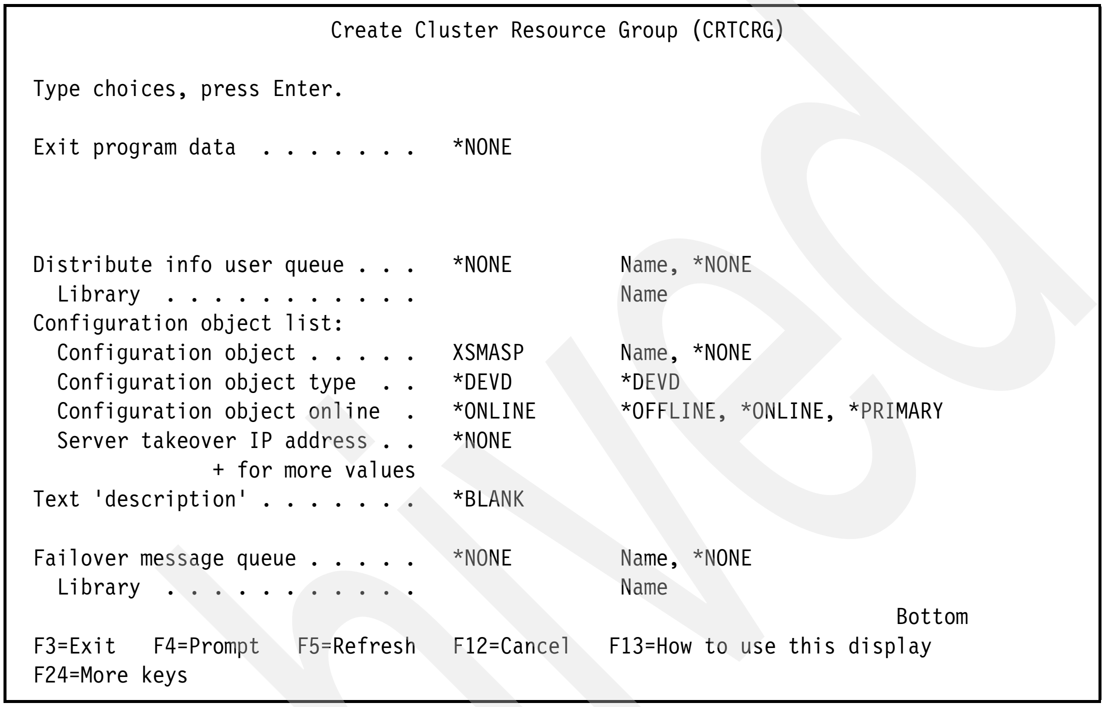


```
                          Specify More Values for Parameter RCYDMN

 Type choices, press Enter.

 Recovery domain node list:
   Node identifier  . . . . . . . > RCHAS10       Name
   Node role  . . . . . . . . . . > *PRIMARY      *BACKUP, *PRIMARY, *REPLICATE
   Backup sequence number . . . .   *LAST         Number, *LAST
   Site name  . . . . . . . . . . > RCHAS10       Name, *NONE
   Data port IP address . . . . . > '10.0.92.16'
                + for more values

   Node identifier  . . . . . . .   RCHAS01b      Name
   Node role  . . . . . . . . . .   *BACKUP       *BACKUP, *PRIMARY, *REPLICATE
   Backup sequence number . . . .   *LAST         Number, *LAST
   Site name  . . . . . . . . . .   RCHAS01B      Name, *NONE
   Data port IP address . . . . .   10.0.92.62
                + for more values


                                                                         More...
 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 5-44 Specify More Values for Parameter RCYDMN panel

5. Press the Page Down (PgDn) key twice.

6. In the last Create Cluster Resource Group (CRTCRG) panel (Figure 5-45), complete the following actions:
   a. For Configuration object, type the name of the IASP device.
   b. For Configuration object type, enter *DEVD.
   c. For Configuration object online, type *ONLINE.
   d. Press Enter.



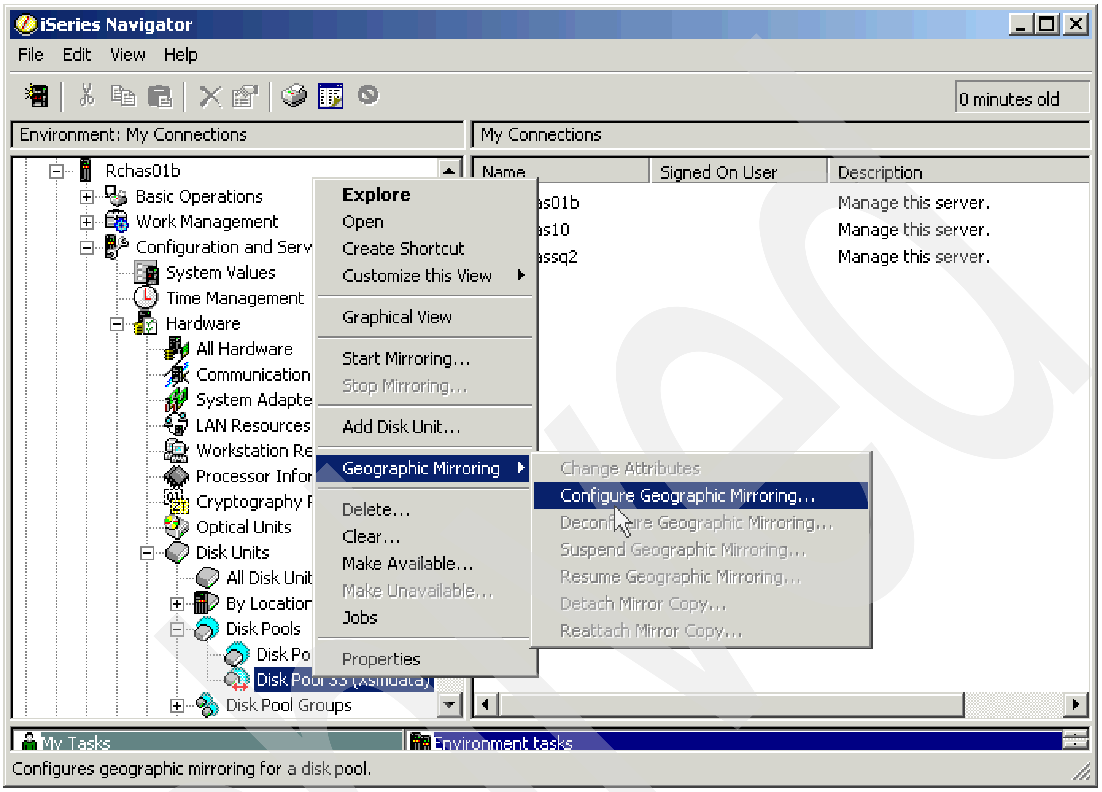
```
                         Create Cluster Resource Group (CRTCRG)

 Type choices, press Enter.

 Exit program data  . . . . . . .   *NONE




 Distribute info user queue . . .   *NONE         Name, *NONE
   Library  . . . . . . . . . . .                 Name
 Configuration object list:
   Configuration object . . . . .   XSMASP        Name, *NONE
   Configuration object type  . .   *DEVD         *DEVD
   Configuration object online  .   *ONLINE       *OFFLINE, *ONLINE, *PRIMARY
   Server takeover IP address . .   *NONE
               + for more values
 Text 'description' . . . . . . .   *BLANK

 Failover message queue . . . . .   *NONE         Name, *NONE
   Library  . . . . . . . . . . .                 Name
                                                                       Bottom
 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 5-45 Last Create Cluster Resource Group (CRTCRG) panel

5.2.4 Adding the geographically mirrored IASP to the cluster configuration

In this section, we add the geographically mirrored IASP to the cluster configuration. The Configure Geographic Mirroring wizard helps you to set up attributes and create the IASP on the second node of your cluster. Because the wizard partially repeats the New Disk Pool wizard steps, we recommend that you review 5.2.2, “Creating a stand-alone IASP” on page 90, and complete the same preparation tasks for the second node. Then, when you are ready, follow these steps in iSeries Navigator:

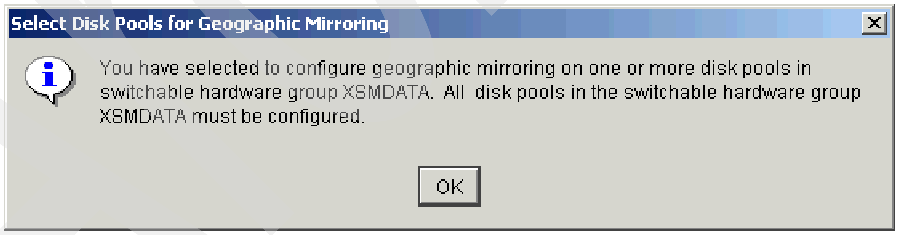
1. Expand your primary node environment. Select **Configuration and Service** → **Hardware** → **Disk Units** → **Disk Pools**, and select the IASP that you created in 5.2.2, “Creating a stand-alone IASP” on page 90. Right-click and select **Geographic Mirroring** → **Configure Geographic Mirroring** as shown in Figure 5-46.



Configures geographic mirroring for a disk pool.

Figure 5-46 Launching the Configure Geographic Mirroring wizard

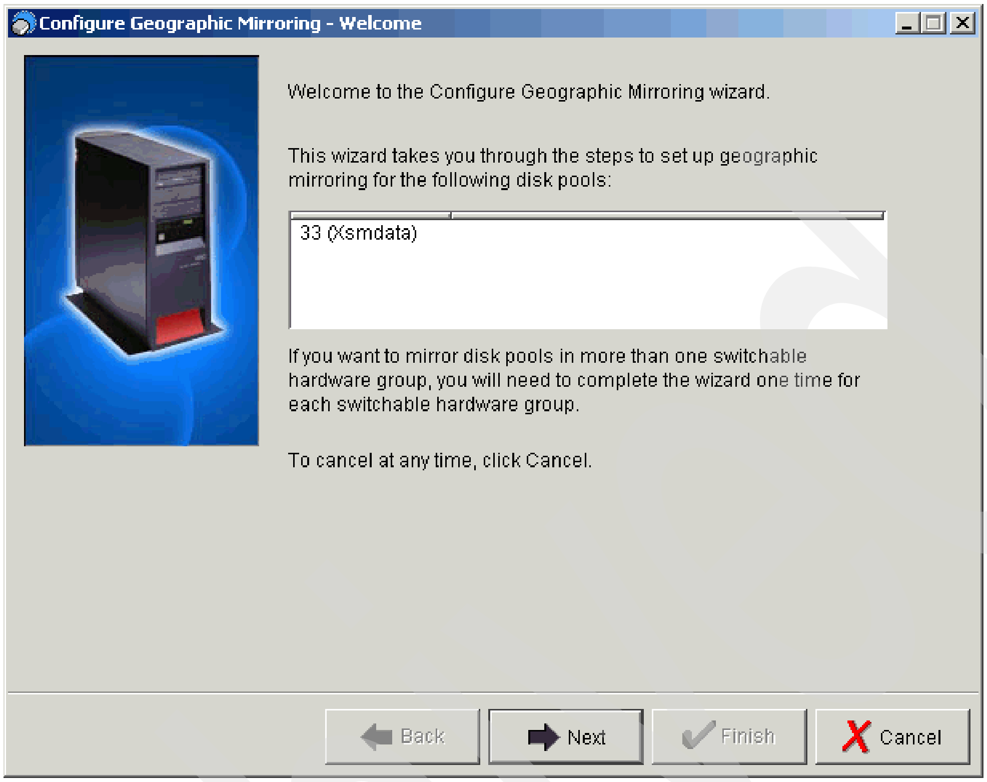
2. In the Select Disk Pools for Geographic Mirroring message window (Figure 5-47), click **OK**.



You have selected to configure geographic mirroring on one or more disk pools in switchable hardware group XSMDATA. All disk pools in the switchable hardware group XSMDATA must be configured.

Figure 5-47 Select Disk Pools for Geographic Mirroring message window

3. In the Configure Geographic Mirroring - Welcome window (Figure 5-48), click **Next**.



Configure Geographic Mirroring - Welcome

Welcome to the Configure Geographic Mirroring wizard.

This wizard takes you through the steps to set up geographic mirroring for the following disk pools:

33 (Xsmdata)

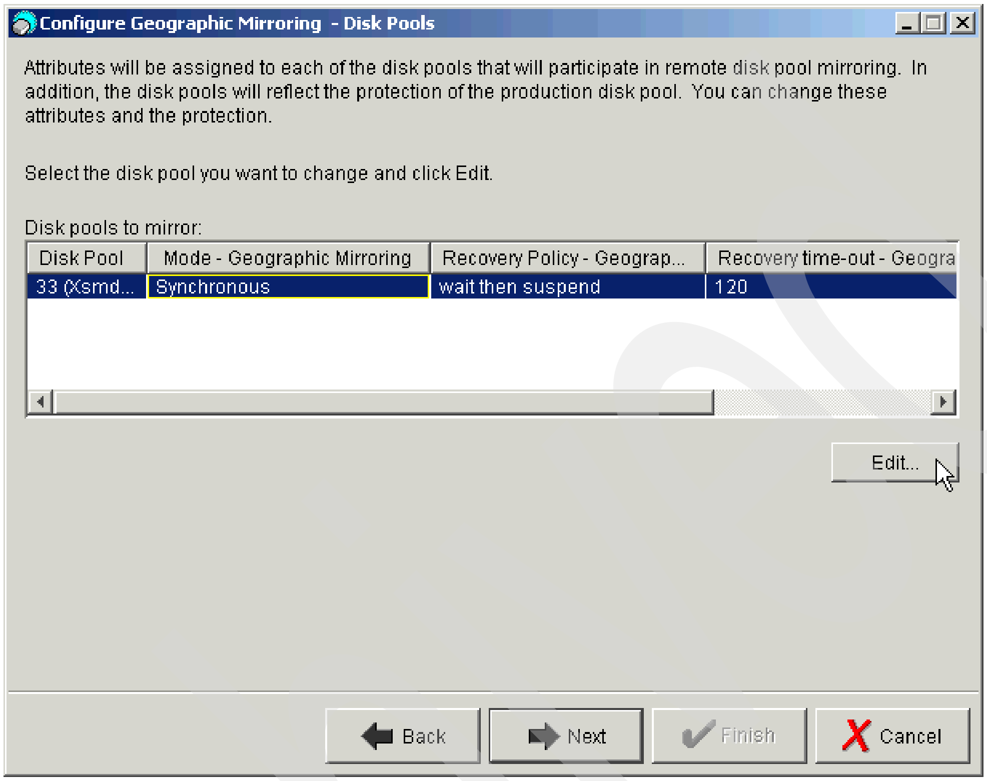
If you want to mirror disk pools in more than one switchable hardware group, you will need to complete the wizard one time for each switchable hardware group.

To cancel at any time, click Cancel.

Figure 5-48 Configure Geographic Mirroring - Welcome window

4. In the Disk Pools window (Figure 5-49), you see the selected disk pool with the default geographic mirroring attributes that are assigned by the wizard.

   a. Click the **Edit** button.



Configure Geographic Mirroring - Disk Pools

Attributes will be assigned to each of the disk pools that will participate in remote disk pool mirroring. In addition, the disk pools will reflect the protection of the production disk pool. You can change these attributes and the protection.

Select the disk pool you want to change and click Edit.

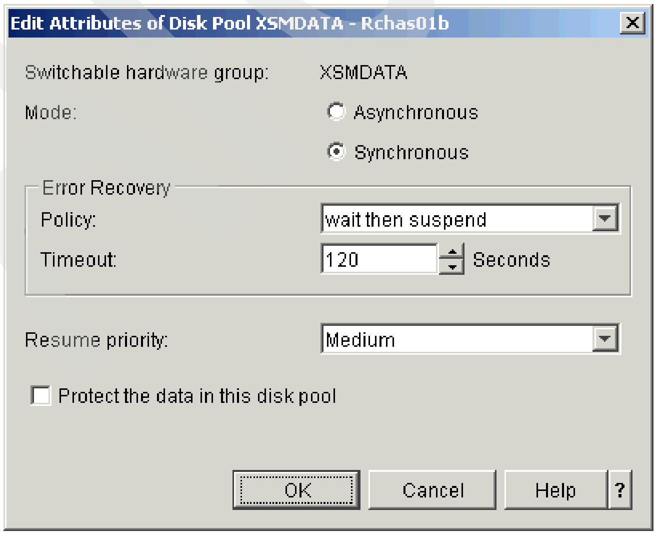
Disk pools to mirror:

Disk Pool	Mode - Geographic Mirroring	Recovery Policy - Geograp...	Recovery time-out - Geogra
33 (Xsmd...	Synchronous	wait then suspend	120

Figure 5-49 Configure Geographic Mirroring - Disk Pools window

   b. In the Edit Attributes of Disk Pool window (Figure 5-50), modify the settings and click **OK**.

      For our scenario, we do not change these attributes and use default values. Therefore, we click **Cancel**.



Edit Attributes of Disk Pool XSMDATA - Rchas01b

Switchable hardware group: XSMDATA

Mode: ○ Asynchronous ● Synchronous

Error Recovery

Policy: wait then suspend

Timeout: 120 Seconds

Resume priority: Medium

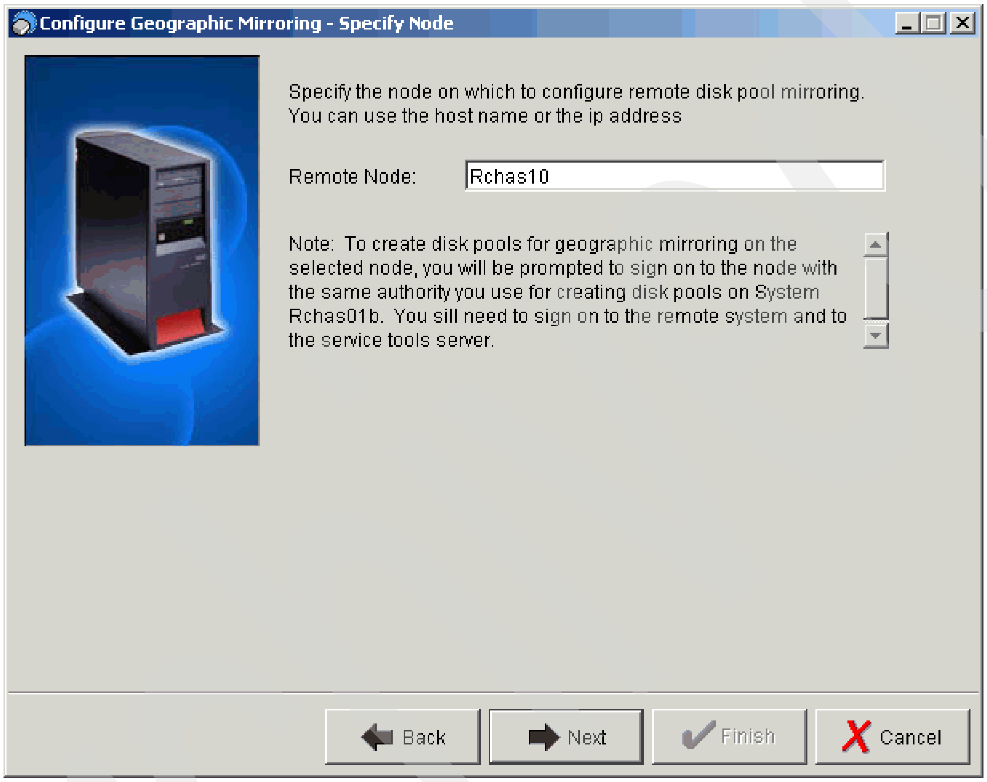
□ Protect the data in this disk pool

Figure 5-50 Edit Attributes of Disk Pool for V5R3

c. When you return to the Disk Pools window (Figure 5-49 on page 115), click **Next**.

If you are running with V5R4, the Edit Attributes of Disk Pool window looks slightly different because you have the option to define a Tracking Space for Source Site Tracking. See Figure 7-3 on page 138. Refer to 7.1.2, “Configuration of the Tracking Space” on page 138, for more information about Tracking Space.

5. In the Specify Node window (Figure 5-51), type the name of the second node of your cluster and click **Next**.



Specify the node on which to configure remote disk pool mirroring. You can use the host name or the ip address

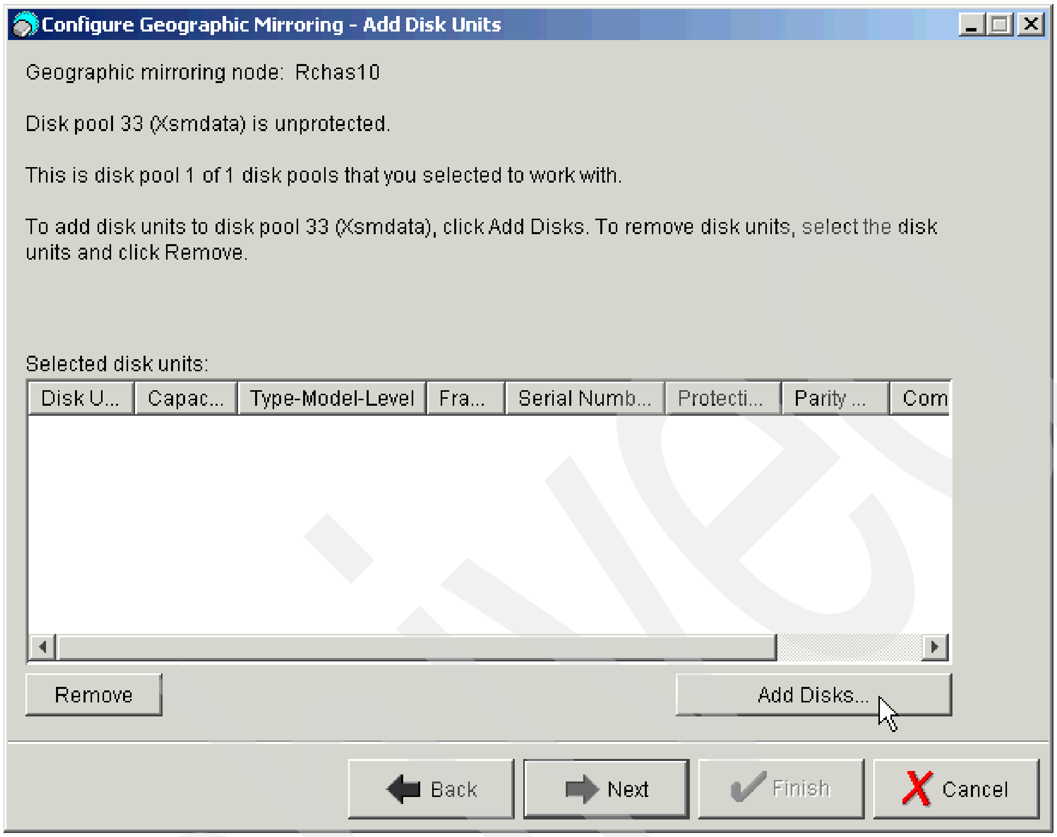
Remote Node: Rchas10

Note: To create disk pools for geographic mirroring on the selected node, you will be prompted to sign on to the node with the same authority you use for creating disk pools on System Rchas01b. You sill need to sign on to the remote system and to the service tools server.

Figure 5-51 Specify Node window

6. In the Add Disk Units window (Figure 5-52), complete the following steps:

   a. Click **Add Disks**.



Geographic mirroring node: Rchas10

Disk pool 33 (Xsmdata) is unprotected.

This is disk pool 1 of 1 disk pools that you selected to work with.

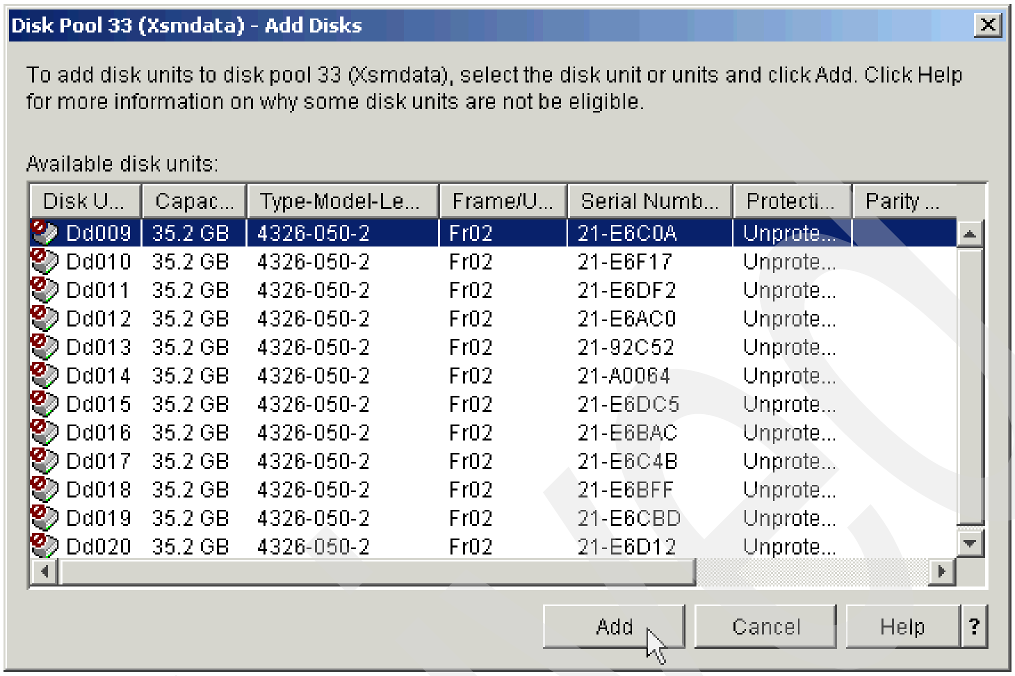
To add disk units to disk pool 33 (Xsmdata), click Add Disks. To remove disk units, select the disk units and click Remove.

Selected disk units:

Disk U...	Capac...	Type-Model-Level	Fra...	Serial Numb...	Protecti...	Parity ...	Com

Figure 5-52 Add Disk Units window

b. In the next window (Figure 5-53), select disk units for the mirror copy in the backup node. Click **Add**.



Disk Pool 33 (Xsmdata) - Add Disks

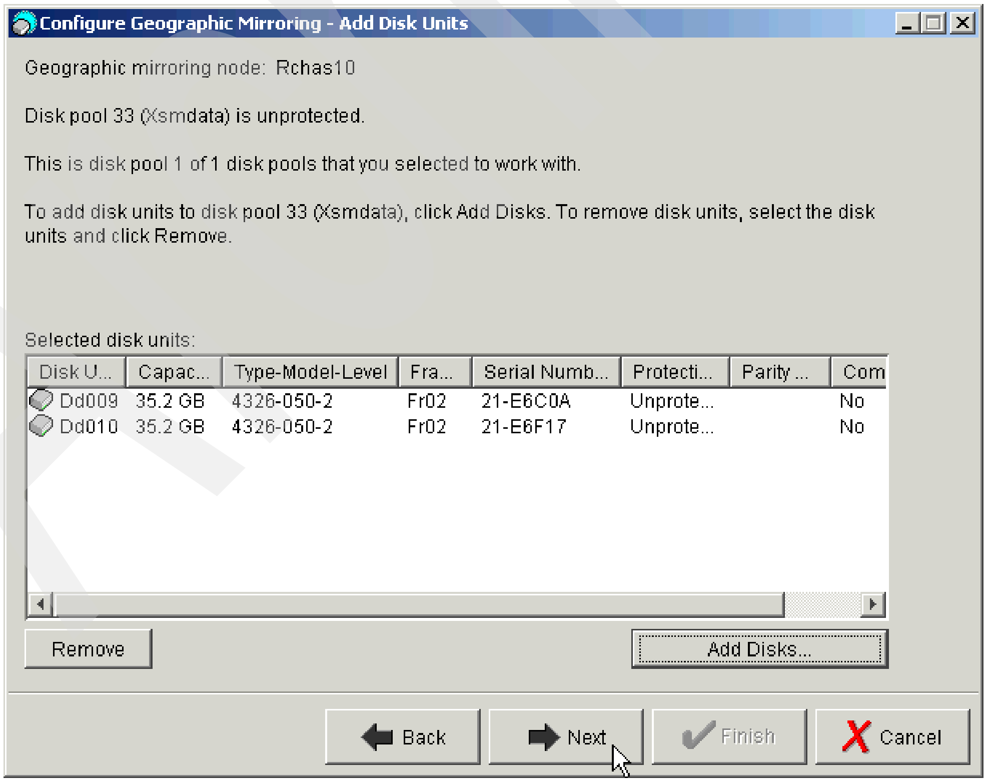
To add disk units to disk pool 33 (Xsmdata), select the disk unit or units and click Add. Click Help for more information on why some disk units are not be eligible.

Available disk units:

Disk U...	Capac...	Type-Model-Le...	Frame/U...	Serial Numb...	Protecti...	Parity ...
Dd009	35.2 GB	4326-050-2	Fr02	21-E6C0A	Unprote...	
Dd010	35.2 GB	4326-050-2	Fr02	21-E6F17	Unprote...	
Dd011	35.2 GB	4326-050-2	Fr02	21-E6DF2	Unprote...	
Dd012	35.2 GB	4326-050-2	Fr02	21-E6AC0	Unprote...	
Dd013	35.2 GB	4326-050-2	Fr02	21-92C52	Unprote...	
Dd014	35.2 GB	4326-050-2	Fr02	21-A0064	Unprote...	
Dd015	35.2 GB	4326-050-2	Fr02	21-E6DC5	Unprote...	
Dd016	35.2 GB	4326-050-2	Fr02	21-E6BAC	Unprote...	
Dd017	35.2 GB	4326-050-2	Fr02	21-E6C4B	Unprote...	
Dd018	35.2 GB	4326-050-2	Fr02	21-E6BFF	Unprote...	
Dd019	35.2 GB	4326-050-2	Fr02	21-E6CBD	Unprote...	
Dd020	35.2 GB	4326-050-2	Fr02	21-E6D12	Unprote...	

Figure 5-53 Selecting disks for the mirrored copy

c. The Add Disk Units window now shows the added disks (see Figure 5-54). Click **Next**.



Configure Geographic Mirroring - Add Disk Units

Geographic mirroring node: Rchas10

Disk pool 33 (Xsmdata) is unprotected.

This is disk pool 1 of 1 disk pools that you selected to work with.

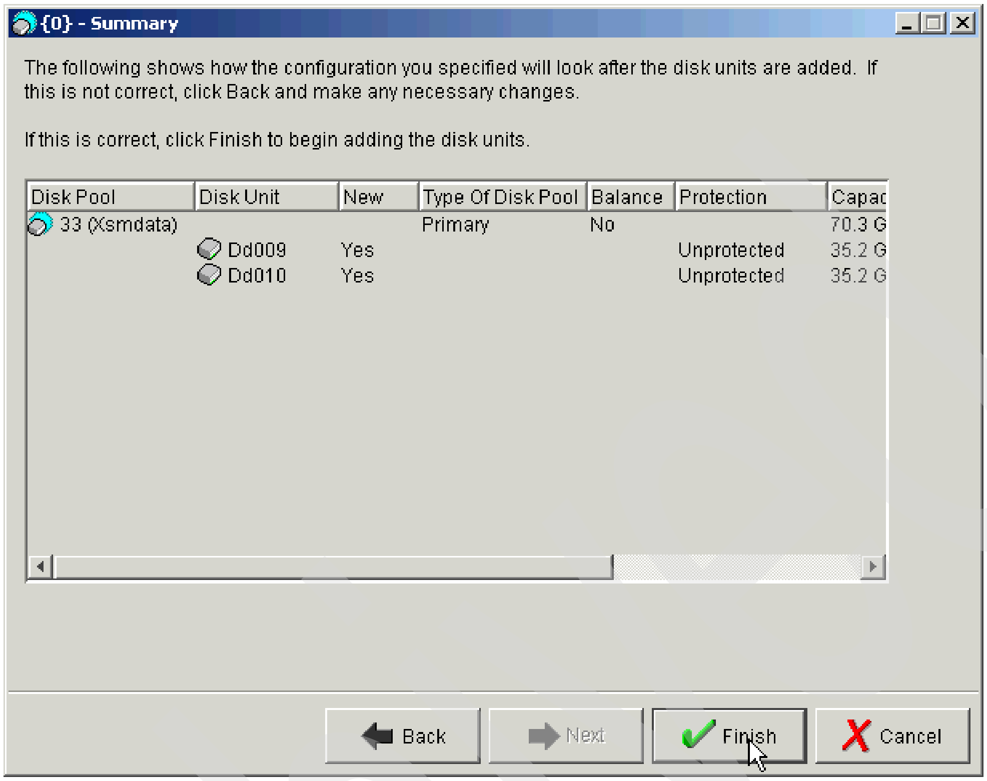
To add disk units to disk pool 33 (Xsmdata), click Add Disks. To remove disk units, select the disk units and click Remove.

Selected disk units:

Disk U...	Capac...	Type-Model-Level	Fra...	Serial Numb...	Protecti...	Parity ...	Com
Dd009	35.2 GB	4326-050-2	Fr02	21-E6C0A	Unprote...		No
Dd010	35.2 GB	4326-050-2	Fr02	21-E6F17	Unprote...		No

Figure 5-54 Add Disk Units window with selected disks

7. In the Summary window (Figure 5-55), click **Finish**.



{0} - Summary

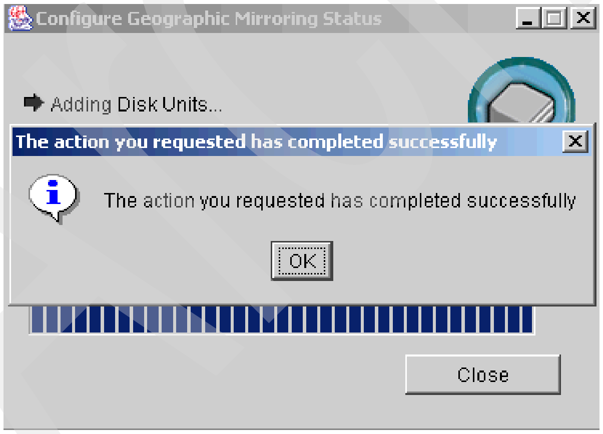
The following shows how the configuration you specified will look after the disk units are added. If this is not correct, click Back and make any necessary changes.

If this is correct, click Finish to begin adding the disk units.

Disk Pool	Disk Unit	New	Type Of Disk Pool	Balance	Protection	Capac
33 (Xsmdata)			Primary	No		70.3 G
	Dd009	Yes			Unprotected	35.2 G
	Dd010	Yes			Unprotected	35.2 G

Figure 5-55 Summary window

8. The Configure Geographic Mirroring wizard shows the Status window (Figure 5-56), which indicates the progress. After the wizard's tasks have completed, click **OK**.



Configure Geographic Mirroring Status

Adding Disk Units...

The action you requested has completed successfully

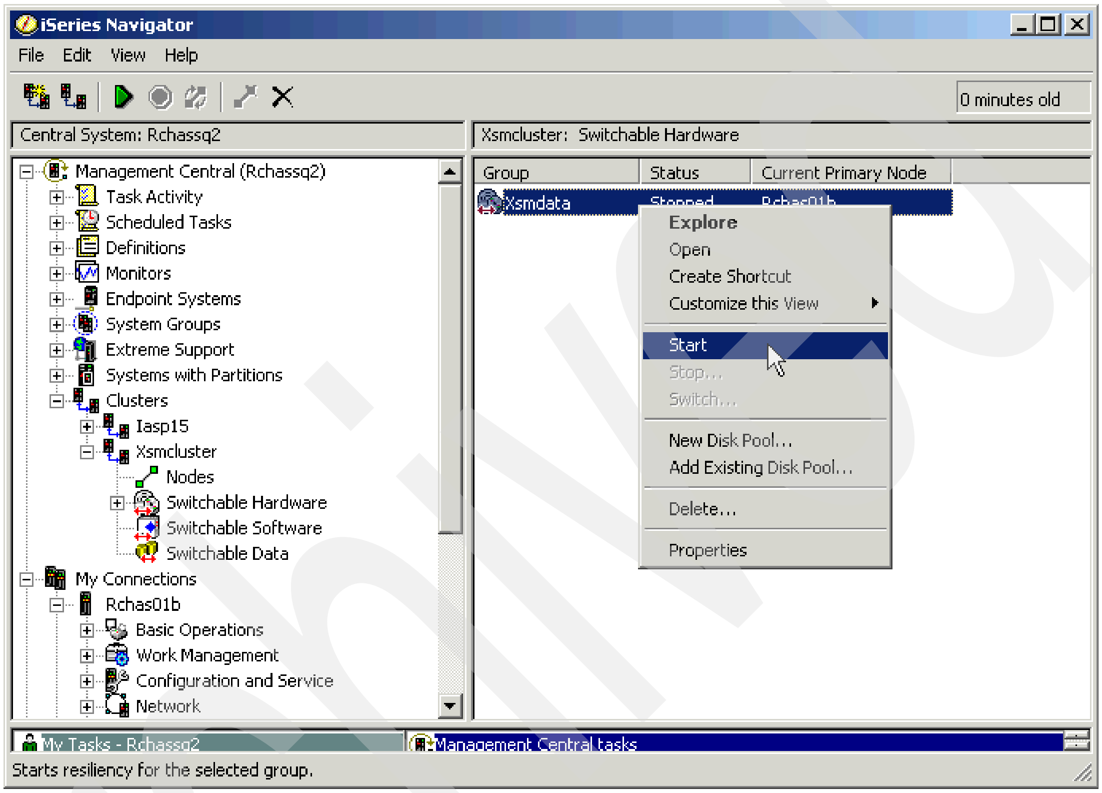
The action you requested has completed successfully

Figure 5-56 The Configure Geographic Mirroring wizard showing a successful task completion

5.2.5 Starting the cross-site mirror environment

To start the XSM environment:

1. Start the switchable hardware group. Open your cluster environment in Management Central, expand **Clusters** → ***your cluster*** → **Switchable Hardware** and select the device CRG that you created earlier. In the right pane, right-click your device CRG and select **Start** as shown in Figure 5-57.

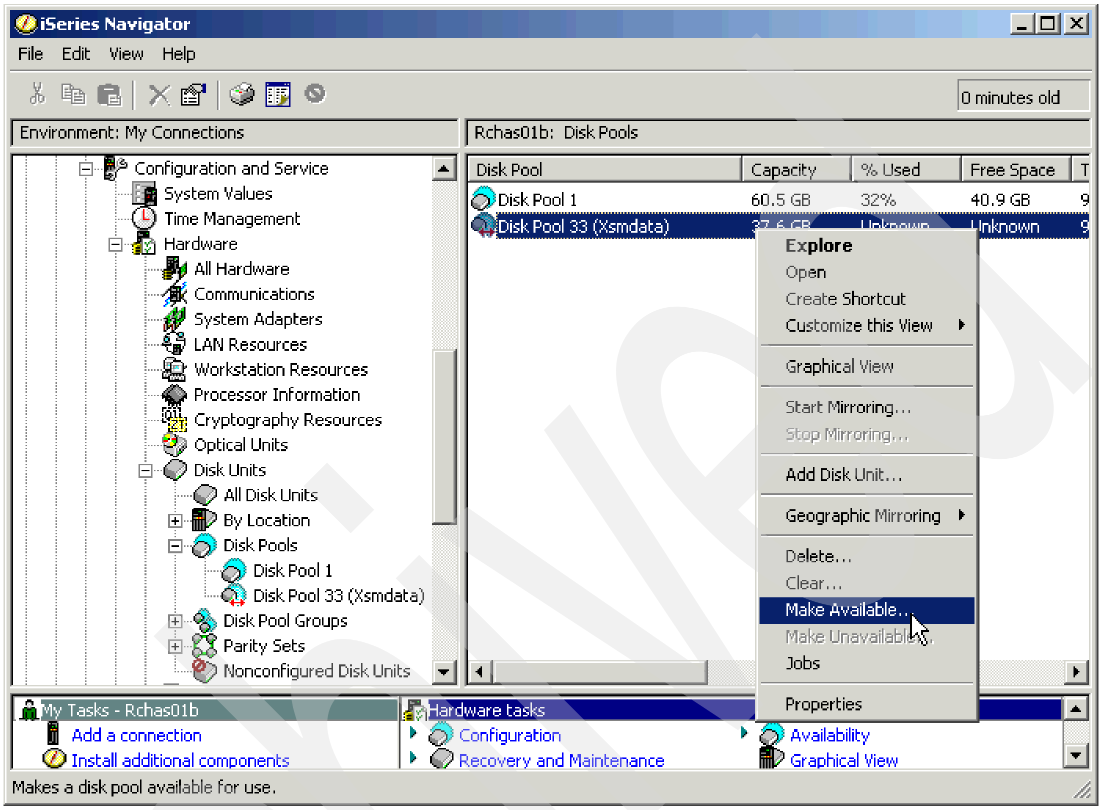


Group	Status	Current Primary Node
Xsmdata	Stopped	Rchas01b

Starts resiliency for the selected group.

Figure 5-57 Starting the CRG

2. When CRG is started, vary on the IASP in the primary node. Using iSeries Navigator to open the primary node, expand **Configuration and Service** → **Hardware** → **Disk Units** → **Disk Pools**. In the right pane, right-click the geographically mirrored disk pool and select **Make Available**, as shown in Figure 5-58.



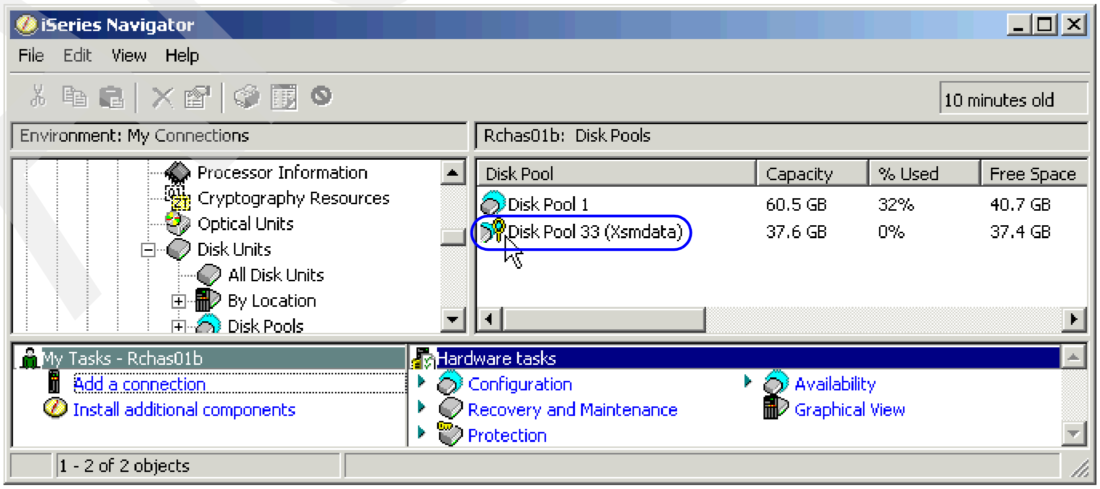
Disk Pool	Capacity	% Used	Free Space	T
Disk Pool 1	60.5 GB	32%	40.9 GB	9
Disk Pool 33 (Xsmdata)	37.6 GB	Unknown	Unknown	9

Makes a disk pool available for use.

Figure 5-58 Varying on the IASP

3. Perform the steps described in the Making Disk Pool Available wizard.

When the wizard completes the vary on task, you have successfully started geographic mirroring for the stand-alone ASP in the two-node cluster scenario. iSeries Navigator uses a different icon for geographically mirrored pools as shown in Figure 5-59.



Disk Pool	Capacity	% Used	Free Space
Disk Pool 1	60.5 GB	32%	40.7 GB
Disk Pool 33 (Xsmdata)	37.6 GB	0%	37.4 GB

1 - 2 of 2 objects

Figure 5-59 Special icon that iSeries Navigator uses for geographically mirrored pools

6

Cluster Administrative Domain

In this chapter, we discuss the new Administrative Domain that is available with i5/OS V5R4. A *Cluster Administrative Domain* monitors and synchronizes changes to selected resources from the system auxiliary storage pool (ASP) within a cluster. The Cluster Administrative Domain provides easier management and synchronization of attributes for resources that are shared within a cluster, such as environment variables or user profiles. The domain can be used to help maintain a consistent environment on the systems that are included in it, thereby ensuring that applications behave the same on all systems that are included in the cluster.

A Cluster Administrative Domain is represented by a peer cluster resource group (CRG). When a Cluster Administrative Domain is created, the peer CRG is created automatically by the system. The name of the Cluster Administrative Domain becomes the name of the peer CRG. Membership in the Cluster Administrative Domain can be changed by adding and removing nodes to the recovery domain of the peer CRG. The nodes that make up the Cluster Administrative Domain are defined by the recovery domain of the peer CRG.

Figure 6-1 on page 124 illustrates a central system with a recovery domain that contains three other systems. A change to a system value is made on the central system. This change is then propagated to the other systems, which results in a change to the Kiosk PC.

All of the nodes are peer nodes (systems or partitions). This means that changes that are made to the objects that are controlled by the peer CRG on *any* of the nodes in the peer CRG are synchronized to *all other* nodes within the peer CRG. Replicate nodes are not allowed in a Cluster Administrative Domain. A cluster node can be defined only in one Cluster Administrative Domain within the cluster.

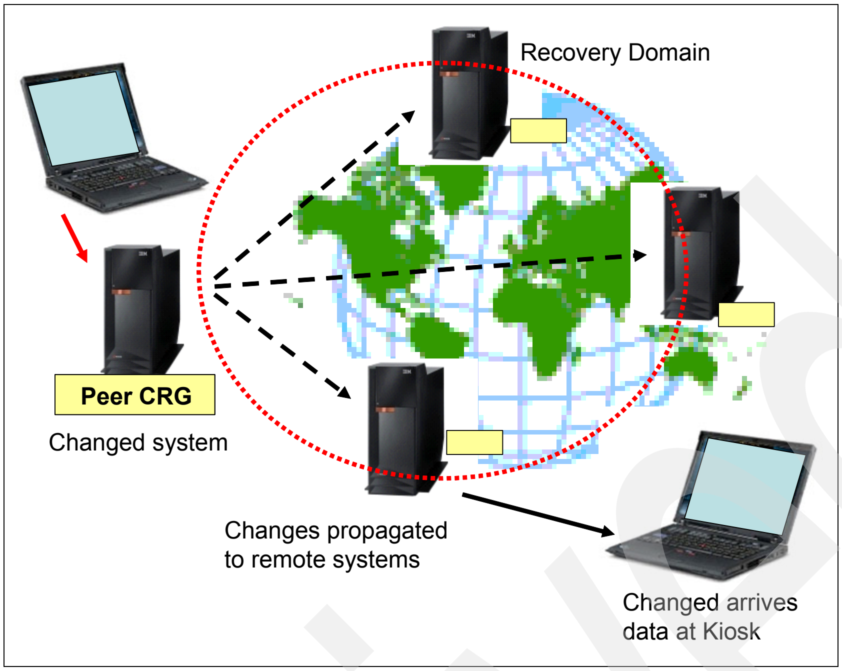
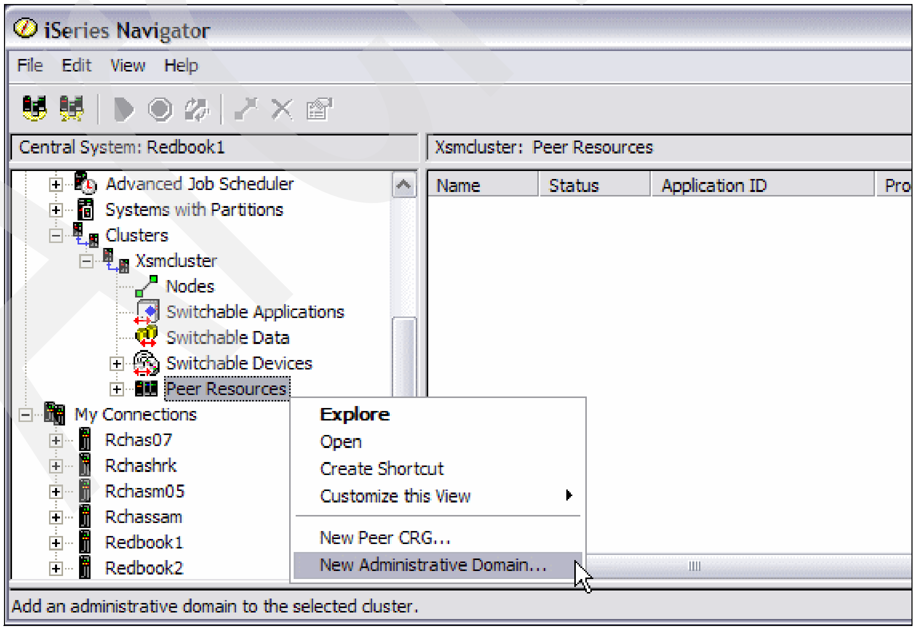


Figure 6-1 Administrative Domain change flow of a monitored resource entry

6.1 Setting up a Cluster Administrative Domain

To set up a Cluster Administrative Domain:

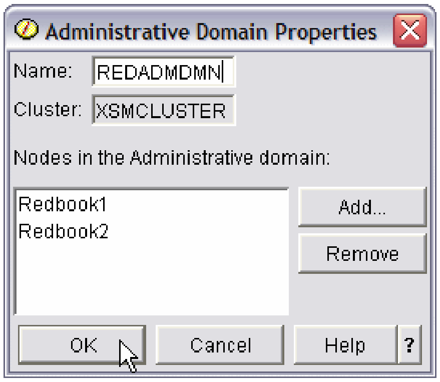
1. Create the Cluster Administrative Domain by using Management Central. In iSeries Navigator, select **Clusters** → ***cluster name***. Right-click **Peer Resources** and select **New Administrative Domain** as shown in Figure 6-2.



Add an administrative domain to the selected cluster.

Figure 6-2 Creating a new Cluster Administrative Domain

2. In the Administrative Domain Properties window (Figure 6-3), enter the name of the new Administrative Domain. This window gives you information about the cluster to which this new Administrative Domain belongs and about the nodes that are part of the Administrative Domain. Click **OK**.



Name: REDADMDMN

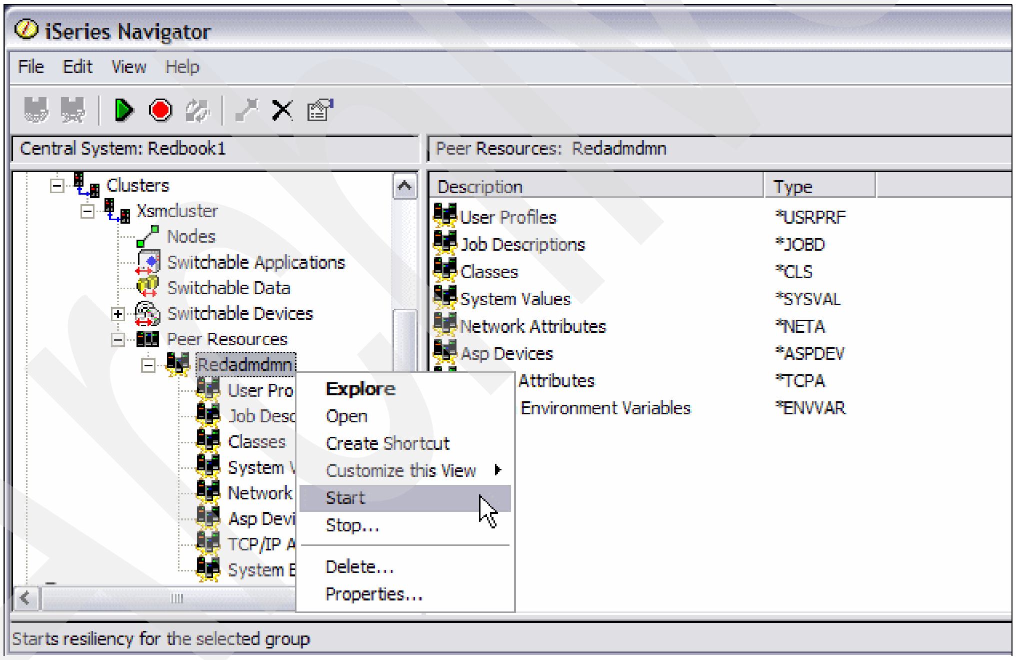
Cluster: XSMCLUSTER

Nodes in the Administrative domain:

Redbook1
Redbook2

Figure 6-3 New Administrative Domain - Properties

3. After the Administrative Domain is created, start it. Again, using Management Central, locate the newly created Administrative Domain by selecting **Cluster** → ***cluster name*** → **Peer Resources** as shown in Figure 6-4. Right-click the ***domain name*** and select **Start**.



Central System: Redbook1

Peer Resources: Redadmdmn

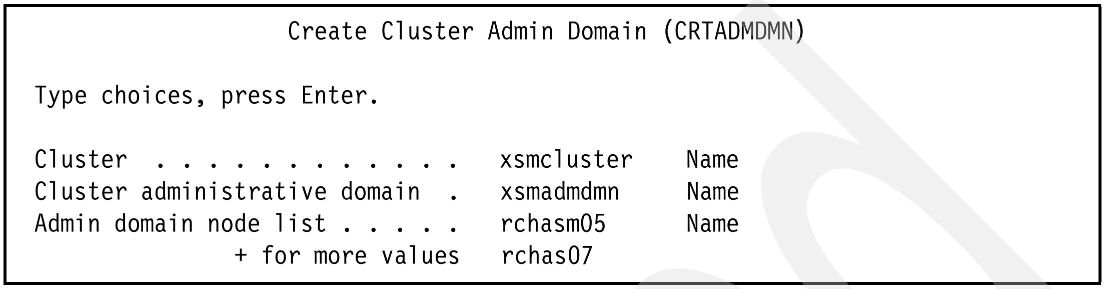
Description	Type
User Profiles	*USRPRF
Job Descriptions	*JOBD
Classes	*CLS
System Values	*SYSVAL
Network Attributes	*NETA
Asp Devices	*ASPDEV
Attributes	*TCPA
Environment Variables	*ENVVAR

Starts resiliency for the selected group

Figure 6-4 Starting the Administrative Domain

You can also create and start the Cluster Administrative Domain by using the 5250 interface. First you enter the Create Cluster Admin Domain (CRTADMDMN) command (Figure 6-5).

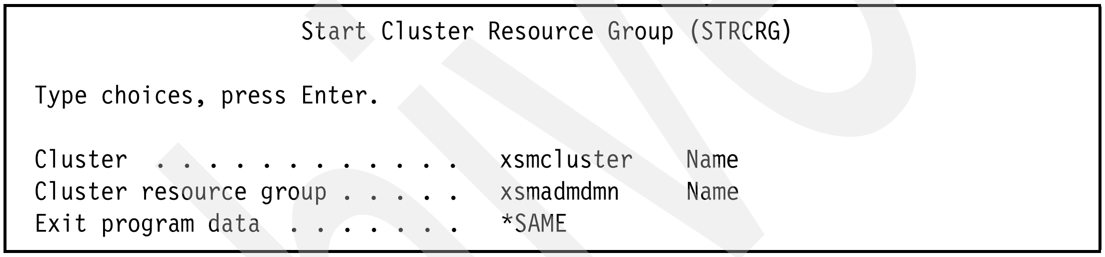
Note: The nodes that you want to add to the Administrative Domain must be active.



```
                   Create Cluster Admin Domain (CRTADMDMN)

Type choices, press Enter.

Cluster  . . . . . . . . . . .   xsmcluster    Name
Cluster administrative domain  .   xsmadmdmn     Name
Admin domain node list . . . . .   rchasm05      Name
               + for more values   rchas07
```

Figure 6-5 Create Cluster Administrative Domain panel

Then you enter the Start Cluster Resource Group (STRCRG) command (Figure 6-6).



```
                    Start Cluster Resource Group (STRCRG)

Type choices, press Enter.

Cluster  . . . . . . . . . . .   xsmcluster    Name
Cluster resource group . . . . .   xsmadmdmn     Name
Exit program data  . . . . . . .   *SAME
```

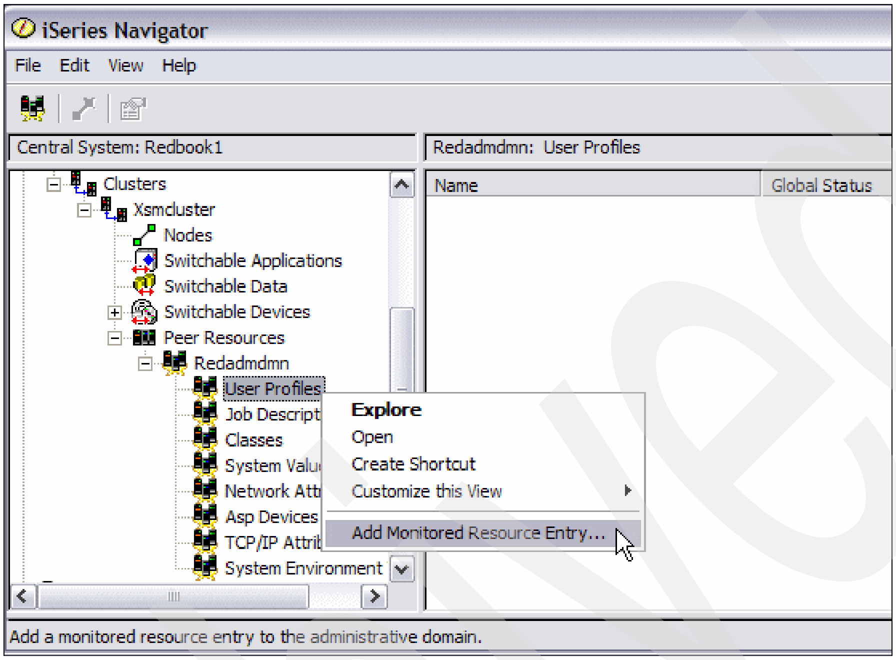
Figure 6-6 Starting the peer CRG

6.2 Adding Monitored Resource Entries

After you start the peer CRG that is associated with your Administrative Domain, you can start adding the resources into it that you want to be monitored and synchronized between the nodes in the Administrative Domain. These are referred to as *Monitored Resource Entries* (MREs). You can either add or remove them by using Management Central or by using specific application programming interfaces (APIs). In addition to the APIs, a set of unsupported CL commands with their command processing programs and sources have been programmed and are available in the QUSRTOOL library. You can find additional information about these commands in the TFPADINFO member, within the QATTINFO file, in the QUSRTOOL library.

The user profile that tries to add Monitored Resource Entries must exist on all nodes in the Administrative Domain. In addition, this user profile must have the appropriate authority to execute the APIs on all nodes in the Administrative Domain.

1. To add a resource that you want to be monitored by the Administrative Domain, select **Management Central** → **Cluster** → ***cluster name*** → **Peer Resources** → ***administrative domain name***. Right-click the type of resource that you want to add. In this example, we select **User Profiles**. Then select **Add Monitored Resource Entry** as shown in Figure 6-7.

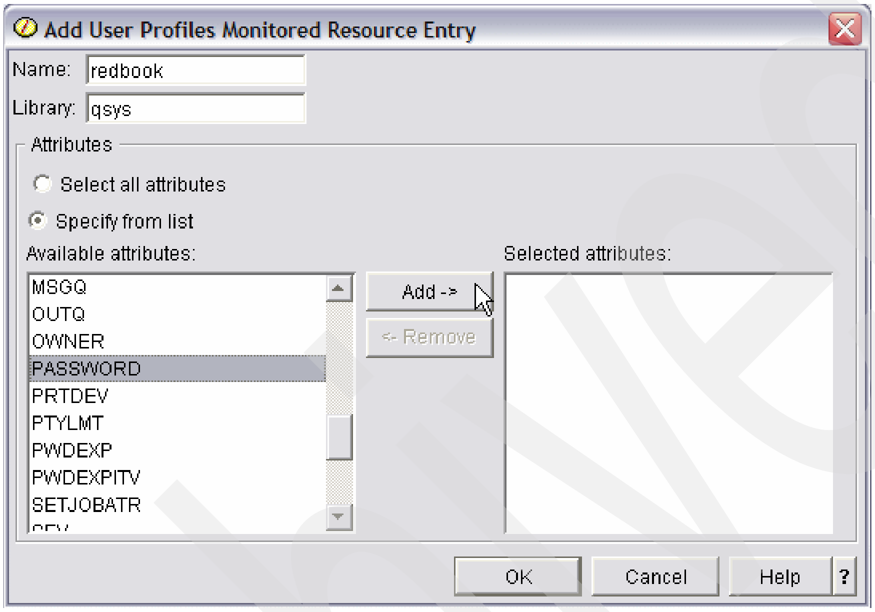


Add a monitored resource entry to the administrative domain.

Figure 6-7 Selecting the Add Monitored Resource Entry option

2. Depending on the object type that you selected, in the window that opens, specify which individual resource you want to monitor. For some resources, such as User Profiles, Job Descriptions, Classes, and ASP Device Descriptions, you can also specify whether you want to monitor and synchronize all (Select all attributes) or just specific attributes for this individual resource (Specify from list).

   Figure 6-8 shows an example of adding the user profile MRE. Note that some attributes, for example text, homedir, and locale, cannot be synchronized using an Administrative Domain of a user profile.


Add User Profiles Monitored Resource Entry

Name: redbook

Library: qsys

Attributes

○ Select all attributes

● Specify from list

Available attributes:

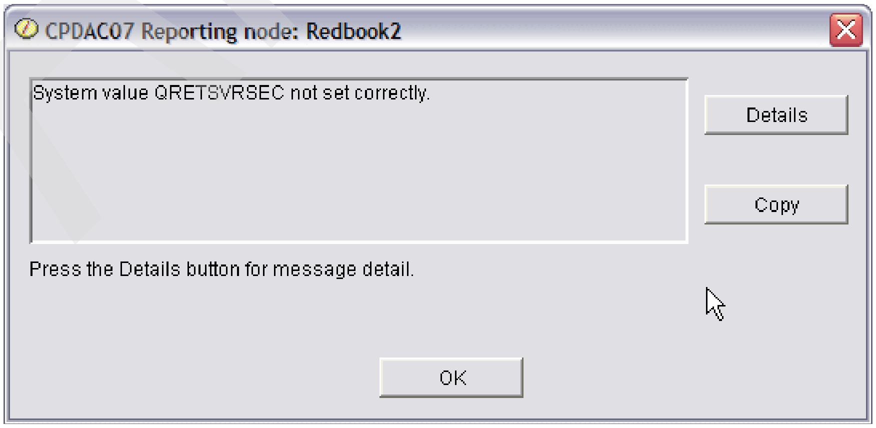
MSGQ
OUTQ
OWNER
PASSWORD
PRTDEV
PTYLMT
PWDEXP
PWDEXPITV
SETJOBATR

Selected attributes:

Figure 6-8 Add User Profiles Monitored Resource Entry window

If you want the Administrative Domain to monitor and synchronize passwords for user profiles, from a command line or in iSeries Navigator, set the system value QRETSVRSEC to 1 (Retain data) on all nodes in the Administrative Domain. Otherwise you receive the error message shown in Figure 6-9.

> **Attention:** You *cannot* synchronize the user profiles, such as QUSER or QSECOFR, that are delivered by IBM by using Admin Domain because it is considered a security risk.


CPDAC07 Reporting node: Redbook2

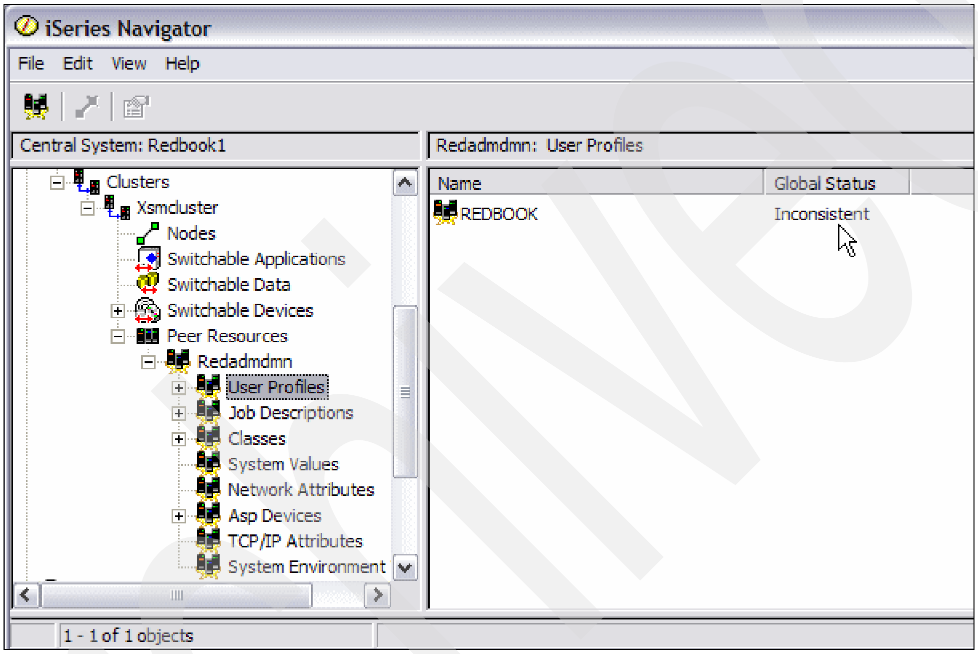
System value QRETSVRSEC not set correctly.

Press the Details button for message detail.

Figure 6-9 Error message for an invalid QRETSVRSEC value

After you add an MRE, a first synchronization occurs. The Administrative Domain uses the concept of a global value for each attribute of an MRE to make sure that resources stay identical on all nodes. It tries to synchronize the MREs to this global value. We explain how this global value is determined in different scenarios in 6.3, “Determining the global value” on page 131.

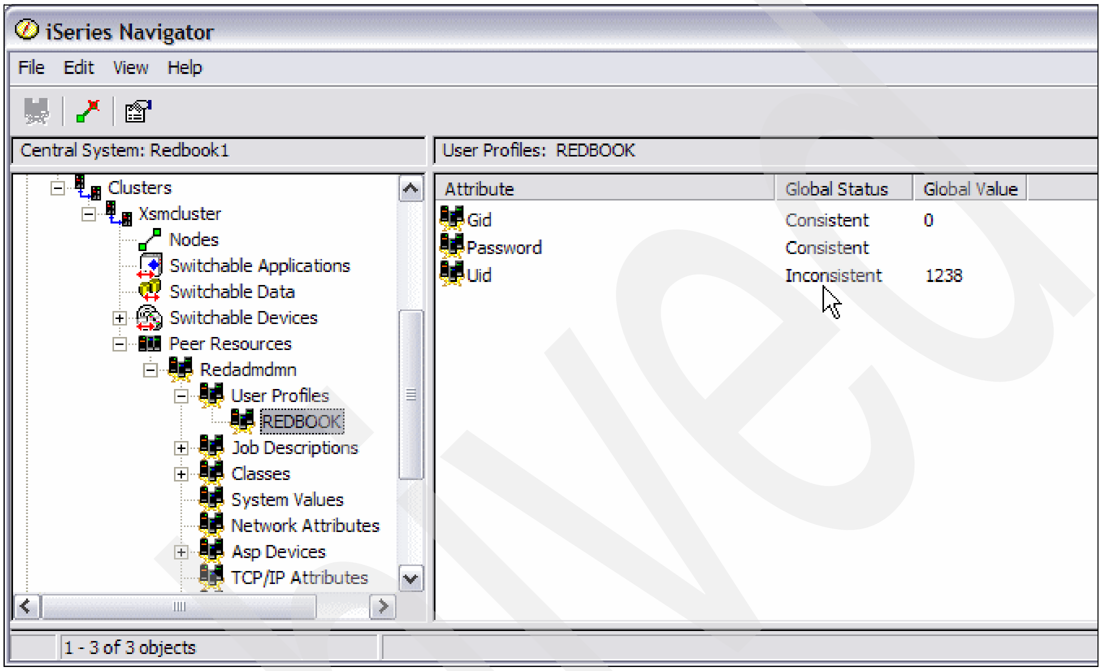
In order to control whether your MREs are consistent across the nodes in the Administrative Domain, you can again use Management Central, the APIs or the CL commands from the QUSRTOOL library. Figure 6-10 shows an example where a user profile that was added as an MRE before shows an *inconsistent* global status. The inconsistent global status is shown on all nodes in the administrative domain, regardless of whether that specific node was synchronized to the global value.



Name	Global Status
REDBOOK	Inconsistent

Figure 6-10   Inconsistent user profile

You can find further information about individual attributes of an MRE and their global status and value by using Management Central as shown in Figure 6-11. In this example, you can see that the Administrative Domain synchronized the GID and the password to their respective global values for all nodes. The global values themselves are also shown here. However, the Administrative Domain was unable to set the UID attribute for the user profile REDBOOK to the global value of 1238 on at least one node in the Administrative Domain.



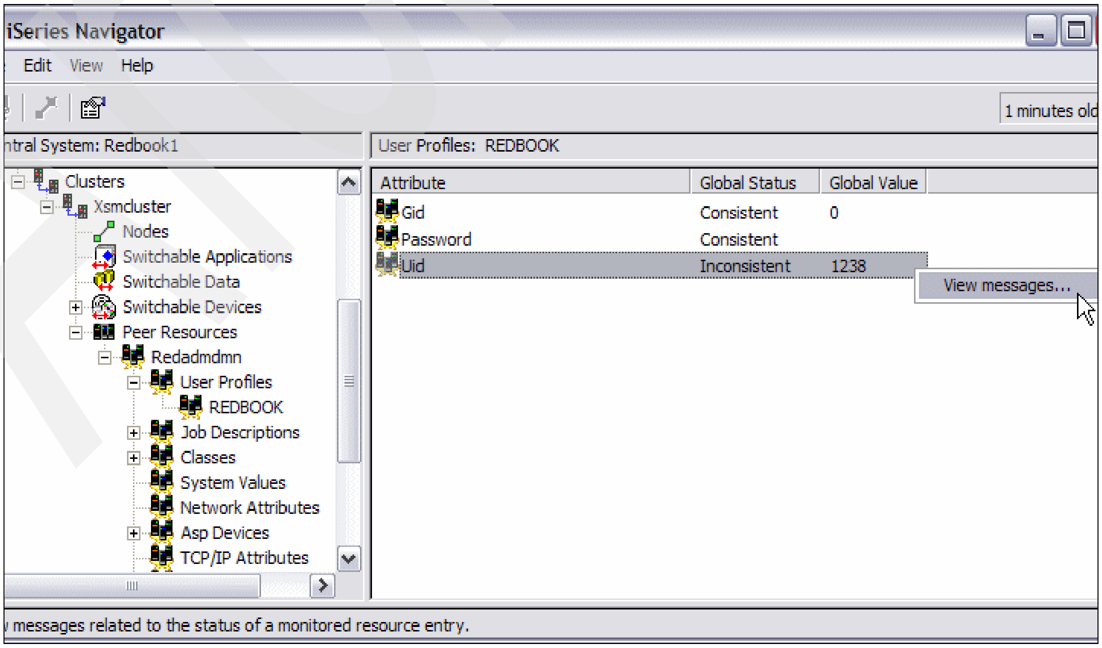
User Profiles: REDBOOK

Attribute	Global Status	Global Value
Gid	Consistent	0
Password	Consistent	
Uid	Inconsistent	1238

1 - 3 of 3 objects

Figure 6-11 Inconsistent user profile attributes

To learn about the reason for this inconsistency, right-click the attribute and select **View messages** as shown in Figure 6-12.



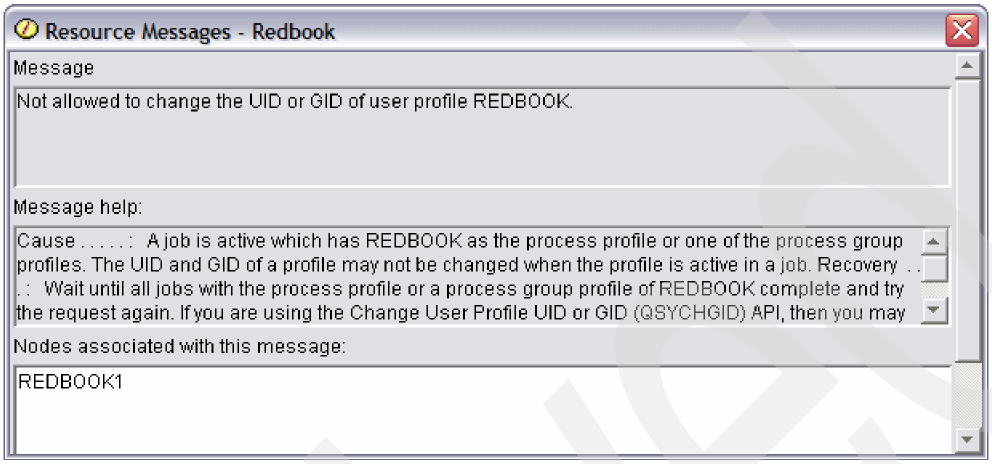
User Profiles: REDBOOK

Attribute	Global Status	Global Value
Gid	Consistent	0
Password	Consistent	
Uid	Inconsistent	1238

messages related to the status of a monitored resource entry.

Figure 6-12 Inconsistent status - View Message

The Resource Messages window opens that gives you the detailed reason for the inconsistency. It also indicates which nodes in the Cluster Administrative Domain cannot be synchronized to the global value. In our example (Figure 6-13), a job was running on the node REDBOOK1 under user profile REDBOOK, which means that the Administrative Domain was unable to change that user profile's UID.



Message

Not allowed to change the UID or GID of user profile REDBOOK.

Message help:

Cause . . . . . : A job is active which has REDBOOK as the process profile or one of the process group profiles. The UID and GID of a profile may not be changed when the profile is active in a job. Recovery . . . : Wait until all jobs with the process profile or a process group profile of REDBOOK complete and try the request again. If you are using the Change User Profile UID or GID (QSYCHGID) API, then you may

Nodes associated with this message:

REDBOOK1

Figure 6-13 Reason for the inconsistent status

Information regarding synchronization failures and their reasons is written to the job log of the peer CRG on the system where the synchronization to the global value failed.

6.3 Determining the global value

In the following paragraphs, we explain different scenarios in regard to the Administrative Domain. The main question for each scenario is: How is the global value for an attribute of an MRE determined?

First synchronization for a new MRE

If you first add a resource to an Administrative Domain using the APIs or the CL commands from the QUSRTOOL library, then this global value is taken from the system where you issued the API or CL command.

If you used Management Central to add the MRE, then the Administrative Domain looks at the recovery domain of the peer CRG. It takes the first system from the recovery domain and populates the global values of the newly added resource with the values from that system.

Adding a new node to an existing Cluster Administrative Domain

If you add a new node to an active Cluster Administrative Domain, by adding the node into the recovery domain of the corresponding peer CRG, then all information about MREs in this Administrative Domain is copied to the new node. The attributes are changed of resources that are represented by the MREs.

Essentially the Administrative Domain overwrites individual values of attributes for monitored resources with the current global values. Resources that do not yet exist on the new node are created automatically. They will be owned by the user profile that issued the ADDMRE command.

Peer CRG corresponding to Administrative Domain is ended

If the peer CRG that corresponds to your Administrative Domain ends while nodes in that domain are still active, then changes to MREs on one node will be pending for all other nodes. All of the monitored resources are considered to be inconsistent because changes to them are not being synchronized. Although changes to monitored resources continue to be tracked, the global value is not changed, and changes are not propagated to the rest of the administrative domain. Any changes that are made to any monitored resource while the Cluster Administrative Domain is inactive are synchronized across all active nodes when the peer CRG is started.

When synchronization is started, you first change each resource with attributes whose values do not match its global value, unless there is a pending change for that resource. A pending change indicates that the attribute was changed on at least one of the nodes while the Administrative Domain was inactive, thereby rendering the global value invalid. Any pending change is distributed to all active nodes in the domain and applied to each affected resource on each node. When the pending changes are distributed, the global value is changed, and the global status of each affected resource is changed to consistent or inconsistent, depending on the outcome of the change operation for the resource on each node. If the affected resource is changed successfully on every active node in the domain, the global status for that resource is consistent. If the change operation failed on any node, the global status is set to inconsistent.

If changes are made to the same resource from multiple nodes while the Cluster Administrative Domain is inactive, all of the changes are propagated to all of the active nodes as part of the synchronization process when the CRG is started. Be aware that, although all pending changes are processed during the activation of the Cluster Administrative Domain, there is no guaranteed order in which the changes are processed. If you make changes to a single resource from multiple cluster nodes while the CRG is inactive, there is no guaranteed order to the processing of the changes during activation.

If you create new resources and add them into the Administrative Domain while the peer CRG is inactive, these resources are created immediately on the other systems in the Administrative Domain. They show a global status of ADDED.

Node in the Administrative Domain is inactive

If you end a cluster node that is part of a Cluster Administrative Domain (by issuing the ENDCLUNOD command), you are still able to make changes to monitored resources on the inactive node. After starting the node again, the changes are resynchronized with the other nodes in the Administrative Domain. All changes from the node that was inactive are applied to the rest of the active nodes in the Administrative Domain, unless changes to the same resource were also made in the active Administrative Domain.

If changes were made to a monitored resource both in the active Administrative Domain and on an inactive node, then the changes that are made in the active Administrative Domain are applied to the node that was inactive and now joins the domain again. This also applies if different attributes where changed on a different system. The Administrative Domain does not trigger that an individual attribute for a resource was changed; it just triggers that the resource was changed.

You might want to end a cluster node that is part of a Cluster Administrative Domain and not allow changes that are made on the inactive node to be propagated back to the active domain when the node is started, for example, when ending the cluster node to do testing on it. In this case, you must remove the node from the Administrative Domain peer CRG before you end the cluster node. You can do this by using the RMVCRGNODE command.

Cluster partitioned status

If a Cluster Administrative Domain is partitioned, changes continue to be synchronized among the active nodes in each partition. When the nodes are merged back together again, the Cluster Administrative Domain propagates all changes that are made in every partition so that the resources are consistent within the active domain. There are several considerations regarding the merge processing for a Cluster Administrative Domain:

- If all partitions were active and changes were made to the same resource in different partitions, the most recent change is applied to the resource on all nodes during the merge. The most recent change is determined using Coordinated Universal Time (UTC) from each node where a change initiated.
- If all partitions were inactive, the global values for each resource are resolved based on the last change that was made while any partition was active. The actual application of these changes to the monitored resources does not happen until the peer CRG that represents the Cluster Administrative Domain is started.
- If some partitions were active and some were inactive prior to the merge, the global values that represent changes that were made in the active partitions are propagated to the inactive partitions. The inactive partitions are then started, causing any pending changes that are made on the nodes in the inactive partitions to propagate to the merged domain.

Save and restore considerations

When you restore a monitored resource on any system that is part of a Cluster Administrative Domain, the resource is resynchronized to the global value that is currently known in the Cluster Administrative Domain when the peer CRG that represents the Cluster Administrative Domain is active.

The RSTLIB, RSTOBJ, RSTUSRPRF, and RSTCFG restore commands result in a resynchronization of system objects. In addition, RSTSYSINF and UPDSYSINF result in a resynchronization of system values and network attributes. In all of these cases, the restored values are overwritten with the current global values. To resynchronize system environment variables after RSTSYSINF or UPDSYSINF, the peer CRG that represents the Cluster Administrative Domain must be ended and started again.

If you want to restore your monitored resources to a previous state, you have to remove the MRE that represents this resource. After restoring the resource, you have to add an MRE for the resource *from the system where the restore was done*. This ensures that the values from the resource that you just restored will be synchronized to the other node in the Administrative Domain.

6.4 Finding changes made by the Administrative Domain

If you want to monitor the changes that were made by the Administrative Domain, you can find information in the job log of the corresponding peer CRG. The job log gives information about the objects that were changed; it does not contain any detailed information about what was changed.

Depending on your setting for the QAUDLVL system, value changes that were made by the Administrative Domain are also shown in the auditing journal. The job that is responsible for doing the change is the peer CRG job.

6.5 APIs for Monitored Resource Entries

The following APIs pertain to Monitored Resource Entries:

- Add Monitored Resource Entry (QfpadAddMonitoredResourceEntry)
- Remove Monitored Resource Entry (QfpadRmvMonitoredResourceEntry)
- Retrieve Monitored Resource Entry (QfpadRtvMonitoredResource Entry)

  You can find a detailed description of these APIs in the i5/OS Information Center on the Web at the following address:

http://publib.boulder.ibm.com/infocenter/iseries/v5r4/index.jsp

6.6 Monitored Resource Entries and the QUSRTOOL commands

The QUSRTOOL library provides three unsupported commands (see the following list) and their sources. You can use them to manage Monitored Resource Entries from a 5250 session or within a CL program.

- ADDMRE
- RMVMRE
- PRTMRE

To create these commands:

1. Enter the following command:

   ```
   CALL QUSRTOOL/UNPACKAGE ('*ALL      ' 1)
   ```

2. Unpack QUSRTOOL.
3. Apply PTF SI21486. If the PTF was applied before, you have to remove it and then re-apply it.
4. Create a library to hold the commands and their command processing programs after you create them:

   ```
   CRTLIB LIB(name)
   ```

   Here *name* is a meaningful name to you. This library holds the command and their command processing programs after you create them.

5. Enter the following command, where *name* is the one of the three programs from QUSERTOOL:

   ```
   CRTCLPGM PGM(name/TFPADCRT) SRCFILE(QUSRTOOL/QATTCL)
   ```

6. Enter the following command, where *name/* is the library that you created, and *name* is the program to call:

   ```
   CALL name/TFPADCRT name
   ```

The PRTMRE command allows you to specify RSCTYPE(*ALL) RESOURCE(*ALL) to generate one report on all Monitored Resource Entries in your Administrative Domain.

> **Important:** Be careful when using the commands. They currently post only error messages to the job log, which might give the impression that everything worked correctly when it did not.

7

Site object tracking

In this chapter, we discuss the new tracking options available with i5/OS V5R4. These options are Source Site Tracking and Target Site Tracking.

Source Site Tracking became available with i5/OS V5R4 in mid 2006. Target Site Tracking became available via a PTF in 2007.

7.1 Source Site Tracking

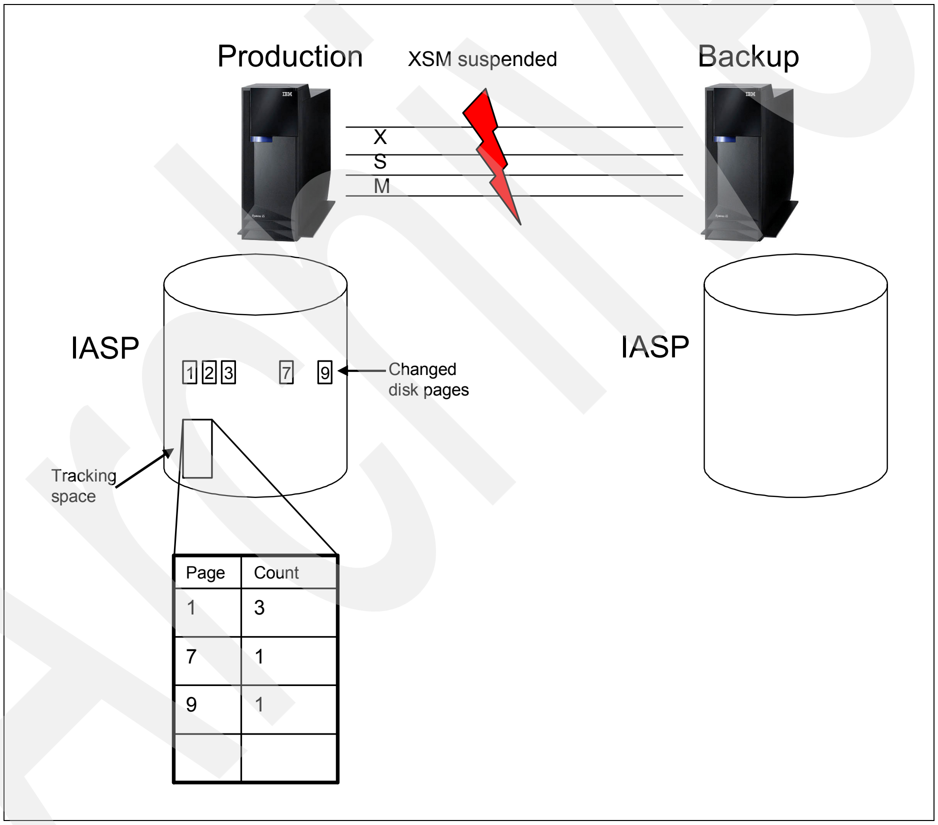
Before the availability of Source Site Tracking with V5R4, you had to fully re-synchronize your backup independent auxiliary storage pool (IASP) to your production IASP. That is you had to send all the data in the IASP over your communication lines. You had to do this whenever the production system was unable to send changes to the backup system, which might occur for the following reasons:

- You needed to take down the backup system, for example to install PTFs.
- The communication link used by cross-site mirroring (XSM) to send data to the backup system was not available.

Source Site Tracking helps to eliminate the need to fully re-synchronize from the production copy to the backup copy in cases where XSM is *suspended*.

7.1.1 How Source Site Tracking works

When you configure Source Site Tracking, a table is created that is used to track which storage pages were changed in your IASP while XSM was suspended. See Figure 7-1.



Page	Count
1	3
7	1
9	1

Figure 7-1 Source Site Tracking

This table does not hold any changed data; it just contains two columns that hold the following data:

- The page number of the first changed page
- The number of pages changed

For example, say that your application changes data in storage pages one, two, and three. This results in one entry in the tracking table that indicates that, starting from page one, three pages were changed. Then when your application changes data on page five, another entry is posted into the table indicating that, starting from page five, one page was changed.

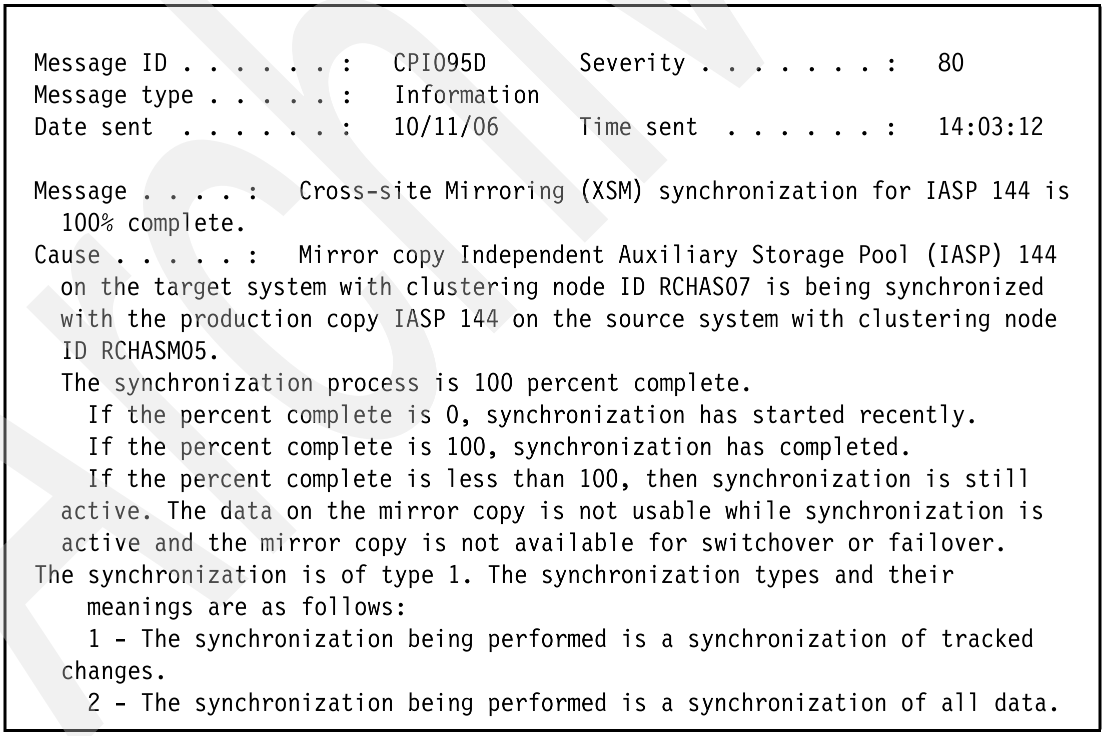
The mechanism that is used to track data in that table is also built in a way that it compresses data in the table whenever possible. If your application changes storage page one, then page three and then page two, the final result of this is still one entry in the tracking table.

The tracking table requires 16 bytes for each entry. Most space in that table is required if your application changes every second page in your IASP.

When XSM is suspended automatically, for example because the lines used were down, the tracking space is automatically used when it is defined. When a user suspends XSM, for example because the user wants to power down the backup system, then the user can choose between *suspend with tracking* and *suspend without tracking*.

If XSM was suspended with tracking and is resumed again, all tracked pages marked in the tracking table are sent to the backup system. While this happens, users can still work normally on the production system. Be aware that you are running unprotected because you cannot switch or fail over to the backup system until synchronization of the tracked data has finished. While synchronization is occurring, your data on the backup system is not in a valid state.

As with full resynchronization, messages are posted to the QSYSOPR message queue to indicate the progress of the partial synchronization. This message also indicates whether a full or a partial resynchronization has taken place as shown in Figure 7-2.



```
Message ID . . . . . . :   CPI095D      Severity . . . . . . . :   80
Message type . . . . . :   Information
Date sent  . . . . . . :   10/11/06     Time sent  . . . . . . :   14:03:12

Message . . . . :   Cross-site Mirroring (XSM) synchronization for IASP 144 is
  100% complete.
Cause . . . . . :   Mirror copy Independent Auxiliary Storage Pool (IASP) 144
  on the target system with clustering node ID RCHAS07 is being synchronized
  with the production copy IASP 144 on the source system with clustering node
  ID RCHASM05.
  The synchronization process is 100 percent complete.
    If the percent complete is 0, synchronization has started recently.
    If the percent complete is 100, synchronization has completed.
    If the percent complete is less than 100, then synchronization is still
  active. The data on the mirror copy is not usable while synchronization is
  active and the mirror copy is not available for switchover or failover.
The synchronization is of type 1. The synchronization types and their
    meanings are as follows:
    1 - The synchronization being performed is a synchronization of tracked
  changes.
    2 - The synchronization being performed is a synchronization of all data.
```

Figure 7-2 Message indicating that synchronization is finished

If your tracking table is not large enough to record all pages changed while XSM is suspended, then a full resynchronization is required to get the production system in synch with the backup system again. However, this is largely unlikely to happen. Looking at the worst case scenario for data changes, for example every second page of the disk is changed while in suspended mode, then the space that is needed to track these changes is calculated by using the following equation:

```
(disk space / page size) / 2 x 16 bytes
```

Looking at a 70 GB drive, this produces the following results:

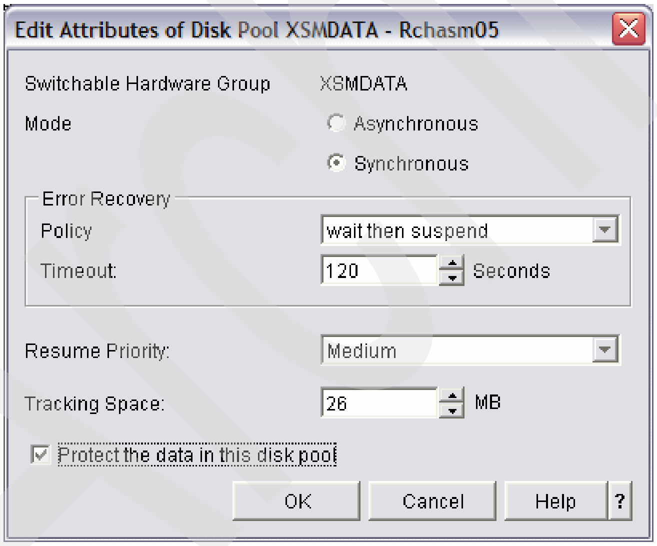
```
(70 x 1024 x 1024 x 1024) / 4096 / 2 x 16 = 146,800,640 bytes = 143,360 KB = 140 MB
```

The result of 140 MB is less than 0.2% of your entire disk space. In addition to this, each pair of object creations and deletions also requires a 16-byte entry in the tracking table.

7.1.2 Configuration of the Tracking Space

Source Site Tracking can be configured easily while setting up your XSM environment. Looking at the attributes that can be defined for XSM, starting with V5R4 of i5/OS, there is a new variable called *Tracking Space*. See Figure 7-3.

Tracking Space is the amount of space (in MB) in the disk pool that the system can use to track changes to the disk pool when geographic mirroring is suspended on the disk pool. The space that you define here is reserved by the system and cannot be used for anything else. The default value is 0.05% of the size of the disk pool. The possible value for this field ranges from 0 to 1% of the size of the disk pool. If the tracking space is set to 0, then the tracking is not used. The range of possible values for the size of the tracking table shows again that its implementation is effective and requires a minimum amount of space. Use the up and down arrows to adjust the amount of MBs in the disk pool that you want to devote to tracking changes to the disk pool when geographic mirroring is suspended on the disk pool.



Edit Attributes of Disk Pool XSMDATA - Rchasm05

Switchable Hardware Group: XSMDATA

Mode: ○ Asynchronous ● Synchronous

Error Recovery

Policy: wait then suspend

Timeout: 120 Seconds

Resume Priority: Medium

Tracking Space: 26 MB

☑ Protect the data in this disk pool

Figure 7-3 Configure Source Site Tracking

If you want to change the Tracking Space value at a later time, you *must* vary off your IASP to do this. Then you can make the change by using iSeries Navigator and selecting **Configuration and Service** → **Hardware** → **All Hardware** → **Disk Units** → **Disk Pool**. Right-click the **disk pool** that you want to change and select **Geographic Mirroring** → **Change Attributes** as shown in Figure 7-4. You see the same window as during the initial setup of the IASP as shown in Figure 7-3.

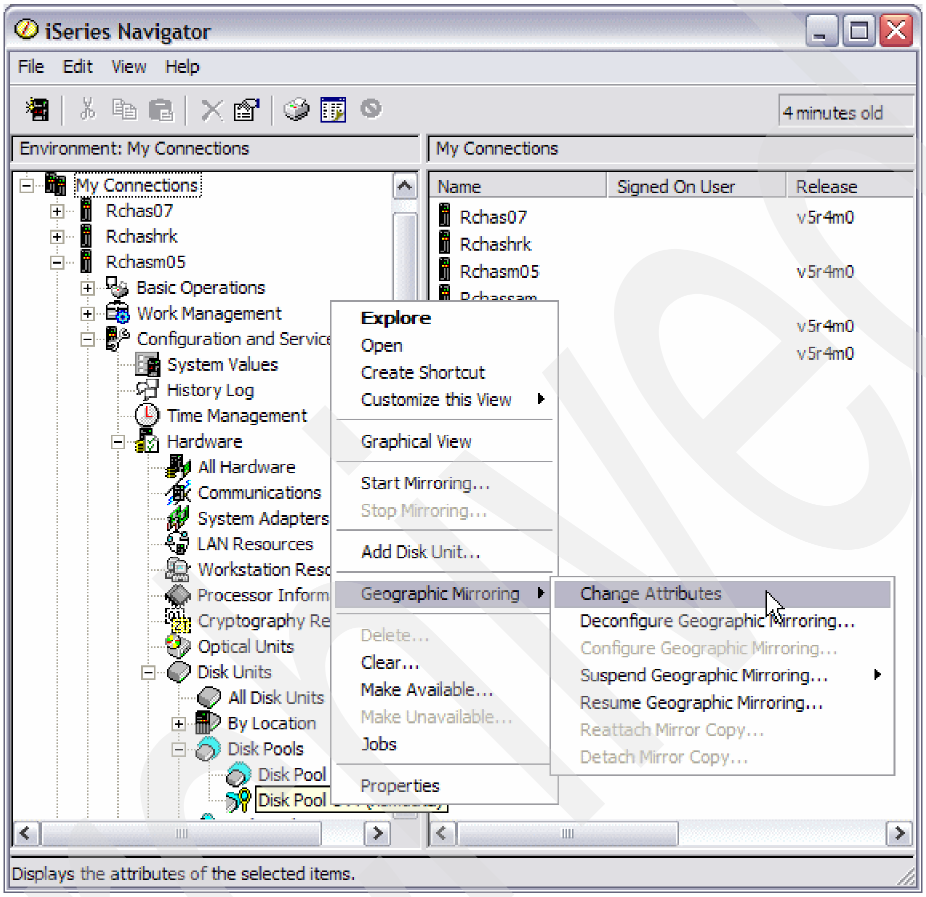


Figure 7-4 Changing the attributes for an existing IASP

7.1.3 Positioning

We recommend that you use Source Site Tracking if you run XSM with V5R4 because it minimizes the need for doing full resynchronizations. Remember that, while resynchronization is running, you are not able to switch over or fail over to your backup system. Therefore, you want to minimize the time required to do this. Additionally the small amount of disk space that is required to use Source Site Tracking should also have little to no impact on your production environment.

Source Site Tracking *only* helps with situations where XSM in suspended. While in suspended status, you are *not* able to access data in the IASP from the backup site.

7.2 Target Site Tracking

Target Site Tracking was introduced with the PTF MF40053 for i5/OS V5R4. Target Site Tracking, together with Source Site Tracking, allows you to perform the following actions:

- Detach XSM
- Vary on the backup copy on the backup system
- Use the copy on the backup system for any activity that you want, such as doing a backup or performing tests
- Re-attach XSM
- Have the systems resynchronize to each other without sending all your IASP data over your communication lines

Be aware that changes made to the backup copy of your data while in detached mode will be overwritten from the production site.

7.2.1 How Target Site Tracking works

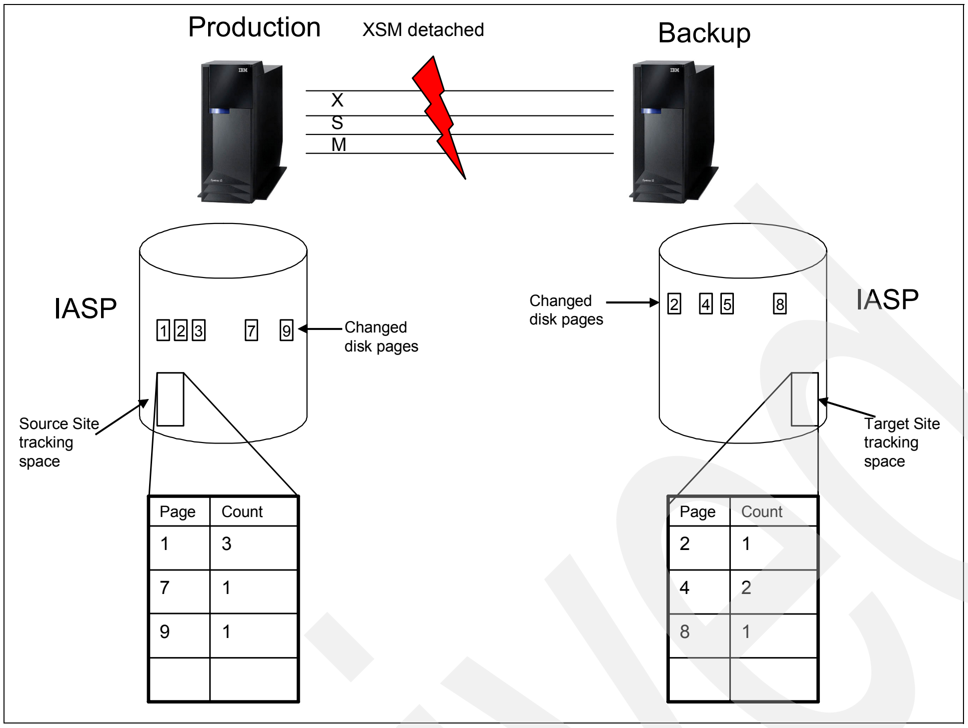
Target Site Tracking reserves some amount of disk space on your backup system to do tracking for changed pages. This is done in a way that is similar to the method that is described in 7.1.1, “How Source Site Tracking works” on page 136.

With the current implementation of Target Site Tracking (Figure 7-5), you need to use the following steps when you want to access data in the IASP of your backup system with Target Site Tracking enabled:

1. Vary off the IASP on your production system.
2. Detach XSM with tracking.
3. Vary on the IASP on your production system and on your backup system.
4. Perform any actions that you want on the backup system.
5. When you are finished with the activities on the backup system, re-attach XSM. You do *not* have to vary off your production IASP again to accomplish this.

While the IASP is detached with Target Site Tracking, the following actions occur:

- On the production site, Source Site Tracking keeps track of the changes that are made to the data in the IASP on that system.
- On the backup site, at the same time, Target Site Tracking keeps track of changes that are made to data in the IASP on the backup system.



Page	Count
1	3
7	1
9	1

Page	Count
2	1
4	2
8	1

Figure 7-5 Target Site Tracking

After XSM is re-attached, these two tables are logically combined. Pages that were changed on the backup system are overwritten with the content from the production copy, thereby eliminating all changes that were made on the backup system while it was detached. Also, all changes that were made to the production copy of the data are sent to the backup system. While this happens, users can work with the applications normally.

7.2.2 Configuration

If you have PTF MF40053 installed, then the size of your Source Site Tracking space is also the size of your Target Site Tracking space.

7.2.3 Positioning and risks

Carefully consider whether you want to access data on your backup copy of the IASP. In a two-node environment, the environment is unprotected from the moment that you detach XSM until the resynchronization process is finished after the re-attach.

At the point in time when you detach XSM, you are in a situation where your backup copy does contain a valid set of data. If you do not make any changes to that data, for example, you do read-only transactions or save operations, then you can still use this data for production purposes should your production system fail while detached. However, you will lose all updates that were made on the production system after XSM was detached.

As soon as you re-attach XSM, then the partial synchronization starts overwriting pages that were changed on the backup system with data from the production system as well as sending changes that were made to pages on the production system while detached. Then the IASP on your backup system goes into an unusable state. This means that until the partial synchronization is finished, you are unable to switch over or fail over to your backup system. The length of this time frame depends on the speed of the communication links that are used for XSM. You must consider whether that is acceptable in your environment. If this is not acceptable, than consider using one of the journal-based High Availability Business Partner (HABP) solutions.

8

Operational considerations of cross-site mirroring

While cross-site mirroring (XSM) can be used to automate some tasks that are normally done by the system administrator, the process still needs to be maintained to ensure XSM is successful. In this chapter, we describe the daily routines and provide precautionary measures to take when implementing XSM. We explain how planned or unplanned actions on the cluster, device cluster resource group (CRG), independent auxiliary storage pool (IASPs), system, or communications pathways can affect XSM in your production environment.

8.1 Daily maintenance

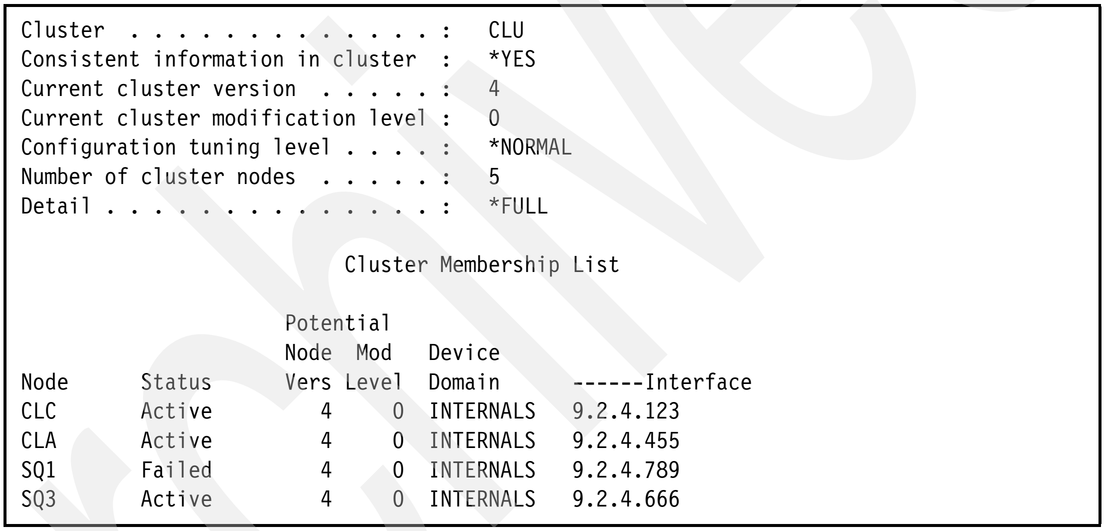
In this section, we discuss the recommended routine actions that a system administrator performs on a daily basis to help ensure that XSM is running successfully. We do not provide a comprehensive list, but include some of the most important tasks.

8.1.1 Cluster node status

If cluster nodes are not active, and participating in XSM, you will have problems, especially if there is a problem with a node that owns the production or mirror copy of the IASP. The best way to view the status of these nodes is to use the following command:

```
DSPCLUINF CLUSTER(cluster_name) DETAIL(*FULL)
```

It is most important that you check Display Cluster Information panel (Figure 8-1) for consistency of the information in the cluster. If the information is consistent, then you must view the status of each node in the cluster to make sure each one is active. If any node is in a Failed or Inactive status, you can try to start the node or investigate further.



```
Cluster  . . . . . . . . . . . . . :   CLU
Consistent information in cluster  :   *YES
Current cluster version  . . . . . :   4
Current cluster modification level :   0
Configuration tuning level . . . . :   *NORMAL
Number of cluster nodes  . . . . . :   5
Detail . . . . . . . . . . . . . . :   *FULL

                           Cluster Membership List

                      Potential
                      Node  Mod   Device
Node      Status      Vers Level  Domain      ------Interface
CLC       Active         4     0  INTERNALS   9.2.4.123
CLA       Active         4     0  INTERNALS   9.2.4.455
SQ1       Failed         4     0  INTERNALS   9.2.4.789
SQ3       Active         4     0  INTERNALS   9.2.4.666
```

Figure 8-1 Display Cluster Information (DSPCLUINF) panel

8.1.2 Device CRG status

Make sure the nodes in the recovery domain of the device CRG are active as well. You can view this information by entering the following command:

```
DSPCRGINF CLUSTER(cluster_name) CRG(device_CRG_name)
```

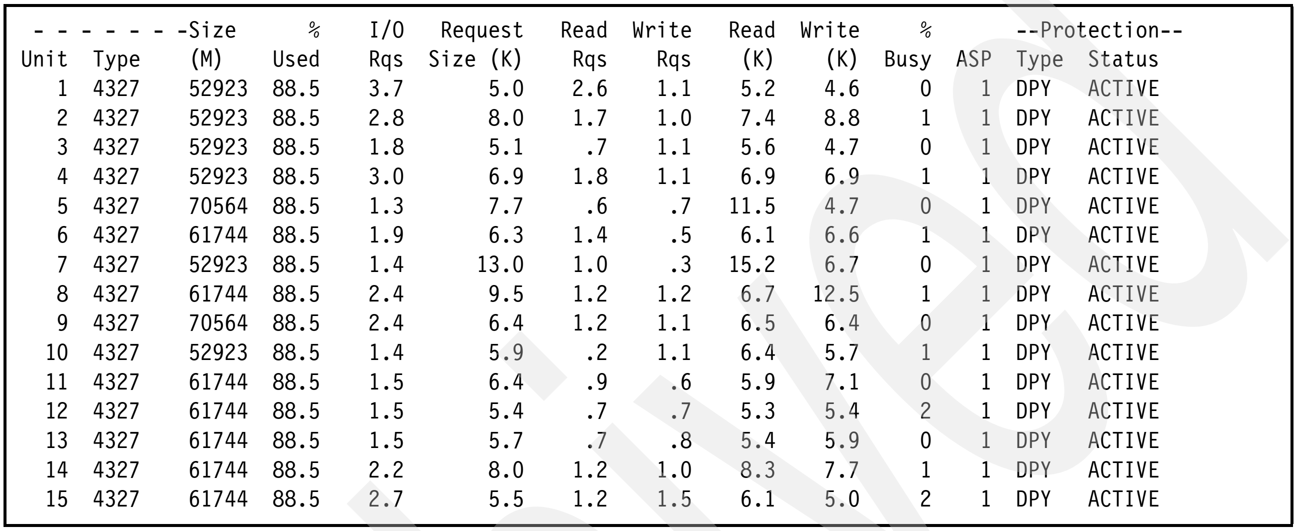
Press Enter to page through the different information panels. The last panel provides information about the recovery domain. If any nodes show a status of Ineligible, the node is not eligible to own the production copy or mirror copy of the IASP.

8.1.3 IASP status

The amount of disk that is being used in the IASP must not be too close to being full on the production or mirror copy. To check this amount, enter the following command on each system:

```
WRKDSKSTS OUTPUT(*PRINT)
```

This Work with Disk Status (WRKDSKSTS) command generates a spooled file, called QPWCDSKS, in your output queue. It is easier to view this file in the output queue than viewing the display, because you can see the ASP number that is associated with each DASD unit, as shown in Figure 8-2. Look for the units that are associated with the IASPs and make sure that the % Used value does not approach 100%. If this happens on your production system, then you will be unable to write any more data to the IASP because an IASP *cannot* overflow to the system ASP. Your applications basically stop running. If the mirrored copy of an IASP becomes full while the production copy still has space left, then XSM is suspended.



```
 - - - - - - -Size      %    I/O   Request   Read  Write   Read  Write     %       --Protection--
Unit  Type    (M)    Used    Rqs  Size (K)    Rqs    Rqs    (K)    (K)  Busy  ASP  Type  Status
   1  4327    52923  88.5    3.7       5.0    2.6    1.1    5.2    4.6     0    1  DPY   ACTIVE
   2  4327    52923  88.5    2.8       8.0    1.7    1.0    7.4    8.8     1    1  DPY   ACTIVE
   3  4327    52923  88.5    1.8       5.1     .7    1.1    5.6    4.7     0    1  DPY   ACTIVE
   4  4327    52923  88.5    3.0       6.9    1.8    1.1    6.9    6.9     1    1  DPY   ACTIVE
   5  4327    70564  88.5    1.3       7.7     .6     .7   11.5    4.7     0    1  DPY   ACTIVE
   6  4327    61744  88.5    1.9       6.3    1.4     .5    6.1    6.6     1    1  DPY   ACTIVE
   7  4327    52923  88.5    1.4      13.0    1.0     .3   15.2    6.7     0    1  DPY   ACTIVE
   8  4327    61744  88.5    2.4       9.5    1.2    1.2    6.7   12.5     1    1  DPY   ACTIVE
   9  4327    70564  88.5    2.4       6.4    1.2    1.1    6.5    6.4     0    1  DPY   ACTIVE
  10  4327    52923  88.5    1.4       5.9     .2    1.1    6.4    5.7     1    1  DPY   ACTIVE
  11  4327    61744  88.5    1.5       6.4     .9     .6    5.9    7.1     0    1  DPY   ACTIVE
  12  4327    61744  88.5    1.5       5.4     .7     .7    5.3    5.4     2    1  DPY   ACTIVE
  13  4327    61744  88.5    1.5       5.7     .7     .8    5.4    5.9     0    1  DPY   ACTIVE
  14  4327    61744  88.5    2.2       8.0    1.2    1.0    8.3    7.7     1    1  DPY   ACTIVE
  15  4327    61744  88.5    2.7       5.5    1.2    1.5    6.1    5.0     2    1  DPY   ACTIVE
```

Figure 8-2 Work with Disk Status (WRKDSKSTS) panel

8.1.4 User profile maintenance

User profiles cannot be kept in an IASP and, therefore, must be maintained separately in the system ASP on both source and target. This is probably not a daily task, but it needs to be managed. If you are running on V5R4 of i5/OS, you can use the Administrative Domain to support you in this task.

Make sure that the user IDs (UID) and group IDs (GID) of users that own objects on the IASPs are the same on both the source and target systems. Because ownership of objects is internally represented by the UID and GID values, mismatches of these values for a user profile can result in longer switchover and failover times. This happens because, during the vary on of an IASP, the system has to touch all objects that are owned by user profiles with mismatches in their UID or GID and change these values within all objects. This process can be lengthy if a large number of objects is involved.

A difference in the authorities can limit the capabilities of successfully running XSM or can prevent successful switchovers or failovers. Adopted authorities cannot be used, because adopted authority only affects the user profile on one node at a time.

8.1.5 Error messages

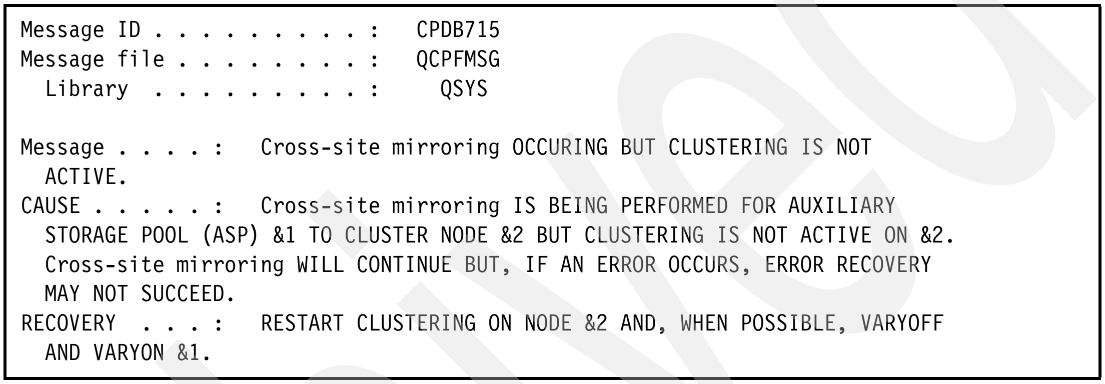
As with any daily maintenance for processes that run on the System i5 platform, it is important to check for error messages. Problems that are related to clustering or XSM cause error messages to be posted in the history log (QHST) or the system operator message queue (QSYSOPR). Check for messages daily and reply to them as appropriate.

8.2 Clustering and IASP considerations

In this section, we discuss operational and runtime considerations that you must be aware of while using XSM. The topics in this section are specific to actions that you might take related to the cluster or IASPs and how those actions affect XSM.

8.2.1 Ending clustering

XSM cannot recover from a communications failure without clustering. XSM relies on cluster messaging services to determine whether there is a communications problem and whether a failover is required. If clustering is ended with XSM active, error message CPDB715 is posted in the QSYSOPR message queue, as shown in Figure 8-3.



```
Message ID . . . . . . . . . :   CPDB715
Message file . . . . . . . . :   QCPFMSG
  Library  . . . . . . . . . :     QSYS

Message . . . . :   Cross-site mirroring OCCURING BUT CLUSTERING IS NOT
  ACTIVE.
CAUSE . . . . . :   Cross-site mirroring IS BEING PERFORMED FOR AUXILIARY
  STORAGE POOL (ASP) &1 TO CLUSTER NODE &2 BUT CLUSTERING IS NOT ACTIVE ON &2.
  Cross-site mirroring WILL CONTINUE BUT, IF AN ERROR OCCURS, ERROR RECOVERY
  MAY NOT SUCCEED.
RECOVERY  . . . :   RESTART CLUSTERING ON NODE &2 AND, WHEN POSSIBLE, VARYOFF
  AND VARYON &1.
```

Figure 8-3 Clustering error with XSM

As a general rule, do not end clustering on nodes while XSM is active. Ending clustering on nodes that own either the production copy or mirror copy of the IASP can have the following results:

- Ending clustering for the node that owns the production copy or mirror copy of the IASP, while the CRG is active causes failover, based on the order in the recovery domain.
- Ending clustering for the node that owns the production copy or mirror copy of the IASP, when failover cannot occur, either because the cluster resource group is inactive or because there is no other active node that can own either copy, suspends XSM for that IASP.

If you end clustering inadvertently, you should restart clustering, ensure that the CRG jobs are active, and make the IASP available (vary it on) at your first opportunity.

8.2.2 Suspend and resume

Remember that suspending XSM affects only the IASP that is specified and not the entire ASP group. If you temporarily suspend XSM while the IASP is active, a complete resynchronization is required when XSM resumes, unless you are using V5R4 and suspended XSM *with tracking*.

If you vary off an ASP group while XSM is synchronizing, then synchronization resumes at the point at which it was suspended when you vary on the ASP group again. The messages in the QSYSOPR message queue that indicate how much of the synchronization process has completed start with a 0% value.

Remember if you suspend XSM, users and applications cannot use the data in the mirror copy. This leaves one copy, the production copy, of the IASP available for use.

8.2.3 Detach and reattach

A detach operation affects the entire ASP group. If you are using XSM and want to detach the mirror copy of the IASP to perform save operations or data mining, you must realize that a complete resynchronization is required when the ASP group is reattached, unless you are running with Target Site Tracking. A full resynchronization can be a lengthy process. Also remember that while the mirror copy is detached, the production copy is left without a node to which a switchover or failover can occur.

Before varying on the detached mirror copy, we recommend that you create a second, unique device description for the IASP to avoid having two instances of the same database in the network. For more information about attach and detach and vary on, refer to the i5/OS Information Center on the Web at the following address:

http://publib.boulder.ibm.com/infocenter/iseries/v5r4/index.jsp

8.2.4 Varying on and varying off an ASP group

Remember that the scope of the VRYCFG command is for the entire ASP group. This command is performed against the device description for the ASP group.

When you specify the devices to use in the CRTCRG command, you have the option to specify the value *ONLINE for the “Configuration object online” portion of the CFGOBJ parameter. By choosing *ONLINE, the system automatically varies on the ASP group as part of a failover or switchover. Therefore, you do not have to issue the VRYCFG command.

On rare occasions, it is possible that an XSM problem occurs during the vary-on process, and the system suspends XSM and completes the vary-on process. While XSM is suspended, a switchover or failover to the mirror copy of the IASP is prohibited because the mirror copy contains back-level data. If the production copy cannot be used, and a switchover or failover is prohibited, you can change the order of the nodes in the recovery domain. You can change the order so that the backup node that owns the back-level mirror copy of the IASP becomes the primary node and owns what becomes the production copy of the IASP.

If XSM is suspended for some IASPs in the ASP group, but not all of the IASPs in the ASP group, you cannot convert the mirror copy of the IASP into a production copy even by changing the order of the recovery domain nodes. If XSM is suspended for all of the IASPs in the ASP group, but it is suspended at different times, do not try to convert the mirror copy of the IASP to a production copy, because the data in the different IASPs is inconsistent.

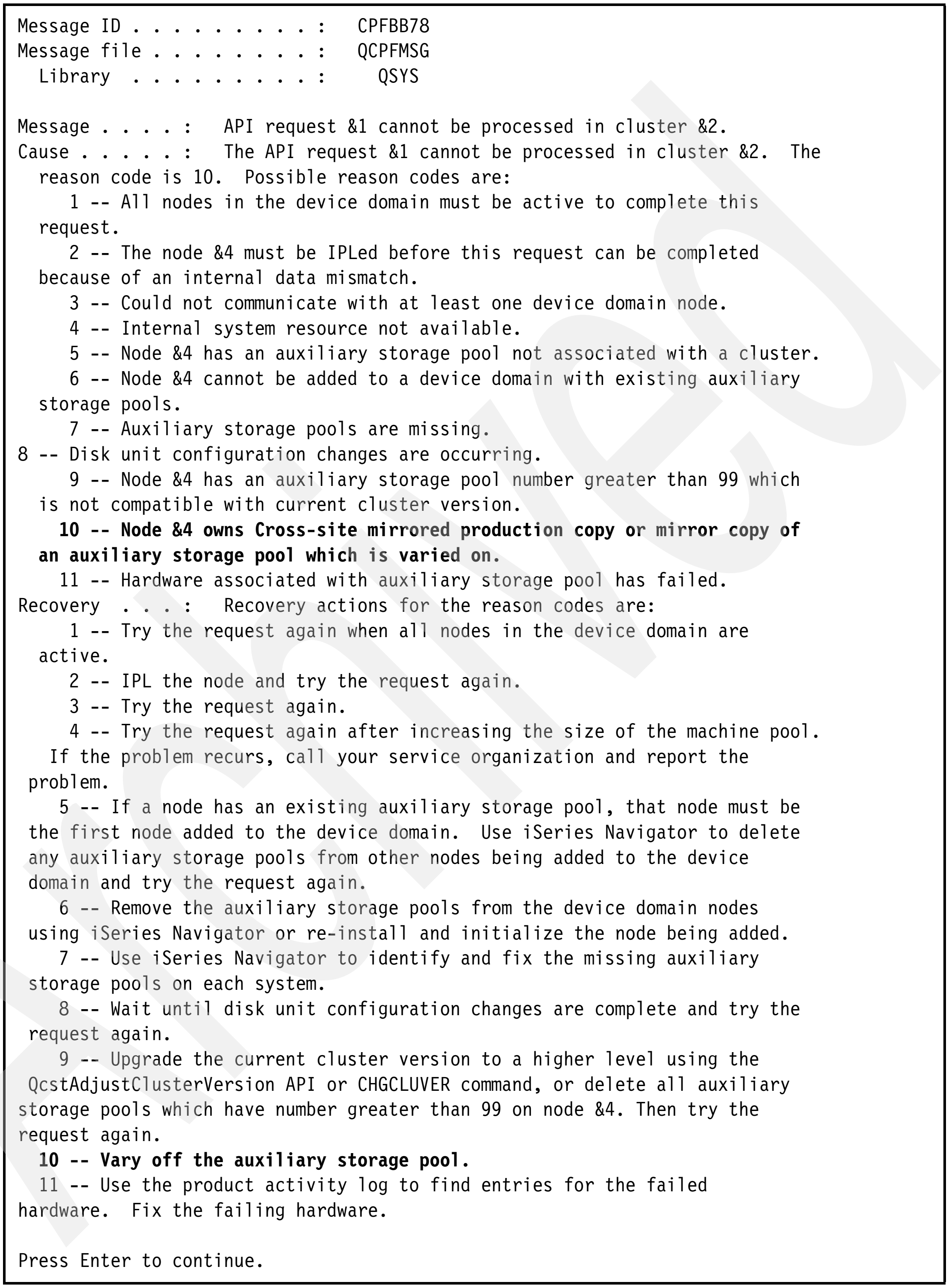
During a switchover or failover, if the vary-on process for the ASP device description fails, clustering services attempts to “undo” the operation and have the original primary node maintain ownership of the production copy of the IASP. This might be one reason to configure device objects with the CFGOBJ(*OFFLINE) command so that the switchover or failover is successful. You can manually vary on the ASP device separately, after the vary-on problem is corrected.

8.2.5 Removing nodes in a cluster

In general, if you do not need to remove a node from a cluster at any time, it is best not to do so. Regardless of whether XSM is active, there are negative effects to removing a node from the cluster.

Removing nodes while cross-site mirroring is active

Do not remove a node from the CRG, device domain, or cluster when that node is used as the primary owner of the production copy of the IASP or the mirror copy of the IASP, and XSM is active. If you try to remove a node from a device domain while XSM is active, the remove operation fails with the error message CPFBB78, reason code 10, as shown in Figure 8-4.



```
Message ID . . . . . . . . . :   CPFBB78
Message file . . . . . . . . :   QCPFMSG
  Library  . . . . . . . . . :     QSYS

Message . . . . :   API request &1 cannot be processed in cluster &2.
Cause . . . . . :   The API request &1 cannot be processed in cluster &2.  The
  reason code is 10.  Possible reason codes are:
     1 -- All nodes in the device domain must be active to complete this
  request.
     2 -- The node &4 must be IPLed before this request can be completed
  because of an internal data mismatch.
     3 -- Could not communicate with at least one device domain node.
     4 -- Internal system resource not available.
     5 -- Node &4 has an auxiliary storage pool not associated with a cluster.
     6 -- Node &4 cannot be added to a device domain with existing auxiliary
  storage pools.
     7 -- Auxiliary storage pools are missing.
8 -- Disk unit configuration changes are occurring.
     9 -- Node &4 has an auxiliary storage pool number greater than 99 which
  is not compatible with current cluster version.
    10 -- Node &4 owns Cross-site mirrored production copy or mirror copy of
  an auxiliary storage pool which is varied on.
    11 -- Hardware associated with auxiliary storage pool has failed.
Recovery  . . . :   Recovery actions for the reason codes are:
     1 -- Try the request again when all nodes in the device domain are
  active.
     2 -- IPL the node and try the request again.
     3 -- Try the request again.
     4 -- Try the request again after increasing the size of the machine pool.
   If the problem recurs, call your service organization and report the
 problem.
    5 -- If a node has an existing auxiliary storage pool, that node must be
 the first node added to the device domain.  Use iSeries Navigator to delete
 any auxiliary storage pools from other nodes being added to the device
 domain and try the request again.
    6 -- Remove the auxiliary storage pools from the device domain nodes
 using iSeries Navigator or re-install and initialize the node being added.
    7 -- Use iSeries Navigator to identify and fix the missing auxiliary
 storage pools on each system.
    8 -- Wait until disk unit configuration changes are complete and try the
 request again.
    9 -- Upgrade the current cluster version to a higher level using the
 QcstAdjustClusterVersion API or CHGCLUVER command, or delete all auxiliary
storage pools which have number greater than 99 on node &4. Then try the
request again.
  10 -- Vary off the auxiliary storage pool.
  11 -- Use the product activity log to find entries for the failed
hardware.  Fix the failing hardware.

Press Enter to continue.
```

Figure 8-4   Error message CPFBB78

In order to remove the node from the cluster, you must vary off the IASP.

Removing nodes while cross-site mirroring is inactive

If you remove a node that owns either the production copy or mirror copy of the IASP when you are *not* performing XSM, then you will not be able to perform XSM using that node. In addition, you will not be able to add the node back into the device domain without deleting IASPs and IPLing the node.

8.2.6 DASD capacity of IASP copies

The disk capacity of the production copy and the mirror copy of the IASP should be about the same. If you exceed the capacity of the mirrored copy, the system suspends XSM.

If one copy has more capacity than the other, we recommend that you set the threshold so that the copy with the lowest capacity determines the threshold value. For example, if the copy with the lowest capacity is 100 GB, and the threshold is set to 90%, and the copy with the higher capacity is 200 GB, then you would want to set the threshold to 45% on the higher copy. That way, the threshold is met on both copies at the same time.

8.2.7 Using RCLSTG on an IASP

If you perform a Reclaim Storage (RCLSTG) command on one IASP in an IASP group, this task occurs on each IASP in the IASP group. In order to perform a RCLSTG on an IASP, you must end all jobs that use the IASP first.

8.2.8 ASP balancing for an IASP

You can use the ASP balancing function, using the STRASPBAL command, to balance the capacity, usage, hierarchical storage management (HSM), and data in an IASP. With an IASP being synchronized through XSM, this function balances only the copy of the IASP for the system on which the command is issued. This command can be issued on the system that owns the mirror copy of the IASP, without promoting the mirror copy to the production copy.

8.3 Synchronization considerations

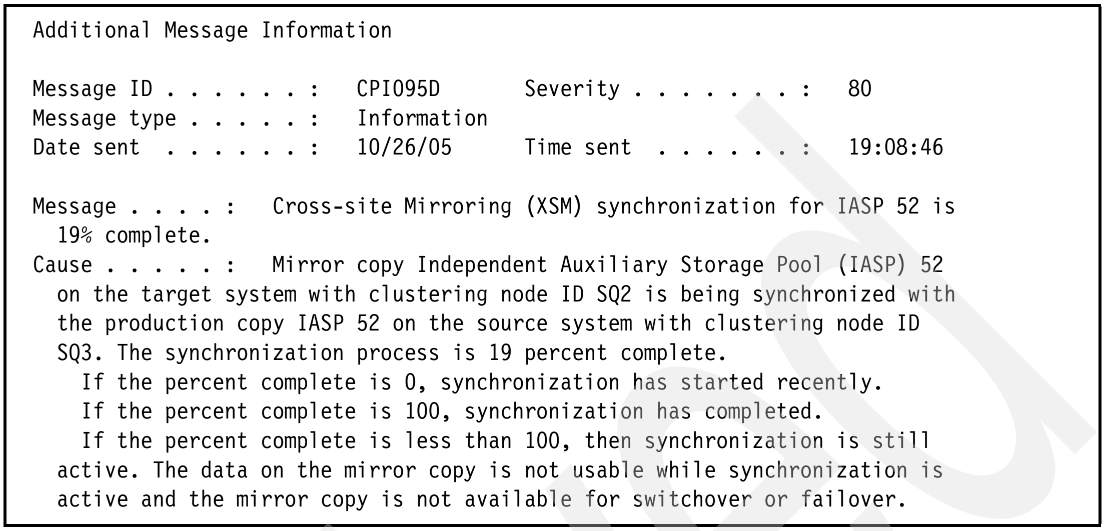
In this section, we discuss how synchronization works, when it is required, and how a switchover or failover affect synchronization.

8.3.1 Full synchronization between production copy and mirror copy

A full synchronization deletes all data on the mirror copy of the IASP and then copies all data from the production copy to the mirror copy. Any changes that are done on the mirror copy while XSM is inactive are *not* preserved.

If you vary off the IASP during synchronization, then synchronization resumes from the point at which it was interrupted, after you vary on the IASP again.

Message CPI095D, shown in Figure 8-5, is posted in the QSYSOPR message queue to let you know the progress of the synchronization.



```
Additional Message Information

Message ID . . . . . . :   CPI095D      Severity . . . . . . . :   80
Message type . . . . . :   Information
Date sent  . . . . . . :   10/26/05     Time sent  . . . . . . :   19:08:46

Message . . . . :   Cross-site Mirroring (XSM) synchronization for IASP 52 is
  19% complete.
Cause . . . . . :   Mirror copy Independent Auxiliary Storage Pool (IASP) 52
  on the target system with clustering node ID SQ2 is being synchronized with
  the production copy IASP 52 on the source system with clustering node ID
  SQ3. The synchronization process is 19 percent complete.
    If the percent complete is 0, synchronization has started recently.
    If the percent complete is 100, synchronization has completed.
    If the percent complete is less than 100, then synchronization is still
  active. The data on the mirror copy is not usable while synchronization is
  active and the mirror copy is not available for switchover or failover.
```

Figure 8-5 CPI095D message during synchronization

8.3.2 Partial synchronization between production copy and mirror copy

A partial synchronization uses information from the source site or target site tracking space to resynchronize data between nodes that was changed while XSM is not active. If you vary off the IASP during synchronization, then synchronization resumes from the point at which it was interrupted, after you vary on the IASP again.

8.3.3 When synchronization is required

Synchronization between the production copy and the mirror copy of the IASP is required when the following user-initiated processes occur:

- The mirror copy of the IASP is detached.
- XSM is suspended while an IASP is still active.

Synchronization between the production and the mirror copy of the IASP is required when the system suspends XSM for any of the following reasons:

- Loss of communication occurs, and the recovery time out is exceeded.
- The mirror copy of the IASP runs out of storage.
- Any other operational failure occurs of the mirror copy of the IASP.

The production copy of the IASP can function normally during the synchronization process. The production copy is available for use, but performance might be impacted. For more information, refer to Chapter 9, “Performance considerations” on page 153.

Note: The mirror copy is unusable until synchronization is complete.

8.3.4 When a switchover or failover occurs

Switchover or failover to the mirrored copy of the IASP is prohibited while XSM is suspended. This is because the mirrored copy contains back-level data and is marked ineligible for switchover or failover. If, however, the production copy of the IASP is lost and unrecoverable while XSM is suspended for all IASPs in an ASP group, you can convert the mirrored copy to the production copy by changing the recovery domain order.

A failover to the mirrored copy of the IASP while the IASP is online results in synchronization. A full synchronization is necessary regardless of how fast the failover occurs and is independent of the recovery time out value. Synchronization is necessary regardless of whether any updates are made to the IASP during the failover.

8.4 System considerations

In the following section, we discuss the implications of powering down the system or upgrading the system and explain how to minimize the effects to XSM.

8.4.1 Shutting down the system

If the system that owns either the production copy or mirror copy of the IASP must be shut down, we recommend that you vary off the IASP at the production copy site to prevent XSM from being suspended. If there is another backup node at either site to which the IASP can be switched, you still want to vary off the IASP at the production copy site first before you perform the switchover.

If you do not vary off the IASP at the production site prior to shutting down the system in question, and another node is at that same site to which the IASP is switched, XSM is suspended temporarily during the switchover. A resynchronization is required once the XSM is resumed.

If there is no other active node at the same site as the system that is being shut down, XSM needs to be suspended while that system is down. If the system being shut down is the node that owns the mirror copy of the IASP, we recommend that you suspend XSM manually first, so that the production copy of the IASP does not wait for the recovery timeout threshold to be reached before the system suspends XSM.

8.5 Communications considerations

XSM depends on some type of communication pathway between the node that owns the production copy of the IASP and the node that owns the mirror copy of the IASP. Therefore, there are many factors that can change that communication pathway and affect XSM.

8.5.1 Ending TCP/IP or the QSYSWRK subsystem

Do not end the QSYSWRK subsystem while XSM is active. This is where the TCP/IP job (QTCPIP) resides. Ending the QSYSWRK subsystem therefore ends TCP/IP. This causes a failover to the next backup node in the recovery domain. The same is true if you end TCP/IP by either using the ENDJOB JOB(QTCPIP) or the ENDTCP command.

8.5.2 Adding TCP/IP filter rules or port restrictions

If you configure filter rules to restrict services over any of the IP addresses on your systems, be sure not to add restrictions to the addresses that are being used for XSM or clustering. If you do, you can prevent the communications jobs from working successfully and inadvertently suspend XSM in the process.

If you modify your network equipment, be sure to adhere to the same set of guidelines.

8.5.3 Ending or changing IP interfaces

If you have to end all IP interfaces or the only IP interfaces that are used for geographic mirroring data port services, consider performing the following actions first:

- Suspend XSM with tracking if you are using V5R4 and have defined a tracking space for you IASP. This ensures that changes to disk pages in the IASP are tracked while the lines for XSM cannot be used. Therefore, you can avoid a lengthy full synchronization when the lines are up and working again.
- If you are running with V5R3 of i5/OS or have not defined a tracking space for your IASP, consider first varying off the IASP. This prevents users and applications from accessing data in the IASP. If you only vary on the IASP again after communication for XSM is restored, then you do not need to run a full synchronization in this scenario.

If you need to change the IP addresses on the system, you should not have to change the IP addresses that are used for XSM if you configured it with dedicated IP addresses and have not changed filter rules or firewall rules to prevent these addresses from still communicating across the network.

9

Performance considerations

In this chapter, we provide an overview of the performance considerations to make when you configure cross-site mirroring (XSM) with or without geographic mirroring. A variety of factors can influence the performance of XSM. Therefore, it is impossible to predict every possible combination of environments. We discuss these factors individually, while keeping other factors constant, to provide reasonable information about how each factor influences geographic mirroring.

The test results in this chapter represent a few possible environments and changes to a few variables in those environments. They are merely meant as a general guideline to help in preparing your cross-site mirror environment.

9.1 System setup considerations

While it is not mandatory to have each system configured exactly the same in an XSM environment, it is important to realize the factors that can impact the performance of XSM. In this section, we provide an outline of the more important factors if you want to try to balance your systems in your environment.

9.1.1 Processor

Processor overhead is associated with running geographic mirroring. However, in the testing that we performed, we saw a minimal impact on processor while running geographic mirroring on our configuration. We had multiple processors and over 10 GB of memory. If your mirror side system is small in comparison to your source side system, in a heavy disk write environment, processor overhead can be noticeable and impact performance. As a general rule, the partitions that you are using to run geographic mirroring need more than a partial processor. In a minimal processor configuration, you might potentially see 5% to 20% processor overhead while running geographic mirroring.

Be aware that you might see the processor overhead even if you are currently not replicating data to the target system. For example, synchronization is not in process. When geographic mirroring is configured, all data that is written is first checked to see if it needs to be sent to the target side.

9.1.2 Machine pool size

For optimal performance of geographic mirroring, particularly during synchronization, increase your machine pool size by the amount given by the following formula. The extra machine pool storage is required on all nodes in the cluster resource group (CRG), so that the target nodes have sufficient storage in case of switchover or failover.

At vary on time of the independent auxiliary storage pool (IASP), the machine pool size is checked. The base amount needed is about 300 MB plus about .3 MB for every arm in the IASP. Given a 90-arm IASP, you need roughly 327 MB free. A 180-arm IASP requires about 354 MB free. See the following calculations:

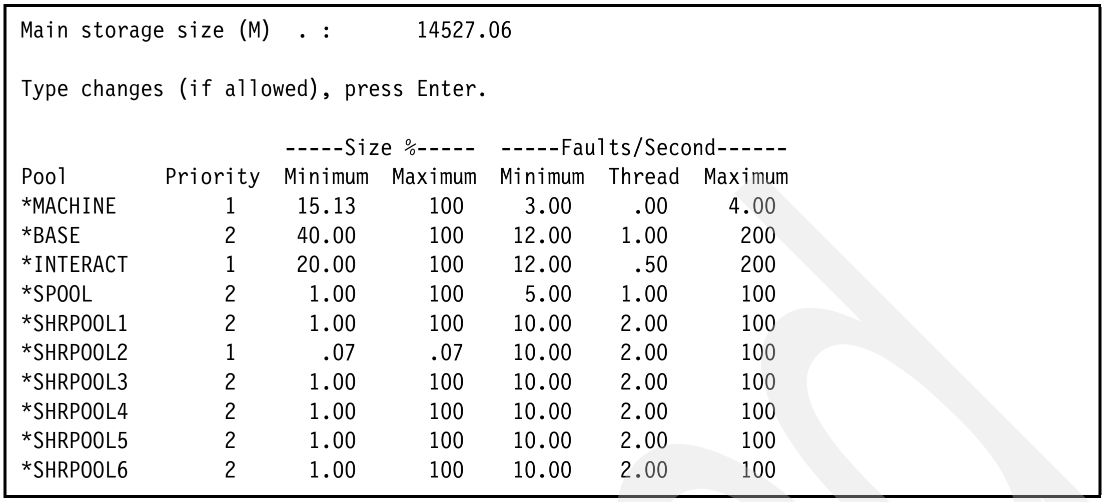
```
300 + (.3 x 90) = 327 MB
300 + (.3 x 180) = 354 MB
```

Having this much memory available in the machine pool helps to guarantee that there is a clear data task for every arm in the IASP. This means the clear data task on all the target IASP disks is done in parallel and is completed as fast as possible.

As always, the more arms there are in the IASP, the better the performance should be, because more processes can be done in parallel. Given the previous two examples, the same size dataset, and enough memory in the machine pool, the clear data task should happen twice as fast in the 180-arm IASP. The speed is doubled because each DASD in the 180-arm IASP is only one half as full as the 90-arm IASP.

To prevent the performance adjuster from reducing the machine pool size, you must perform the following steps:

1. Enter the Work With Shared Pools (WRKSHRPOOL) command.
2. In the Work with Shared Pools panel (Figure 9-1), set the machine pool minimum size to the calculated amount (the current size plus the extra size for geographic mirroring from the previous equation).



```
Main storage size (M)  . :       14527.06

Type changes (if allowed), press Enter.

                      -----Size %-----  -----Faults/Second------
Pool        Priority  Minimum  Maximum  Minimum  Thread  Maximum
*MACHINE          1     15.13      100     3.00     .00     4.00
*BASE             2     40.00      100    12.00    1.00      200
*INTERACT         1     20.00      100    12.00     .50      200
*SPOOL            2      1.00      100     5.00    1.00      100
*SHRPOOL1         2      1.00      100    10.00    2.00      100
*SHRPOOL2         1       .07      .07    10.00    2.00      100
*SHRPOOL3         2      1.00      100    10.00    2.00      100
*SHRPOOL4         2      1.00      100    10.00    2.00      100
*SHRPOOL5         2      1.00      100    10.00    2.00      100
*SHRPOOL6         2      1.00      100    10.00    2.00      100
```

Figure 9-1 Setting the minimum machine pool size as a percentage of the total amount of storage

If your environment does not require automatic performance adjustment, you can set the system value QPFRADJ to zero (0), which prohibits the performance adjuster from changing the size of the machine pool. As long as you configure the minimum size with the method shown previously, having the performance adjuster active (QPFRADJ set to 2 or 3) is still going to be acceptable.

9.1.3 Disk units

Disk unit and input/output adapter (IOA) performance can impact overall geographic mirroring performance. This is especially true when the disk subsystem is slower on the target side of the mirror. When geographic mirroring is in synchronous mode, all writes on the source system are gated by target system writes to disk. Therefore, a slow target disk subsystem can impact source side performance. You can remedy this problem somewhat by running geographic mirroring in asynchronous mode. Running in asynchronous mode alleviates the wait for the disk subsystem on the target side and sends confirmation back to the source side when the changed memory page is in memory on the target side.

For example, a typical user might upgrade an 830 production machine with a new 570 model with new expansion towers. The application performance can increase substantially with the new system. The user can then employ the 830 machine as a target of the geographic mirror configuration. In this case, all writes to the IASP become gated by the performance of the 830 disk subsystem.

Ensure that DASD cache batteries are in working order on both systems. If they are non-functional, the system runs in degraded mode and severely impacts DASD subsystem performance. You can quickly check the Product Activity Log in Service Tools for a *xxxx*8008 SRC (where *xxxx* is the model of your DASD controller) indicating that you need to replace a DASD cache battery. There is also an entry in the QSYSOPR message queue (CPPEA13) indicating that a battery needs to be replaced.

If possible, we recommend that you separate configured disks into an IASP at an IOA boundary. This means that you should have all disks from an IOA in either SYSBASE (auxiliary storage pool (ASP) 1-32) or an IASP. If you share disks across an IOA, DASD utilization from SYSBASE can affect IASP performance or vice versa.

9.1.4 System ASP versus IASP configuration

Similar to any system disk configuration, the number of disk arms that are available to the application can significantly impact the performance of XSM. Putting additional workload on a limited number of disk arms might result in longer disk waits and ultimately longer response times to the application.

This consideration is particularly important when it comes to temporary storage in a system that is configured with IASPs. All temporary storage is written to the SYSBAS ASP. This means that if your application uses a lot of temporary storage, more disk arms will potentially be required in the SYSBAS ASP. They might also be required if you are using Structured Query Language (SQL) applications that have to create temporary indexes or joins or if you are creating large database objects in QTEMP.

If your application does not use a lot of temporary storage, then you can get by with fewer disk arms in the SYSBAS ASP. You must also remember that the operating system and basic functions occur in the SYSBAS ASP. A general rule says that the ratio between SYSBAS and IASP disks must be one to three (1:3), but with an SAP environment, for example, the ratio can be reversed.

9.1.5 Communication lines

We recommend that geographic mirroring to have its own communication line or lines for both performance and availability. The geographic mirroring function is tied to important I/O, just like I/O performance to local disk.

Without its own line or lines, geographic mirroring can contend with other applications using the same communication line and affect user network performance and throughput. This includes the ability to affect cluster heartbeating, resulting in a cluster partition state.

If your configuration is such that multiple applications or services require the use of the same communication line, some of these problems can be alleviated by implementing quality of service (QoS) through the TCP/IP functions of i5/OS. Overall we recommend that TCP/IP communication lines are dedicated for geographic mirroring. These are important considerations because geographic mirroring performance is directly impacted by communications throughput.

One TCP/IP line from each node must connect the two sites. We recommend that you have a second TCP/IP line to provide redundancy and better performance. You can configure up to four TCP/IP communication lines. Geographic mirroring distributes changes over multiple lines in a round-robin approach for optimal performance. The round-robin approach means that the Data Port Services layer sends data out from each of the configured communication lines in turn, from 1 to 4, over and over again.

Four communication lines allow for the highest performance, but lab tests show that you can obtain relatively good performance with two lines. If you use more than one communication line between the nodes for geographic mirroring, it is best to separate those lines into different subnets, so that the usage of those lines is balanced on both systems.

Given normal TCP/IP functionality on i5/OS, if multiple lines exist in the same network, it is not possible to guarantee, without routing, where the communication traffic will exit the system. For example, given a 255.255.255.0 netmask, if the source system, has IP addresses 10.1.1.1 and 10.1.1.2, and the target system has IP addresses 10.1.1.3 and 10.1.1.4, the source system does not balance the traffic over both the 10.1.1.1 and 10.1.1.2 addresses. Instead, the system sends all traffic out one of the addresses, 10.1.1.1 for example, and leaves the other one underutilized.

To force the system to use a one-to-one relationship, you can place the multiple IP address pairs in different networks. Given the previous example, a better way to configure the IP addresses is to have the source system have the IP addresses, 10.1.1.1 and 10.1.2.1, and the target system have 10.1.1.2 and 10.1.2.2. This places a pair of addresses (source and target) in the same network, so the system has only one path to send data over for each line that Data Port Services uses. This guarantees equal utilization for each line that the system uses for Data Port Services.

Along with the number of lines, the type of lines that are used for Data Port Services can impact performance. For example, if two lines are used, it is important to have the same speed across the lines. Due to the round-robin architecture for sending data, the slowest line becomes the gating factor. Often, having two unequal lines (one fast and one slow) is slower than using one fast line.

Latency can also play a factor for geographic mirroring. The farther the source and target systems are apart, the more latency there is for every communication transaction between the systems. Normal latency estimates should apply for geographic mirroring. However, in basic testing that we ran, given a highly write-intensive application, we saw a 3.77% increase in time for the application that ran in asynchronous mode from a 10-ft. length of CAT6 UTP cable, to a 10-km. length of single mode (9 micron) fiber. We saw a 3.29% increase in time for the application running in synchronous mode from a 10-ft. length of CAT6 UPT cable, to a 10-km. length of 9 micron fiber.

The effect that distance induced latency can have on an application depends on the type of application that you are running. In an interactive application where transactions are started by a user on an interactive panel, a small increase in time (latency) might not even be seen. However, a batch type application that is write-intensive might be severely impacted. The only way to *accurately* predict the impact of latency on your application is to run some tests with simulated distance between the source and target systems.

9.1.6 Network configuration considerations

During our test, we found that network cabling and configuration were crucial to geographic mirroring performance. Besides guaranteeing that network addressing is set up in a “point-to-point” manner (different subnets for each set of dataport IPs) as mentioned previously, the network topology must also be cabled and configured in a “point-to-point” manner. In this section, we present specific technical examples of network configurations that yield drastically different performance characteristics.

When we first started testing, the source and target systems were each plugged into the same switch as shown in Figure 9-2. We did this with 1 Gb to 4 Gb Ethernet connections from each system. We achieved good performance in this configuration, topping 1 TB per hour throughput with four dataport connections from each system. The system configurations that we used were sizable. This kind of throughput might not be achievable on a smaller system. All four data port services connections used IP addresses in different subnets.

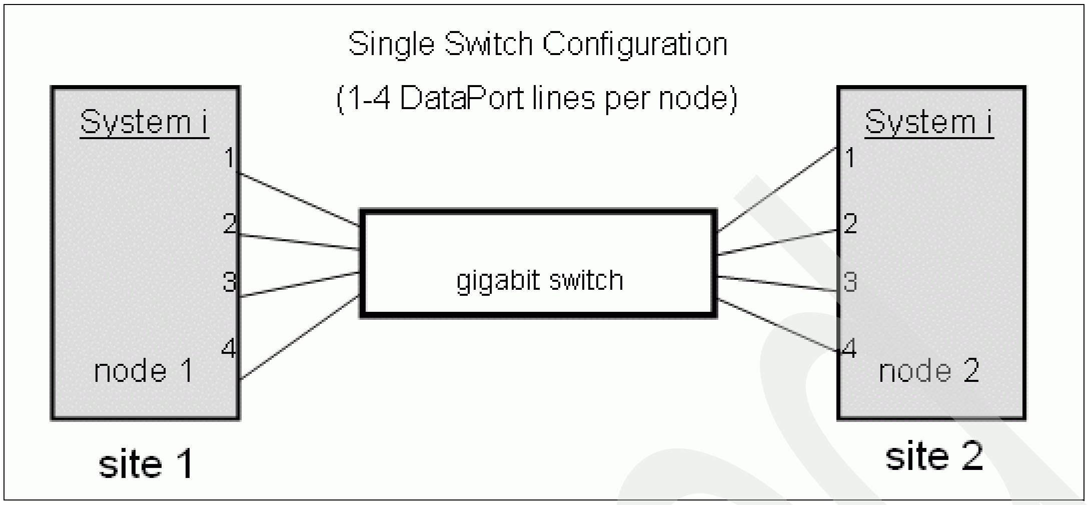


Figure 9-2 Single switch configuration

The next configuration that we tested was to have the source and target plugged into different switches with one Gb Ethernet connection from each system as shown in Figure 9-3. The switches were then connected with a single UTP CAT6 Ethernet cable. This gave us good performance, and we achieved above 98% utilization on the Ethernet connections in the network.

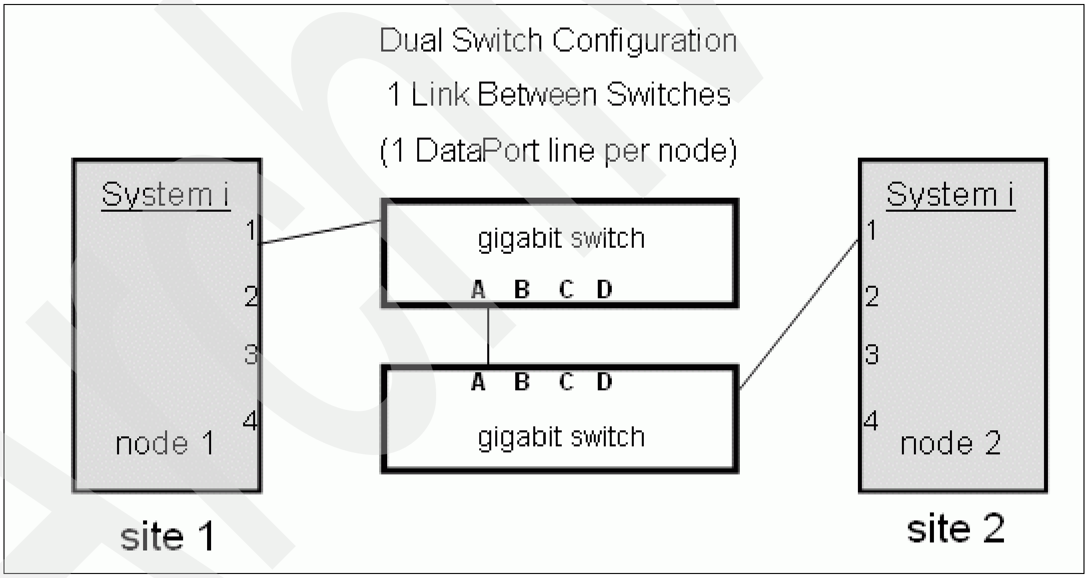


Figure 9-3 Dual switch configuration

We then took the dual switch configuration and added more dataport Ethernet lines from the source and target, but left only one Ethernet connection between the switches. We expected to see the utilization between the switches stay at about 98%, and each of the links coming from the source and target spread the 98% utilization equally (48% for two connections, 32% for three connections and 24% for four connections). This would account for the round-robin balancing algorithm that data port services uses across the configured lines.

However, these results are not what we experienced. As we added the second line, we saw the overall throughput go from 98% down to (26% x 2) or 52%. When we added the third line, the throughput went down to (11% x 3) or 33%. The fourth line dropped the throughput to

(6% x 4) or 24%. We were unable to determine the cause of the slowdown, but we alleviated the problem by changing the network topology to accommodate the point-to-point traffic that dataport services produces.

We show the recommended network configuration in Figure 9-4.

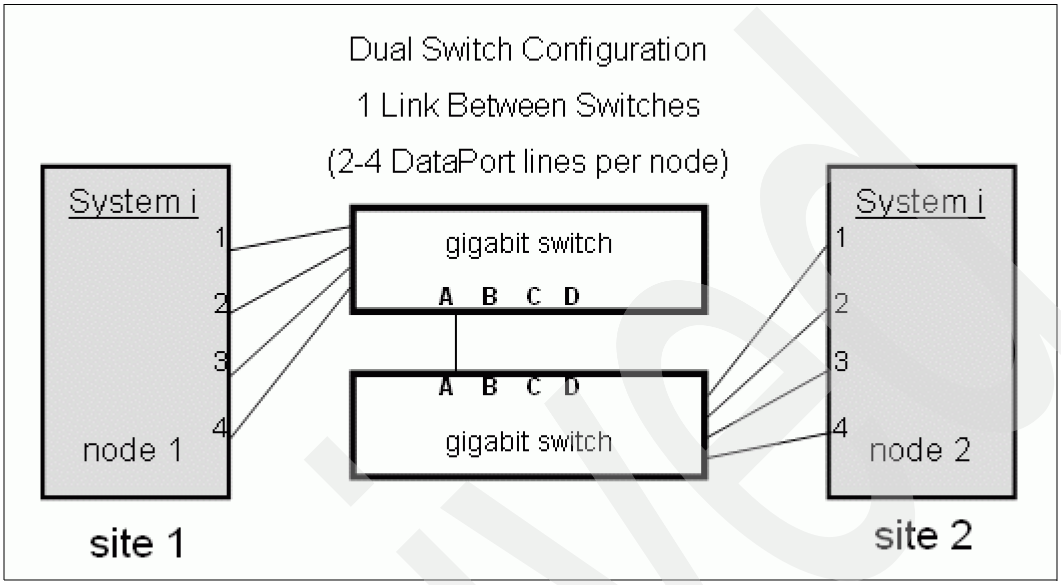


Figure 9-4 Dual switch configuration with switch link

The next test that we ran was to include more connections between the switches as shown in Figure 9-5.

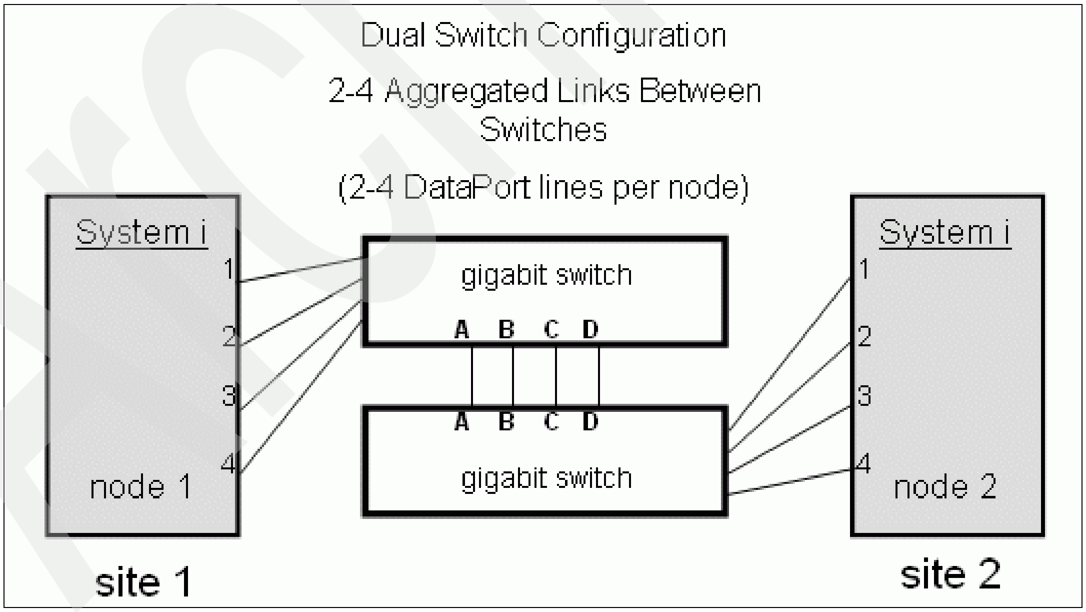

Dual Switch Configuration
2-4 Aggregated Links Between Switches
(2-4 DataPort lines per node)
System i
1
2
3
4
node 1
site 1
gigabit switch
A B C D
A B C D
gigabit switch
System i
1
2
3
4
node 2
site 2

Figure 9-5 Dual switch configuration with multiple switch links

As we expected, with one dataport line from each system to the switches, and multiple connections (aggregated link) between the switches, we again saw good performance (~98% utilization). When we added more dataport connections, the throughput went down again. We learned that this was due to the way that the switches distribute traffic on the connections that make up the aggregated link. From our experience, we saw that the switches do not “share” the bandwidth from each IP address over multiple links in the trunk (aggregated link). They typically look at the MAC address or the IP address that the traffic is coming from and decide on which link in the trunk to send it over. The balancing over the links in the trunk probably happens very well with hundreds of IPs or MAC addresses that are trying to use the trunk. However, with dataport services, we only have four IPs. Many times we saw three or four of the dataport lines sending traffic over the same link in the trunk. When this happened, we then saw the same slowdown exhibited in the configuration with multiple dataport lines and only one connection between the switches.

Given the performance slowdown, it is important for the user to understand the path that data from each dataport line is using to get from the source to the target. In the next test, we explain how to configure your network to maximize throughput.

Because the switches did not spread the dataport traffic across all four lines, the next test that we ran was to force the traffic to do so. There are a couple different ways to achieve the desired balancing. The first way that we achieved this balance was to use a virtual local area network (VLAN) for the switches as shown in Figure 9-6. This technique basically separates all traffic by port, by assigning each port to a specific VLAN. Each pair of dataport connections was assigned a different VLAN, including one of the connections between the switches. This way, all traffic from each dataport IP on the source only had one possible way to reach the target machine. This gave us four parallel streams of data. When we did this, testing again showed very good performance, letting us achieve over 1 TB of data throughput per hour as we did in a single switch configuration.

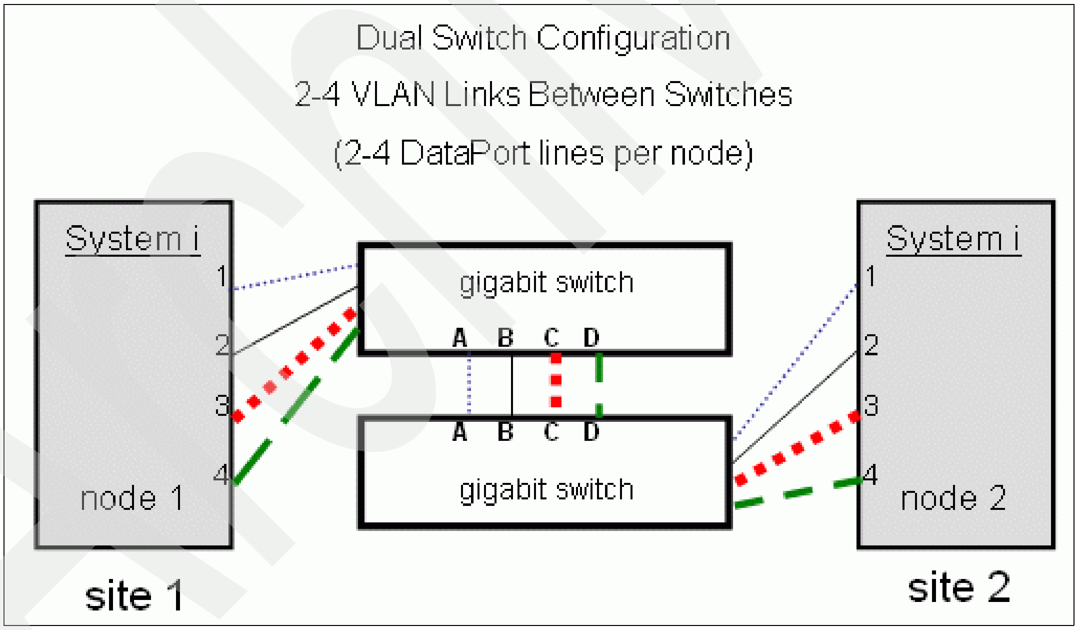


Figure 9-6 Dual switch configuration with multiple VLAN links

We also achieved the same performance without using VLANs by changing the IP addresses that the dataport was using so that the switch algorithm was forced to route the four different dataport streams over four separate links in the trunk. We were unable to determine how the switch algorithm works for balancing data over a trunk, so we kept changing IP addresses until we found the correct combination. We do not recommend using this technique to tune a user’s environment. We mention it here to show that you can achieve very good performance if the network is configured in such a way that allows for a point-to-point connection for each pair of dataport IP addresses.

Configuring the switches to use VLANs to maintain a point-to-point network topology seemed to be the most logical choice to yield the best performance for multiple dataport lines.

9.1.7 Normal production runtime impacts

An overhead is associated with running geographic mirroring. While results can vary, typically asynchronous mode offers better performance than synchronous mode. Sometimes in a “bursty” environment, running in synchronous mode can make the environment run more smoothly and offer better performance than asynchronous mode.

Synchronous mode

In synchronous mode, the client waits until the write operation is complete to the disk cache (IOA) on the source system and to the disk cache (IOA) on the target system. This provides the highest level of data integrity because the data is guaranteed to be on disk, or at least on the IOA cache on the target system before control is released to the user or application.

Asynchronous mode

In asynchronous mode, the client must wait until the write operation is complete to system memory on the source system and has been received into system memory on the target system for processing.

9.1.8 Impact on other job response times

While geographic mirroring is active, other jobs can be affected. In general, interactive jobs are impacted significantly because the added time for the remote system DASD writes is minimal. For example, a person at a terminal might be able to put through 10 transactions each minute. Each of these transactions might have a small increase in time, but the increase can easily go unnoticed by the operator. A batch job might react a bit differently to a geographic mirrored environment. During a batch job, there might be hundreds of thousands of transactions. If each of these transactions has a little bit of time added because of the remote disk write, the cumulative batch run time can increase and become significant.

9.2 Mirror copy synchronization

Geographic mirroring activity can be stopped by a suspend or detach operation. By following certain guidelines, a full synchronization should not be needed.

A *suspend operation* can be user-initiated or initiated automatically by the system. The system can suspend geographic mirroring due to communication failures, DASD failures, the source or mirror copy of the IASP reaching 100% ASP capacity, or other reasons. When the system suspends the mirror because of some kind of failure, the suspend is done automatically with tracking. Given the CRG is active, the system auto-resumes the mirroring process when the problem is alleviated. This places the mirror into a partial synchronization.

When a user suspends the mirror, the user has the option of suspending with tracking or without tracking. If the user chooses to suspend *without* tracking, a full synchronization is needed when a resume is performed. If the user chooses to suspend *with* tracking, the system starts to track all the changes that are made to the source IASP. When a resume is performed after a suspend with tracking, only the data that has changed is sent to the target IASP. This is referred to as a *partial synchronization*.

A *detach operation* is user initiated and severs the link between the source and target IASPs. This allows you to vary on the IASP on the target system so that the data can be accessed.

The best way to use the detach function is to first vary off the source side IASP. This in turn brings the target side IASP from a Varied On state to a Varied Off state. With both sides in a varied of state, all memory resident data is flushed to disk, which ensures the data was replicated to the target system.

The detach can be done while the source side IASP is available. However, since geographic mirroring sends data only to the target when an actual DASD write happens, all memory resident IASP data might not make it to the target side. This can be considered a “dirty detach” because there is no guarantee that *all* data made it to the target IASP. The only way to guarantee a synchronization point between the source and target when a detach is performed is to first vary off the IASP on the source side. After the IASP vary off is done to flush data to disk, and a detach is performed, the IASP can immediately be varied on, and made available again to be used.

Note: While the mirror is detached, your data is not highly available and changes are *not* being replicated to the target system.

A reattach is performed to re-establish the mirroring of data between the source and target IASPs. A reattach can be done while the source IASP is in an Available or Varied Off state but the target IASP must be varied off. A reattach requires a full synchronization from the source IASP to the target.

As mentioned in 7.2, “Target Site Tracking” on page 140, a PTF is now available that enables a user to detach with tracking. This means that the target IASP can be made available and used, and data can be changed. While changes are happening, the addresses of all changed data on the source and target are kept in a tracking space on each side. When the reattach occurs, the tracking spaces are joined together and *only* the data that was changed on either side gets replicated from the source side to the target. Another way to look at a reattach after a detach with tracking is that *all* changes on the target side that were made since the detach with tracking are erased, or rolled back to match the source side.

9.3 Partial synchronization

Partial synchronization time is largely dependent on the amount of data that was changed in the IASP while it was suspended with tracking or detached with tracking. The IASP data is not cleared on the target node when starting a partial synchronization, and the data can start flowing immediately, thus shortening the synchronization time. Partial synchronization time is also shorter because only the changed data is tracked and sent to the target system. This data is usually only a small subset of the data in the IASP. Partial synchronization is much different from full synchronization where the entire contents of the IASP are sent to the target system after the IASP data on the target is cleared.

Because full and partial synchronization use different algorithms, the data synchronization step in partial synchronization should be slightly faster. In addition, the communication line utilization in a partial synchronization should be higher than for full synchronization.

9.4 Resume priority options

With synchronization, three resume priority options exist that might affect synchronization performance and time. Resume priorities of low, medium, and high for geographic mirroring affect the processor utilization, disk subsystem performance, and the speed at which data is transferred. The default value is set at medium priority. A system set at high priority transfers data faster, is more disk intensive, and consumes more processor than a lower setting. Your choice depends on how much time, disk performance, and processor you want to allocate for this synchronization function.

The priority at which a user should run the resume depends on how their application performs. The only way to *accurately* see how an application is affected is to run a synchronization at different priorities and test application performance.

> **Note:** Keep in mind that this change is not dynamic. Resume priority must be changed when the source side IASP is varied off.

9.5 Synchronization considerations

During the first phase of a full synchronization, a number of clear data tasks are run on the target system. As we mentioned in 9.1.2, “Machine pool size” on page 154, the size of the machine pool is an important factor during this phase. This ensures that there are enough tasks to clear the IASP on the target system as quickly as possible.

To a certain extent, the number of disk arms in the IASP affect synchronization speed. During synchronization, you can use the production copy, but application performance might be negatively affected because the user applications and the synchronization process both use the same system resources.

Another factor that can impact the speed of the synchronization is the type of data that is contained in the IASP. The system uses different block sizes of memory depending on the type of object that is in the IASP. For instance Journal Receivers can be very large objects, and can allocate very large chunks of data on disk. When multiple chunks of data get stacked on the same DASD, the synchronization process can be somewhat gated by the performance of a single disk arm trying to read all of those contiguous chunks of data. If you have many small objects, the sheer volume of objects can slow down the synchronization process. The synchronization process is tuned for the average size object.

9.6 Case studies using synchronization

We performed tests using basic synchronization in a geographic mirrored environment as explained in the following section. We also performed testing over WAN connections as explained in 9.6.2, “V5R4 case study over a WAN” on page 167.

9.6.1 Case study for V5R4

For our test of basic synchronization in a geographic mirrored environment, we used two System i5 model 520s, each with one partition and two dedicated processors, in our environment. Each partition had 15 GB of memory, and the *SYSBASE had eight DASD units, each one a model 4326 DASD with 36 GB. The IASP on the source and target consisted of a 24-arm IASP. All IASP DASD were model 4326 (36 GB). The IASP DASD controllers were model 2780 and were controlled by 2844 IOPs.

In our testing, we tried to use data that was representative of a typical customer library. The library that we used contained object types such as commands, programs, database files, journals, and journal receivers.

Both partial and full synchronization tests were run. Full synchronization tests were run with a 10 GB library (9.778 GB). The IASP was suspended without tracking, and the library was restored to the IASP. The geographic mirror was then resumed, and the messages that geographic mirroring logs in the QSYSOPR message queue were used to determine the start and finish time of the synchronization. This process was repeated on low, medium, and high priority with one, two, and three instances of the library to produce roughly a 10, 20, and 30 GB synchronization.

All of our synchronization times were measured without any additional disk I/O happening on the system, simulating a quiesced system. The synchronization time in each specific user environment depends on how many additional disk writes are happening in the source IASP, which geographic mirroring needs to send to the target system. For example, given the same set of data, synchronization should happen faster during the night when there is little activity on the system versus running synchronization in the middle of the day with a large workload. During the day, the system has the geographic mirror synchronization traffic in addition to the normal disk writes that XSM needs to mirror to the target system.

Figure 9-7 shows a graphical view of the results.

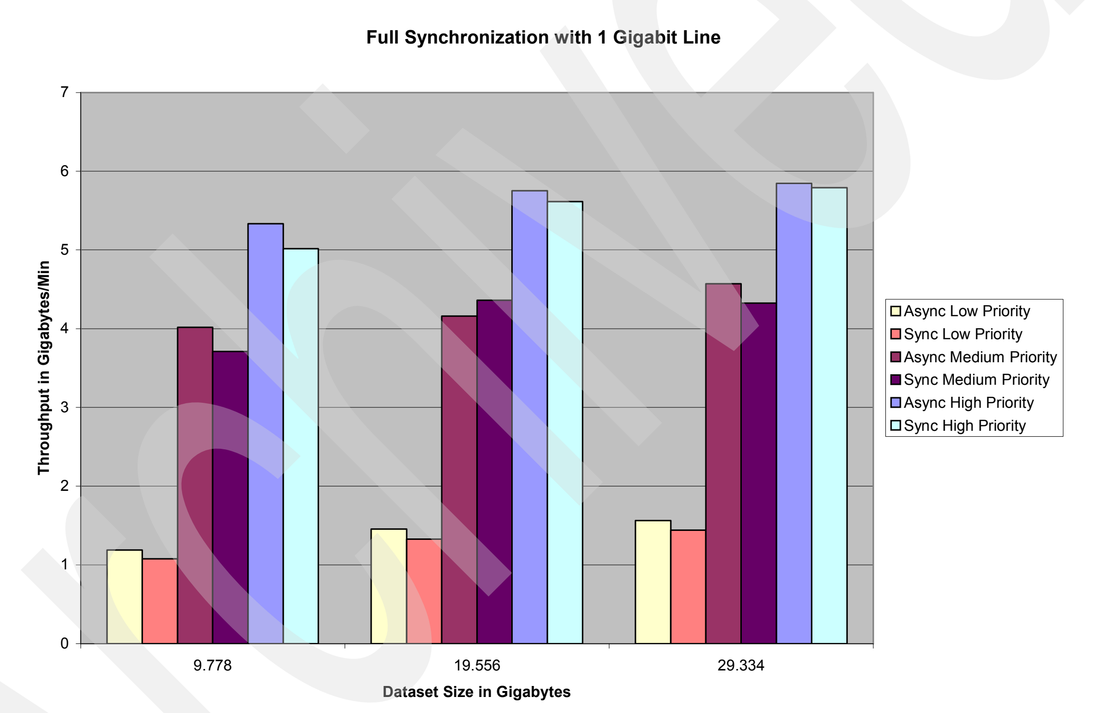


Figure 9-7 Full resynchronization with a 1 GB line

Table 9-1 shows the results of this test with different mirroring modes and resume priorities.

Table 9-1 Results for full resynchronization with a 1 GB line

Mirroring mode	Resume priority	Dataset 9.77 GB	Dataset 19.556 GB	Dataset 29.334 GB	Average throughput (GB/min)
Async	High	5.33	5.75	5.84	5.64
	Medium	4.02	4.16	4.57	4.25
	Low	1.19	1.46	1.56	1.40
Sync	High	5.01	5.61	5.79	5.47
	Medium	3.71	4.36	4.32	4.13
	Low	1.07	1.33	1.44	1.28

In our next test, we increased the number of communications lines to two 1 GB lines. The results are different, but not as radical as expected. See Figure 9-8.

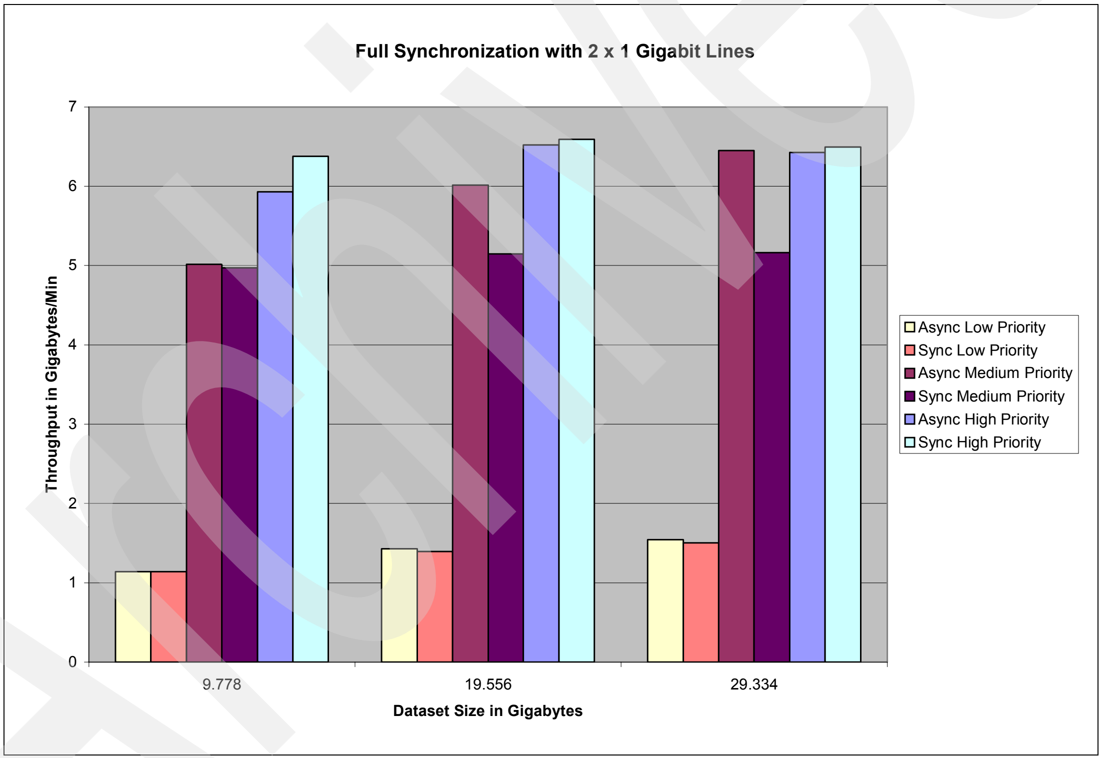


Figure 9-8 Full resync with two 1 GB lines

As you can see in Figure 9-9 the Async Medium Priority numbers are close to the High Priority numbers. We believe this can be attributed to the system being I/O bound. Adding more tasks for synchronization from medium to high priority did not cause the throughput to go up because the disk subsystem was already fully used. This is a good example to show that adding communication lines does not necessarily increase overall throughput. It is important for a user to determine the bottleneck for their specific environment.

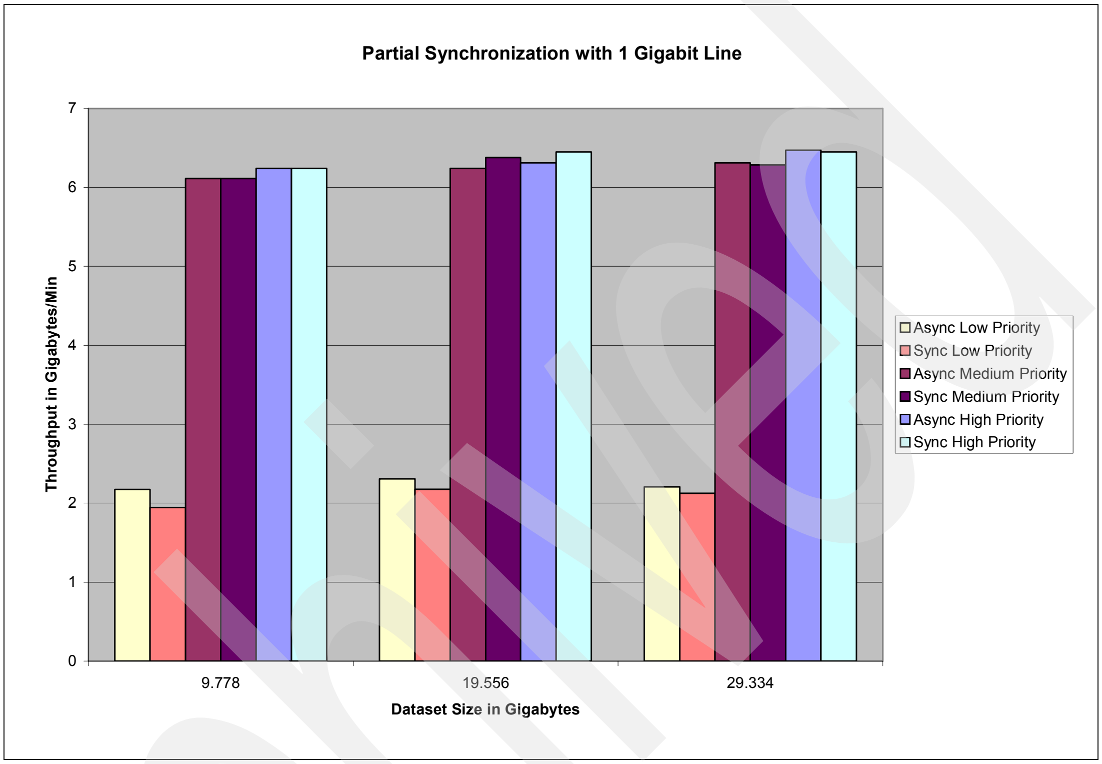


Figure 9-9   Partial resynchronization with a 1 GB line

Table 9-2 shows more details about the results for this test.

Table 9-2   Results for full resynchronization with two 1 GB lines

Mirroring mode	**Resume priority**	**Dataset 9.778 GB**	**Dataset 19.556 GB**	**Dataset 29.334 GB**	**Average throughput (GB/min)**
Async	High	5.93	6.52	6.42	6.29
	Medium	5.01	6.02	6.45	5.83
	Low	1.14	1.43	1.55	1.37
Sync	High	6.38	6.59	6.49	6.49
	Medium	4.97	5.15	5.16	5.09
	Low	1.14	1.40	1.50	1.35

In Table 9-3, we see similar throughput from medium and high priority. This is likely due to the fact that the disk subsystem is very busy.

Table 9-3 Results for partial resynchronization with a 1 GB line

Mirroring mode	Resume priority	Dataset 9.778 GB	Dataset 19.556 GB	Dataset 29.334 GB	Average throughput (GB/min)
Async	High	6.24	6.31	6.47	6.34
	Medium	6.11	6.24	6.31	6.22
	Low	2.17	2.31	2.21	2.23
Sync	High	6.24	6.45	6.45	6.37
	Medium	6.11	6.38	6.29	6.25
	Low	1/94	2.17	2.12	2.08

Figure 9-10 shows the results of testing with a partial resynchronization running across two 1 GB Ethernet lines.

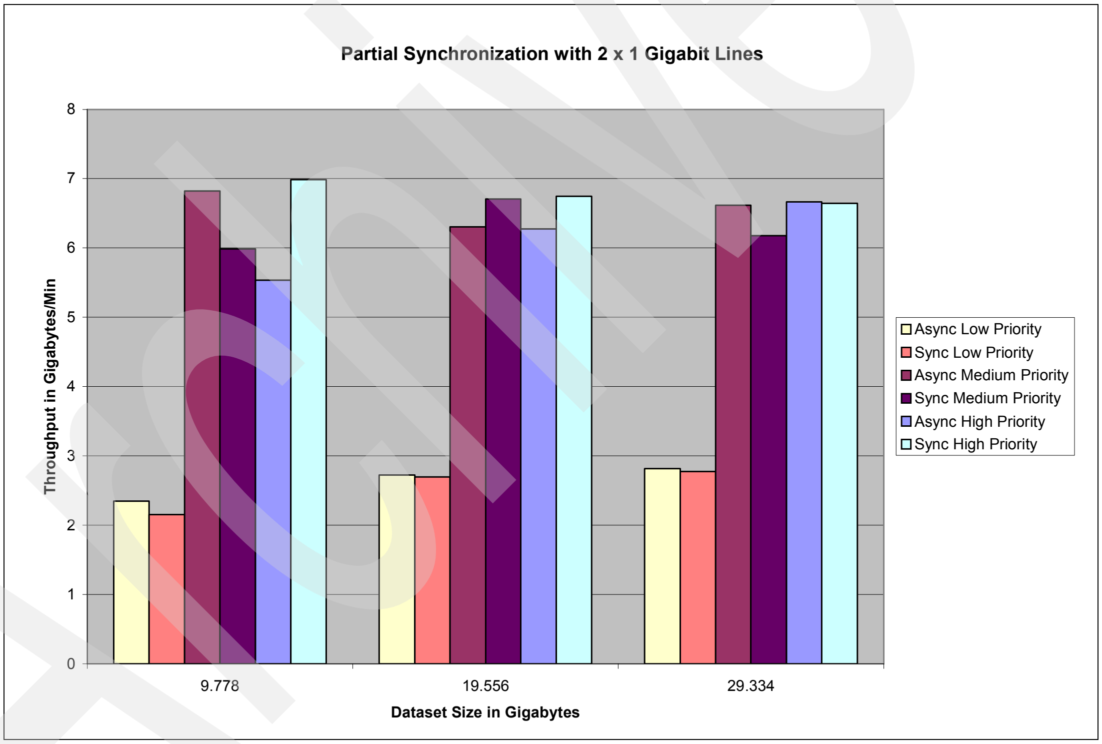


Figure 9-10 Partial resynchronization with two 1 GB lines

9.6.2 V5R4 case study over a WAN

We also ran some tests over WAN connections. The WAN configuration consisted of four model 2742 WAN cards on each partition. Each WAN line simulated a T1 Frame Relay line, which has a theoretical throughput of 1.5 Mbps. We were able to push roughly 1.3 to 1.4 Mbps per line. The number of lines that we used showed a linear trend for the throughput that geographic mirroring was able to send to the target system. Therefore, in this configuration, if the dataset takes one hour to synchronize with a single T1 line, a second T1 would cut the time in half, a third line would bring the time down to 20 minutes, and four lines would bring the time to 15 minutes. We believe this means that the bottleneck of the configuration is in the communication lines, not in any resources on the systems.

For slower communication lines, like a T1, the bottleneck probably ends up being the communication lines rather than the system. The difficult part is knowing when the configuration crosses over from communication bound to some other kind of bottleneck, whether it is a processor, disk subsystem, or something else. Because of the greatly varying workloads that can be run on the system, and the large number of system configurations, it is almost impossible to predict where the cutover will be from being communication bound to being system performance bound.

9.7 Switchover and failover

A *switchover* occurs when the mirror copy assumes the role of the production copy. A switchover is user-initiated, using the CHGCRGPRI command or by performing a switch through iSeries Navigator. A *failover* happens automatically if the CRG is active and the primary system in the CRG recovery domain fails.

9.8 Cluster recovery domain impacts

When running a cluster with more than two nodes, it is important to understand the impact of the different cluster commands. We use a four-node cluster with switchable IASPs and XSM to explain what happens when different operations are performed on the device CRG.

Figure 9-11 shows a four-node cluster with a switchable IASP at site A between nodes 1 and 2. The IASP is mirrored to another switchable IASP at site B between nodes 3 and 4. In this section, we give a number of different examples with a specific CRG recovery domain, an operation that was performed, and the resultant cluster action. Overall in the following examples, it is important to understand that the IASP, which is the target of the geographic mirror, is always located on the node that is the highest node listed in the recovery domain for a different site than the primary node. Because the mirrored IASP (target) is always located on the highest node on a different site than the primary, in the recovery domain, such operations as CHGCRG can trigger a target IASP switchover even though the primary node in the recovery domain has not changed.

In Figure 9-11, the geographic mirroring configuration is set up with the source system indicated by a star (upper left side of the figure) and the target system indicated by the blue star (upper right side of the figure). Even though the recovery domain has Node 2 listed before Node 3, we are still mirroring between Node 1 and Node 3. This happens because geographic mirroring is always performed between sites.

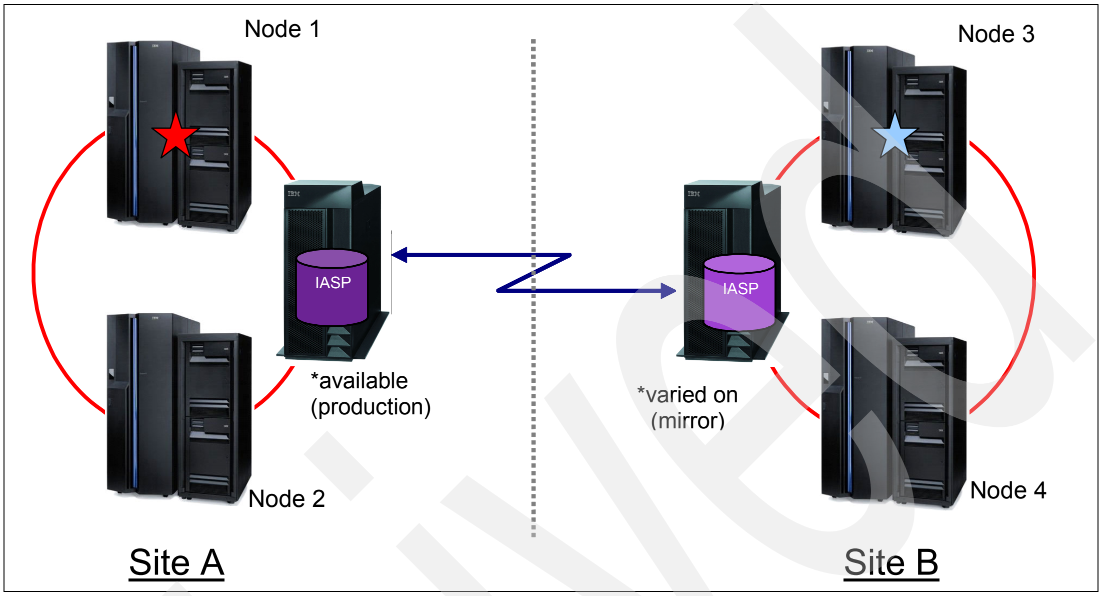


Figure 9-11 Two site cluster recovery domain

Table 9-4 shows the current role of each node in the cluster.

Table 9-4 Recovery domain CRG role

Node	**Site**	**Role**
Node 1	Site A	Primary
Node 2	Site A	Backup1
Node 3	Site B	Backup2
Node 4	Site B	Backup3

When the CHGCRGPRI command is executed with the defined recovery domain, the following actions occur:

1. CHGCRGPRI is executed on any node in the cluster.
2. Geographic mirroring is suspended with tracking.
3. IASP 33 is varied off on Node 1 and Node 3.
4. A disk switch occurs on Site A from Node 1 to Node 2.
5. IASP 33 is varied on to an Available state on Node 2 and a Varied On state on Node 3.
6. CRG recovery domain is reordered to reflect the changes.
7. Control is released back to the user.

In Figure 9-12 the geographic mirroring configuration is set up with the source system indicated by the red star (lower left side) and the target system indicated by the blue star (upper right side).

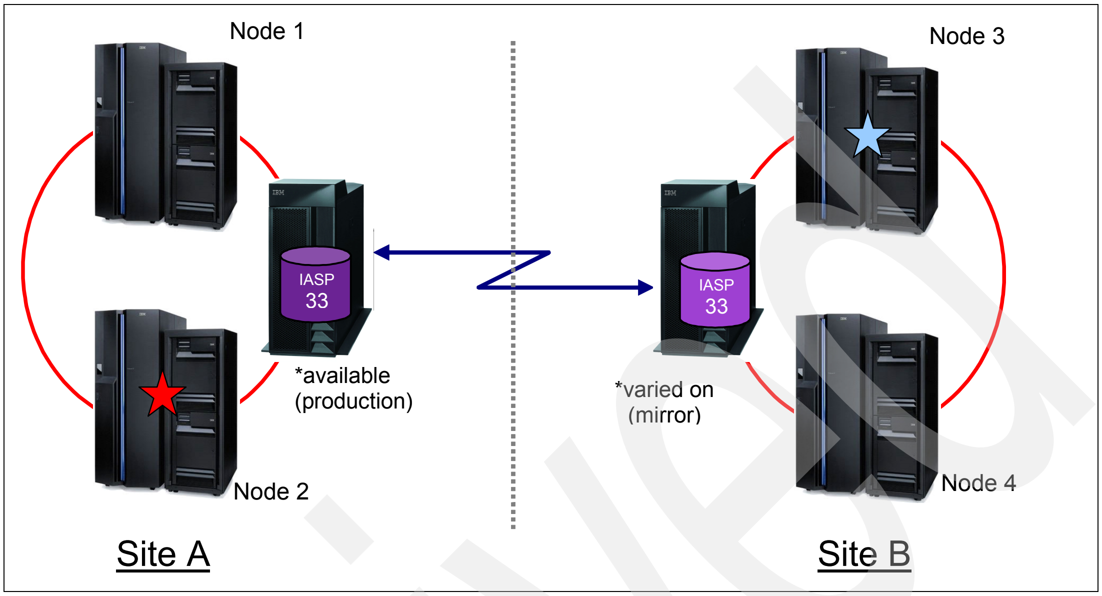


Figure 9-12 Recovery domain state 1

Table 9-5 shows the current role of each node in the cluster.

Table 9-5 CRG recovery domain

Node	Site	Role
Node 2	Site A	Primary
Node 3	Site B	Backup1
Node 4	Site B	Backup2
Node 1	Site A	Backup3

When the CHGCRGPRI command is executed with the defined recovery domain, the following actions occur:

1. CHGCRGPRI is executed on any node in the cluster.
2. IASP 33 is varied off on Node 2 and Node 3.
3. A disk switch occurs between Node 2 and Node 1.
4. A replication (geographic mirror) switch occurs between Node 3 and Node 1.
5. IASP 33 is varied on to an Available state on Node 3 and a Varied On state on Node 1.
6. CRG recovery domain is reordered to reflect the changes.
7. Control is released back to the user.

In Figure 9-13, the geographic mirroring configuration is set up with the source system indicated by the red star (on the right) and the target system indicated by the blue star (on the left). When the command CHGCRGPRI is executed with the defined recovery domain, the following actions occur:

1. Geographic mirroring is suspended with tracking.
2. IASP 33 is varied off on Node 2.
3. Node 3 is promoted to primary.
4. Switch disk occurs on site A to Node 1.

5. IASP 33 is varied on to an available state on Node 1.
6. Geographic mirroring resumes from Node 3 to Node 1.
7. The recovery domain is rebuilt.

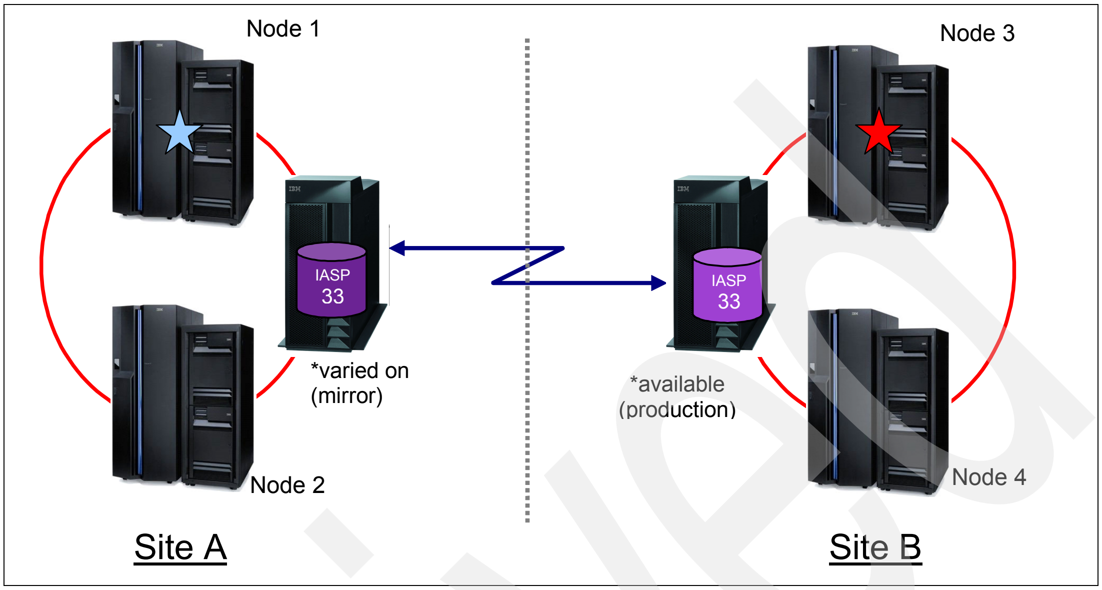


Figure 9-13 Recovery domain state 2

Table 9-6 shows the current role of each node in the cluster.

Table 9-6 CRG recovery domain

Node	Site	Role
Node 3	Site B	Primary
Node 4	Site B	Backup1
Node 1	Site A	Backup2
Node 2	Site A	Backup3

Starting with the previous state, the CHGCRGPRI command is executed for the environment in Figure 9-14 with the defined recovery domain and the star indicating the current owner of the IASPs. The red star on the lower right side is for the primary node, and the blue star on the upper left side is for the backup Node 1. It also indicates that geographic mirroring is between Node 3 and Node 1. The following changes result:

1. CHGCRGPRI is executed on any node in the cluster.
2. IASP 33 is varied off on Node 3 and Node 1.
3. A disk switch occurs between Node 3 and Node 4.
4. IASP 33 is varied on to an Available state on Node 4 and a Varied On state on Node 1.
5. CRG recovery domain is reordered to reflect the changes.
6. Control is released back to the end user.
7. Geographic mirroring is suspended with tracking.
8. IASP 33 is varied off on Node 3.
9. Switch disk occurs on site B to Node 4.
10. IASP 33 is varied on to an available state on Node 4.
11. Geographic mirroring resumes from Node 4 to Node 1.
12. The recovery domain is rebuilt.

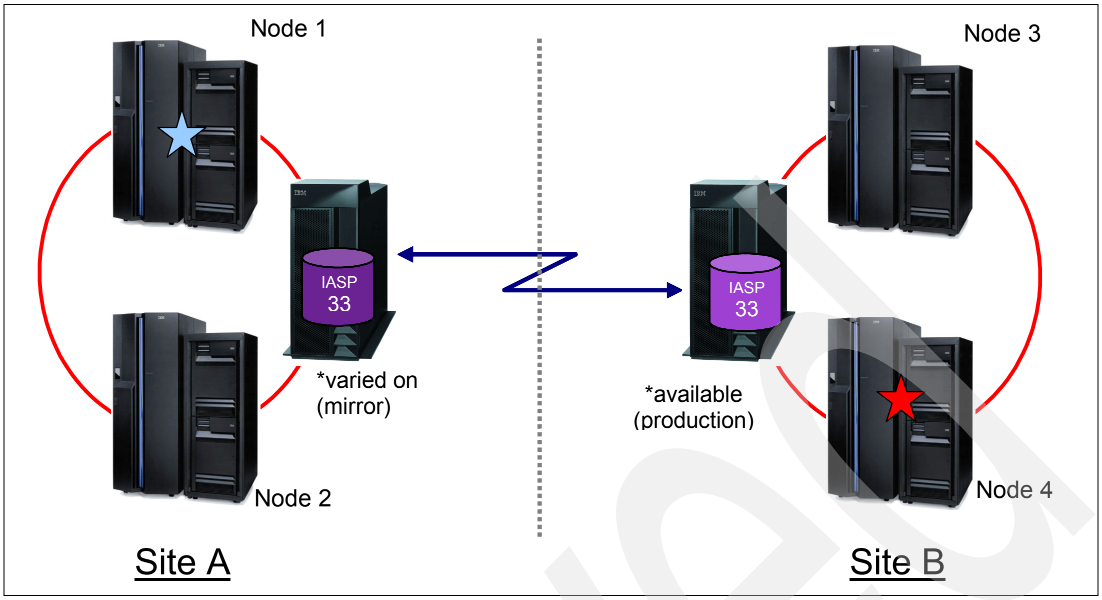


Figure 9-14 Recovery domain state 3

Table 9-7 shows the resulting roles of nodes within the cluster.

Table 9-7 CRG recovery domain

Node	Site	Role
Node 4	Site B	Primary
Node 1	Site A	Backup1
Node 2	Site A	Backup2
Node 3	Site B	Backup3

Note: Geographic mirroring must always occur between defined sites. The primary node is always the production mirror, and the next node in the recovery domain in a different site than the primary is the mirror node.

The flow chart in Figure 9-15 describes the CHGCRGPRI algorithm and allows the user to understand the resources that will be moved when a CHGCRGPRI command is issued. Use this chart with the provided recovery domain examples to walk through a CHGCRGPRI command. Make sure that you understand the resultant action of running switchovers in your high availability environment. The next CHGCRGPRI command should return the previous example to its initial configuration with the primary node being Node 1 mirroring to Node 3.

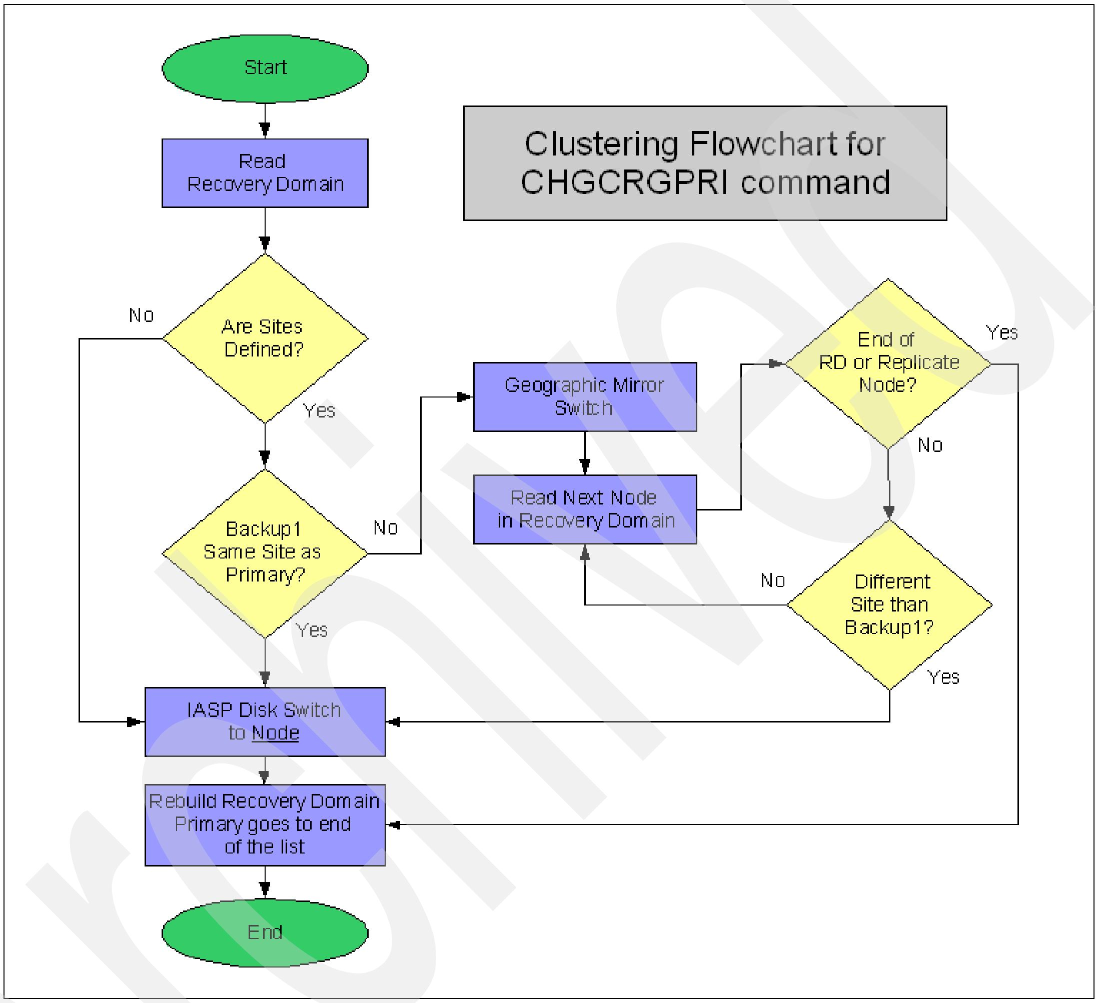

Start
Read Recovery Domain
Clustering Flowchart for CHGCRGPRI command
No
Are Sites Defined?
Yes
Geographic Mirror Switch
End of RD or Replicate Node?
Yes
No
Backup1 Same Site as Primary?
No
Read Next Node in Recovery Domain
No
Different Site than Backup1?
Yes
Yes
IASP Disk Switch to Node
Rebuild Recovery Domain Primary goes to end of the list
End

Figure 9-15 Recover decision flow chart

Notes for Figure 9-15:

- The "End of recovery domain or replicate" question asks if the previous step to read in a node from the recovery domain was successful. The "No" decision is not reached until the reading of the next node in the recovery domain does not produce a node. That is, it is off the end of the recovery domain node list.
- When following the flow, the nodes are not reordered in the recovery domain until the last step.

9.9 User ID and group ID

It is important to synchronize user IDs (UIDs) and group IDs (GIDs) of user profiles that own objects in the IASP that is switchable between systems. During a switchover or failover, the mirror copy of the IASP is promoted to primary and changed to an Available state. When this happens, all objects in the IASP are mapped to an owning user profile. If the profile's UID or GID does not match from the source system to the target system. All objects that have a mismatch need to go through a change of object ownership to map to the object to the UID or GID on the new system.

This mapping happens every time a switchover is performed until the UIDs and GIDs for the profiles are manually synchronized. Management Central has the ability to synchronize the UID and GID for profiles across multiple systems. The clustering framework also has the ability to synchronize these user profile attributes through an Administrative Domain.

9.10 Summary and recommendations

In summary, an ideal environment for geographic mirroring begins with a balanced environment between the two nodes that own the production and mirror copy of the IASP. We make the following recommendations:

- Enough processor is required to handle the additional workload of geographic mirroring.
- Both systems must have a correctly sized machine pool.
- A sufficient number of disk arms is required on each system. Check their utilization before you configure geographic mirroring. Remember that geographic mirroring adds an extra workload to the disk subsystem during synchronization.
- Make sure that the system ASP size is large enough to handle the temporary storage that is required by the applications that use the IASPs.
- Using two communication lines of the same type yields acceptable performance in most cases.
- Ensure that there are no network bottlenecks between the source and target systems as indicated earlier.

For additional information about IASP and geographic mirroring performance, refer to the *V5R3 Performance Capabilities Reference - July 2004*, which is in the Information Center on the Web at the following address:

http://publib.boulder.ibm.com/infocenter/iseries/v5r3/topic/rzahx/sc410607.pdf

10

Troubleshooting

In order to set up cross-site mirroring (XSM), a cluster environment must be set up. In this chapter, we present the order that we recommend for troubleshooting any issues or problems that can arise with respect to XSM. We discuss ways to troubleshoot the cluster environment first, since problems with the cluster can and will cause problems with XSM. We also discuss the communications pathway used with XSM and ways to avoid problems in that area. Our example shows an XSM environment that has also implemented geographic mirroring.

10.1 Release levels and recommended fixes

As with any product or function on the System i platform, it is important to maintain the most current fixes for clustering, independent auxiliary storage pools (IASPs), and XSM. If your system remains at an older PTF level, it is possible to encounter known and corrected problems that can be prevented by maintaining a current PTF level.

In this section, we outline considerations for XSM with regard to the cluster release level and explain how to check for the latest PTFs for these products.

10.1.1 Cross-site mirror release considerations

XSM can run only in a cluster that is set to cluster version 4, which requires that all participating nodes be at V5R3M0 or higher of OS/400 or i5/OS. If the cluster is not at cluster version 4, and you attempt to configure XSM, error message CPFBB70 is posted.

10.1.2 Recommended fixes

Recommended fix pages were established to help customers stay current with the latest fixes. You can find the main recommended fixes Web page at the following address:

http://www-912.ibm.com/s_dir/slkbase.nsf/recommendedfixes

> **Note:** The recommended fixes page lists only PTFs that are not on a cumulative PTF package.

The following recommended fixes are the primary ones that affect XSM:

- Server Firmware: Update Policy Set to Hardware Management Console (HMC; if using System i5 hardware)
- Server Firmware: Update Policy Set to Operating System (if using non-System i5 hardware)
- Cluster
- XSM
- Switchable IASP (if more than two nodes)

> **Important:** If your system is using System i5 hardware, then check the following Web addresses for *firmware updates* at regular intervals:
>
> - If you have the firmware update policy set to *HMC*
>
>   http://www-912.ibm.com/s_dir/slkbase.nsf/ibmscdirect/E58D7BBF0EAC9A2786256EAD005F54D8
>
> - If you have the firmware update policy set to *operating system*
>
>   http://www-912.ibm.com/s_dir/slkbase.nsf/ibmscdirect/604992740F846A4986256FD3006029B5
>
> In addition, check the Software Knowledge Base for recommended fixes at regular intervals:
>
> http://www-912.ibm.com/s_dir/slkbase.NSF/wHighAv?OpenView&Start=1

10.2 Troubleshooting clusters

In the following sections, we focus on issues that specifically affect device cluster resource groups (CRGs) and are important in a cross-site mirror environment.

10.2.1 Cluster and CRG status

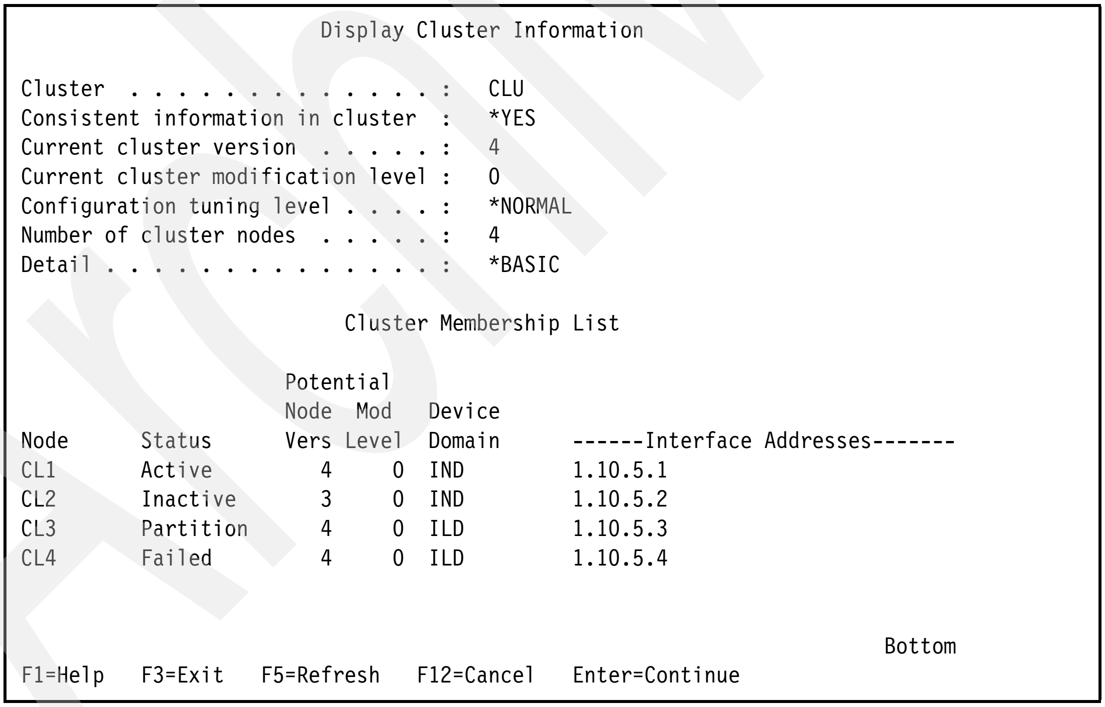
It is important to know the health of your cluster. If there is a problem with XSM, we recommend that you start with the status of the cluster to help you to understand if the problem is with the cluster itself or specific to XSM.

In the following sections, we explain how to check the health of your cluster by viewing the status of the cluster and the CRG by using a command line and iSeries Navigator.

Checking the cluster status by using command line

To check the status of the cluster from the command line, use the Display Cluster Information (DSPCLUINF) command.

On the Display Cluster Information panel (Figure 10-1), check the value of the Consistent information in cluster parameter. If the value is set to *NO, then the rest of the information is not current. This node only has the last information that it had when it was part of the cluster. If the value is set to *YES, then you know that the rest of the information is current. Always check this parameter first.



```
                          Display Cluster Information

Cluster  . . . . . . . . . . . . . :   CLU
Consistent information in cluster  :   *YES
Current cluster version  . . . . . :   4
Current cluster modification level :   0
Configuration tuning level . . . . :   *NORMAL
Number of cluster nodes  . . . . . :   4
Detail . . . . . . . . . . . . . . :   *BASIC

                            Cluster Membership List

                      Potential
                      Node  Mod   Device
Node      Status      Vers Level  Domain      ------Interface Addresses-------
CL1       Active         4     0  IND         1.10.5.1
CL2       Inactive       3     0  IND         1.10.5.2
CL3       Partition      4     0  ILD         1.10.5.3
CL4       Failed         4     0  ILD         1.10.5.4


                                                                        Bottom
F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 10-1 Display Cluster Information panel

Next, check the Current cluster version and Current cluster modification level. Below the Cluster Membership List in the lower half of the panel, each cluster node is listed. For each node, the potential cluster version and release are listed. Remember that this version is only the potential version and release at which the cluster can be. It is not necessarily its current version and release level.

For each cluster node, the device domain is listed. Each node can be only in one device domain. The IP interface addresses are listed for each cluster node as well. These addresses are the interfaces that are configured for cluster communications and do not represent all of the IP interfaces that exist on each system.

The panel in Figure 10-1 also lists the status of each node and shows each of the most common status types. An *Active* status means that node is currently part of the cluster. An *Inactive* status means that node is not currently part of the cluster and left the cluster in a normal fashion. The Active and Inactive status types are related to the status of Active P and Inactive P, which indicate a pending status, which indicate that the node is in a transition state.

A *Partition* status means that the node has lost contact with the cluster, and its status is unknown. A *Failed* status means that node has lost contact and has disconnected in some abnormal fashion. Other status types are possible; you can see more information about these status types by using the help function on the panel.

You can obtain a lot of other good configuration information by entering the following command:

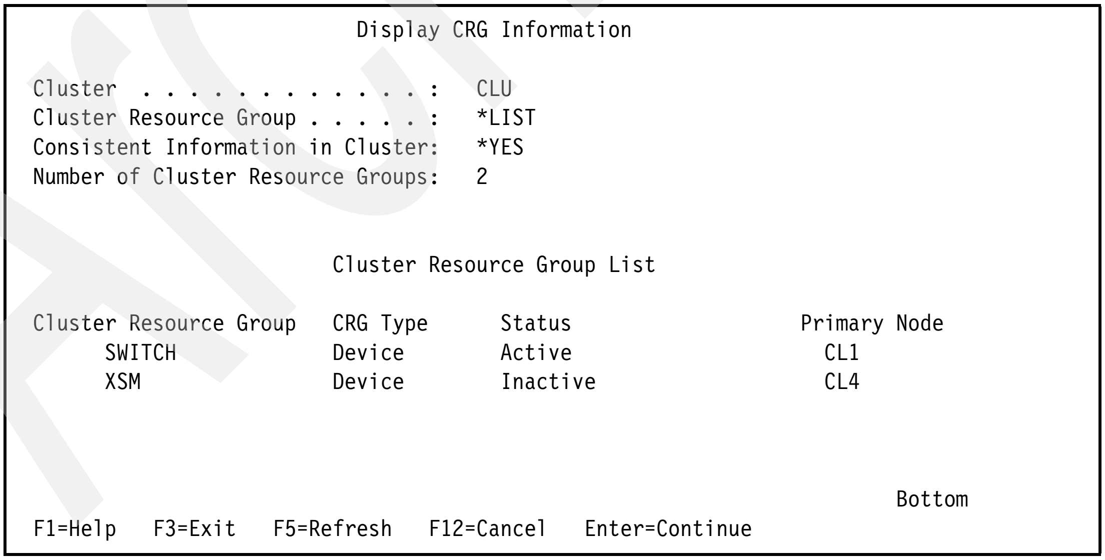
```
DSPCLUINF CLUSTER(cluster_name) DETAIL(*FULL)
```

The example in Figure 10-1 shows the information that is seen with the DETAIL(*BASIC) parameter.

Checking the CRG status by using a command line

After you determine the status of the cluster, it is important to check the status of the CRGs. Use the Display CRG Information (DSPCRGINF) command to do this.

In the Display CRG Information panel (Figure 10-2), as with the DSPCLUINF command, you must first check the Consistent Information in Cluster parameter. If the value is *YES, then the rest of the information is current. If the value is *NO, then the rest of the information is the last status information that this particular cluster node had.



```
                             Display CRG Information

 Cluster  . . . . . . . . . . . . :   CLU
 Cluster Resource Group . . . . . :   *LIST
 Consistent Information in Cluster:   *YES
 Number of Cluster Resource Groups:   2


                           Cluster Resource Group List

 Cluster Resource Group   CRG Type      Status                   Primary Node
       SWITCH             Device        Active                     CL1
       XSM                Device        Inactive                   CL4




                                                                         Bottom
 F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 10-2 Display CRG Information panel

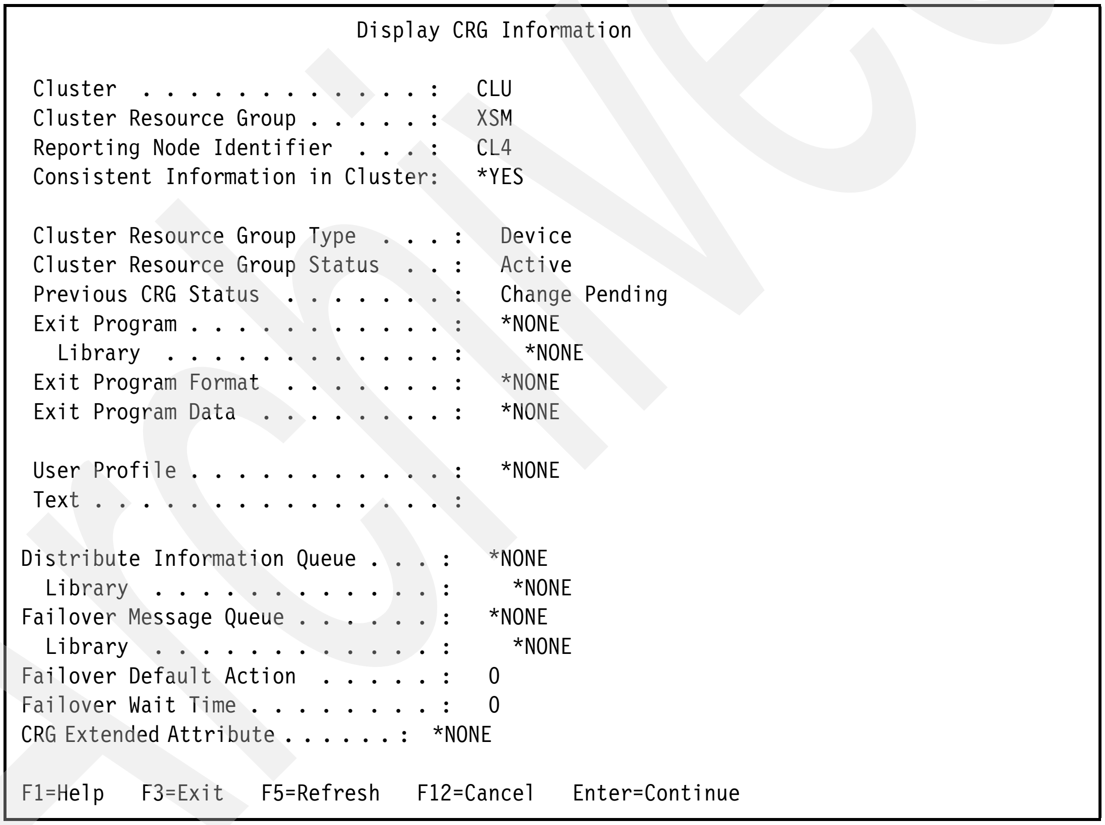
The two most common status types are Active and Inactive. An *Active* status indicates that the CRG is up and running and is able to perform switchovers and failovers. An Inactive status indicates that the CRG is not active and is unable to perform switchovers and failovers. You can use the Help function on this panel to view the many other status types and their meanings. Keep in mind that, with a device CRG in an Inactive status, the IASP can still be in use, and XSM can still be running. Of course, this is not an ideal status, since switchovers and failovers cannot be performed.

The Display CRG Information panel also indicates the current active primary node and each of the CRG types. It is important to know which node in the CRG is currently the primary node. For XSM, the CRG type is always a device CRG.

For a more detailed view of each CRG as shown in Figure 10-3, use the following command for the specific CRG in question:

```
DSPCRGINF CLUSTER(cluster_name) CRG(CRG_name)
```

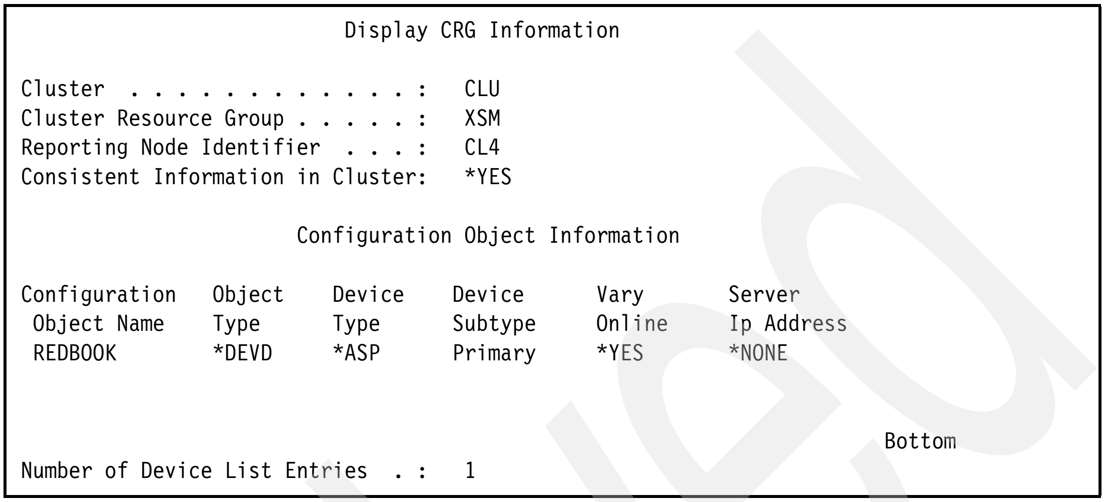
The first panel of information (Figure 10-3) indicates much of the same information that is seen with the generic summary panel. It also indicates the previous CRG status, whether an exit program is used, whether a failover message queue is defined, and other helpful information.



```
                              Display CRG Information

  Cluster  . . . . . . . . . . . . . :   CLU
  Cluster Resource Group . . . . . . :   XSM
  Reporting Node Identifier  . . . . :   CL4
  Consistent Information in Cluster:   *YES

  Cluster Resource Group Type  . . . :   Device
  Cluster Resource Group Status  . . :   Active
  Previous CRG Status  . . . . . . . :   Change Pending
  Exit Program . . . . . . . . . . . :   *NONE
    Library  . . . . . . . . . . . . :     *NONE
  Exit Program Format  . . . . . . . :   *NONE
  Exit Program Data  . . . . . . . . :   *NONE

  User Profile . . . . . . . . . . . :   *NONE
  Text . . . . . . . . . . . . . . . :

 Distribute Information Queue . . . :   *NONE
   Library  . . . . . . . . . . . . :     *NONE
 Failover Message Queue . . . . . . :   *NONE
   Library  . . . . . . . . . . . . :     *NONE
 Failover Default Action  . . . . . :   0
 Failover Wait Time . . . . . . . . :   0
 CRG Extended Attribute . . . . . . :  *NONE

 F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 10-3 Display CRG Information panel for a specific CRG (Part 1 of 3)

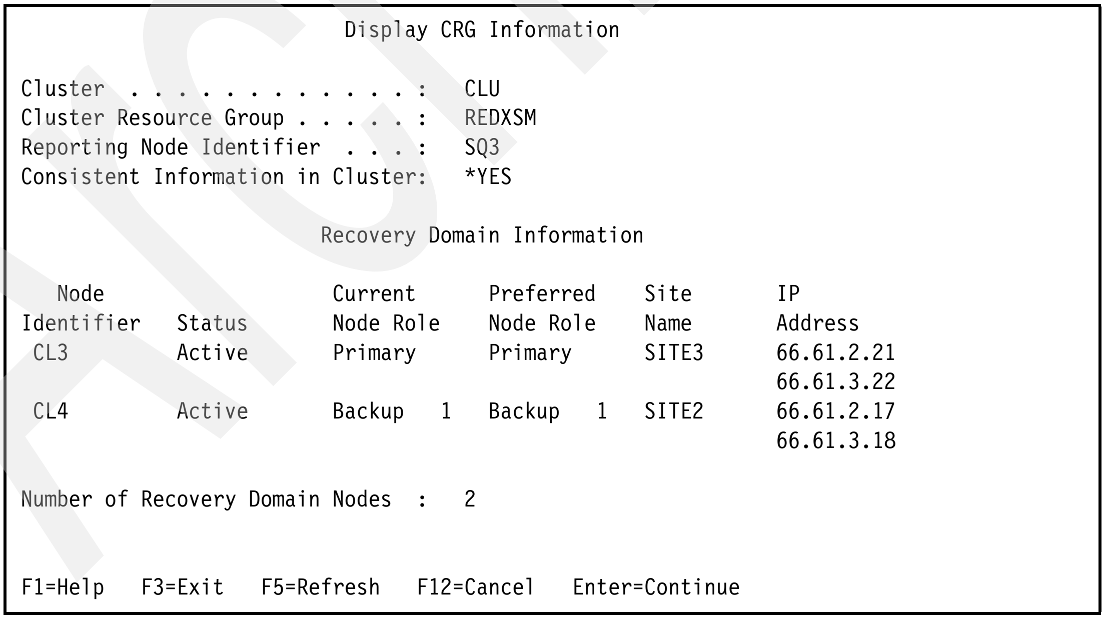
You press the Enter key to see the second panel, which is shown in Figure 10-4. The second panel indicates which IASPs are associated with the CRG. It also indicates if the IASP is a primary or secondary, if it has a takeover IP address, and if the CRG should try to vary on the IASP when it performs a switchover or failover.



```
                            Display CRG Information

Cluster  . . . . . . . . . . . . :   CLU
Cluster Resource Group . . . . . :   XSM
Reporting Node Identifier  . . . :   CL4
Consistent Information in Cluster:   *YES

                        Configuration Object Information

Configuration   Object    Device    Device      Vary       Server
 Object Name    Type      Type      Subtype     Online     Ip Address
 REDBOOK        *DEVD     *ASP      Primary     *YES       *NONE



                                                                        Bottom
Number of Device List Entries  . :   1
```

Figure 10-4 Display CRG Information panel for a specific CRG (Part 2 of 3)

You press the Enter key again to view the third panel of information shown in Figure 10-5. In the third panel, the recovery domain information is listed. Each node is listed with its status, current and preferred role, the site name, and IP addresses. This information is useful in order to check if the mirror copy of the IASP is in synch with the production copy of the IASP. If both copies are in synch, then the node with the mirror copy has a status of Active. If not, the status is Ineligible.

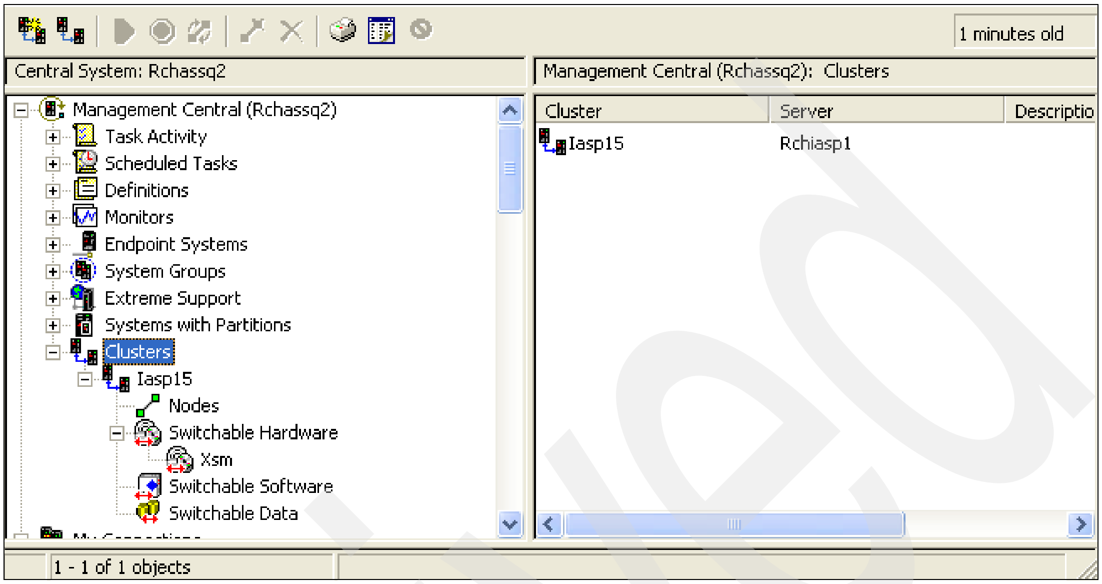


```
                            Display CRG Information

Cluster  . . . . . . . . . . . . :   CLU
Cluster Resource Group . . . . . :   REDXSM
Reporting Node Identifier  . . . :   SQ3
Consistent Information in Cluster:   *YES

                          Recovery Domain Information

   Node                   Current      Preferred    Site       IP
Identifier   Status       Node Role    Node Role    Name       Address
 CL3         Active       Primary      Primary      SITE3      66.61.2.21
                                                               66.61.3.22
 CL4         Active       Backup   1   Backup   1   SITE2      66.61.2.17
                                                               66.61.3.18

Number of Recovery Domain Nodes  :   2



F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 10-5 Display CRG Information panel for a specific CRG (Part 3 of 3)

Checking the cluster status using iSeries Navigator

In iSeries Navigator, clustering is a subcategory of Management Central, as shown in Figure 10-6.



Management Central (Rchassq2): Clusters

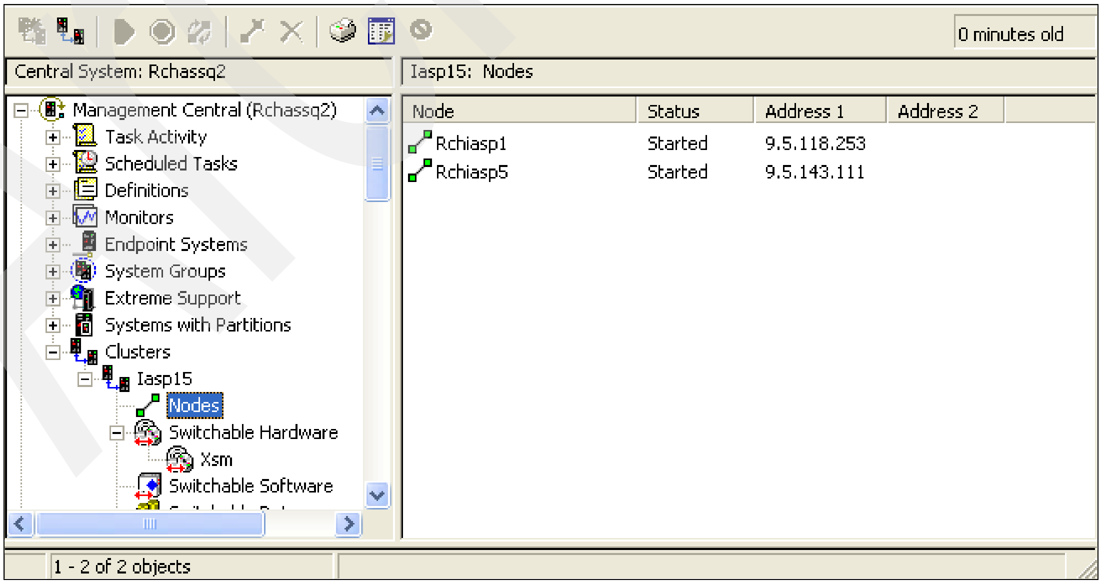
Cluster	Server	Descriptio
Iasp15	Rchiasp1	

1 - 1 of 1 objects

Figure 10-6 Clustering in Management Central

In this example, the name of the cluster is Iasp15. The GUI refers to the various cluster resource groups differently than what is seen on a 5250 panel. A device CRG is called *Switchable Hardware*. An application CRG is called *Switchable Software*. A data CRG is called *Switchable Data*. It is important to remember these differences when using either the command line or the GUI.

To see the status of the nodes in the cluster, click **Nodes** under the cluster name, as shown in Figure 10-7. The status of each node in the cluster is listed in the right pane.



Iasp15: Nodes

Node	Status	Address 1	Address 2
Rchiasp1	Started	9.5.118.253	
Rchiasp5	Started	9.5.143.111	

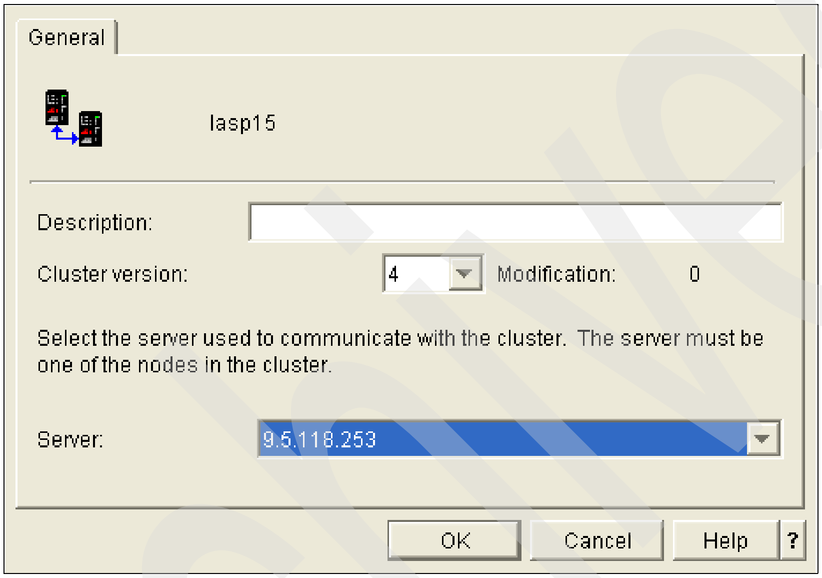
1 - 2 of 2 objects

Figure 10-7 Cluster node status

The GUI also refers to the types of node status differently. A status of *Started* indicates that the node is currently part of the cluster. A status of *Stopped* indicates that the node is not currently part of the cluster and left the cluster in a normal fashion. The statuses of Started and Stopped also have transition states of Starting and Stopping, respectfully.

A status of *Not communicating* indicates that the node lost contact with the cluster and its status is unknown. A status of *Failed* indicates the node has lost contact and has disconnected in some abnormal fashion. There are other status types as well. You can see the full list of status types and meanings by using the Help function.

When you right-click the cluster and choose the **Properties** option, you see an additional window (Figure 10-8) that shows you the cluster version. In this window, using your iSeries Navigator session, you also define through which node of the cluster you are connected.



Iasp15

Description:

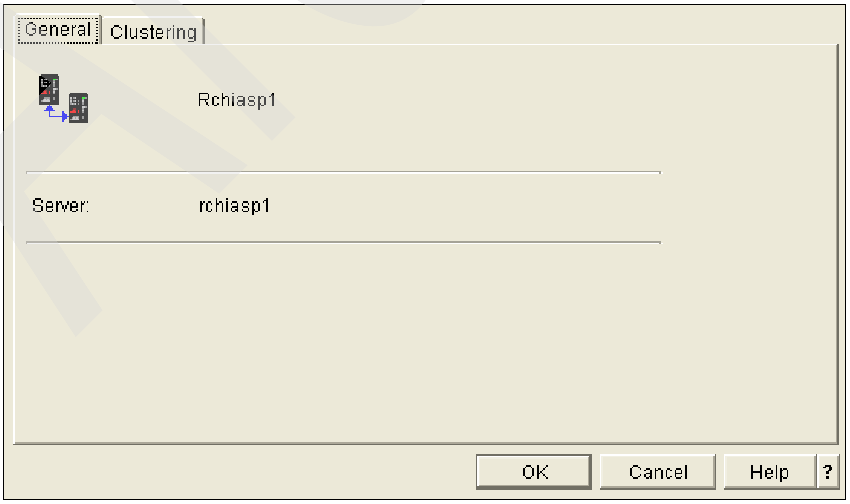
Cluster version: 4 Modification: 0

Select the server used to communicate with the cluster. The server must be one of the nodes in the cluster.

Server: 9.5.118.253

Figure 10-8 Cluster properties

When you right-click the cluster node and choose the **Properties** option, you see an additional window. Under the General tab (Figure 10-9), you see the name of the server and the name of the cluster node.

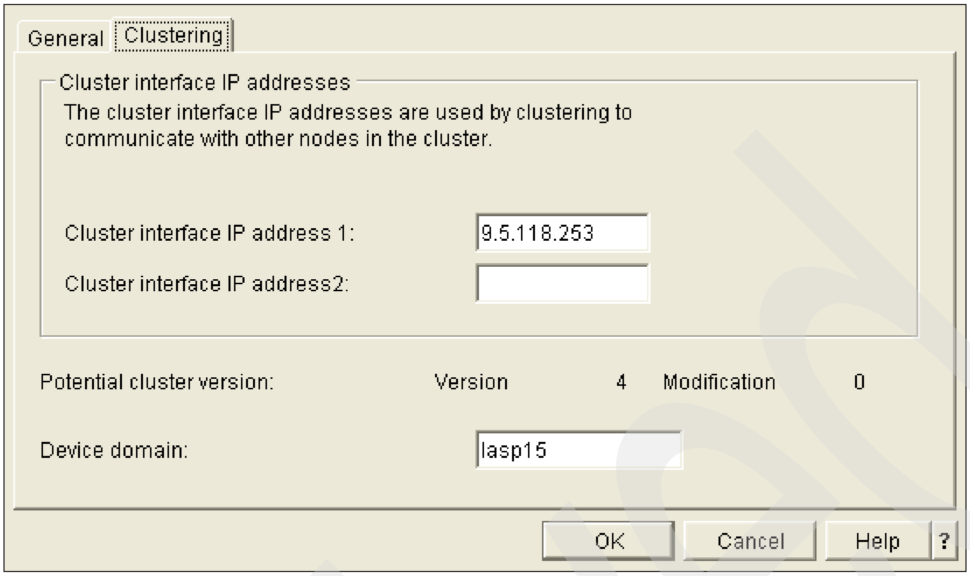


Rchiasp1

Server: rchiasp1

Figure 10-9 Cluster node - General tab

Under the Clustering tab (Figure 10-10), you see the IP addresses and the device domain.



Cluster interface IP addresses

The cluster interface IP addresses are used by clustering to communicate with other nodes in the cluster.

Cluster interface IP address 1: 9.5.118.253

Cluster interface IP address2:

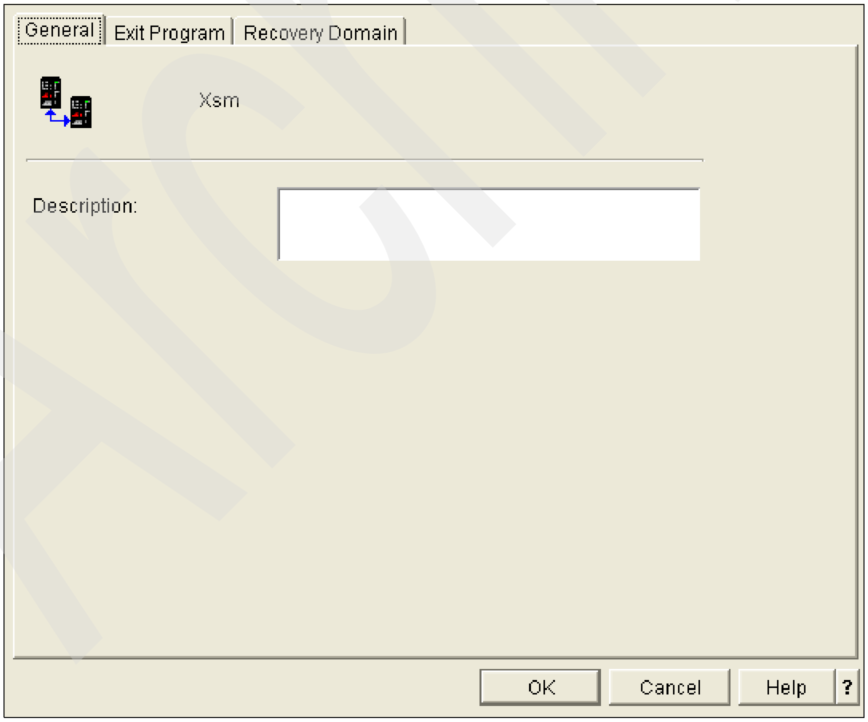
Potential cluster version: Version 4 Modification 0

Device domain: Iasp15

Figure 10-10 Cluster node - Clustering tab

Checking the device CRG status using iSeries Navigator

To see the status of the device CRG, you select **Switchable Hardware**, select the ***name of the CRG***, right-click, and choose the **Properties** option. The General tab shown in Figure 10-11 shows the name of the CRG and any description that is associated with it.

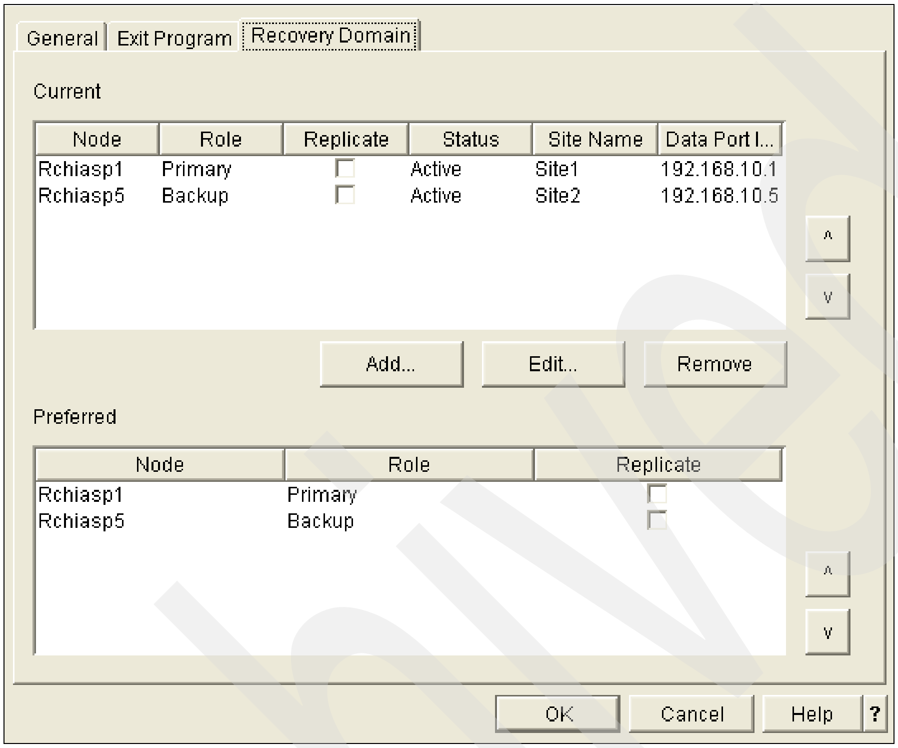


Xsm

Description:

Figure 10-11 Device CRG General tab

The Recovery Domain tab (Figure 10-12) shows the cluster nodes between which the device CRG can switch, as well as the current role of each node and the status of each. For XSM configurations, this window also shows the site name and data port IP addresses. To change the node order in the recovery domain, highlight the node and use the up and down arrows on the right side of the window.



Current

Node	Role	Replicate	Status	Site Name	Data Port I...
Rchiasp1	Primary	□	Active	Site1	192.168.10.1
Rchiasp5	Backup	□	Active	Site2	192.168.10.5

Preferred

Node	Role	Replicate
Rchiasp1	Primary	□
Rchiasp5	Backup	□

Figure 10-12 Recovery Domain tab

10.2.2 Cluster jobs

In addition to the status of the cluster and the CRG, it is important to know which jobs you should expect to find when clustering is active and how to correct errors. In this section, we list those jobs that are important to XSM and the related cluster.

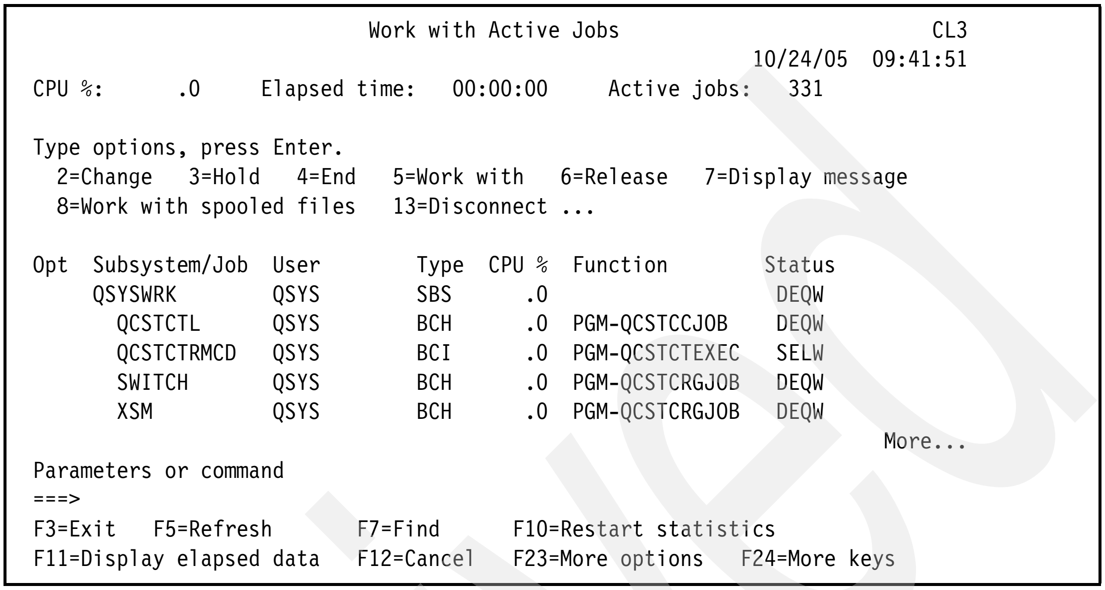
Finding cluster jobs

Table 10-1 lists the cluster jobs that are most important for an XSM environment.

Table 10-1 Cluster job information

Server name	Job description	Subsystem	Job name
Cluster Resource Services	QGPL/QDFTJOBD	QSYSWRK	QCSTCTL
Cluster Resource Services	QGPL/QDFTJOBD	QSYSWRK	QCSTCRGM
Cluster Resource Services	QGPL/QDFTJOBD	QSYSWRK	CRG name

You can see that each CRG has an associated job with the same name, which runs under the QSYSWRK subsystem. The QCSTCTL and QCSTCRGM jobs also run in QSYSWRK, as shown in Figure 10-13.

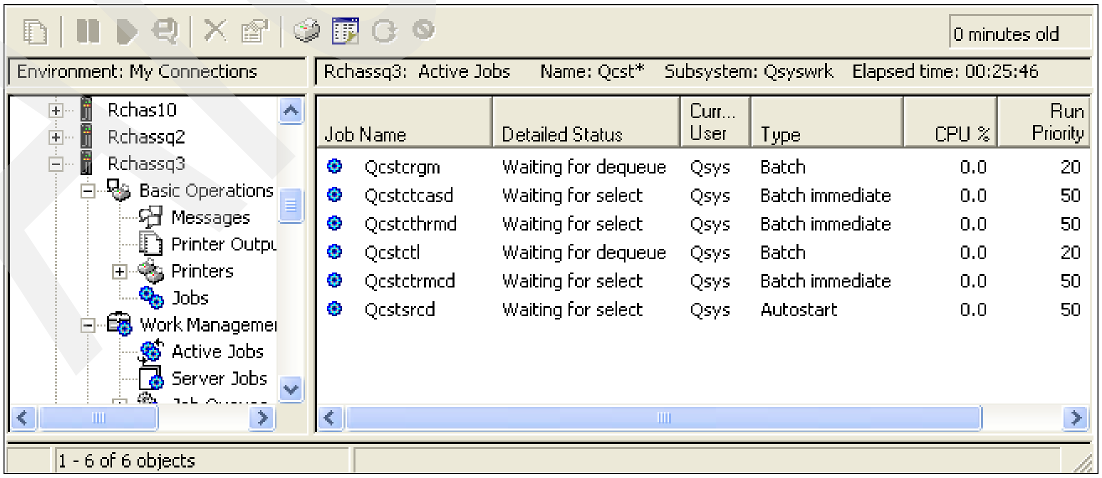


```
                              Work with Active Jobs                        CL3
                                                             10/24/05  09:41:51
CPU %:      .0     Elapsed time:   00:00:00     Active jobs:   331

Type options, press Enter.
  2=Change   3=Hold   4=End    5=Work with   6=Release   7=Display message
  8=Work with spooled files    13=Disconnect ...

Opt  Subsystem/Job  User        Type  CPU %  Function        Status
     QSYSWRK        QSYS        SBS      .0                  DEQW
       QCSTCTL      QSYS        BCH      .0  PGM-QCSTCCJOB    DEQW
       QCSTCTRMCD   QSYS        BCI      .0  PGM-QCSTCTEXEC   SELW
       SWITCH       QSYS        BCH      .0  PGM-QCSTCRGJOB   DEQW
       XSM          QSYS        BCH      .0  PGM-QCSTCRGJOB   DEQW
                                                                       More...
Parameters or command
===>
F3=Exit   F5=Refresh       F7=Find      F10=Restart statistics
F11=Display elapsed data   F12=Cancel   F23=More options   F24=More keys
```

Figure 10-13 Example of cluster jobs that are important for XSM

In this example, two CRGs, called SWITCH and XSM, are defined on the system. Each node in the cluster has jobs with the same name running in the QSYSWRK subsystem. It does not matter if that node is not part of that particular CRG. If it is in the cluster, these jobs will exist.

It is important to note that other OS/400 jobs exist that begin with the letters QCST. These jobs are related to Resource Monitoring and Control (RMC) and are independent of clustering. RMC is related to specific reporting done through the Hardware Management Console (HMC). The names are similar, which can sometimes make it more difficult to find the cluster jobs and separate them from the RMC jobs. RMC jobs are started if you have System i5 hardware or earlier models, regardless of whether an HMC is used. These jobs can also be seen in iSeries Navigator, as shown in Figure 10-14.



Rchassq3: Active Jobs   Name: Qcst*   Subsystem: Qsyswrk   Elapsed time: 00:25:46

Job Name	Detailed Status	Curr... User	Type	CPU %	Run Priority
Qcstcrgm	Waiting for dequeue	Qsys	Batch	0.0	20
Qcstctcasd	Waiting for select	Qsys	Batch immediate	0.0	50
Qcstcthrmd	Waiting for select	Qsys	Batch immediate	0.0	50
Qcstctl	Waiting for dequeue	Qsys	Batch	0.0	20
Qcstctrmcd	Waiting for select	Qsys	Batch immediate	0.0	50
Qcstsrcd	Waiting for select	Qsys	Autostart	0.0	50

1 - 6 of 6 objects

Figure 10-14 Cluster and RMC jobs as seen in iSeries Navigator

For more information about RMC, refer to the IBM Systems Hardware Information Center, which you can find on the Web at the following address:

http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp?lang=en

Problems with cluster and CRG jobs

If a cluster job is missing or has ended, the cluster node needs to be ended, by using the ENDCLUNOD command, and then restarted, by using the STRCLUNOD command. To troubleshoot why the job ended, look for the job logs for any of these jobs.

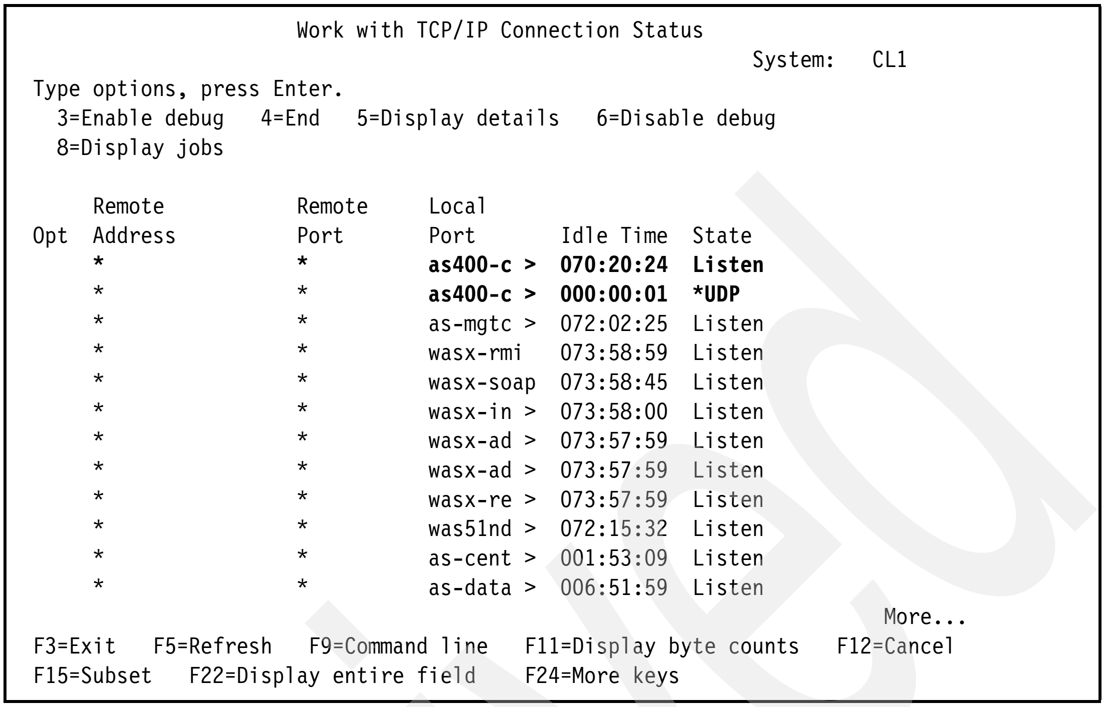
When a CRG job is ended, clustering is still active. When the CRG job is ended for the primary node of the CRG, a failover is automatically initiated. To restart the CRG job, the cluster node needs to be ended by using the ENDCLUNOD command and then restarted by using the STRCLUNOD command. However, during the failover process, these commands cannot be processed until the failover has completed. Users can initiate the command but must wait for the failover to complete before the command is processed.

10.2.3 Cluster communications

In order for clustering to work successfully, the nodes in the cluster must be able to communicate with one another through TCP/IP. The easiest way to check the ability for the nodes to communicate is to use the PING command between nodes. Keep in mind that the PING command uses Internet Control Messaging Protocol (ICMP). If ICMP is not allowed between the nodes, PING is not a good test of connectivity.

The Internet daemon (INETD) server must be started for a node in the cluster to be added or started. The server job is QTOGINTD and runs in the QSYSWRK subsystem. We recommend that you set this server to autostart by using the STRTCP command. When this job is active, TCP port 5550 (as400-cluster) can be seen through NETSTAT. If you run the NETSTAT command and choose option 3 (Work with TCP/IP connection status), you can see two instances of as400-c > under the Local Port column, as shown in Figure 10-15. The full name of the first one is `as400-cluster` and is used for INETD.

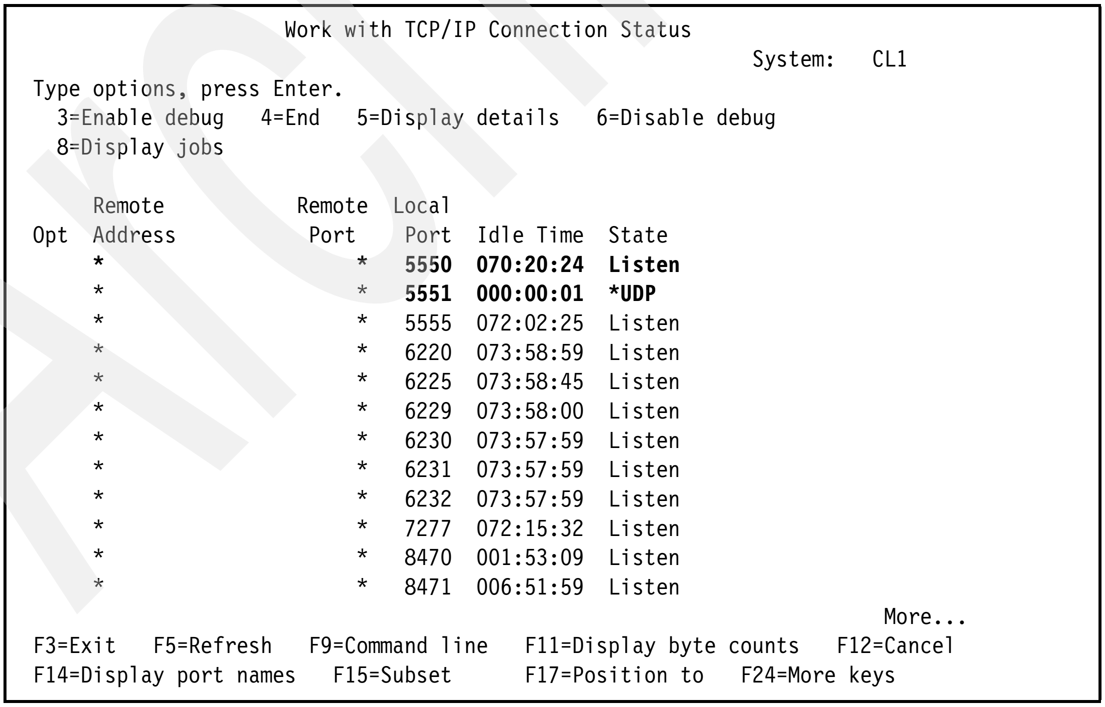
Clustering uses its own User Datagram Protocol (UDP) port for heartbeat messaging. UDP is another transport-layer protocol, such as TCP, but it uses a different format for its header information. When clustering is started on the node, UDP port 5551 (as400-clusterbase) can be seen through NETSTAT. If you run the NETSTAT command, and choose option 3 (Work with TCP/IP connection status), you can see the two instances of as400-c > under the Local Port column as shown in Figure 10-15. The full name of the second one is `as400-clusterbase` and is used for clustering messaging. If this port is in use by anything else before clustering on that node is started, it prevents clustering from starting successfully.



```
                      Work with TCP/IP Connection Status
                                                            System:   CL1
Type options, press Enter.
  3=Enable debug   4=End   5=Display details   6=Disable debug
  8=Display jobs

     Remote           Remote     Local
Opt  Address          Port       Port       Idle Time  State
     *                *          as400-c >  070:20:24  Listen
     *                *          as400-c >  000:00:01  *UDP
     *                *          as-mgtc >  072:02:25  Listen
     *                *          wasx-rmi   073:58:59  Listen
     *                *          wasx-soap  073:58:45  Listen
     *                *          wasx-in >  073:58:00  Listen
     *                *          wasx-ad >  073:57:59  Listen
     *                *          wasx-ad >  073:57:59  Listen
     *                *          wasx-re >  073:57:59  Listen
     *                *          was51nd >  072:15:32  Listen
     *                *          as-cent >  001:53:09  Listen
     *                *          as-data >  006:51:59  Listen
                                                                       More...
F3=Exit   F5=Refresh   F9=Command line   F11=Display byte counts   F12=Cancel
F15=Subset   F22=Display entire field    F24=More keys
```

Figure 10-15 TCP/IP connection status - port names

From the NETSTAT (Option 3) panel, you can press F14 to view the port numbers that are shown in Figure 10-16. TCP port 5550 is used for INETD, while UDP port 5551 is used for clustering.



```
                     Work with TCP/IP Connection Status
                                                            System:   CL1
Type options, press Enter.
  3=Enable debug   4=End   5=Display details   6=Disable debug
  8=Display jobs

     Remote           Remote  Local
Opt  Address           Port    Port  Idle Time  State
     *                     *   5550  070:20:24  Listen
     *                     *   5551  000:00:01  *UDP
     *                     *   5555  072:02:25  Listen
     *                     *   6220  073:58:59  Listen
     *                     *   6225  073:58:45  Listen
     *                     *   6229  073:58:00  Listen
     *                     *   6230  073:57:59  Listen
     *                     *   6231  073:57:59  Listen
     *                     *   6232  073:57:59  Listen
     *                     *   7277  072:15:32  Listen
     *                     *   8470  001:53:09  Listen
     *                     *   8471  006:51:59  Listen
                                                                       More...
F3=Exit   F5=Refresh   F9=Command line   F11=Display byte counts   F12=Cancel
F14=Display port names   F15=Subset      F17=Position to   F24=More keys
```

Figure 10-16 TCP/IP connection status - port numbers

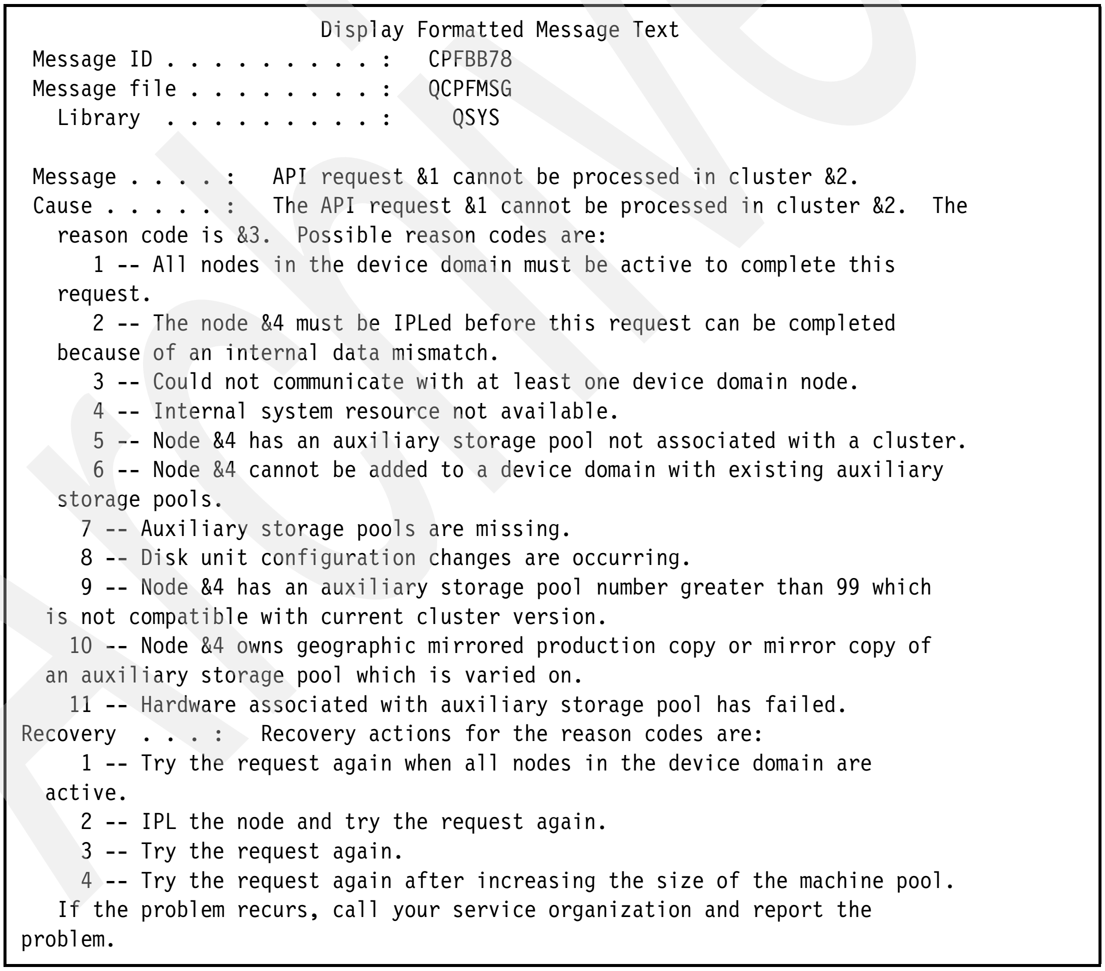
You can use the WRKSRVTBLE command to see the exact port mappings. NETSTAT Option 1 (Work with TCP/IP interface status) shows the IP addresses that are used for clustering and should be in a status of Active. The *LOOPBACK interface (127.0.0.1) must also be active.

10.2.4 Cluster domains

Auxiliary storage pool (ASP) numbers are divided among the nodes in the device domain, and only the node that owns a particular ASP number can create an IASP with that number. As a result, an ASP number that is shown on systems that do not own the IASP is the same IASP as the one on its owning system.

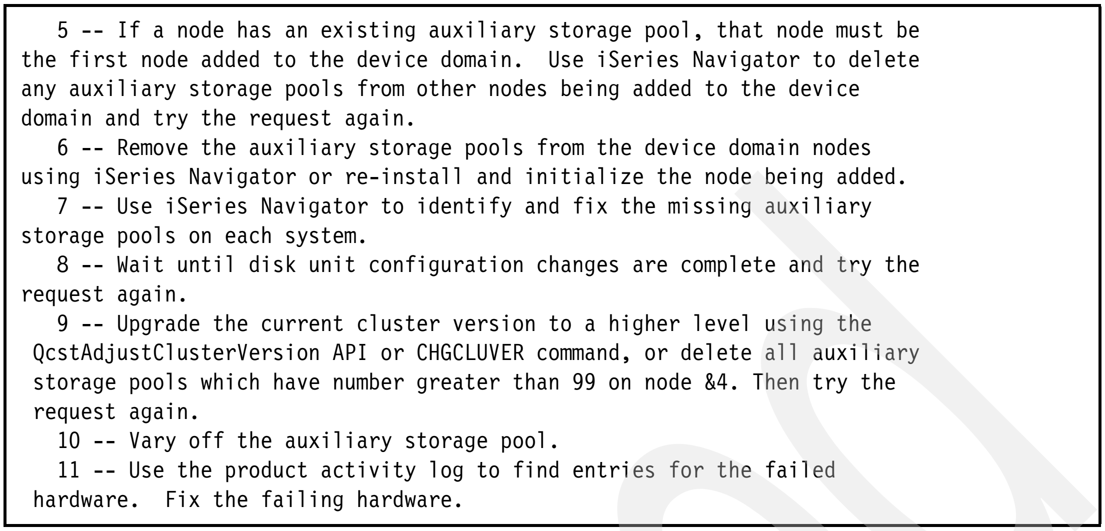
A lot of care has to be taken in future planning of which systems and nodes might be included in a cluster in the future. At some point, if you decide to add a new system to a cluster and device domain, and it already has an IASP created, there will be problems. An IASP on a system must be deleted before that system can be added to a device domain in a cluster of which it had not previously been a part.

When creating an IASP, deleting an IASP, or trying to create a cluster, you might receive error message CPFBB78 (Figure 10-17).



```
                        Display Formatted Message Text
 Message ID . . . . . . . . . :   CPFBB78
 Message file . . . . . . . . :   QCPFMSG
   Library  . . . . . . . . . :     QSYS

 Message . . . . :   API request &1 cannot be processed in cluster &2.
 Cause . . . . . :   The API request &1 cannot be processed in cluster &2.  The
   reason code is &3.  Possible reason codes are:
      1 -- All nodes in the device domain must be active to complete this
   request.
      2 -- The node &4 must be IPLed before this request can be completed
   because of an internal data mismatch.
      3 -- Could not communicate with at least one device domain node.
      4 -- Internal system resource not available.
      5 -- Node &4 has an auxiliary storage pool not associated with a cluster.
      6 -- Node &4 cannot be added to a device domain with existing auxiliary
   storage pools.
     7 -- Auxiliary storage pools are missing.
     8 -- Disk unit configuration changes are occurring.
     9 -- Node &4 has an auxiliary storage pool number greater than 99 which
  is not compatible with current cluster version.
    10 -- Node &4 owns geographic mirrored production copy or mirror copy of
  an auxiliary storage pool which is varied on.
    11 -- Hardware associated with auxiliary storage pool has failed.
Recovery  . . . :   Recovery actions for the reason codes are:
     1 -- Try the request again when all nodes in the device domain are
  active.
     2 -- IPL the node and try the request again.
     3 -- Try the request again.
     4 -- Try the request again after increasing the size of the machine pool.
   If the problem recurs, call your service organization and report the
problem.
```

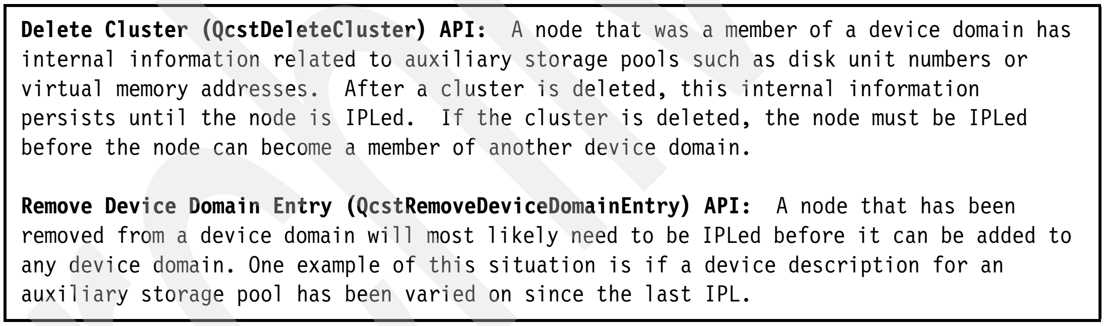
Figure 10-17 Error message CPFBB78 (Part 1 of 2)



```
   5 -- If a node has an existing auxiliary storage pool, that node must be
the first node added to the device domain.  Use iSeries Navigator to delete
any auxiliary storage pools from other nodes being added to the device
domain and try the request again.
   6 -- Remove the auxiliary storage pools from the device domain nodes
using iSeries Navigator or re-install and initialize the node being added.
   7 -- Use iSeries Navigator to identify and fix the missing auxiliary
storage pools on each system.
   8 -- Wait until disk unit configuration changes are complete and try the
request again.
   9 -- Upgrade the current cluster version to a higher level using the
 QcstAdjustClusterVersion API or CHGCLUVER command, or delete all auxiliary
 storage pools which have number greater than 99 on node &4. Then try the
 request again.
   10 -- Vary off the auxiliary storage pool.
   11 -- Use the product activity log to find entries for the failed
 hardware.  Fix the failing hardware.
```

Figure 10-18 Error message CPFBB78 (Part 2 of 2)

More information is in the API reference, but the excerpts shown in Figure 10-19 are for the Delete Cluster (QcstDeleteCluster) and Remove Device Domain Entry (QcstRemoveDeviceDomainEntry) APIs. If you receive this message, an IPL is required on the affected node.



```
Delete Cluster (QcstDeleteCluster) API:  A node that was a member of a device domain has
internal information related to auxiliary storage pools such as disk unit numbers or
virtual memory addresses.  After a cluster is deleted, this internal information
persists until the node is IPLed.  If the cluster is deleted, the node must be IPLed
before the node can become a member of another device domain.

Remove Device Domain Entry (QcstRemoveDeviceDomainEntry) API:  A node that has been
removed from a device domain will most likely need to be IPLed before it can be added to
any device domain. One example of this situation is if a device description for an
auxiliary storage pool has been varied on since the last IPL.
```

Figure 10-19 API information

10.2.5 Clustering subsystems and user profiles

When you set up clustering, make sure that you follow all the steps in the cluster configuration checklist, which is found in the iSeries Information Center, available on the Web at the following address:

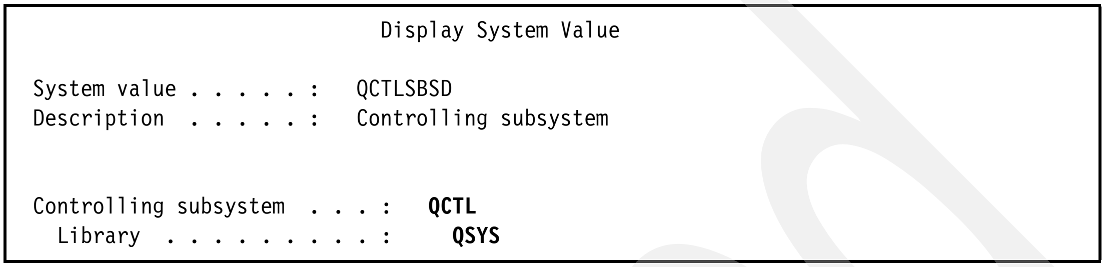
http://publib.boulder.ibm.com/infocenter/iseries/v5r3/index.jsp

> **Tip:** The Cluster configuration checklist is a helpful place to learn about what is necessary for clustering to work correctly. In the Information Center, it is under **Systems Management → Clusters → Plan for Clusters → Cluster configuration checklist**.

One item on the cluster configuration checklist is to verify that the QUSER profile does not have *SECADM or *ALLOBJ special authority. If it does, clustering commands can appear to hang as a result.

It is also important to make sure that all of the cluster jobs can run successfully when needed. To make sure there are enough resources for cluster jobs to run:

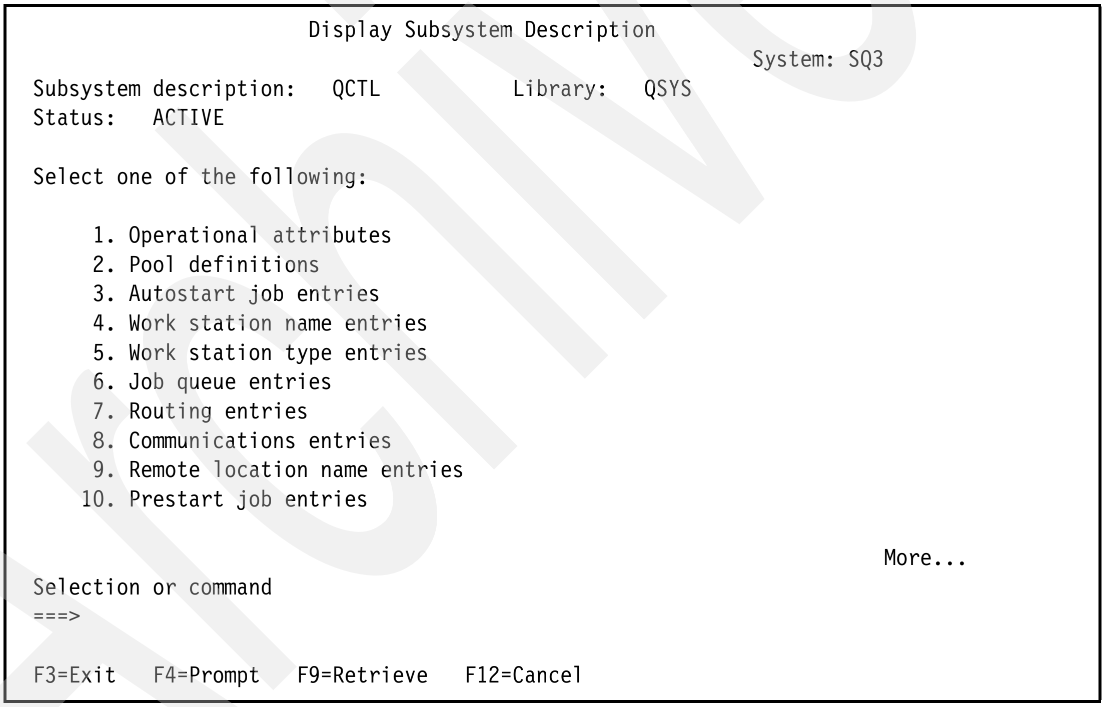
1. Run the DSPSYSVAL command for the QCTLSBSD system value.
2. In the Display System Value panel (Figure 10-20), run the DSPSBSD command for the subsystem that was found in step 1.



```
                              Display System Value

 System value . . . . . :   QCTLSBSD
 Description  . . . . . :   Controlling subsystem


 Controlling subsystem  . . . :   QCTL
   Library  . . . . . . . . . :     QSYS
```

Figure 10-20 Display System Value panel for QCTLSBSD

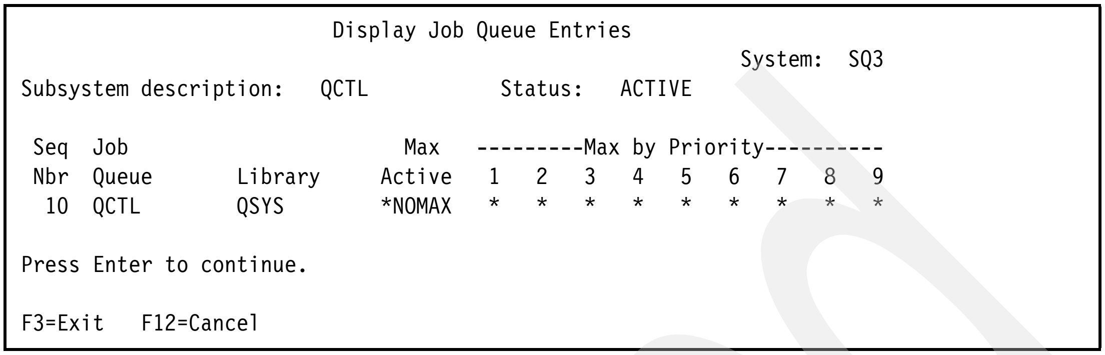
3. In the Display Subsystem Description panel (Figure 10-21), select option **6** (Job queue entries).



```
                        Display Subsystem Description
                                                             System: SQ3
 Subsystem description:   QCTL           Library:   QSYS
 Status:   ACTIVE

 Select one of the following:

      1. Operational attributes
      2. Pool definitions
      3. Autostart job entries
      4. Work station name entries
      5. Work station type entries
      6. Job queue entries
      7. Routing entries
      8. Communications entries
      9. Remote location name entries
     10. Prestart job entries

                                                                        More...
 Selection or command
 ===>

 F3=Exit   F4=Prompt   F9=Retrieve   F12=Cancel
```

Figure 10-21 Display Subsystem Description panel

In the Display Job Queue Entries panel (Figure 10-22), the value for Max Active (maximum active jobs) for the job queue listed should be `*NOMAX`. If this is limited, cluster jobs can appear to hang, because clustering requires that multiple jobs run at the same time.



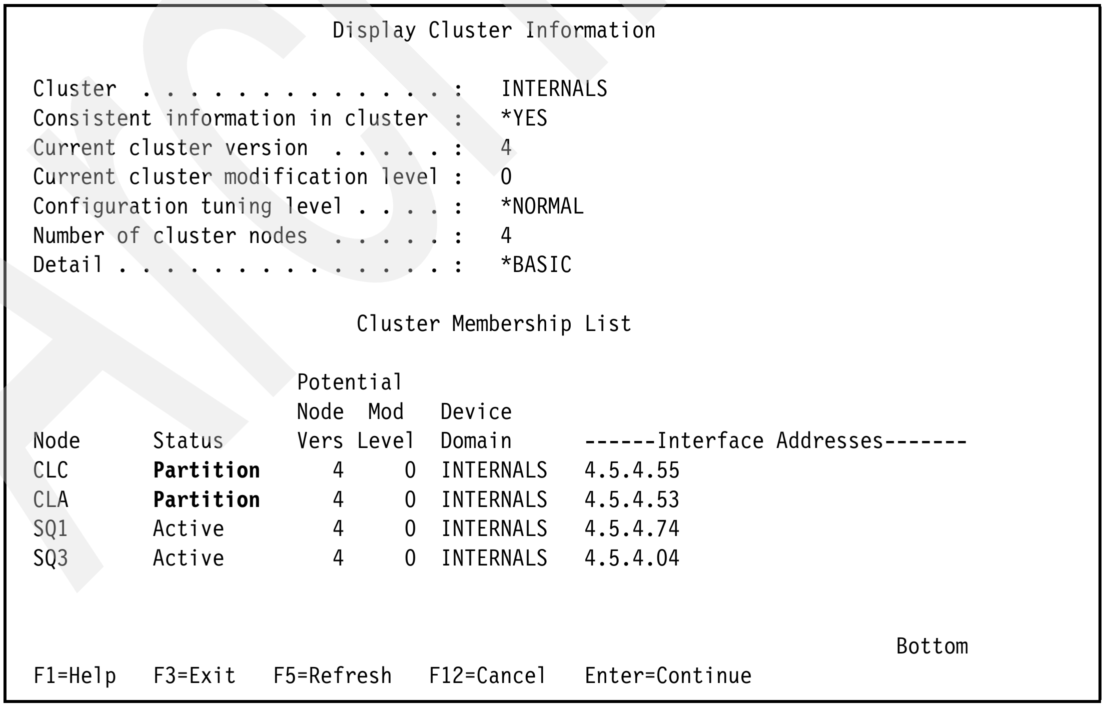
```
                          Display Job Queue Entries
                                                            System:   SQ3
Subsystem description:   QCTL          Status:   ACTIVE

 Seq  Job                     Max   ---------Max by Priority----------
 Nbr  Queue       Library     Active   1   2   3   4   5   6   7   8   9
  10  QCTL        QSYS        *NOMAX   *   *   *   *   *   *   *   *   *

Press Enter to continue.

F3=Exit   F12=Cancel
```

Figure 10-22   Display Job Queue Entries panel

Jobs are submitted by clustering to do the work and actions that are needed. We suggest that you consider using a unique job description and unique job queue to separate the cluster jobs from other jobs that you submit. When the subsystem or jobq limits the amount of jobs that can run at one time, performance can be affected, or the clustering can appear to hang since a job that needs to run is still in the job queue.

10.2.6 Cluster partitions

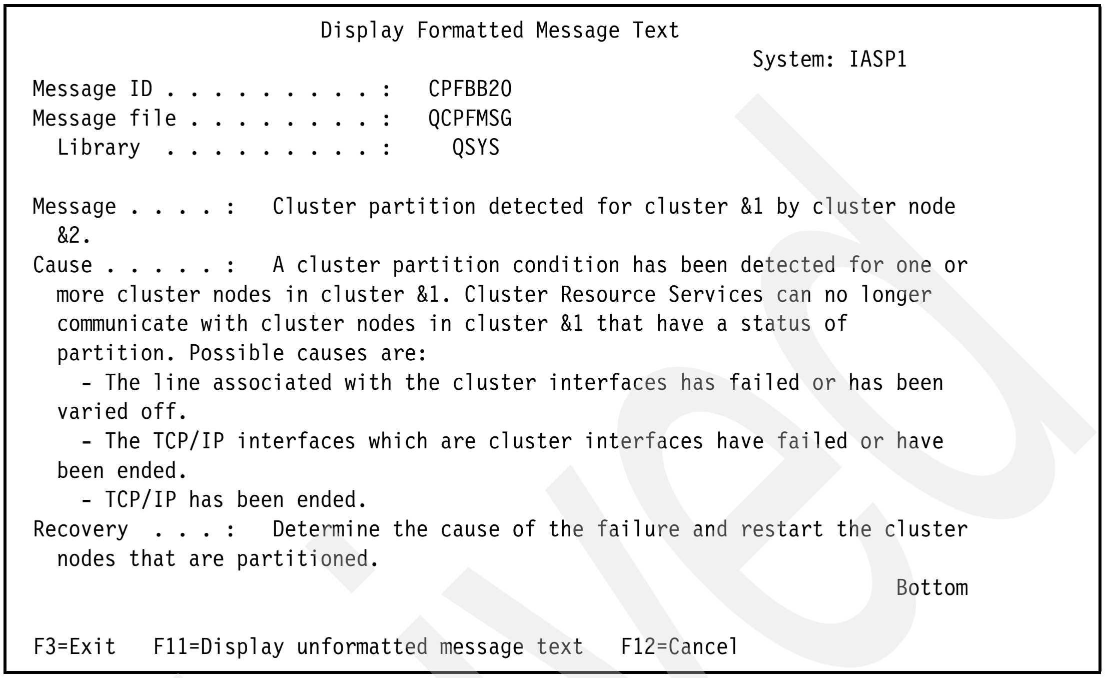
In the Display Cluster Information panel (Figure 10-23), it is possible that a node shows a status of `Partition`. Figure 10-23 shows both nodes CLC and CLA as having a Partition status. Nodes SQ1 and SQ3 do not know what happened to CLC and CLA.



```
                          Display Cluster Information

 Cluster  . . . . . . . . . . . . . :   INTERNALS
 Consistent information in cluster  :   *YES
 Current cluster version  . . . . . :   4
 Current cluster modification level :   0
 Configuration tuning level . . . . :   *NORMAL
 Number of cluster nodes  . . . . . :   4
 Detail . . . . . . . . . . . . . . :   *BASIC

                            Cluster Membership List

                      Potential
                      Node  Mod    Device
 Node      Status     Vers  Level  Domain      ------Interface Addresses-------
 CLC       Partition     4     0   INTERNALS   4.5.4.55
 CLA       Partition     4     0   INTERNALS   4.5.4.53
 SQ1       Active        4     0   INTERNALS   4.5.4.74
 SQ3       Active        4     0   INTERNALS   4.5.4.04



                                                                          Bottom
 F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 10-23   Display Cluster Information panel showing a status of Partition

In such a case where a node has a status of Partition, error message CPFBB20 is displayed (see Figure 10-24).



```
                          Display Formatted Message Text
                                                                   System:  IASP1
 Message ID . . . . . . . . . :   CPFBB20
 Message file . . . . . . . . :   QCPFMSG
   Library  . . . . . . . . . :     QSYS

 Message . . . . :   Cluster partition detected for cluster &1 by cluster node
   &2.
 Cause . . . . . :   A cluster partition condition has been detected for one or
   more cluster nodes in cluster &1. Cluster Resource Services can no longer
   communicate with cluster nodes in cluster &1 that have a status of
   partition. Possible causes are:
     - The line associated with the cluster interfaces has failed or has been
   varied off.
     - The TCP/IP interfaces which are cluster interfaces have failed or have
   been ended.
     - TCP/IP has been ended.
 Recovery  . . . :   Determine the cause of the failure and restart the cluster
   nodes that are partitioned.
                                                                         Bottom

 F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 10-24 Error message CPFBB20

Cluster partitions occur when the heartbeat function no longer can communicate with the remote system. One possibility for this inability to communicate is that the communication link between those nodes has become disconnected (see Figure 10-25).

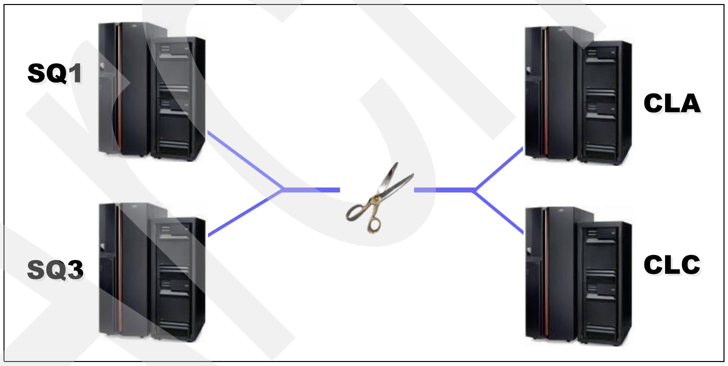
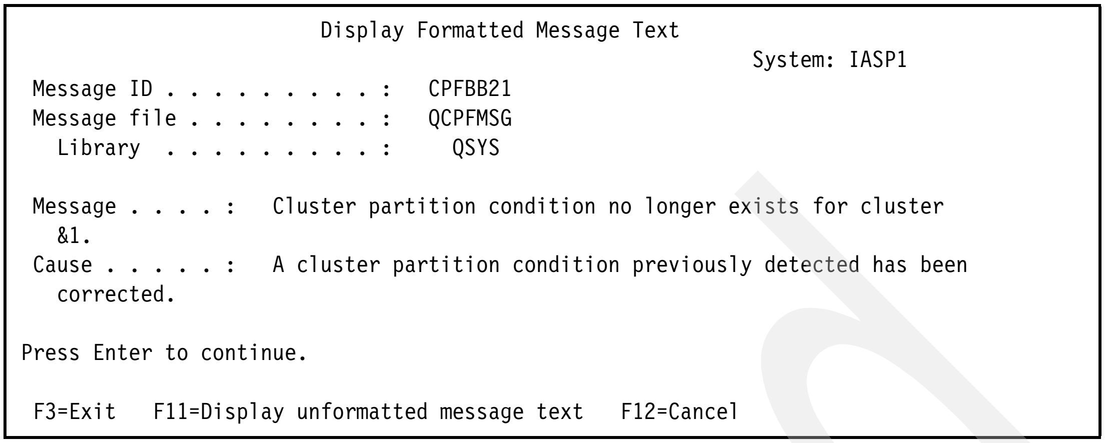


Figure 10-25 Cluster partition with a disconnected communication link

If the cause of the cluster partition is due to an external communication link issue, after the communication problem is fixed and the link is re-established, the cluster detects the re-established connection and issues message CPFBB21 (see Figure 10-26) in either the history log or the QCSTCTL job log. Within a few minutes, clustering resumes among all nodes.



```
                          Display Formatted Message Text
                                                                    System: IASP1
 Message ID . . . . . . . . . :   CPFBB21
 Message file . . . . . . . . :   QCPFMSG
   Library  . . . . . . . . . :     QSYS

 Message . . . . :   Cluster partition condition no longer exists for cluster
   &1.
 Cause . . . . . :   A cluster partition condition previously detected has been
   corrected.

Press Enter to continue.

 F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 10-26 Message for a resolved cluster partition issue

Another possibility for the cluster partition condition is that CLC and CLA are down, as shown in Figure 10-27. Losing multiple nodes at the same time can be a result of a hardware problem where both nodes are partitions in the same managed system. Even if you use the PWRDWNSYS OPTION(*IMMED) command to bring down the two nodes, those nodes still send a notification to the other nodes in the cluster that they are going down in an intended manner. A cluster partition condition indicates a much more abrupt and severe problem than a planned system outage.

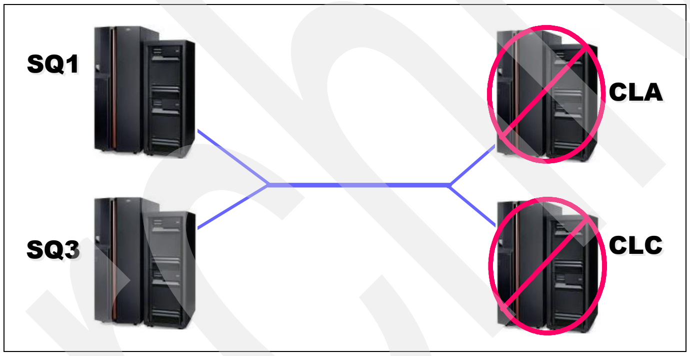
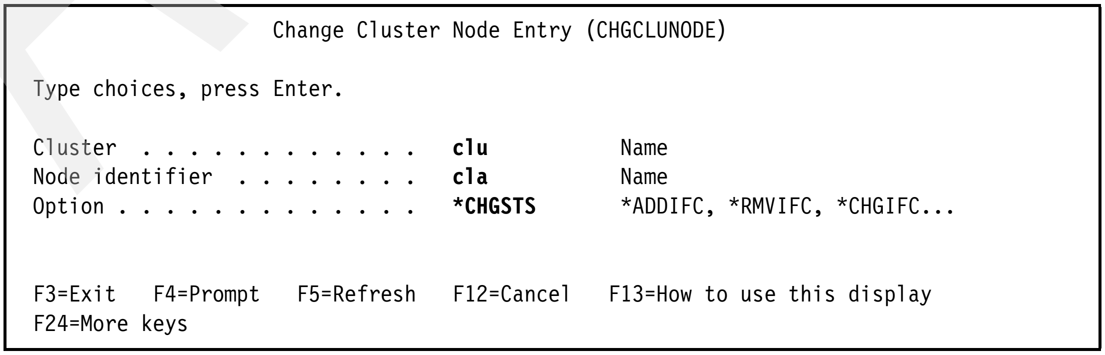


Figure 10-27 Cluster partition when systems are down

If you determine that the cluster partition condition is a result of the nodes going down unexpectedly, you can use the CHGCLUNODE OPTION(*CHGSTS) command to change the cluster node status from Partition to Failed (see Figure 10-28).



```
                      Change Cluster Node Entry (CHGCLUNODE)

 Type choices, press Enter.

 Cluster  . . . . . . . . . . . .   clu           Name
 Node identifier  . . . . . . . .   cla           Name
 Option . . . . . . . . . . . . .   *CHGSTS       *ADDIFC, *RMVIFC, *CHGIFC...


 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 10-28 Change Cluster Node Entry (CHGCLUNODE) panel

In either case, if XSM is used, and the node that owns either the production copy or the mirror copy of the IASP goes into Partition status, the two copies are no longer in synch if XSM uses the same communication link as clustering, and the problem is due to an external communication failure. XSM is suspended once the time-out threshold is reached, and a resume and full resynchronization are required when the problem is fixed.

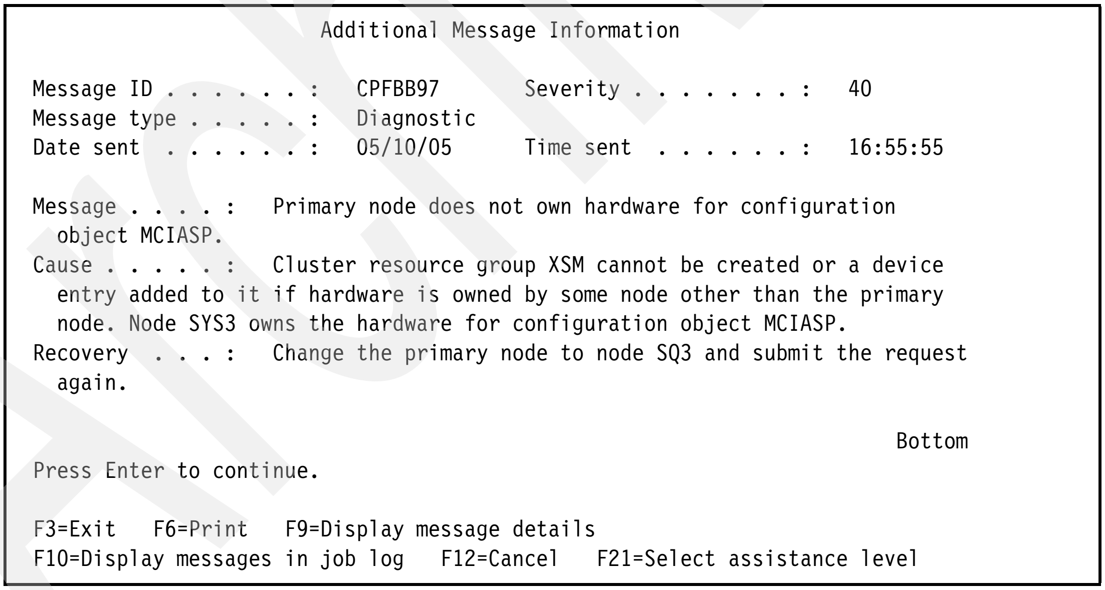
If the communications failure does not affect XSM, it continues even though the cluster is not active. A message of CPDB715 is issued in this case, as shown in Figure 10-45 on page 206.

If the node that owns the mirror copy of the IASP is affected, the CHGCLUNODE OPTION(*CHGSTS) command marks that node as *Failed*. After this command is issued, the recovery time-out threshold for XSM is no longer used if it has not been reached yet. The production copy of the IASP is still usable and can resynch with the mirror copy after the issue is resolved.

If the node that owns the production copy of the IASP is affected, the CHGCLUNODE OPTION(*CHGSTS) command marks that node as *Failed*. The mirror copy is promoted to the production copy as a result. After the problem with the affected node is resolved, and after XSM resumes, that node owns the mirror copy of the IASP and begins to resynch with the production copy, that is, the former mirror copy.

10.2.7 Hardware issues with a device CRG

When a cluster is restarted from a node that was not the last active node in the cluster, it can send older information to the rest of the cluster nodes. In that case, error message CPFBB97 (Figure 10-29) can occur with the STRCRG command.



```
                          Additional Message Information

 Message ID . . . . . . :   CPFBB97       Severity . . . . . . . :   40
 Message type . . . . . :   Diagnostic
 Date sent  . . . . . . :   05/10/05      Time sent  . . . . . . :   16:55:55

 Message . . . . :   Primary node does not own hardware for configuration
   object MCIASP.
 Cause . . . . . :   Cluster resource group XSM cannot be created or a device
   entry added to it if hardware is owned by some node other than the primary
   node. Node SYS3 owns the hardware for configuration object MCIASP.
 Recovery  . . . :   Change the primary node to node SQ3 and submit the request
   again.

                                                                         Bottom
 Press Enter to continue.

 F3=Exit   F6=Print   F9=Display message details
 F10=Display messages in job log   F12=Cancel   F21=Select assistance level
```

Figure 10-29 CPFBB97 error message regarding hardware issues

In an environment with a switchable IASP and no XSM involved, see which node on which the IASP is reporting in, and enter the CHGCRGPRI command to correct the problem.

For an environment that uses XSM, this problem can happen only with three or four node configurations. In that case, you need to find the node that owns the production copy of the IASP and the node that owns the mirror copy of the IASP.

Figure 10-30 shows a configuration for a four-node geographic mirroring environment. Here, we see that SYS4 owns the production copy of the IASP, and SYS1 owns the mirror copy of the IASP. Using iSeries Navigator, you can determine which copy is which, as shown in Figure 10-35 on page 198.

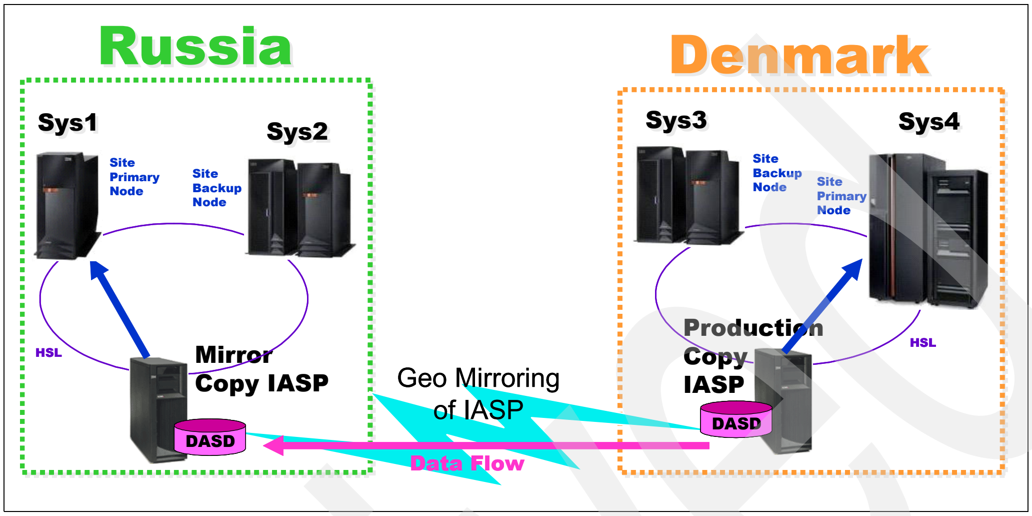


Figure 10-30 Current hardware example

If error message CPFBB97 is issued on the STRCRG command, and SYS4 is the node that owns the production copy of the IASP, you can use the following command to let the cluster know which node is the primary and which nodes are backups. This command enables the STRCRG command to function successfully.

```
CHGCRG CLUSTER(cluster_name) CRG(CRG_name) CRGTYPE(*DEV) RCYDMNACN(*CHGCUR) RCYDMN((SYS4
*PRIMARY) (SYS3 *BACKUP) (SYS1 *BACKUP) (SYS2 *BACKUP))
```

> **Note:** The node that owns the mirror copy of the IASP does not have to be listed first in sequence in the command, as long as it is listed as a backup.

10.3 Cross-site mirroring basics

XSM relies on clustering to be active. Therefore, it is important to always check the status of the cluster and CRG prior to troubleshooting errors that are specifically related to XSM. In the following sections, we focus on the primary issues to check for when troubleshooting XSM, although clustering functionality is used quite often.

10.3.1 Cross-site mirroring status

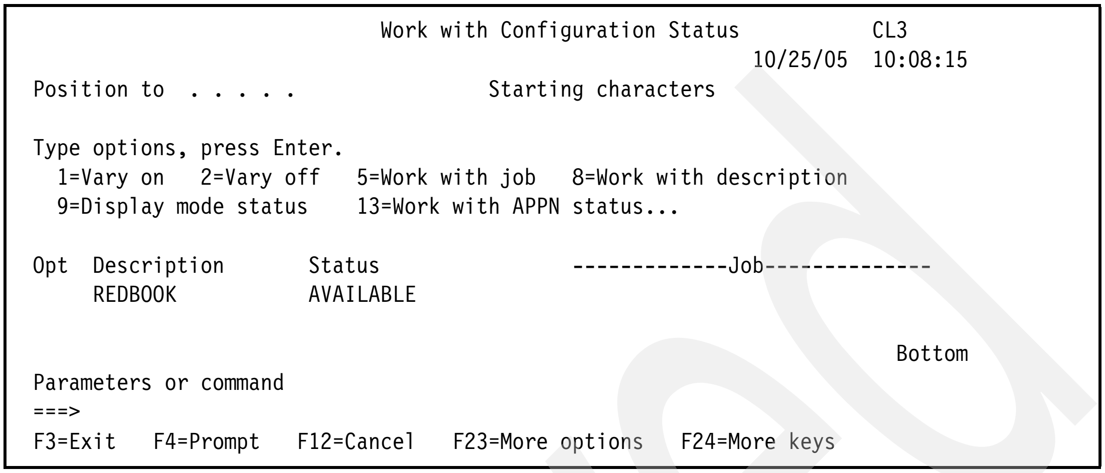
After you check the cluster and CRG status, you need to check the XSM status. You do part of this check by looking at the cluster CRG. You can use iSeries Navigator to give an additional perspective on the XSM status.

Checking the status of XSM by using a command line

Earlier in this chapter, we used the DSPCRGINF command to review the status of a device CRG used for XSM (see Figure 10-5 on page 180). Both of these node are in Active status, meaning that XSM is active between both nodes in the cluster. If the mirror copy is not in synch with the production copy, the status is Ineligible. In this case, XSM cannot switch over or fail over to the Ineligible node until that node is no longer in the Ineligible status. This is the best way to see if the IASP copies are in synch.

When the IASP is ready to be used, use the following command to view the production copy in Available status (see Figure 10-31):

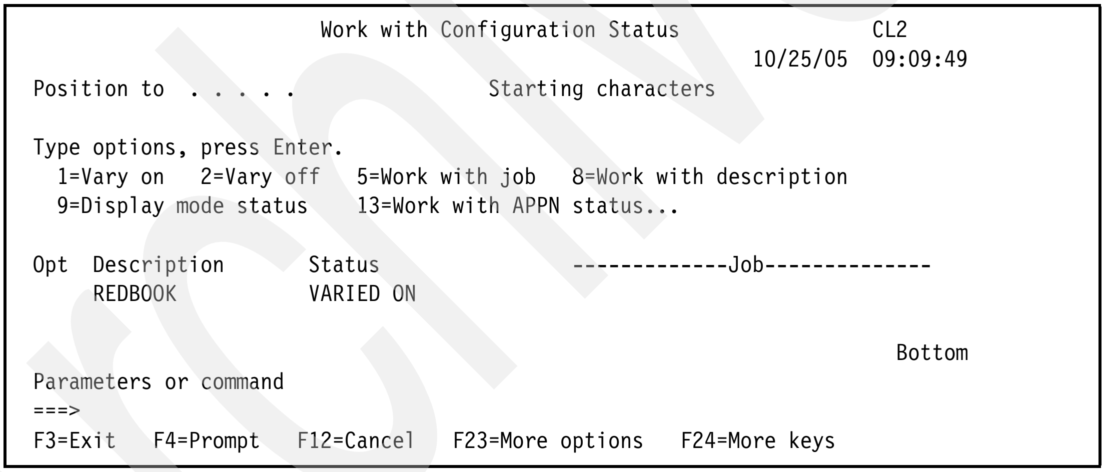
```
WRKCFGSTS CFGTYPE(*DEV) CFGD(*ASP)
```



```
                         Work with Configuration Status           CL3
                                                         10/25/05  10:08:15
Position to  . . . . .                Starting characters

Type options, press Enter.
  1=Vary on   2=Vary off   5=Work with job   8=Work with description
  9=Display mode status    13=Work with APPN status...

Opt  Description      Status               -------------Job--------------
     REDBOOK          AVAILABLE

                                                                    Bottom
Parameters or command
===>
F3=Exit   F4=Prompt   F12=Cancel   F23=More options   F24=More keys
```

*Figure 10-31 WRKCFGSTS CFGTYPE(*DEV) CFGD(*ASP) - production copy*

The same command also shows the mirror copy in Varied On status (see Figure 10-32).



```
                    Work with Configuration Status               CL2
                                                         10/25/05  09:09:49
Position to  . . . . .                Starting characters

Type options, press Enter.
  1=Vary on   2=Vary off   5=Work with job   8=Work with description
  9=Display mode status    13=Work with APPN status...

Opt  Description      Status               -------------Job--------------
     REDBOOK          VARIED ON

                                                                    Bottom
Parameters or command
===>
F3=Exit   F4=Prompt   F12=Cancel   F23=More options   F24=More keys
```

*Figure 10-32 WRKCFGSTS CFGTYPE(*DEV) CFGD(*ASP) - mirror copy*

Checking the status of XSM by using iSeries Navigator

Earlier in our discussion about clustering, we showed a device CRG used for XSM, where both nodes are in the Active status (see Figure 10-12 on page 184). This status means that XSM is active between the two nodes in the cluster.

By default, iSeries Navigator does not necessarily show the XSM status, although you can configure it to show the status:

1. Start an iSeries Navigator session.
2. In the left navigation pane, expand **My Connections** → ***node system***. Sign into the system with your user profile and password.

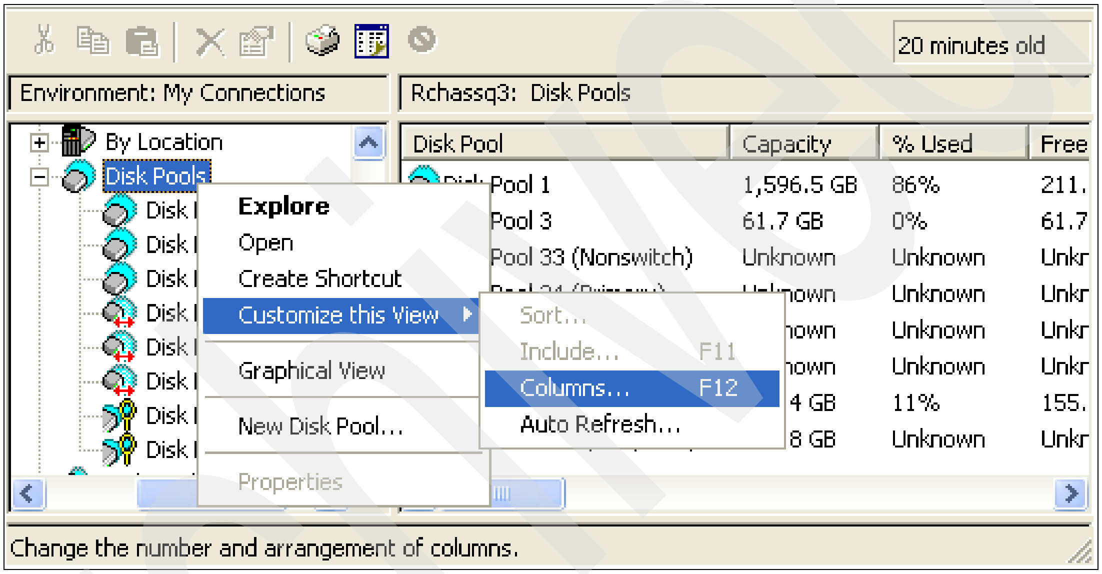
3. Expand **Hardware** → **Disk Units**.

> **Note:** To access this view, you need a valid System Service Tools (SST) user profile and password.

4. Click **Disk Pools**.

   The names of the IASPs are listed in the right pane. In iSeries Navigator, these IASPs are called *disk pools*. There is a list of a variety of columns for each IASP, such as capacity, threshold, status, type, and so on.

   Right-click **Disk Pools** (left navigation pane) and choose **Customize this View** → **Columns** as shown in Figure 10-33. You can add the XSM information as additional columns in this view.



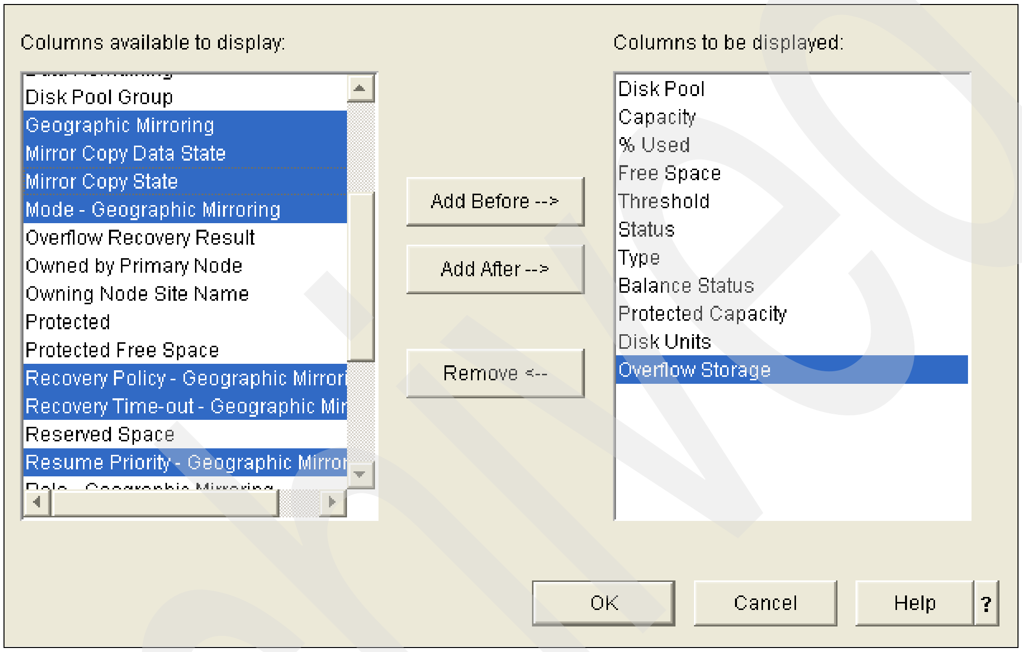
Disk Pool	Capacity	% Used	Free
Pool 1	1,596.5 GB	86%	211.
Pool 3	61.7 GB	0%	61.7
Pool 33 (Nonswitch)	Unknown	Unknown	Unkr
	nown	Unknown	Unkr
	nown	Unknown	Unkr
	nown	Unknown	Unkr
	4 GB	11%	155.
	8 GB	Unknown	Unkr

Change the number and arrangement of columns.

Figure 10-33 Adding columns

In the next window (Figure 10-34), you can choose additional columns and place them in your preferred order. The following available columns are related to geographic mirroring:

- Geographic Mirroring
- Mode - Geographic Mirroring
- Mirror Copy State
- Mirror Copy Data State
- Auto resume Geographic Mirroring
- Recovery Policy - Geographic Mirroring
- Recovery Time-out Geographic Mirroring
- Resume Priority - Geographic Mirroring



Columns available to display:

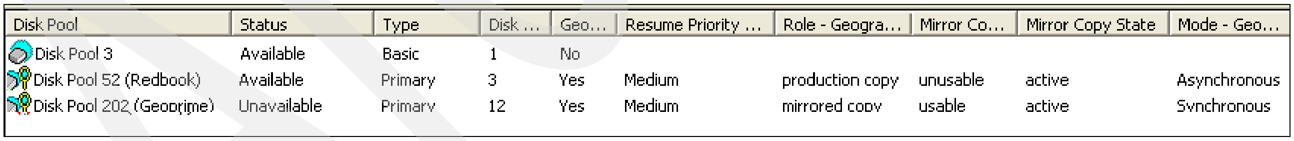
Disk Pool Group
Geographic Mirroring
Mirror Copy Data State
Mirror Copy State
Mode - Geographic Mirroring
Overflow Recovery Result
Owned by Primary Node
Owning Node Site Name
Protected
Protected Free Space
Recovery Policy - Geographic Mirrori
Recovery Time-out - Geographic Mir
Reserved Space
Resume Priority - Geographic Mirror
Role - Geographic Mirroring

Columns to be displayed:

Disk Pool
Capacity
% Used
Free Space
Threshold
Status
Type
Balance Status
Protected Capacity
Disk Units
Overflow Storage

Figure 10-34 Adding columns

After you add the columns, you see the additional columns that are associated with each IASP, as shown in Figure 10-35.



Disk Pool	Status	Type	Disk ...	Geo...	Resume Priority ...	Role - Geogra...	Mirror Co...	Mirror Copy State	Mode - Geo...
Disk Pool 3	Available	Basic	1	No					
Disk Pool 52 (Redbook)	Available	Primary	3	Yes	Medium	production copy	unusable	active	Asynchronous
Disk Pool 202 (Geoprime)	Unavailable	Primary	12	Yes	Medium	mirrored copy	usable	active	Synchronous

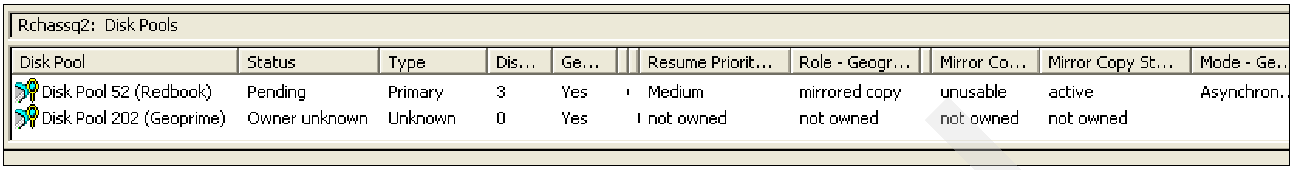
Figure 10-35 Example of GUI showing geographic mirroring status - production copy

In the view shown in Figure 10-35, the IASP (disk pool) is shown, along with its status, type, number of disks, whether it is geographically mirrored, the resume priority, the role of the IASP copy, the mode of geographic mirroring used, and whether the mirror copy can be used.

In asynchronous mode, the data is not forced to be written to DASD prior to receiving an acknowledgement. Therefore, the mirror copy always is displayed as unusable at any particular time when running in asynchronous mode, because it is not guaranteed that the data on the mirror copy is usable at any given time.

An IASP with a status of Available, as for the IASP called *Redbook* in Figure 10-35, has the same status as though the IASP device shows Available from the command line interface

(CLI). An IASP with a status of unavailable, or Owner unknown as for the IASP called Geoprime in Figure 10-36, corresponds to a status of Varied Off from the CLI.



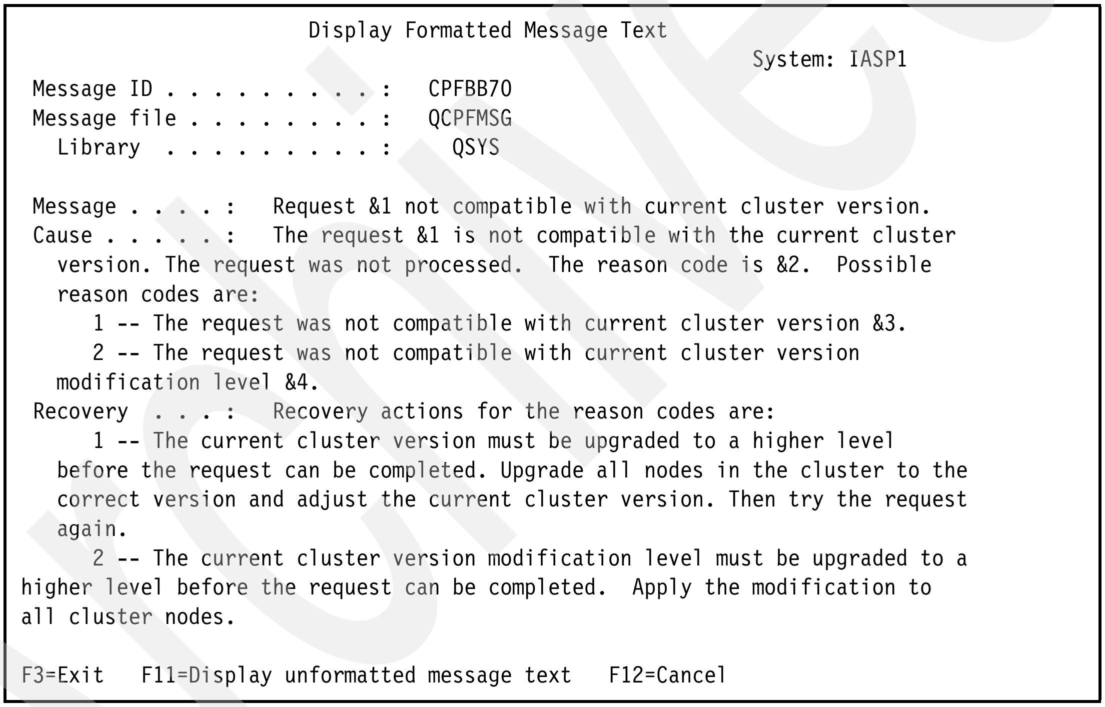
Rchassq2: Disk Pools

Disk Pool	Status	Type	Dis...	Ge...	Resume Priorit...	Role - Geogr...	Mirror Co...	Mirror Copy St...	Mode - Ge...
Disk Pool 52 (Redbook)	Pending	Primary	3	Yes	Medium	mirrored copy	unusable	active	Asynchron...
Disk Pool 202 (Geoprime)	Owner unknown	Unknown	0	Yes	not owned	not owned	not owned	not owned	

Figure 10-36 iSeries Navigator showing geographic mirroring status - mirror copy

10.3.2 Geographic mirroring release considerations

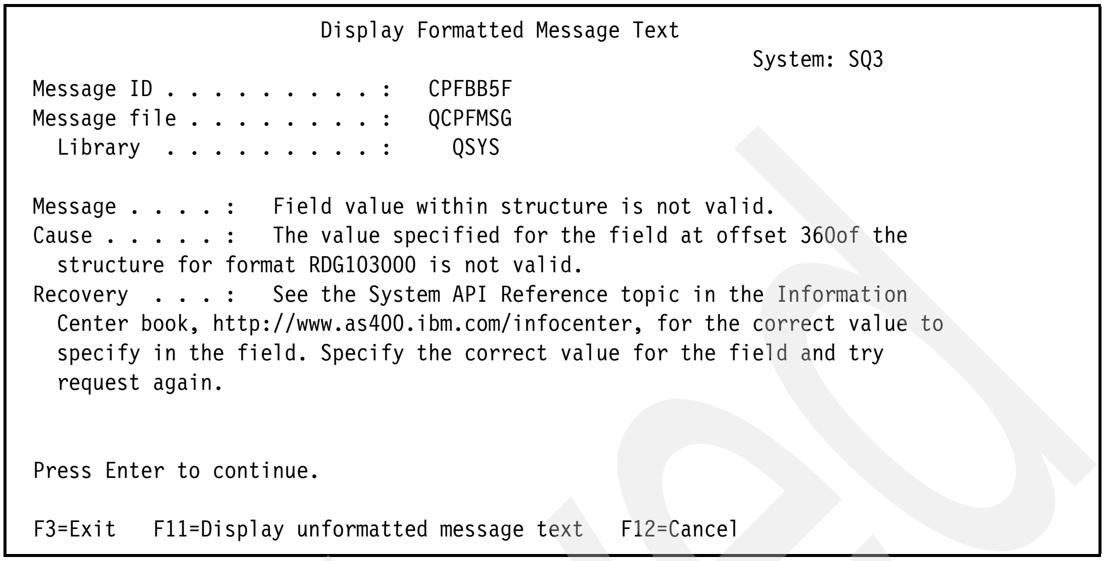
Geographic mirroring can only run in a cluster that is set to cluster version 4, which requires that all participating nodes are at V5R3M0 or later of OS/400 or i5/OS. If the cluster is not at cluster version 4, and you attempt to configure geographic mirroring, error message CPFBB70 is displayed, as shown in Figure 10-37.



```
                         Display Formatted Message Text
                                                                     System: IASP1
 Message ID . . . . . . . . . :   CPFBB70
 Message file . . . . . . . . :   QCPFMSG
   Library  . . . . . . . . . :     QSYS

 Message . . . . :   Request &1 not compatible with current cluster version.
 Cause . . . . . :   The request &1 is not compatible with the current cluster
   version. The request was not processed.  The reason code is &2.  Possible
   reason codes are:
      1 -- The request was not compatible with current cluster version &3.
      2 -- The request was not compatible with current cluster version
   modification level &4.
 Recovery  . . . :   Recovery actions for the reason codes are:
      1 -- The current cluster version must be upgraded to a higher level
   before the request can be completed. Upgrade all nodes in the cluster to the
   correct version and adjust the current cluster version. Then try the request
   again.
      2 -- The current cluster version modification level must be upgraded to a
higher level before the request can be completed.  Apply the modification to
all cluster nodes.

F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 10-37 Error message CPFBB70

Alternatively, error message CPFBB5F is displayed, as shown in Figure 10-38.



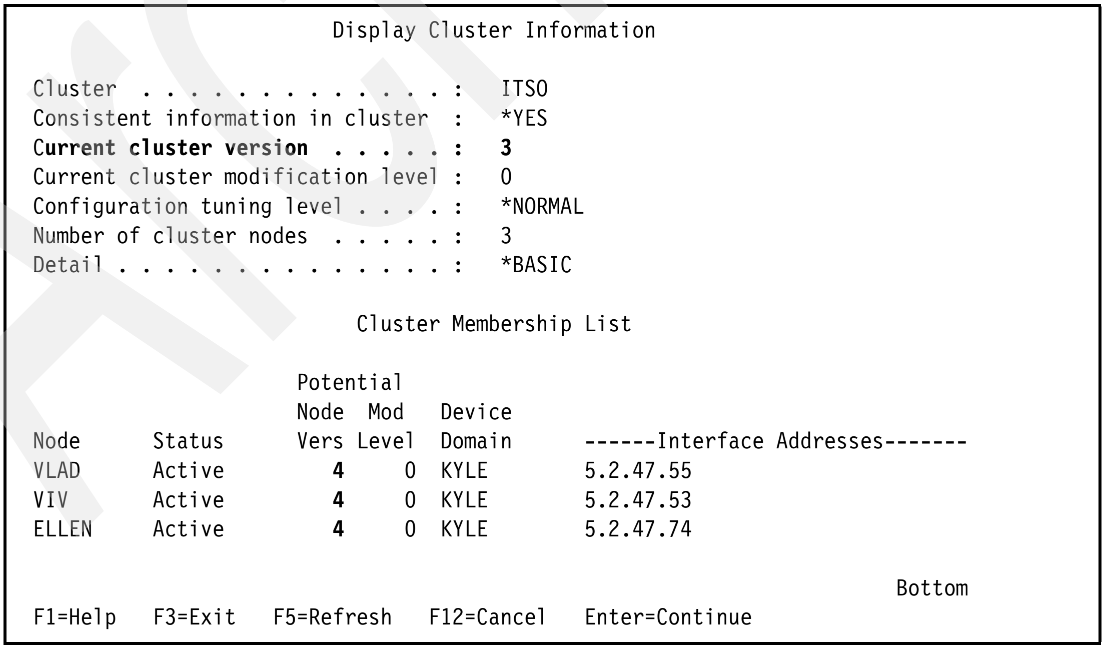
```
                        Display Formatted Message Text
                                                            System: SQ3
Message ID . . . . . . . . . :   CPFBB5F
Message file . . . . . . . . :   QCPFMSG
  Library  . . . . . . . . . :     QSYS

Message . . . . :   Field value within structure is not valid.
Cause . . . . . :   The value specified for the field at offset 360of the
  structure for format RDG103000 is not valid.
Recovery  . . . :   See the System API Reference topic in the Information
  Center book, http://www.as400.ibm.com/infocenter, for the correct value to
  specify in the field. Specify the correct value for the field and try
  request again.


Press Enter to continue.

F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 10-38 Error message CPFBB5F

Important: For more information about cluster releases, refer to the “Cluster version” section in the iSeries Information Center on the Web at the following address:

http://publib.boulder.ibm.com/infocenter/iseries/v5r3/index.jsp?topic=/rzaig/rzaigplanclusterversions.htm

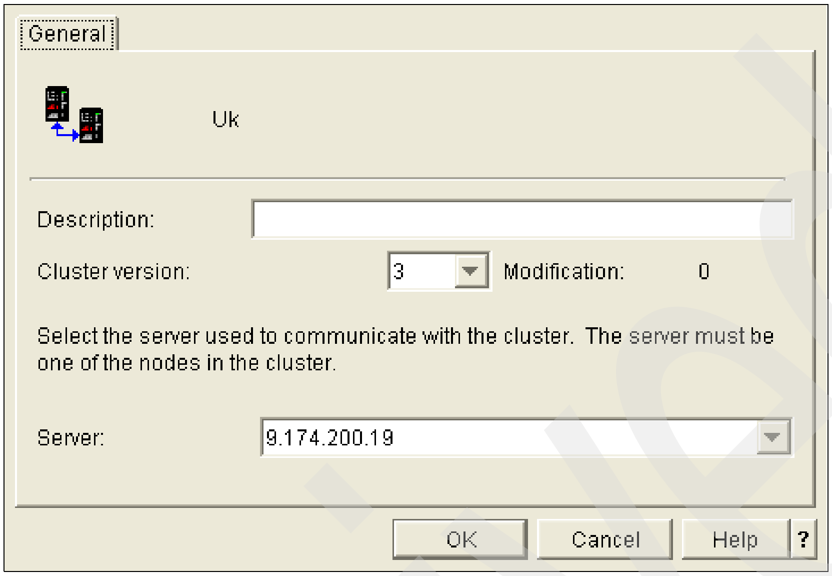
Remember that you can use the DSPCLUINF command to check the cluster version. See Figure 10-39.



```
                         Display Cluster Information

Cluster  . . . . . . . . . . . . . :   ITSO
Consistent information in cluster  :   *YES
Current cluster version  . . . . . :   3
Current cluster modification level :   0
Configuration tuning level . . . . :   *NORMAL
Number of cluster nodes  . . . . . :   3
Detail . . . . . . . . . . . . . . :   *BASIC

                           Cluster Membership List

                      Potential
                      Node  Mod   Device
Node      Status      Vers Level  Domain      ------Interface Addresses-------
VLAD      Active         4     0  KYLE        5.2.47.55
VIV       Active         4     0  KYLE        5.2.47.53
ELLEN     Active         4     0  KYLE        5.2.47.74

                                                                        Bottom
F1=Help   F3=Exit   F5=Refresh   F12=Cancel   Enter=Continue
```

Figure 10-39 Display Cluster Information panel

In iSeries Navigator, right-click the cluster and choose **Properties** to see the same information. Figure 10-40 shows the Properties window. In this example, the cluster version is 3, but each node in the cluster has a potential version of 4. This means that each of the nodes must be at V5R3M0 or later of OS/400 or i5/OS.



Uk

Description:

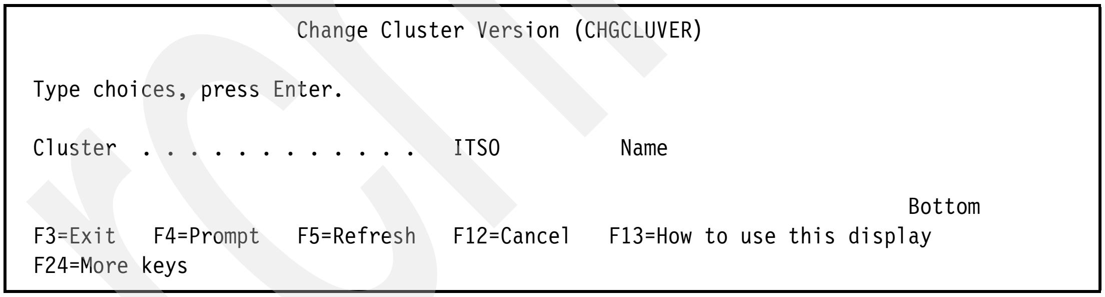
Cluster version: 3 Modification: 0

Select the server used to communicate with the cluster. The server must be one of the nodes in the cluster.

Server: 9.174.200.19

Figure 10-40 Cluster version Properties window in iSeries Navigator

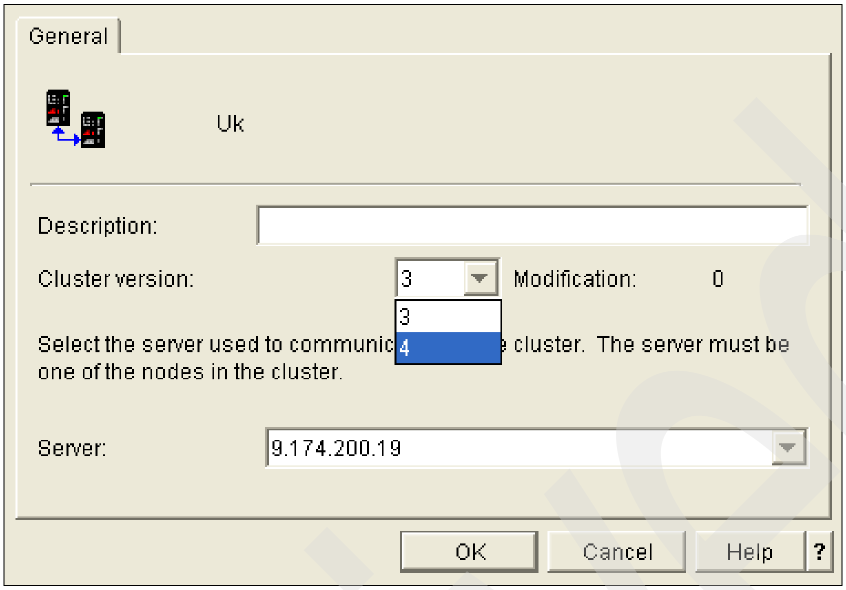
To change the cluster to its potential version, use the CHGCLUVER command. See Figure 10-41.



```
                      Change Cluster Version (CHGCLUVER)

 Type choices, press Enter.

 Cluster  . . . . . . . . . . . .   ITSO          Name

                                                                    Bottom
 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 10-41 Chance Cluster Version (CHGCLUVER) panel

Alternately, you can change the cluster version by using iSeries Navigator as shown in Figure 10-42.



Uk

Description:

Cluster version: 3

3
4

Modification: 0

Select the server used to communic… cluster. The server must be one of the nodes in the cluster.

Server: 9.174.200.19

Figure 10-42 Changing cluster version using iSeries Navigator

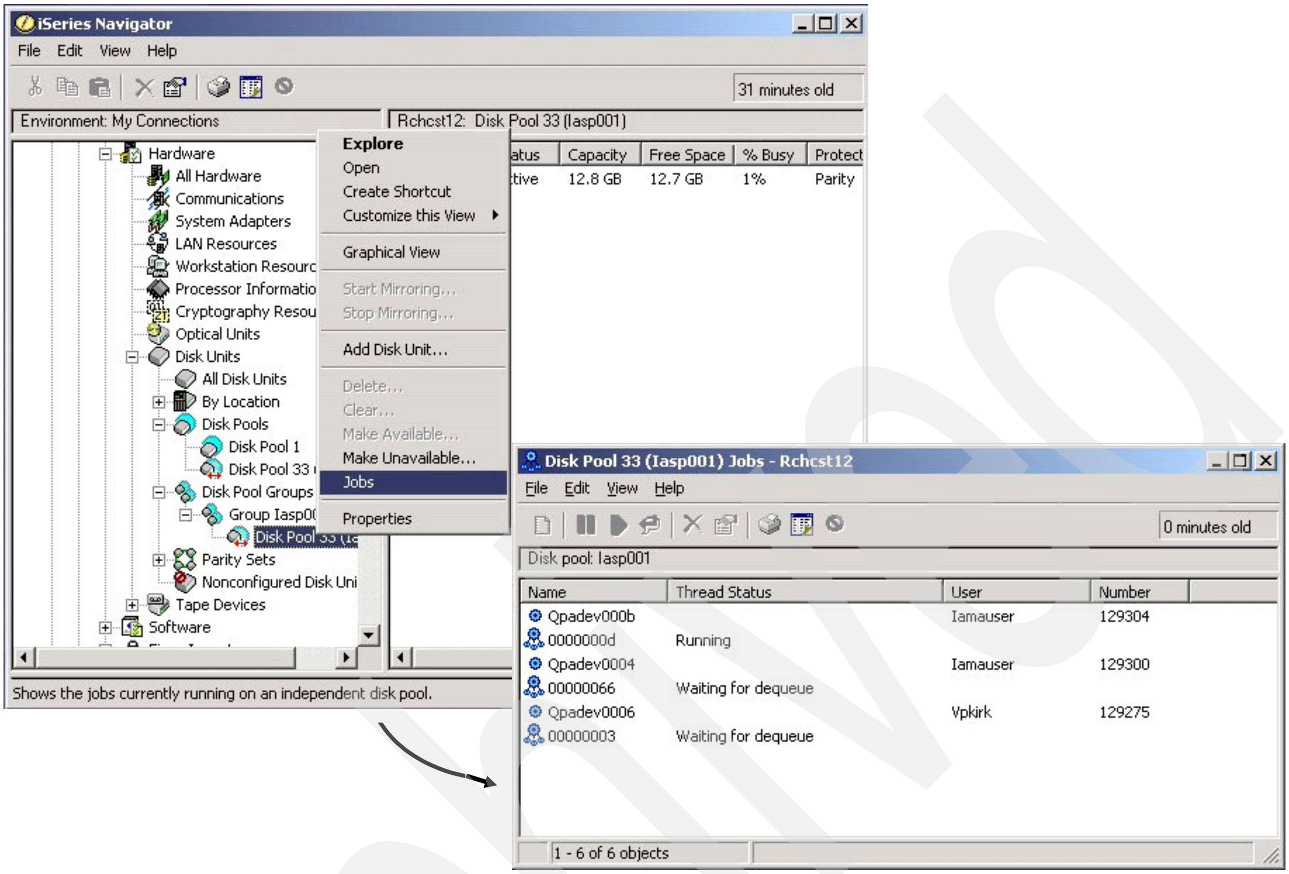
10.3.3 IASP jobs

We recommend that you configure separate main storage pools for user jobs that access IASPs in order to prevent those jobs from contending with other jobs on the system and using more main storage than desired. More specifically, IASP jobs should not use the machine pool or base pool. If IASP jobs use the same memory as jobs that are not accessing the IASPs, IASP jobs can monopolize the memory pool, lock out other jobs, and in extreme situations deadlock the system. Exposure for this situation is greater when using geographic mirroring.

To check which jobs are active for an IASP, you can use iSeries Navigator:

1. Expand **My Connections** → ***your system (node) name*** → **Configuration and Service** → **Hardware** → **Disk Units** → **Disk Pool Groups**.
2. Select the group with your IASP name.

3. Right-click the name of the IASP (disk pool) and choose **Jobs**. A new window (Figure 10-43) opens that shows all of the jobs that are using the IASP.

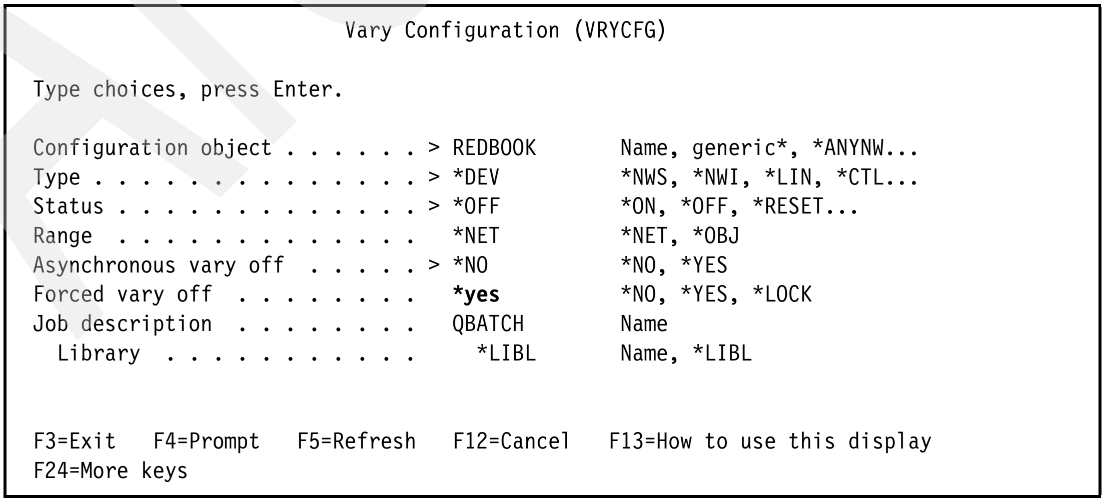


atus	Capacity	Free Space	% Busy	Protect
tive	12.8 GB	12.7 GB	1%	Parity

Name	Thread Status	User	Number
Qpadev000b		Iamauser	129304
0000000d	Running		
Qpadev0004		Iamauser	129300
00000066	Waiting for dequeue		
Qpadev0006		Vpkirk	129275
00000003	Waiting for dequeue		

Figure 10-43 Checking for jobs using the IASP via iSeries Navigator

If you need to end a job that is using the IASP, you can use the VRYCFG FRCVRYOFF(*YES) command against the device description that corresponds to the IASP. See Figure 10-44. This command ends all jobs that are using the IASP. Keep in mind that the application might end abnormally as a result and might not be designed to handle the abnormal end.



```
                          Vary Configuration (VRYCFG)

 Type choices, press Enter.

 Configuration object . . . . . . > REDBOOK       Name, generic*, *ANYNW...
 Type . . . . . . . . . . . . . . > *DEV          *NWS, *NWI, *LIN, *CTL...
 Status . . . . . . . . . . . . . > *OFF          *ON, *OFF, *RESET...
 Range  . . . . . . . . . . . . .   *NET          *NET, *OBJ
 Asynchronous vary off  . . . . . > *NO           *NO, *YES
 Forced vary off  . . . . . . . .   *yes          *NO, *YES, *LOCK
 Job description  . . . . . . . .   QBATCH        Name
   Library  . . . . . . . . . . .     *LIBL       Name, *LIBL


 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 10-44 Forcing an IASP device to vary off

In addition, depending on how the job handles the abnormal end, when the IASP device is varied on again, the vary on process can follow a different code path. This is similar to doing an IPL of a system while jobs are active, rather than bringing the system down to a restricted state first. The VRYCFG FRCVRYOFF(*YES) command affects jobs in a similar way that the PWRDWNSYS OPTION(*IMMED) command does, although VRYCFG FRCVRYOFF(*NO) does not vary off the IASP if jobs are active and using it.

Use iSeries Navigator to find the jobs that are using the IASP and end those jobs individually before you vary off the IASP device. This is the best and recommended method to avoid any further problems at the job or application level.

10.3.4 Geographic mirroring tasks

Geographic mirroring does not use jobs to keep the IASPs in synch. Instead, a variety of tasks are used to process geographic mirroring. These tasks all run in the machine pool. It is always a good idea when troubleshooting to make sure these tasks are running successfully.

On the node that owns the *production copy* of the IASP, the following tasks are used:

- CSTCDATAPORT0000

  This task runs on both the node that owns the production copy of the IASP and the node that owns the mirror copy of the IASP. It is the data port task that allows the two nodes to communicate with each other.

- SMRMSYNCFDIR_*nnn*

  One of these tasks runs for each IASP that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number. This task is used to synchronize directory information for the IASP.

- SMRMSYNCFDATA*nnn*

  One of these tasks runs for each IASP that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number. This task is used to synchronize the data for the IASP.

- SMRMSYNC_*nnn_xxx*

  There are 16, 24, or 32 of these tasks for each IASP that is being synchronized through geographic mirroring. The number depends on the sync priority that is being used. The first three-digit number *(nnn)* at the end of the task name corresponds to the IASP number. The second three-digit number *(xxx)* at the end of the task name corresponds to a task number for the IASP, from 000 to 031. These tasks are used to synchronize the data for the IASP.

- SMRMSYNCSTS_*nnn*

  One of these tasks runs for each IASP that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number. This task is used to monitor and report the status and progress of the synchronization for the IASP.

On the node that owns the *mirror copy* of the IASP, the following tasks are used:

- CSTCDATAPORT0000

  This task runs on both the node that owns the production copy of the IASP and the node that owns the mirror copy of the IASP. It is the data port task that allows the two nodes to communicate with each other.

- SMRMIRMAIN_*nnn*

  One of these tasks runs for each IASP group that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number of the primary IASP in the IASP group.

- SMRMIRGEN_*nnn*

  One of these tasks runs for each IASP group that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number of the primary IASP in the IASP group.

- SMRMIRIASP_*nnn*

  One of these tasks runs for each IASP that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number.

- SMRMIR_*nnn_xx_yy*

  Several of these tasks run for each IASP group that is being synchronized through geographic mirroring. The first three-digit number (*nnn*) at the end of the task name corresponds to the IASP number of the primary IASP in the IASP group. The other two numbers at the end of the task name (*xx* and *yy*) are variables to differentiate between the various tasks.

- SMRMIRPOOL_*nnn*

  Several of these tasks run for each IASP group that is being synchronized through geographic mirroring. The three-digit number (*nnn*) at the end of the task name corresponds to the IASP number of the primary IASP in the IASP group.

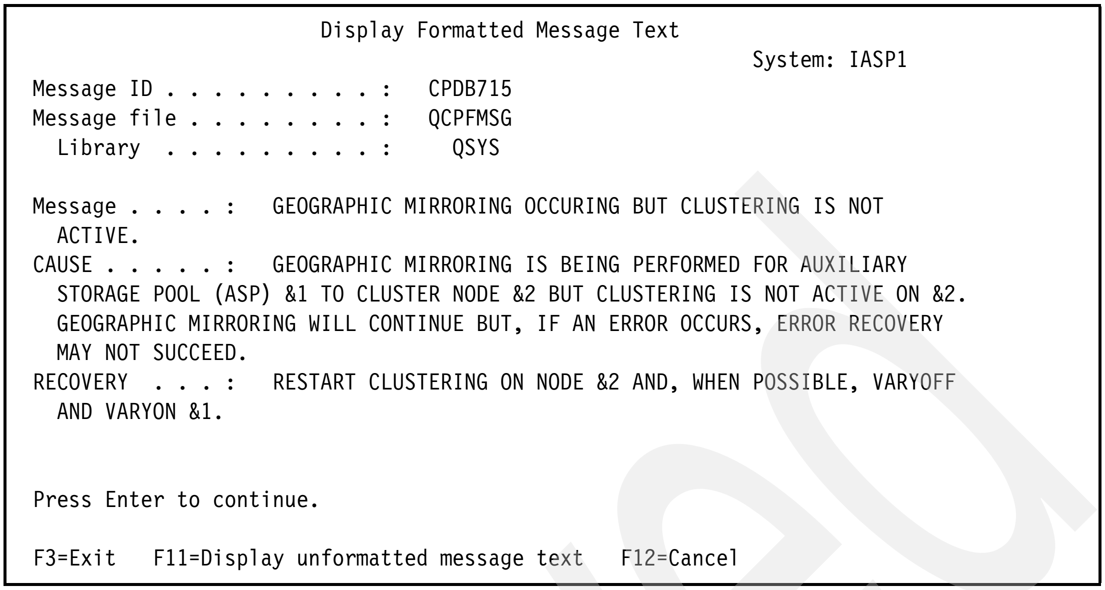
- SMASMDRTASKMGR

  A variable number of these tasks can be running on the system. These tasks are used during the preparation phase of synchronization.

10.3.5 Communication problems

Most errors with communications require network analysis. If the communication problem is caused by a user on the system that is ending the objects, jobs, or subsystems that are required for communications, then a variety of problems can occur.

If clustering communications uses a different communications link than geographic mirroring, it is possible to end clustering, while geographic mirroring remains active. The problem with this scenario is that error detection and recovery are not available on the geographically mirrored IASP copies. If this occurs, error message CPDB715 is logged as a warning (see Figure 10-45 on page 206).



```
                          Display Formatted Message Text
                                                            System:   IASP1
Message ID . . . . . . . . . :   CPDB715
Message file . . . . . . . . :   QCPFMSG
  Library  . . . . . . . . . :     QSYS

Message . . . . :   GEOGRAPHIC MIRRORING OCCURING BUT CLUSTERING IS NOT
  ACTIVE.
CAUSE . . . . . :   GEOGRAPHIC MIRRORING IS BEING PERFORMED FOR AUXILIARY
  STORAGE POOL (ASP) &1 TO CLUSTER NODE &2 BUT CLUSTERING IS NOT ACTIVE ON &2.
  GEOGRAPHIC MIRRORING WILL CONTINUE BUT, IF AN ERROR OCCURS, ERROR RECOVERY
  MAY NOT SUCCEED.
RECOVERY  . . . :   RESTART CLUSTERING ON NODE &2 AND, WHEN POSSIBLE, VARYOFF
  AND VARYON &1.


Press Enter to continue.

F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 10-45 Example of CPDB715

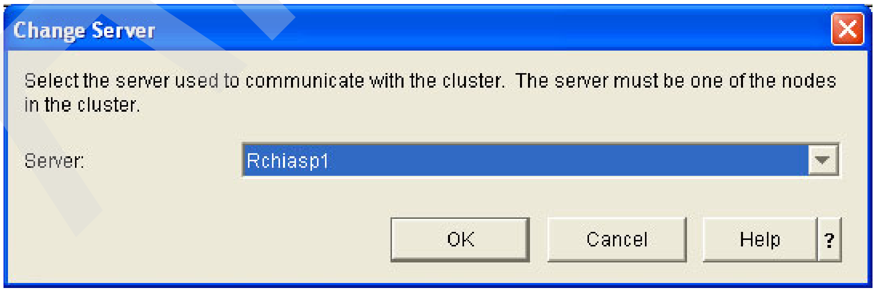
If a particular line or interface needs to be inoperable for an extended period of time, remove the IP address from the CRG to eliminate any issues when trying to establish connections through it.

10.4 Geographic and iSeries Navigator

Some errors can be seen only by using iSeries Navigator. In this section, we list some of these errors and explain how to correct them.

10.4.1 GUI preferred server

A *preferred server* is the term that is used in the GUI to refer to the server that is used to communicate with the clustering functions. When a preferred server is down, the preferred server can be changed to another node and used to communicate with the cluster from that preferred server. This usually means that the server is down. In this case, right-click the name of the cluster, and select **Change Server**. In the Change Server window (Figure 10-46), for Server, change the server and click **OK**.



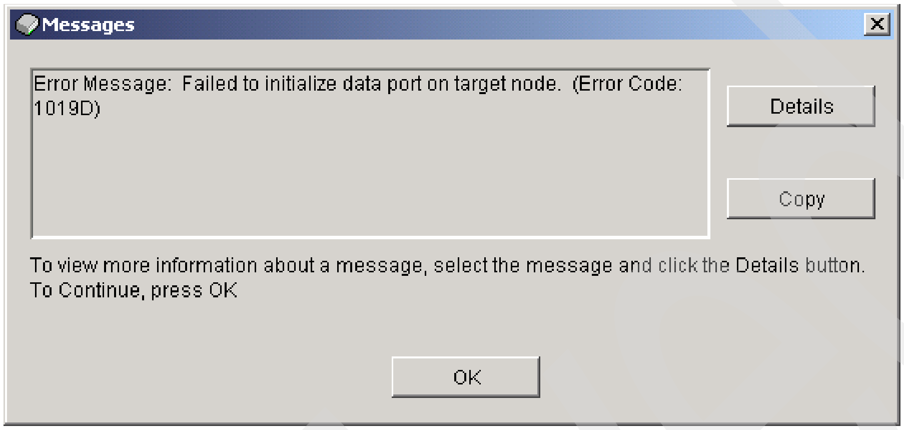
Select the server used to communicate with the cluster. The server must be one of the nodes in the cluster.

Server: Rchiasp1

Figure 10-46 Changing the central server

10.4.2 Data port initialization problems

If geographic mirroring is suspended and begins to resume, you might see an error message like the example shown in Figure 10-47. This error can occur because the data ports are still initializing. You can correct the error by waiting a few minutes and then retrying the resume option.


Messages

Error Message: Failed to initialize data port on target node. (Error Code: 1019D)

To view more information about a message, select the message and click the Details button.
To Continue, press OK

Figure 10-47 Error message indicating a failure to initialize the data port

10.4.3 Missing iSeries Navigator options

In some instances, certain options are not enabled through iSeries Navigator, as shown in Figure 10-48. This happens if a change is made to the geographic mirroring environment that has not been properly updated by the GUI. In most cases, refreshing the view corrects the issue. In rare instances, it is necessary to restart the GUI session.

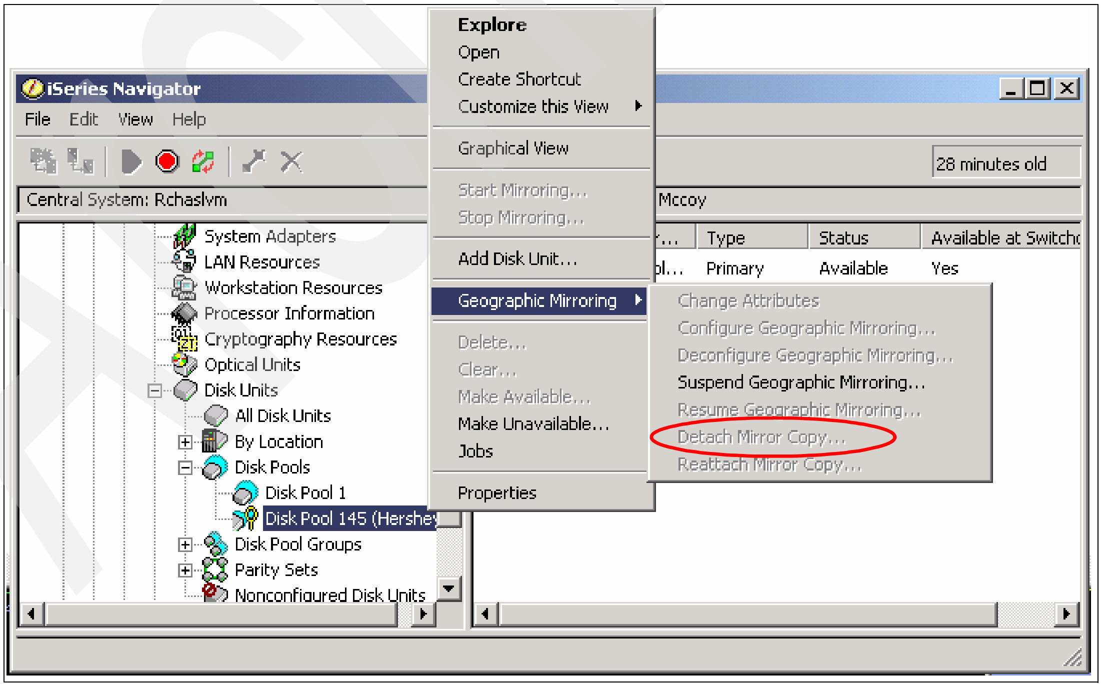


Figure 10-48 Missing options in iSeries Navigator

10.4.4 GUI session hangs

As with any emulator session, it is possible to have an iSeries Navigator session hang. To correct this, restart your session.

If you are in the Disk Units section of the GUI at the time that the session hangs, ending and restarting the GUI does not exit your session from SST. If you try to sign into SST again, you are not allowed to do so.

One method to correct this problem that sometimes works is to change the system values shown in Table 10-2.

Table 10-2 System value changes to exit SST after a hung GUI session

System value	Option
QDEVRCYACN	*ENDJOB
QDSCJOBITV	5 min

If you wait up to 10 minutes after the changes are made, and you are still unable to re-establish a connection to the Disk Units section or SST, the only recovery is to perform an IPL at your next convenience. Be sure to return these system values to their previous settings after the condition is corrected.

10.4.5 Detach operation not allowed

iSeries Navigator does not allow a detach operation when the IASP is varied off and geographic mirroring is suspended, even if the mirror copy of the IASP is in synch with the production copy. In this case, the Detach option is unavailable.

To recover, resume geographic mirroring. The resume changes the mirror copy state to Active because the mirror copy is in synch. Then the Detach option is shown. For an ASP group, you might need to do this for each IASP in the ASP group.

In this situation, no synchronization is performed because the system knows that the copies are already in synch even though they are suspended.

10.4.6 Error message 10386 during mirror copy rebuild in iSeries Navigator

If an error occurs that causes both copies of the IASP to behave as though they are the production copy, the system detects that problem during the resume process, and the resume is unsuccessful. The system reverts the second production copy back to the mirror copy, but you must issue the resume request again to restart the synchronization process.

If this situation occurs, iSeries Navigator posts a 10386 error message:

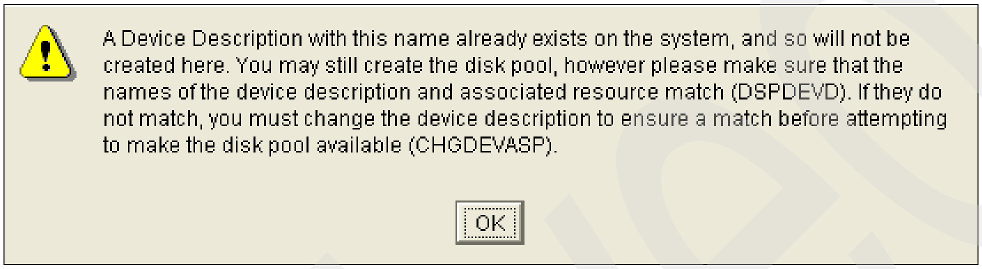
```
GUI Error 10386 — The system rebuilt the mirror copy of the IASP with the correct
configuration source.
```

If you attempt to resume geographic mirroring before you vary on the IASP, no message is generated.

10.4.7 Device descriptions

If you follow the recommendations that are provided in the Information Center to create a new device description for a detached mirror copy of the IASP, you must use VRYCFG from the command line. You cannot use iSeries Navigator to vary on the new device description.

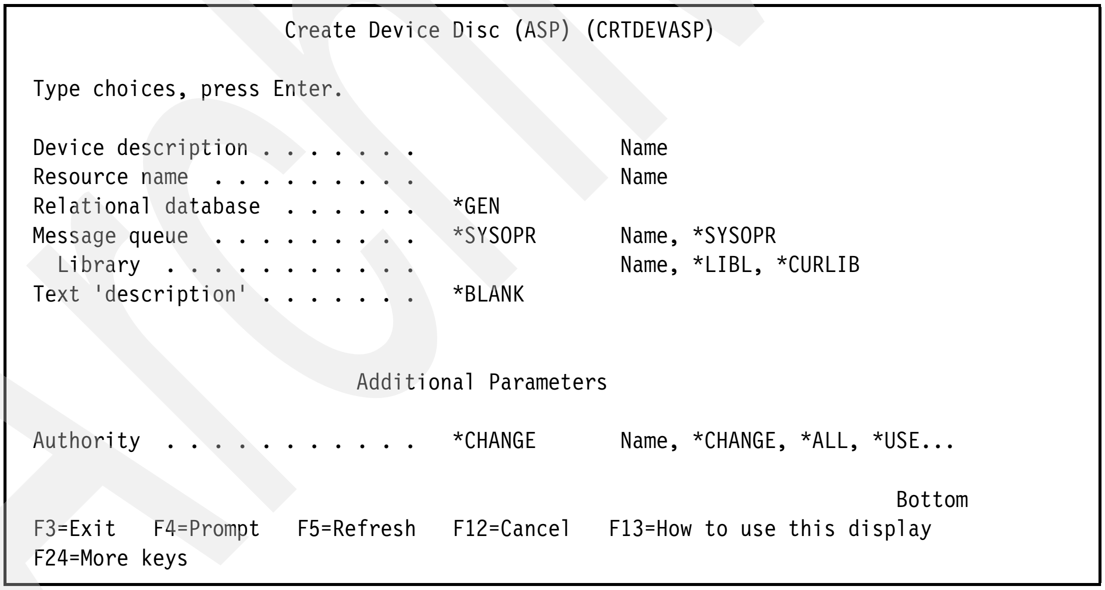
When recreating an IASP or creating a device CRG environment using the command line for many of the steps, you might receive a message in iSeries Navigator indicating that the device description is already created (see Figure 10-49), if the device description has the same name of an existing IASP. In this case, the device description is not changed. If the device description name and associated resource do not match, this becomes a problem.



A Device Description with this name already exists on the system, and so will not be created here. You may still create the disk pool, however please make sure that the names of the device description and associated resource match (DSPDEVD). If they do not match, you must change the device description to ensure a match before attempting to make the disk pool available (CHGDEVASP).

Figure 10-49 DEVD reuse message

If a new DEVD needs to be created, use the CRTDEVASP command, as shown in Figure 10-50, and *not* the CRTDEVD command. Devices that are created using CRTDEVD cannot be used with an IASP.



```
                       Create Device Disc (ASP) (CRTDEVASP)

 Type choices, press Enter.

 Device description . . . . . . .                Name
 Resource name  . . . . . . . . .                Name
 Relational database  . . . . . .   *GEN
 Message queue  . . . . . . . . .   *SYSOPR      Name, *SYSOPR
   Library  . . . . . . . . . . .                Name, *LIBL, *CURLIB
 Text 'description' . . . . . . .   *BLANK


                             Additional Parameters

 Authority  . . . . . . . . . . .   *CHANGE      Name, *CHANGE, *ALL, *USE...

                                                                       Bottom
 F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
 F24=More keys
```

Figure 10-50 Create Device Disc (ASP) (CRTDEVASP) panel

10.5 Other related cross-site mirroring topics

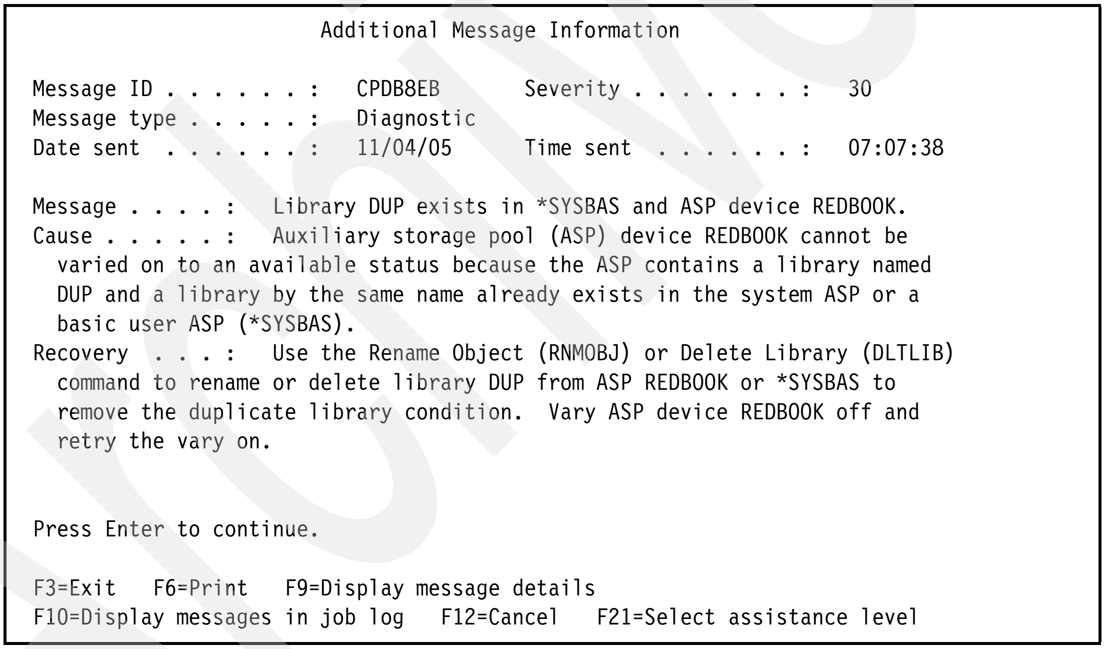
In this section, we describe other potential problems, troubleshooting methods, and recovery actions for errors that are related to XSM.

10.5.1 Duplicate libraries and IASPs

It is possible for IASPs that exist in separate ASP groups to have the same library name. However, problems arise if the IASP has the same library name as an IASP in the system ASP.

When an IASP is varied off, the operating system does not know the names of the libraries that are contained in the IASP. For example, an IASP might be varied off to one system because it was switched to a different system. While it is offline to that system, users are *not* prevented from creating a library by the same name as on the IASP.

If the IASP is then switched back to that system, it contains a duplicate library, and error message CPDB8EB is posted, as shown in Figure 10-51. In this case, the duplicate library needs to be renamed or deleted on the system. The IASP can then be varied off and varied back on to resolve the situation.

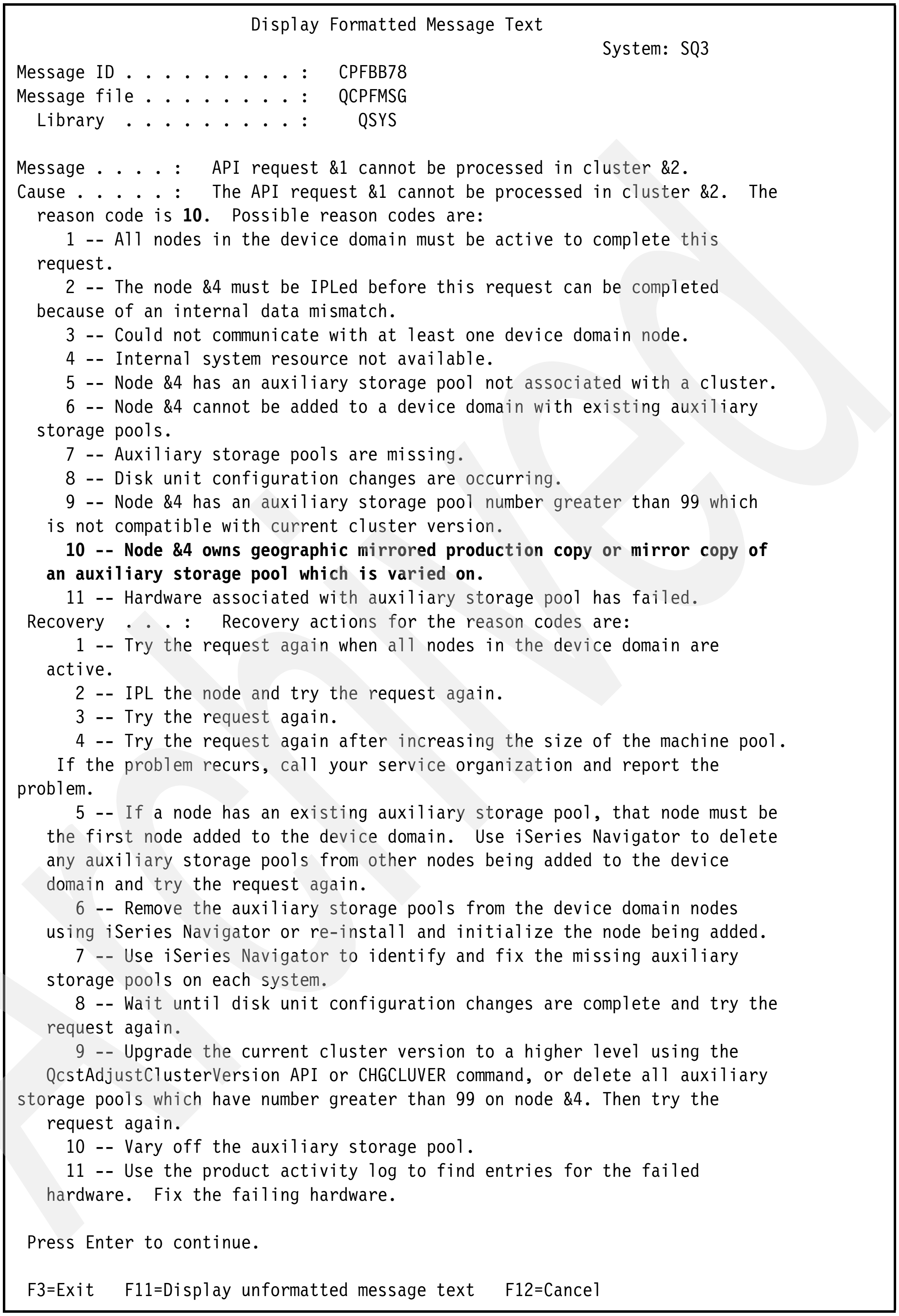


```
                         Additional Message Information

Message ID . . . . . . :   CPDB8EB       Severity . . . . . . . :   30
Message type . . . . . :   Diagnostic
Date sent  . . . . . . :   11/04/05      Time sent  . . . . . . :   07:07:38

Message . . . . :   Library DUP exists in *SYSBAS and ASP device REDBOOK.
Cause . . . . . :   Auxiliary storage pool (ASP) device REDBOOK cannot be
  varied on to an available status because the ASP contains a library named
  DUP and a library by the same name already exists in the system ASP or a
  basic user ASP (*SYSBAS).
Recovery  . . . :   Use the Rename Object (RNMOBJ) or Delete Library (DLTLIB)
  command to rename or delete library DUP from ASP REDBOOK or *SYSBAS to
  remove the duplicate library condition.  Vary ASP device REDBOOK off and
  retry the vary on.


Press Enter to continue.

F3=Exit   F6=Print   F9=Display message details
F10=Display messages in job log   F12=Cancel   F21=Select assistance level
```

Figure 10-51 Example of a duplicate library error

10.5.2 Removing a node from a device domain

If you try to remove a node from a device domain while that node owns the mirror copy of the IASP, and XSM is active, the procedure fails with error message CPFBB78, reason code 10, as shown in Figure 10-52. However if the node is removed when XSM is not active, the mirror copy of the IASP becomes a stand-alone IASP.



```
                        Display Formatted Message Text
                                                           System: SQ3
Message ID . . . . . . . . . :   CPFBB78
Message file . . . . . . . . :   QCPFMSG
  Library  . . . . . . . . . :     QSYS

Message . . . . :   API request &1 cannot be processed in cluster &2.
Cause . . . . . :   The API request &1 cannot be processed in cluster &2.  The
  reason code is 10.  Possible reason codes are:
     1 -- All nodes in the device domain must be active to complete this
  request.
     2 -- The node &4 must be IPLed before this request can be completed
  because of an internal data mismatch.
     3 -- Could not communicate with at least one device domain node.
     4 -- Internal system resource not available.
     5 -- Node &4 has an auxiliary storage pool not associated with a cluster.
     6 -- Node &4 cannot be added to a device domain with existing auxiliary
  storage pools.
     7 -- Auxiliary storage pools are missing.
     8 -- Disk unit configuration changes are occurring.
     9 -- Node &4 has an auxiliary storage pool number greater than 99 which
   is not compatible with current cluster version.
     10 -- Node &4 owns geographic mirrored production copy or mirror copy of
   an auxiliary storage pool which is varied on.
     11 -- Hardware associated with auxiliary storage pool has failed.
 Recovery  . . . :   Recovery actions for the reason codes are:
      1 -- Try the request again when all nodes in the device domain are
   active.
      2 -- IPL the node and try the request again.
      3 -- Try the request again.
      4 -- Try the request again after increasing the size of the machine pool.
    If the problem recurs, call your service organization and report the
problem.
      5 -- If a node has an existing auxiliary storage pool, that node must be
   the first node added to the device domain.  Use iSeries Navigator to delete
   any auxiliary storage pools from other nodes being added to the device
   domain and try the request again.
      6 -- Remove the auxiliary storage pools from the device domain nodes
   using iSeries Navigator or re-install and initialize the node being added.
      7 -- Use iSeries Navigator to identify and fix the missing auxiliary
   storage pools on each system.
      8 -- Wait until disk unit configuration changes are complete and try the
   request again.
      9 -- Upgrade the current cluster version to a higher level using the
   QcstAdjustClusterVersion API or CHGCLUVER command, or delete all auxiliary
storage pools which have number greater than 99 on node &4. Then try the
   request again.
     10 -- Vary off the auxiliary storage pool.
     11 -- Use the product activity log to find entries for the failed
   hardware.  Fix the failing hardware.

 Press Enter to continue.

 F3=Exit   F11=Display unformatted message text   F12=Cancel
```

Figure 10-52   Error message CPFBB78

At this point, you can no longer perform XSM, and the system is unable to add it back to the device domain unless you perform the following steps first:

1. Vary off the IASP on both systems.
2. Delete the IASP on the node that owns the mirror copy.
3. Delete the geographic configurations for this IASP on the system that owns the production copy of the IASP.
4. Try to add the backup node back into the device domain. This step should be unsuccessful, but it is required before you perform an IPL on that system.
5. Perform an IPL on the backup node.
6. Add the backup node into the device domain.
7. Reconfigure geographic mirroring.

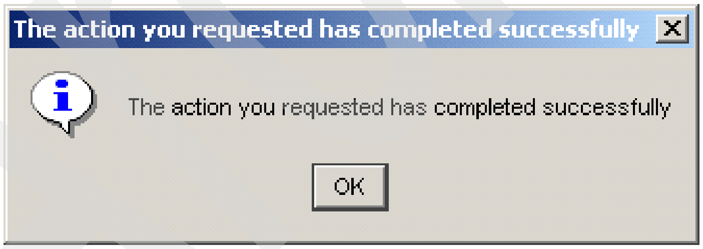
10.5.3 Duplicate production copies of IASPs

In a failover scenario while performing XSM, the two copies of the IASP typically keep consistent information. During the resume or next vary on of the IASPs, the production copy automatically becomes the mirror copy, and the mirror copy becomes the production copy. The two copies are then resynchronized in normal fashion.

However, if a user makes the two production copies available independently during the time the nodes are not communicating by suspending XSM, it is possible to have a situation where both copies are treated as production copies. To resolve the inconsistency, change the recovery domain order, and select the node that owns the production copy. After the change is made, the duplicate production copy reverts back to a mirror copy.

10.5.4 Synchronization

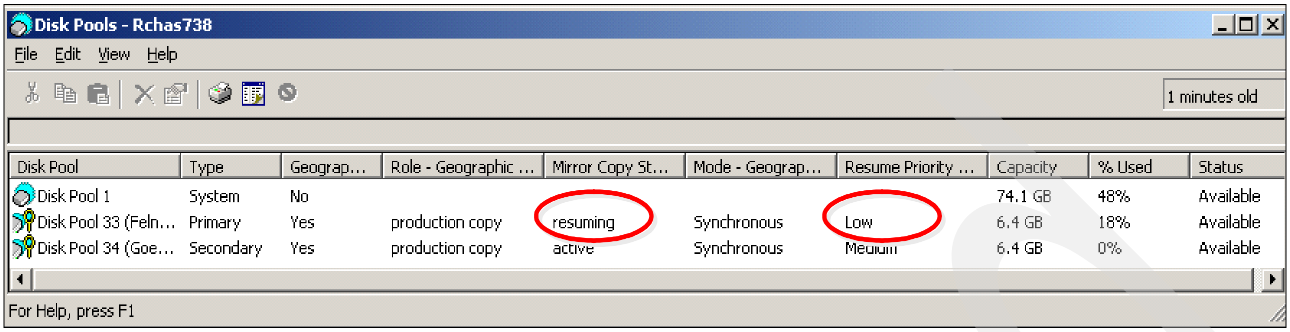
When you perform a resume or reattach operation, you see a message like the one shown in Figure 10-53 in iSeries Navigator. This message indicates that the operation is complete, but synchronization is still required and not necessarily complete yet.


The action you requested has completed successfully

The action you requested has completed successfully

Figure 10-53 Resume or reattach completion message

To check the synchronization process by using iSeries Navigator, note the status of *resuming* (see Figure 10-54). When it is no longer in resuming status, the synchronization process is complete.



Disk Pools - Rchas738

Disk Pool	Type	Geograp...	Role - Geographic ...	Mirror Copy St...	Mode - Geograp...	Resume Priority ...	Capacity	% Used	Status
Disk Pool 1	System	No					74.1 GB	48%	Available
Disk Pool 33 (Feln...	Primary	Yes	production copy	resuming	Synchronous	Low	6.4 GB	18%	Available
Disk Pool 34 (Goe...	Secondary	Yes	production copy	active	Synchronous	Medium	6.4 GB	0%	Available

Figure 10-54 Checking synchronization status from the GUI

From the CLI, use the DSPLOG MSGID(CPI095D) command to determine how much of the process is complete (see Figure 10-55). The status is reported upon starting, completing, and at every 15-minute interval throughout the process.

```
                            Display History Log Contents

Cross-site Mirroring (XSM) synchronization for IASP 89 is 0% complete.
Cross-site Mirroring (XSM) synchronization for IASP 89 is 3% complete.
Cross-site Mirroring (XSM) synchronization for IASP 89 is 6% complete.
Cross-site Mirroring (XSM) synchronization for IASP 89 is 9% complete.

Press Enter to continue.

F3=Exit   F10=Display all   F12=Cancel
```

Figure 10-55 Checking the synchronization status

If you need to change the synchronization priority, you must vary off the IASP first. If you need to change the synchronization priority while the synchronization is already in progress, you can vary off the IASP, change the priority, and vary back on the IASP. The synchronization restarts at the point at which it stopped. No progress is lost, other than the amount of time it took to vary off the IASP, vary back on the IASP, and restart synchronization.

1. Vary off the IASP. While the IASP is varied off, you cannot use it.
2. In the left navigation area in iSeries Navigator, select **Hardware** → **Disk Units** → **Disk Pools**. Right-click your disk pool and select **Geographic Mirroring** → **Change Attributes**. See Figure 10-56.

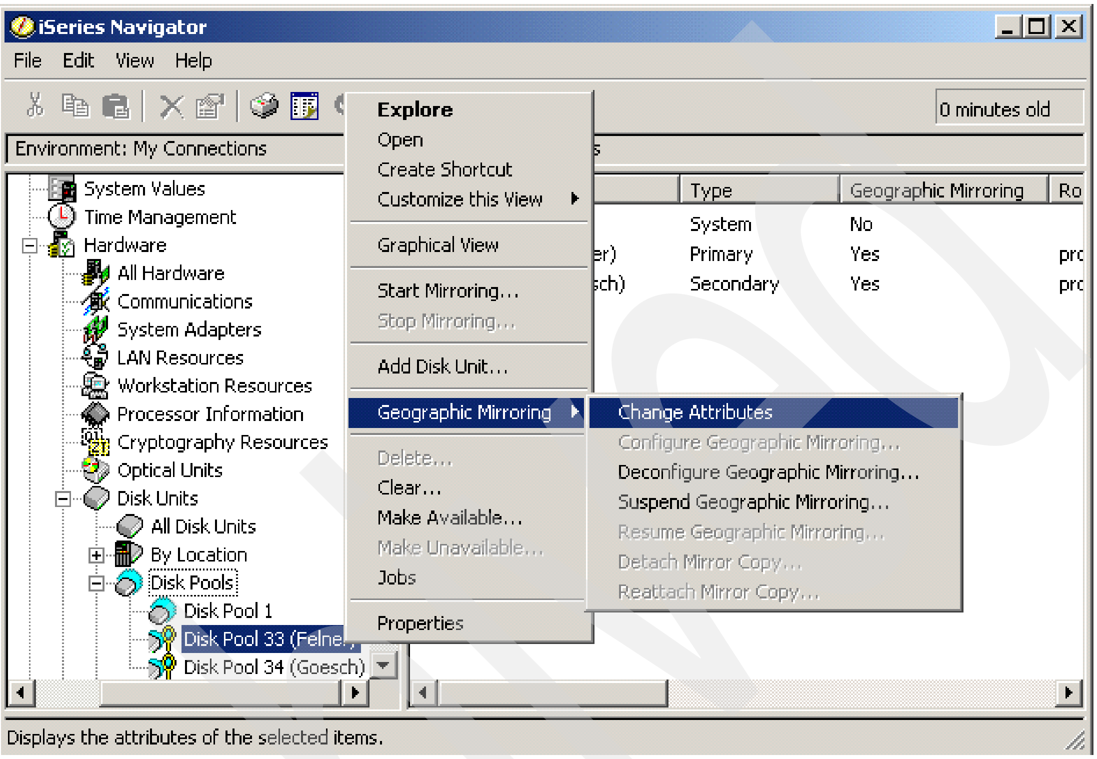
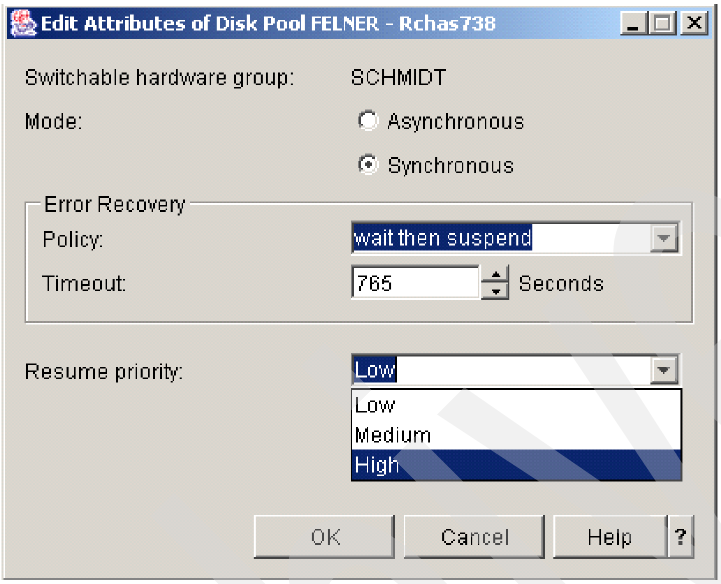


Figure 10-56 Changing synchronization priority

3. In the Edit Attributes of Disk Pool window (Figure 10-57), complete the following tasks:
   a. For Mode, select **Synchronous**.
   b. Under Error Recovery, for Policy, select **wait then suspend**. For Timeout, specify **765** seconds.
   c. For Resume priority, select the priority (low, medium, or high) that you want.
   d. Click **OK**.



Edit Attributes of Disk Pool FELNER - Rchas738

Switchable hardware group: SCHMIDT

Mode: ○ Asynchronous ● Synchronous

Error Recovery

Policy: wait then suspend

Timeout: 765 Seconds

Resume priority: Low

Figure 10-57 Edit Attributes of Disk Pool window

If you vary off the IASP while synchronization is in progress, the next vary on process takes longer than usual. Normally, when the production copy of the IASP is varied on, and synchronization is required, the first tasks that are performed during the synchronization process involve clearing the mirror copy of the IASP. During this phase, no data is transmitted between the nodes, and the production copy is not reading any data. Therefore, the production copy of the IASP can vary on quickly, because there is no contention between the vary on task and the synchronization task.

After you vary off the IASP during synchronization and then vary it on again, the production copy immediately begins reading the data to send to the mirror copy. The vary on task competes for DASD access with the synchronization task, which can cause the vary on process to take longer than in normal situations.

If you choose to suspend geographic mirroring prior to varying off the IASP to prevent this from occurring, synchronization needs to start from the beginning, and all progress to that point is lost.

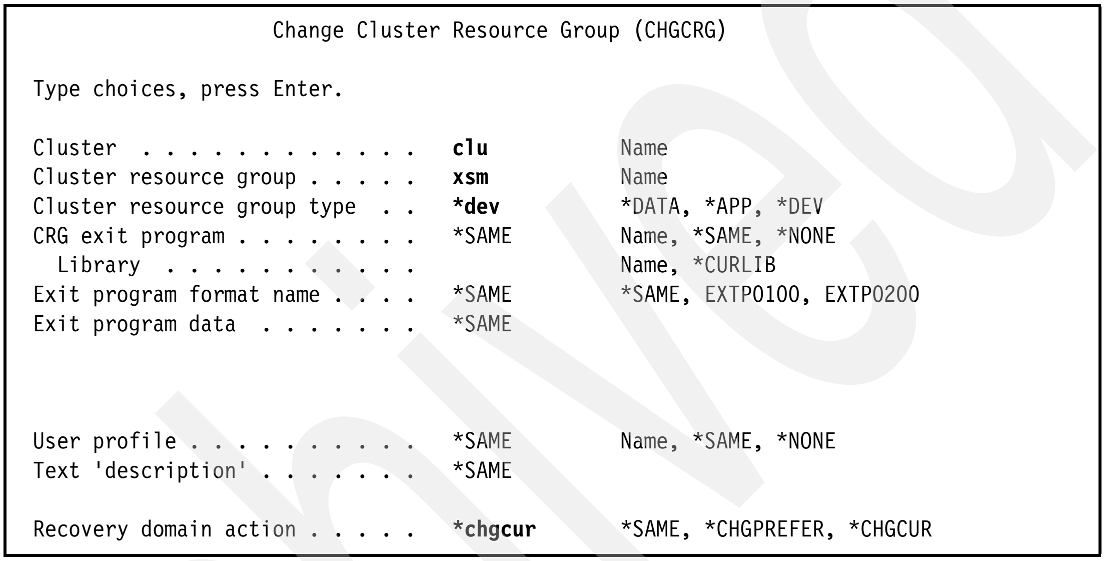
10.5.5 Mirror copy promotion

In some situations, it is necessary to promote the mirror copy of the IASP to the production copy. For example, if XSM is suspended, and during that time, a failure occurs that prevents the production copy from being usable, the mirror copy can be promoted to the production copy. You can convert this copy by using the Change Cluster Resource Group (CHGCRG) command, as shown in Figure 10-58 on page 216, Figure 10-59 on page 216, and Figure 10-60 on page 217. You can accomplish this by using Management Central within the iSeries Navigator product.

The mirror copy of the IASP might not have the most current information, but in an emergency, it might be quicker to use that copy than to restore from a backup, which might have even older data than the mirror copy.

Figure 10-58 through Figure 10-60 on page 217 demonstrate how to change the CRG by using the CHGCRG command. After prompting the command, in the first Change Cluster Resource Group display (Figure 10-58), complete the following fields:

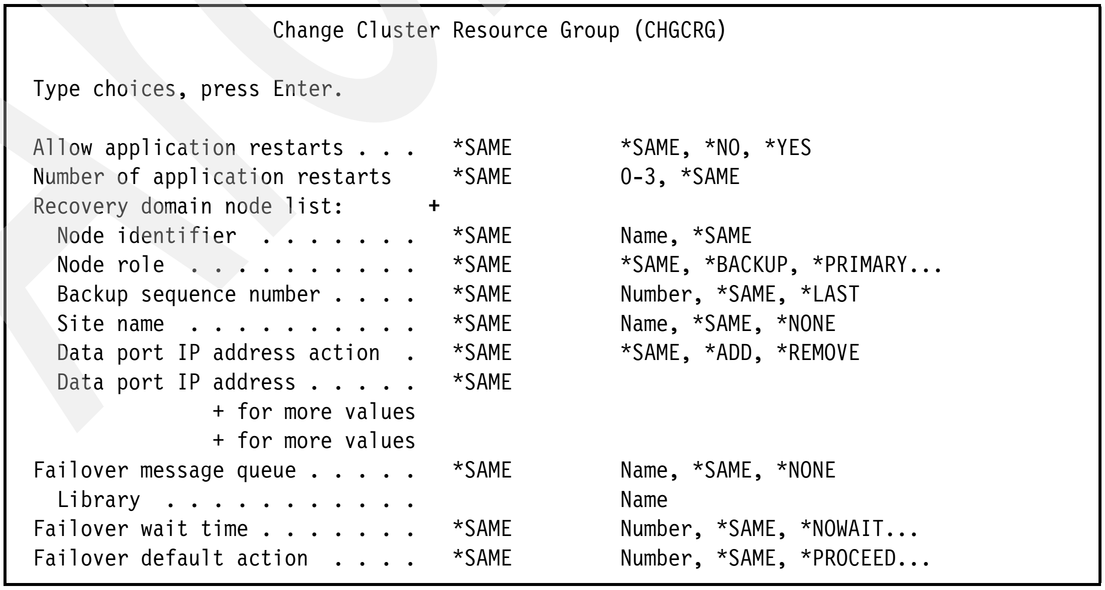
- For Cluster, type `clu`.
- For Cluster Resource Group, type `xsm`.
- For Cluster resource group type, type `*dev`.
- For Recovery domain action, type `*chgcur`.



```
                    Change Cluster Resource Group (CHGCRG)

 Type choices, press Enter.

 Cluster  . . . . . . . . . . .   clu           Name
 Cluster resource group . . . . . xsm           Name
 Cluster resource group type  . . *dev          *DATA, *APP, *DEV
 CRG exit program . . . . . . . . *SAME         Name, *SAME, *NONE
   Library  . . . . . . . . . . .               Name, *CURLIB
 Exit program format name . . . . *SAME         *SAME, EXTP0100, EXTP0200
 Exit program data  . . . . . . . *SAME



 User profile . . . . . . . . . . *SAME         Name, *SAME, *NONE
 Text 'description' . . . . . . . *SAME

 Recovery domain action . . . . . *chgcur       *SAME, *CHGPREFER, *CHGCUR
```

Figure 10-58 Change Cluster Resource Group (CHGCRG) panel (Part 1 of 3)

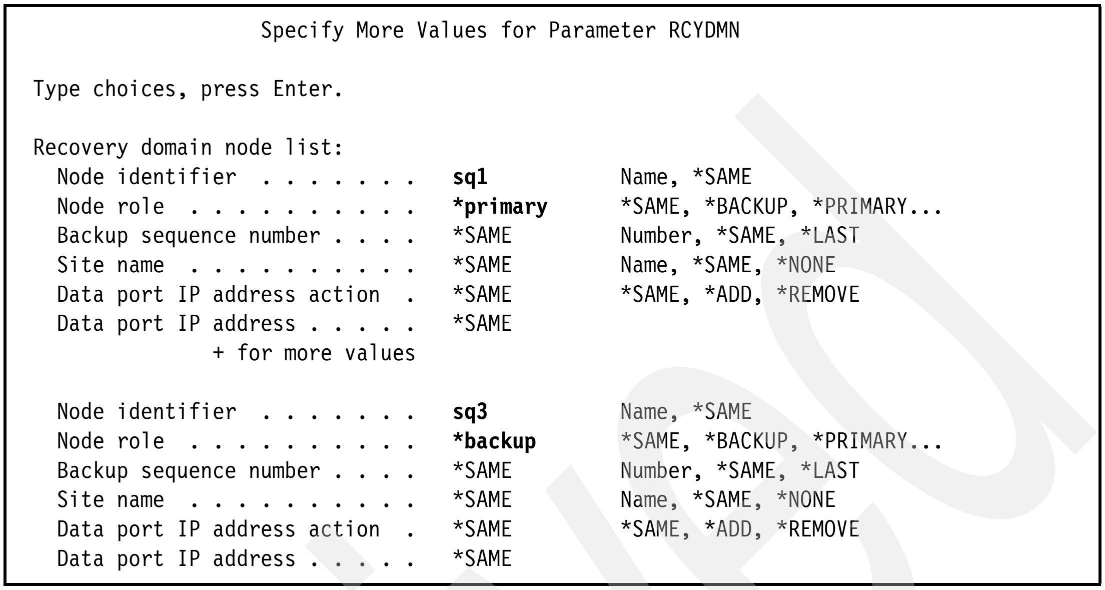
Page down and you see the display shown in Figure 10-59. Modify the Recovery domain node list appropriately.



```
                    Change Cluster Resource Group (CHGCRG)

 Type choices, press Enter.

 Allow application restarts . . .  *SAME         *SAME, *NO, *YES
 Number of application restarts    *SAME         0-3, *SAME
 Recovery domain node list:      +
   Node identifier  . . . . . . .  *SAME         Name, *SAME
   Node role  . . . . . . . . . .  *SAME         *SAME, *BACKUP, *PRIMARY...
   Backup sequence number . . . .  *SAME         Number, *SAME, *LAST
   Site name  . . . . . . . . . .  *SAME         Name, *SAME, *NONE
   Data port IP address action  .  *SAME         *SAME, *ADD, *REMOVE
   Data port IP address . . . . .  *SAME
               + for more values
               + for more values
 Failover message queue . . . . .  *SAME         Name, *SAME, *NONE
   Library  . . . . . . . . . . .                Name
 Failover wait time . . . . . . .  *SAME         Number, *SAME, *NOWAIT...
 Failover default action  . . . .  *SAME         Number, *SAME, *PROCEED...
```

Figure 10-59 Change Cluster Resource Group (CHGCRG) panel (Part 2 of 3)

Figure 10-60 shows how to edit the Recover domain node list to promote the mirror copy (Backup 1 node) to the production copy.



```
                     Specify More Values for Parameter RCYDMN

 Type choices, press Enter.

 Recovery domain node list:
   Node identifier  . . . . . . .   sq1           Name, *SAME
   Node role  . . . . . . . . . .   *primary      *SAME, *BACKUP, *PRIMARY...
   Backup sequence number . . . .   *SAME         Number, *SAME, *LAST
   Site name  . . . . . . . . . .   *SAME         Name, *SAME, *NONE
   Data port IP address action  .   *SAME         *SAME, *ADD, *REMOVE
   Data port IP address . . . . .   *SAME
                + for more values

   Node identifier  . . . . . . .   sq3           Name, *SAME
   Node role  . . . . . . . . . .   *backup       *SAME, *BACKUP, *PRIMARY...
   Backup sequence number . . . .   *SAME         Number, *SAME, *LAST
   Site name  . . . . . . . . . .   *SAME         Name, *SAME, *NONE
   Data port IP address action  .   *SAME         *SAME, *ADD, *REMOVE
   Data port IP address . . . . .   *SAME
```

Figure 10-60 Change Cluster Resource Group (CHGCRG) panel (Part 3 of 3)

In this example, we assume that node sq3 was previously the primary node in the cluster and that node sq1 was the first backup node. You can review this prior to running the CHGCRG command by viewing the current CRG recovery domain.

If more than two nodes are in your recovery domain, you might want to explicitly list the new node role for each node. In our example, only nodes sq1 and sq3 are modified.

Another reason to promote the mirror copy to the production copy is if you detach the mirror copy and mistakenly make production changes to the mirror copy, rather than to the production copy. In this case, the data on the mirror copy is the most current, and you can promote the mirror copy to the production copy by using the CHGCRG command.

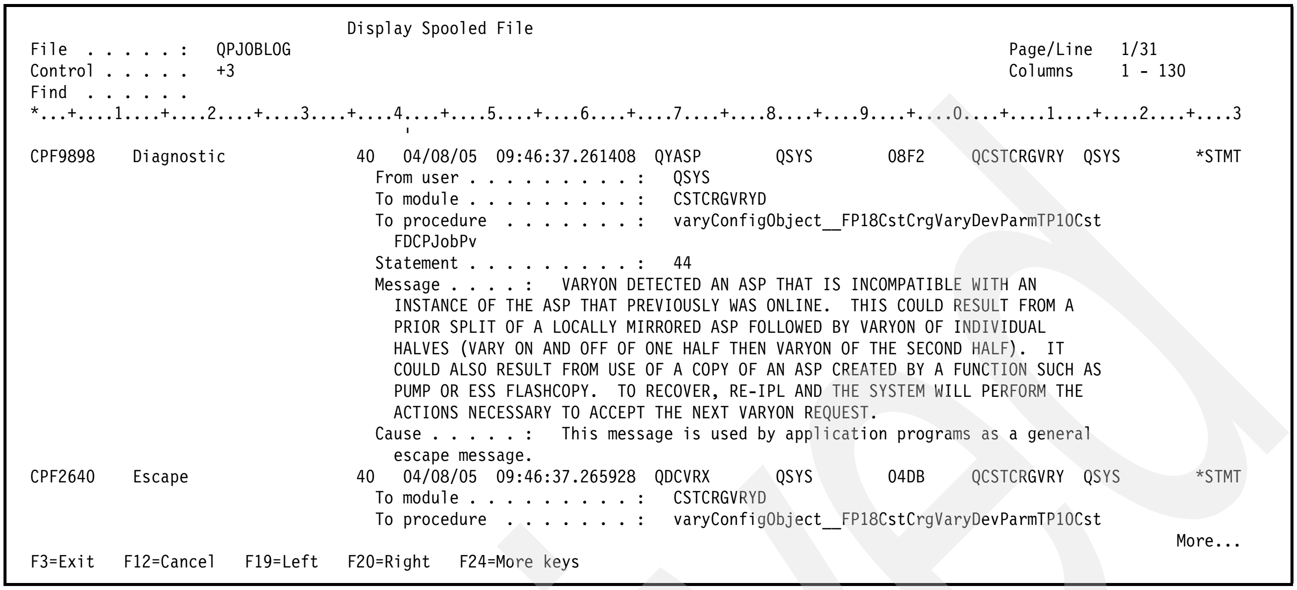
You cannot use the CHGCRG command to promote a mirror copy if geographic mirroring is suspended for some IASPs, but not all of them, in an ASP group. Use of this command can cause the IASPs in the ASP group to have an inconsistent level of data. If geographic mirroring is suspended for all IASPs in the group, you can use the CHGCRG command to promote a mirror copy. However, if you suspend geographic mirroring at different times for each IASP, the IASPs *will* have inconsistent data.

10.5.6 Signature violation

Signatures are written to the system and to an IASP when you vary off an IASP. When you vary on the IASP, signatures are compared and the vary on fails if they do not match. An error message of CPF9898, as shown in Figure 10-61 on page 218, is posted. The purpose of the signatures is to detect whether a system using a copy of the IASP is using a copy that does not contain the most current data.

While the system uses the data in the IASP, it assigns virtual addresses to system objects. These addresses can change as the data keeps updating, and previously-used addresses can be assigned to other objects. If an older copy of the IASP is used, the addresses that are assigned to objects in the older copy can be used currently for other objects in the current copy.

In this case, an IPL is required to enable the system to nullify all pointers in SYSBAS that point to the IASP to prevent integrity exposures to the data in the IASP.



```
                                  Display Spooled File
File  . . . . . :   QPJOBLOG                                                                    Page/Line   1/31
Control . . . . .   +3                                                                          Columns     1 - 130
Find  . . . . . .
*...+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8....+....9....+....0....+....1....+....2....+....3
                                         '
CPF9898    Diagnostic             40   04/08/05  09:46:37.261408  QYASP       QSYS        08F2     QCSTCRGVRY  QSYS        *STMT
                                     From user . . . . . . . . . :   QSYS
                                     To module . . . . . . . . . :   CSTCRGVRYD
                                     To procedure  . . . . . . . :   varyConfigObject__FP18CstCrgVaryDevParmTP10Cst
                                       FDCPJobPv
                                     Statement . . . . . . . . . :   44
                                     Message . . . . :   VARYON DETECTED AN ASP THAT IS INCOMPATIBLE WITH AN
                                       INSTANCE OF THE ASP THAT PREVIOUSLY WAS ONLINE.  THIS COULD RESULT FROM A
                                       PRIOR SPLIT OF A LOCALLY MIRRORED ASP FOLLOWED BY VARYON OF INDIVIDUAL
                                       HALVES (VARY ON AND OFF OF ONE HALF THEN VARYON OF THE SECOND HALF).  IT
                                       COULD ALSO RESULT FROM USE OF A COPY OF AN ASP CREATED BY A FUNCTION SUCH AS
                                       PUMP OR ESS FLASHCOPY.  TO RECOVER, RE-IPL AND THE SYSTEM WILL PERFORM THE
                                       ACTIONS NECESSARY TO ACCEPT THE NEXT VARYON REQUEST.
                                     Cause . . . . . :   This message is used by application programs as a general
                                       escape message.
CPF2640    Escape                 40   04/08/05  09:46:37.265928  QDCVRX      QSYS        04DB     QCSTCRGVRY  QSYS        *STMT
                                     To module . . . . . . . . . :   CSTCRGVRYD
                                     To procedure  . . . . . . . :   varyConfigObject__FP18CstCrgVaryDevParmTP10Cst
                                                                                                                         More...
F3=Exit   F12=Cancel   F19=Left   F20=Right   F24=More keys
```

Figure 10-61 Signature violation

If there is a scenario that produces two production copies of the IASP, there is a 50% chance of getting a signature violation on the next switchover.

Suppose that XSM is active in a two-node cluster, and the CHGCRGPRI command is used to perform a switchover. XSM reverses direction and the backup node becomes the primary and owns the production copy of the IASP. An ENDSYS command is issued on the node that owns the mirror copy of the IASP. The system recovers and is added back into the cluster. XSM restarts, and resynchronization occurs. Later, the CHGCRGPRI command is used again and error message CPF9898 occurs for a signature violation. An IPL is required.

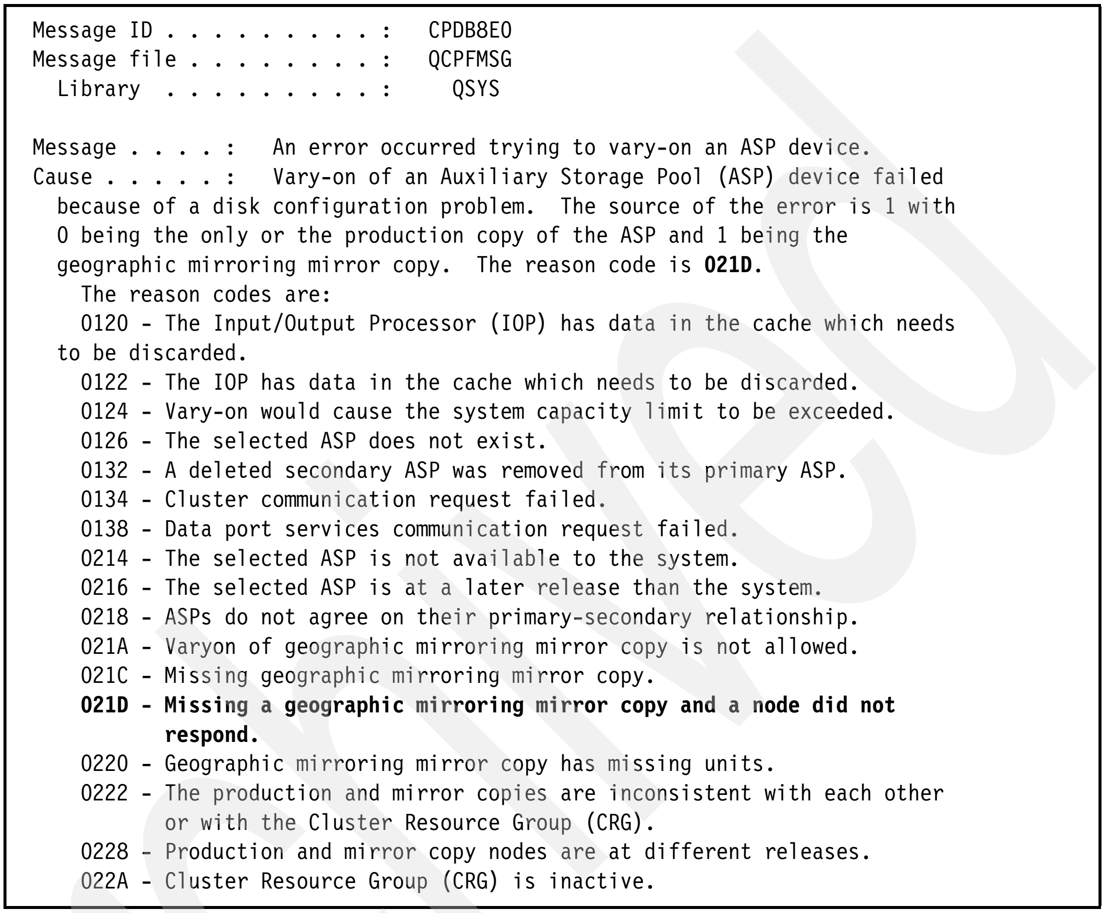
Why does this happen? Unlike clustering, IASP support considers ENDSYS to be a normal termination because it can write changed pages to DASD. In this scenario, these signatures do not make it back to the failed-to system, and the signature error is generated.

10.5.7 Failover scenario with vary online IASP *NO

When setting up the device CRG for your IASP in your cluster environment, you have to decide on whether you want to system to automatically vary on your IASP on the backup node in case of a switchover or of a failover. Both settings carry both advantages and disadvantages with them.

If you set the vary online IASP parameter in the device CRG to *YES, whenever a switchover or a failover to the backup node occurs, the IASP is automatically varied on on the backup node during this process. However, there might be reasons why the IASP cannot be varied on, for example if someone created a library in the system ASP of the backup node that also exists in the IASP. If this is the case, then with a failover, the IASP is not varied on. In case of a switchover however, the whole process is rolled back. You end up with your original production system being the primary node and the IASP being varied on on that node. It requires less time if you switch over to the backup node, try to vary on the IASP manually, and then encounter and solve the problem.

If you set the vary online IASP parameter in the device CRG to *NO, in case of a switchover or failover to the backup node, the IASP is not automatically varied on. The vary on must be done manually. Be aware that, in case of a failover, you are presented with error message CPDB8E0 (see Figure 10-62).



```
Message ID . . . . . . . . . :   CPDB8E0
Message file . . . . . . . . :   QCPFMSG
  Library  . . . . . . . . . :     QSYS

Message . . . . :   An error occurred trying to vary-on an ASP device.
Cause . . . . . :   Vary-on of an Auxiliary Storage Pool (ASP) device failed
  because of a disk configuration problem.  The source of the error is 1 with
  0 being the only or the production copy of the ASP and 1 being the
  geographic mirroring mirror copy.  The reason code is 021D.
    The reason codes are:
    0120 - The Input/Output Processor (IOP) has data in the cache which needs
  to be discarded.
    0122 - The IOP has data in the cache which needs to be discarded.
    0124 - Vary-on would cause the system capacity limit to be exceeded.
    0126 - The selected ASP does not exist.
    0132 - A deleted secondary ASP was removed from its primary ASP.
    0134 - Cluster communication request failed.
    0138 - Data port services communication request failed.
    0214 - The selected ASP is not available to the system.
    0216 - The selected ASP is at a later release than the system.
    0218 - ASPs do not agree on their primary-secondary relationship.
    021A - Varyon of geographic mirroring mirror copy is not allowed.
    021C - Missing geographic mirroring mirror copy.
    021D - Missing a geographic mirroring mirror copy and a node did not
           respond.
    0220 - Geographic mirroring mirror copy has missing units.
    0222 - The production and mirror copies are inconsistent with each other
           or with the Cluster Resource Group (CRG).
    0228 - Production and mirror copy nodes are at different releases.
    022A - Cluster Resource Group (CRG) is inactive.
```

Figure 10-62 Error message CPDB8E0

This message indicates that XSM is trying to communicate with the mirrored copy, which it cannot because the other node is down. To resolve that situation, use iSeries Navigator to suspend XSM before varying on the IASP.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

For information about ordering these publications, see "How to get Redbooks" on page 222. Note that some documents referenced here might be available in softcopy only.

- *AS/400 Remote Journal Function for High Availability and Data Replication*, SG24-5189
- *Clustering and IASPs for Higher Availability on the IBM eServer iSeries Server*, SG24-5194
- *Data Resilience Solutions for IBM i5/OS High Availability Clusters*, REDP-0888
- *High-speed Link Loop Architecture for the IBM eServer iSeries Server: OS/400 Version 5 Release 2*, REDP-3652
- *IBM @server i5 and iSeries System Handbook: IBM i5/OS Version 5 Release 3 October 2004*, GA19-5486
- *IBM eServer iSeries Independent ASPs: A Guide to Moving Applications to IASPs*, SG24-6802
- *Independent ASP Performance Study on the IBM @server iSeries Server*, REDP-3771
- *iSeries and IBM TotalStorage: A Guide to Implementing External Disk on eServer i5*, SG24-7120

Other publications

The publication *IBM eServer iSeries OptiConnect for OS/400 Version 5,* SC41-5414-04, is also relevant as a further information source.

Online resources

These Web sites are also relevant as further information sources:

- *Clustering for High Availability* by Steve Finnes

  `http://www-03.ibm.com/servers/eserver/iseries/ha/pdf/ClusteringForHA.pdf`

- "HA 101" by Finnes, Steve; Gintowt, Bob; and Snyder, Mike in *IBM Systems Magazine,* System i5 edition, July 2003

  `http://www.ibmsystemsmag.com/i5/july03/coverstory/7974p1.aspx`

- i5/OS Information Center

  `http://publib.boulder.ibm.com/infocenter/iseries/v5r4/index.jsp`

How to get Redbooks

You can search for, view, or download Redbooks, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

IBM Redbooks

Availability Management: Planning and Implementing Cross-Site Mirroring on IBM System i5

(0.2”spine)
0.17”<->0.473”
90<->249 pages

IBM®

Availability Management

Planning and Implementing Cross-Site Mirroring on IBM System i5

Redbooks

Understand availability technology for IBM i5/OS using cross-site mirroring

Learn about planning considerations for cross-site mirroring

Explore the new functions that are available at V5R4

In this IBM Redbooks publication, we introduce the concept of cross-site mirroring (XSM) as implemented in IBM OS/400 or i5/OS at V5R3M0 and later. XSM describes the replication of data at multiple sites. It involves the use of clustering, cluster resource groups (CRGs), independent auxiliary storage pools (IASPs), and other components.

In this updated version, we include the new i5/OS V5R4 functions of Administrative Domain and Source Site Tracking. We also include preview information of the Target Site Tracking function.

An additional component of this highly available technology is that XSM keeps two identical copies of an independent disk pool at two sites to provide high availability and disaster recovery. These sites can be geographically close to one another or far apart, depending on the needs of the business.

This IBM Redbook is written for IBM technical professionals, Business Partners, and customers who are considering, planning, and implementing a highly available solution on the IBM System i5 platform.

SG24-6661-01 ISBN 0738489905