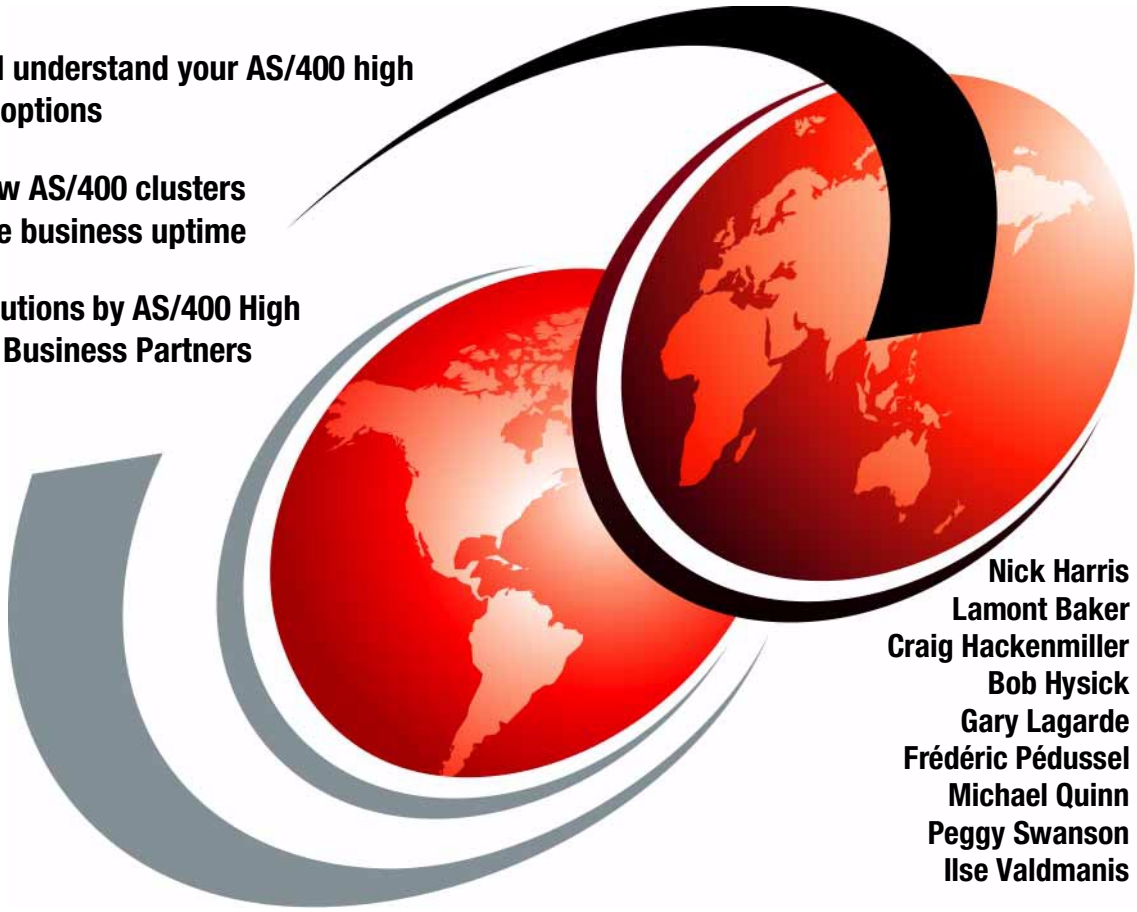
IBM

AS/400 Clusters

A Guide to Achieving Higher Availability

- Explore and understand your AS/400 high availability options
- Find out how AS/400 clusters can improve business uptime
- Preview solutions by AS/400 High Availability Business Partners



Nick Harris
Lamont Baker
Craig Hackenmiller
Bob Hysick
Gary Lagarde
Frédéric Pédussel
Michael Quinn
Peggy Swanson
Ilse Valdmanis

Redbooks

ibm.com/redbooks

SG24-5194-00

International Technical Support Organization

AS/400 Clusters: A Guide to Achieving Higher Availability

August 2000

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix D, “Special notices” on page 157.

First Edition (August 2000)

This edition applies to OS/400 Version 4, Release 4.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. JLU Building 107-2
3605 Highway 52N
Rochester, Minnesota 55901-7829

Contents

Preface ix
The team that wrote this redbook ix
Comments welcome xi

Part 1. Clusters and the IBM AS/400 system 1

Chapter 1. Introduction 3
1.1 Cluster 3
1.2 High availability 3
1.3 The chapters in overview 4
1.3.1 Downtime issues 4
1.3.2 Availability technology 5
1.3.3 AS/400 clusters explained 6
1.3.4 ClusterProven 6
1.3.5 Planning for AS/400 clusters 6
1.4 High Availability Business Partners 7
1.4.1 DataMirror iCluster 7
1.4.2 Lakeview Technology availability solutions 7
1.4.3 Vision Solutions OMS/400 Cluster Manager 7
1.5 Appendices 7
1.5.1 AS/400 high availability functions 8
1.5.2 Problem determination 8

Chapter 2. Downtime issues 9
2.1 Basic backup models 9
2.1.1 Factors influencing availability 11
2.1.2 Financial impact of an outage 13
2.2 Activities that cause downtime 14
2.3 Example of business impact analysis 15

Chapter 3. Availability technology 19
3.1 Single system availability 19
3.2 The AS/400 system and 99.9+% availability 20
3.3 Standby secondary 22
3.4 Active secondary 22
3.5 Replication technology 23
3.6 Switched disk 24
3.7 Shared disk 25
3.8 Separate server 27
3.9 The AS/400 system and its availability 28
3.9.1 AS/400 availability options 28

3.9.2 Clusters 30
3.9.3 Logical partitioning 32
3.10 AS/400 high availability middleware 34

Chapter 4. AS/400 clusters explained 37
4.1 What an AS/400 cluster is 37
4.1.1 How a cluster is used 38
4.1.2 Four-node mutual takeover cluster 39
4.1.3 Application and data resilience 40
4.2 Cluster Resource Services structure 41
4.2.1 Underlying technologies 44
4.2.2 Partition state 48
4.2.3 Versioning 49
4.2.4 Conclusion 50

Chapter 5. ClusterProven 51
5.1 Overview of the ClusterProven components 51
5.1.1 OS/400 Cluster Resource Services 52
5.1.2 Cluster middleware data resiliency 52
5.1.3 Cluster middleware cluster management 52
5.1.4 ClusterProven application resiliency 52
5.2 ClusterProven for AS/400 53
5.2.1 Basic ClusterProven for AS/400 53
5.2.2 Advanced ClusterProven for AS/400 53

Chapter 6. Planning for AS/400 clusters 55
6.1 Cluster planning steps 55
6.1.1 Simple cluster 55
6.1.2 Full cluster deployment 55
6.2 Cluster planning 56
6.2.1 Business impact costs 56
6.2.2 Level of availability 57
6.2.3 Configuration of a cluster 58
6.2.4 Replication environment 59
6.3 Applications 60
6.3.1 Application information 60
6.3.2 Application object inventory 60
6.3.3 Resilient data 61
6.3.4 Resilient applications 61
6.3.5 Switchover 62
6.3.6 Failover 62
6.3.7 Restart 62
6.3.8 Maintenance 63
6.3.9 Database performance 67

6.3.10 HABP selection 68
6.4 Systems management 68
6.4.1 Service level agreements 68
6.4.2 Operations management 68
6.4.3 Problem and change management 69
6.4.4 Capacity 70
6.4.5 Performance 70
6.4.6 Security 70
6.5 Hardware 71
6.5.1 Redundancy 71
6.5.2 Network planning 72
6.6 Cluster testing 72
6.6.1 General system management-related tests 74
6.6.2 Cluster management-related tests 74

Part 2. High Availability Business Partners 75

Chapter 7. DataMirror iCluster 77
7.1 Getting started with iCluster 78
7.2 Creating a cluster 79
7.2.1 Adding a node to the cluster 80
7.2.2 Activating and de-activating nodes in the cluster 82
7.3 Creating and using Cluster Resource Groups (CRGs) 83
7.3.1 Creating data CRGs 83
7.3.2 Selecting objects for a data CRG for high availability 85
7.3.3 Creating application CRGs 87
7.3.4 Changing a CRG recovery domain 87
7.3.5 Activating or starting a data CRG 89
7.3.6 De-activating or ending a data CRG 90
7.3.7 Switching over a data CRG 90
7.4 Using ClusterProven applications 91
7.4.1 Setting up a resilient application 91
7.4.2 Selecting objects to a resilient application 93
7.4.3 Changing or updating a resilient application 93
7.4.4 Changing a resilient application's recovery domain 94
7.4.5 Activating or starting a resilient application 95
7.4.6 De-activating or ending a resilient application 95
7.4.7 Switching over a resilient application 95
7.5 Removing the cluster and its components 96
7.5.1 Removing a resilient application 96
7.5.2 Removing a data CRG 96
7.5.3 Removing a node from the cluster 96
7.5.4 Removing the entire cluster 97

7.6 Using iCluster commands to access Cluster Services operations 97

Chapter 8. Lakeview Technology availability solutions 99
8.1 MIMIX ClusterServer 99
8.1.1 The need for availability 99
8.1.2 MIMIX ClusterServer for AS/400 solution 100
8.2 MIMIX FastPath 102
8.2.1 The need for ClusterReady applications 102
8.2.2 Why MIMIX FastPath 102
8.2.3 The MIMIX FastPath process 103

Chapter 9. Vision Solutions 105
9.1 Vision Solutions OMS/400 Cluster Manager 105
9.1.1 Implementation goals 105
9.2 Getting started with OMS/400 Cluster Manager 105
9.2.1 Installing the client 106
9.2.2 Starting the product 106
9.2.3 Defining host systems 106
9.2.4 Auto-detection of clustered nodes 107
9.2.5 IP interface selection 108
9.2.6 Working with ClusterProven applications 108
9.3 OMS/400 Cluster Manager sample displays 109
9.3.1 Working with clusters and CRGs 109
9.3.2 Creating new clusters 109
9.3.3 Viewing cluster information 110
9.3.4 Adding a node to the cluster 111
9.3.5 Activating and de-activating nodes in the cluster 112
9.3.6 Creating and using Cluster Resource Groups (CRGs) 113
9.3.7 Changing a CRG recovery domain 114
9.3.8 Activating or starting a data or application CRG 115
9.3.9 De-activating or ending a data or application CRG 116
9.3.10 Creating an application CRG recovery domain 118
9.3.11 Removing a data or application CRG 119
9.3.12 Removing a node from the cluster 119
9.3.13 Removing the entire cluster 120
9.4 Working with applications 121
9.4.1 ISV data area contents 121
9.4.2 Creating ISV data areas for application CRGs 122
9.4.3 Changing or updating data areas 123
9.4.4 Changing a resilient application's data area contents 124
9.4.5 Working with object specifiers 125
9.4.6 Object Selection Results. 127
9.4.7 Creating a list of objects for high availability 127

9.4.8 Viewing OMS/400 links and statistics 128

Part 3. Appendices 131

Appendix A. AS/400 cluster resources 133

Appendix B. AS/400 high availability functions 135
B.1 Journaling 135
B.2 Access path protection 136
B.3 Auxiliary storage pools 137
B.4 Device parity protection 138
B.5 Mirrored protection 140
B.6 Separate servers. 141
B.7 Clusters. 141
B.8 Logical partitioning 143
B.9 Additional information 146

Appendix C. Problem determination 147
C.1 Monitoring for problems 147
C.2 Common cluster questions 147
C.2.1 Why won't my cluster start? 148
C.2.2 Why is my CRG hung up? 148
C.2.3 Is my cluster up and running? 148
C.2.4 Why do I have two clusters after fixing my cluster partition? 148
C.3 Recovering from a clustered partition 148
C.3.1 Cluster partition tips 150
C.3.2 Merging a cluster partition example 152

Appendix D. Special notices 157

Appendix E. Related publications 161
E.1 IBM Redbooks 161
E.2 IBM Redbooks collections. 161
E.3 Other resources 162
E.4 Referenced Web sites. 162

How to get IBM Redbooks 163
IBM Redbooks fax order form 164

Index 165

IBM Redbooks review 171

Preface

Gain a broad understanding of the new cluster architecture available with OS/400 Version 4 Release 4. In this era of e-commerce, availability is of the utmost importance for business survival. This new cluster architecture provides support for customers who want to make their businesses continuously available.

This redbook presents an overview of a generic cluster and the basic terminology surrounding clusters. It also examines the AS/400 cluster and its implementation. It introduces you to the new brand initiative ClusterProven for AS/400 and explains how it applies to AS/400 customers and independent software vendors.

This redbook targets IBM customers, technical representatives, and Business Partners who are planning business solutions and systems that are continuously available.

The team that wrote this redbook

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

There are many names on the cover of this book, not because the subject is difficult, but because the information is disparate. We have brought together this group to give you the best of all worlds, an IBM perspective, an IBM High Availability Business Partner perspective, and a customer perspective.

Nick Harris is an Senior Systems Specialist for the AS/400 system in the International Technical Support Organization, Rochester Center. He specializes in server consolidation, AS/400 hardware, and the Integrated Netfinity Server. He also writes and teaches IBM classes worldwide on areas of AS/400 system design, business intelligence, and database technology. He spent 12 years as a System Specialist in the United Kingdom AS/400 Business and has experience in S/36, S/38, and the AS/400 system.

Lamont Baker worked for four years in the Rochester ITSO shortly after the AS/400 system was announced. He concentrated mainly on distributing knowledge in the database, application development, and performance areas of the AS/400 system. Lamont left IBM in 1995 when he was a senior systems engineer in the IBM Toronto AS/400 Lab. He began Development Support and continued to work in the AS/400 teaching and application development

area. Lamont works closely with Application Solutions Inc., an IBM business partner. Lamont is available at LamontBaker@attglobal.net.

Craig Hackenmiller is a Software Developer in the United States. He has 10 years of experience in AS/400 programming. He has worked at Lakeview Technology for over three years. His area of expertise is AS/400 DB2 high availability software.

Bob Hysick is a product architect in the United States. He has 25 years of experience as a software developer, including 13 years with the AS/400 system. He has worked at Lakeview Technology for four years.

Gary Lagarde is Senior Technologist at Reynolds Metals Company in Richmond, Virginia, USA. He has 18 years of experience in the computer field; 16 of which have been spent concentrating on the S/38 and AS/400 system. He has been with Reynolds Metals since 1989 and has participated in every AS/400 beta program since V1R3. He has a wide range of expertise in using AS/400 hardware, software, communications, and client/server in the business environment, including server consolidation, Web serving, and ERP. He has given presentations at COMMON on single system availability and is chairman of the AS/400 Large Users Group. He is also an adjunct professor at Virginia Commonwealth University where he teaches courses featuring AS/400 technology.

Michael Quinn is a Senior Instructional Designer based in Irvine, California. He has 12 years of experience in the technical training and documentation field. He has worked at Vision Solutions for three years. His areas of expertise include developing Web-based knowledge management systems for customers, business partners, and employees, and designing interactive courseware on selling, using and implementing software products. He has made presentations on the benefits of developing online learning initiatives to conferences, universities, and Fortune 500 companies.

Frédéric Pédussel is a Frenchman working as a Research Scientist in Irvine, California. He has ten years of experience in the computer technology field. He has worked for Vision Solutions for three years. His areas of expertise include clustering and TCP/IP communications on both the AS/400 and Unix platforms. Frédéric has co-written IBM Redbooks on multi-threading.

Peggy Swanson is a technical writer for the AS/400 system in User Technologies in Rochester, Minnesota. She has 12 years of experience in writing about the AS/400 system. She has written many publications and online help for various AS/400 functions. Some of these functions include CL commands, DDM, cryptography, work management, system operations, and

clustering. She has also written several marketing brochures describing the new features in each AS/400 release.

Ilse Valdmanis is a senior product developer at DataMirror Corporation in Toronto, Canada. She developed and is responsible for the AS/400 portion of the DataMirror iCluster product. Before joining DataMirror, Ilse was a developer in the IBM Canada laboratory, working on the ILE C/400 compiler and the XL Fortran compiler for AIX. Her areas of expertise are compilers, symbolic debuggers, and AS/400 High Availability.

Thanks to the following people in the IBM Rochester team for their invaluable contributions to this project:

Lou Antoniolli
Tim Block
Amit Dave
Jenny Dervin
Clint Laschkewitsch
Vicki Morey
Ron Peterson
Mike Snyder
Chuck Stupca
Kiswanto Thayib
Laurie Williams

Comments welcome

Your comments are important to us!

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

- Fax the evaluation form found in “IBM Redbooks review” on page 171 to the fax number shown on the form.
- Use the online evaluation form found at http://www.redbooks.ibm.com/
- Send your comments in an Internet note to redbook@us.ibm.com

Part 1. Clusters and the IBM AS/400 system

This part covers general clustering information as it applies to high availability. It also discusses AS/400 cluster implementation and the new IBM Enterprise Systems Group initiative ClusterProven. Plus, it reviews the planning considerations required to implement an AS/400 cluster.

Chapter 1. Introduction

This IBM redbook is for customers who are interested in implementing very highly available solutions. This redbook guides you through high availability, as it applies to an AS/400 cluster solution, and explains what the word *cluster* means. It provides general information on clusters and the new AS/400 implementation of cluster support. There is some information on planning for AS/400 clusters. The final chapters of this redbook review the three Cluster Management tools available from the IBM AS/400 High Availability Business Partners.

1.1 Cluster

For many years, various computer manufacturers have praised the virtues of their cluster solution. Most of these clusters were brought about to solve the limited horizontal growth in distributed systems. They maintained the premise that a number of systems closely coupled together could provide the capacity required for a growing business. The limitations of horizontal growth are now less of a problem as systems begin to demonstrate enormous scalability.

While horizontal growth is still a reason to attempt to cluster processors, there is a new imperative, high availability. The new reason to cluster is to provide businesses with resilient processes. Behind these business processes are the applications that must have the ability to recover and restart. Businesses of all sizes are finally starting to understand the criticality of their systems.

In recent months, there have been many cases of companies losing millions of dollars through Web site crashes. In the past, a business would still have lost large amounts of money. However, such losses were not as visible and measuring processes were not mature enough to place an accurate value on the disaster.

As you read this redbook, do not think that this is the only way to produce a continuously available solution. There is a myriad of possible permutations for customers to provide highly available solutions for their businesses.

1.2 High availability

This section defines availability as we refer to it throughout this redbook. If you discover that system availability is critical to the success and growth of your business, you need to familiarize yourself with some basic availability terms. Some of these terms include:

Outage A period when the system is not available to users. During a scheduled outage, you deliberately make your system unavailable to users. You might use a scheduled outage to run batch work, save your system, or apply program temporary fixes (PTFs). An unscheduled outage is usually caused by a failure of some type.

High availability
The system has no unscheduled outages.

Continuous operations
The system has no scheduled outages.

Continuous availability
The system has no scheduled or unscheduled outages.

The AS/400 system can achieve all of these levels of availability. There are alternative definitions of availability levels. The level of availability that you choose simply depends on whether you want to define availability from an I/T or business perspective.

This is not the only method of describing high availability, with an I/T department focus. There are other methods that view availability from the business transaction perspective.

1.3 The chapters in overview

We preview the main topics in this redbook to allow you focus on those that are of particular interest to you.

1.3.1 Downtime issues

Chapter 2, “Downtime issues” on page 9, looks at the various types of downtime experienced. In the past, most disaster recovery focused on unscheduled downtime. This type of disaster includes a fire, storm, flood, and plane crash. The normal solution to this was that the failure occurred, and the business stopped and then moved to a remote recovery site. The business interruption could be measured in many hours or even days. This disaster scenario is well known and well documented.

There are also many unscheduled downtimes that are not a disaster. These can vary from simple acts like someone inadvertently pushing the Emergency Power Off (EPO) button in the machine room, to deliberate acts of sabotage. Both of these actions may crash applications and possibly the systems. Other examples of unscheduled downtimes are:

- An ASP overflowing (see B.3, “Auxiliary storage pools” on page 137)
- An application error (see 4.1.3, “Application and data resilience” on page 40)
- Power or environmental system loss (see 6.5, “Hardware” on page 71)
- Hardware or Network failure (see 6.5, “Hardware” on page 71)

The emerging requirement in businesses today is protection from scheduled downtimes. Examples of these downtimes are:

- Hardware upgrades
- Software upgrades
- Fix applies

Scheduled downtimes are more problematic than the remote chance of a disaster. In this new era of e-commerce, it is more important that systems are available to the thousands of unknown and unforgiving Internet users. Even short periods of server unavailability give these people the excuse to point and click elsewhere.

1.3.2 Availability technology

Chapter 3, “Availability technology” on page 19, discusses the description of availability and the options provided by the AS/400 system. As one of the best platforms for single system availability, the AS/400 system has a great story to tell.

There is guidance on the terminology of high availability or continuous availability. Which do you want? How much is your system’s uptime worth to your business?

The AS/400 system has had some great recovery features for many years. Some date back to the System/38. Most were forgotten by programmers early in the AS/400 system’s life because of the performance overhead of running them. Programmers were scared that applications that used these functions would not perform, or the cost of a system with enough resources to enable acceptable performance would be too high. See B.1, “Journaling” on page 135.

This paradigm has changed. The cost of scheduled downtime is so great that businesses demand zero downtime, and are prepared to pay the premium price of large systems with more resources. Application providers can now implement recovery features with little worry about resource cost.

The appendices offer information on the existing AS/400 availability technology. This includes journal support, commitment control, and hardware availability support.

1.3.3 AS/400 clusters explained

Chapter 4, “AS/400 clusters explained” on page 37, reviews the new support available on the AS/400 system with OS/400 Version 4 Release 4. It describes our concept of a cluster. The underlying cluster technology is covered. This technology controls and maintains the cluster. There is also a description of the relationship between the application, data replication middleware, and the cluster APIs.

Different failure scenarios are described, as well. In these descriptions, we tell you how the Cluster Resource Services react and coordinate with the replication middleware and your application.

1.3.4 ClusterProven

Chapter 5, “ClusterProven” on page 51, introduces ClusterProven, a new IBM brand initiative. It gives application providers the ability to label their product to say it conforms to a certain level of resilience on a particular product.

This branding enables customers to select application software appropriate to their availability needs. We cover the ClusterProven process so that customers are aware of the level of resilience their ISVs have attained and how this level was achieved.

1.3.5 Planning for AS/400 clusters

Chapter 6, “Planning for AS/400 clusters” on page 55, provides the information you need to start planning for AS/400 clusters. It gives information on the areas that need focus and possible changes to make the business highly available.

Planning for clusters depends on from where your I/T infrastructure is coming. If you have already implemented dual systems and have an HABP solution, your tasks will be greatly reduced. Those who are planning their cluster from scratch will have a more exhaustive planning process to go through. You need to review the various aspects of systems management. Your hardware planning not only covers the systems involved. You also need to extend out of the internal network and beyond to suppliers and customers.

1.4 High Availability Business Partners

The AS/400 High Availability Business Partners (HABP) have a great selection of products to help application providers and customers with their availability needs. We review the Cluster Management Solutions provided by DataMirror, Lakeview Technology, and Vision Solutions.

1.4.1 DataMirror iCluster

DataMirror's iCluster product provides a set of easy-to-use interfaces to set up and manage an IBM AS/400 cluster for high availability. iCluster also provides replication support for resilient data and applications using the DataMirror HA Suite, a real-time replication utility for AS/400 data and objects. iCluster is presented in Chapter 7, "DataMirror iCluster" on page 77.

1.4.2 Lakeview Technology availability solutions

Lakeview Technology provides availability management for the IBM AS/400 with its MIMIX solutions suite. MIMIX ClusterServer and MIMIX FastPath components, combined with the IBM AS/400 system, provide a clustering environment for data and application resiliency. Learn more about them in Chapter 8, "Lakeview Technology availability solutions" on page 99.

1.4.3 Vision Solutions OMS/400 Cluster Manager

With OMS/400 Cluster Manager, recovery from unplanned failovers or planned switchovers can now be both seamless and rapid. Building upon the Vision Suite of middleware HA software products, OMS/400 Cluster Manager extends customer's abilities to create highly available and resilient data, application, and user environments.

1.5 Appendices

In the appendices, we cover the availability functions that have been provided as standard with AS/400 hardware and OS/400 system software. Many of these functions have been around for years. Most IBMers and Business Partners are aware of them, but customers and developers may have never implemented them. We provide this detail as supporting information.

We also cover the functions available to perform problem determination on a cluster. Most of this problem determination is performed using OS/400 commands, which are included in an appendix.

1.5.1 AS/400 high availability functions

Appendix A, “AS/400 cluster resources” on page 133, discusses the basic AS/400 hardware and OS/400 software availability options. Hardware and software recovery have been available with the AS/400 system for some years. Some customers are well aware of these functions and use them to improve the resilience of their business systems. To others, these may be new concepts.

1.5.2 Problem determination

Appendix B, “AS/400 high availability functions” on page 135, covers the basics of cluster problem determination available in OS/400. It includes the commands that can be used to view the cluster information, plus scenarios that could arise and need more intervention than others.

Chapter 2. Downtime issues

In discussing downtime issues, we may generalize in comparing systems, not specifically the AS/400 system. The potential business improvements to be gained from clusters are significant. Some of these benefits have already been seen by many customers. The AS/400 system has realized many of these benefits with its single system availability model. Now, with support for clusters provided in OS/400 V4R4, a business can increase its system uptime even further.

2.1 Basic backup models

To provide some background on the existing availability options, we look at some basic backup models. Downtime planning has filtered into larger companies in recent years in an effort to improve availability. This planning was primarily aimed at disaster recovery models that focused on hot or cold sites or redundant computer centers. Some companies also used AS/400 symmetric multi-processing techniques to give horizontal growth and workload balancing.

Figure 1 on page 10 illustrates the hot and cold backup scenarios. A cold backup has no applications running on the backup system. This cold backup system is loaded with the most recent backup tapes from the production system in the event of a failure or disaster. The I/T department then manually re-synchronizes the backup system with the business. When the production system becomes available, the most recent save tapes from the backup site are used to bring the production system close to the state of the backup, and then another manual re-synchronization occurs.

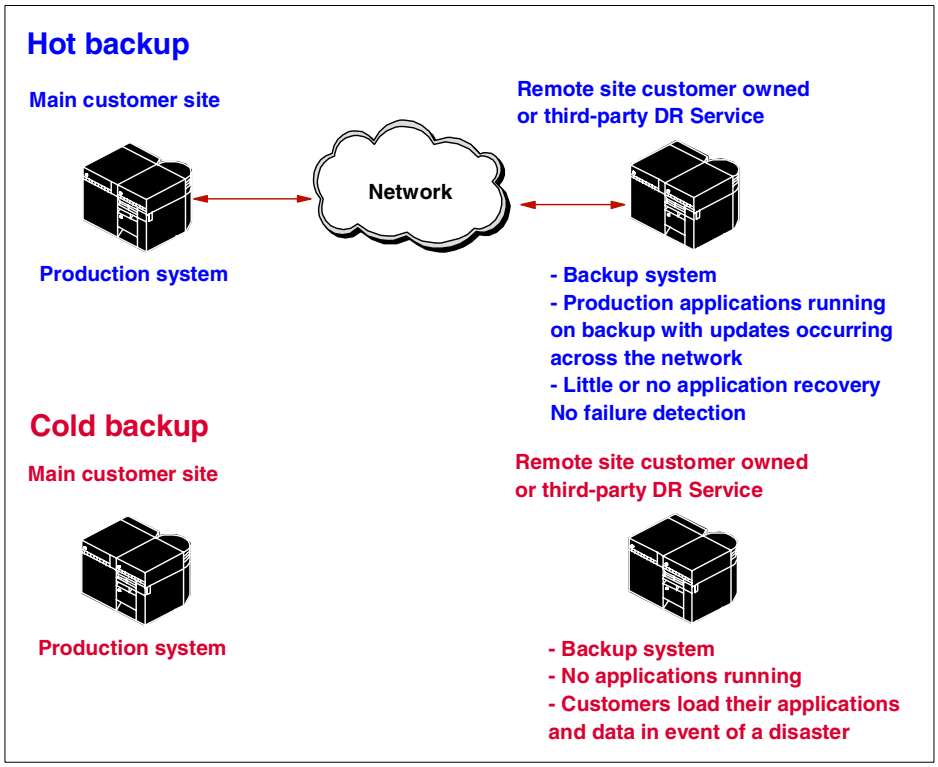


Figure 1. Hot and cold backups

The hot backup scenario is similar to a cold backup if no AS/400 advanced functions have been implemented. The only difference is that the databases are maintained in sync with library or file transfers across the communications link outside normal business hours. There is still a significant amount of manual update required after a failure.

In a more advanced hot backup solution, the backup database is maintained in tight synchronization through the communications links and the backup is immediately ready to assume the role of the primary. Synchronicity between the applications and database is paramount and is achieved though a combination of journalling, commitment control, and application design. The downtime is greatly reduced, and with good application design only the incomplete transactions need to be manually re-keyed. When the production system is available, the journalling process can be reversed to re-synchronize the two machines. However, at some stage, the production and backup must be switched. This will involve another outage to achieve the switch.

2.1.1 Factors influencing availability

Many larger companies have some level of disaster backup. The hardware and software costs, specifically processors, memory, disk, and communication bandwidth, are considered very expensive. This has tended to prohibit small and medium companies from entering into the disaster recovery market. Instead, companies place their emphasis on programming applications with an eye on limiting hardware and software costs. This is because manpower seems relatively inexpensive. Many of the applications being used today were built using inexpensive labor, but expensive hardware resources. Today, this paradigm has reversed, and manpower is the limiting factor.

On the AS/400 system, there are techniques to improve data and application recovery, namely journalling and commitment control. Typically, these techniques were not used because of the cost in application and system performance. Companies viewed recovery in many hours or even days as acceptable or manageable. Manual systems were still available to maintain the business flow. Today, this is not be case. Business processes are too complicated and have too many external links to run in manual mode.

Processes and methodology for application and infrastructure development were learned when these technologies were limiting factors. With the large up-front investment, many companies elected to improve single system availability rather than buy into redundant systems. The AS/400 system demonstrated the essence of this over the years by supplying the highest single system availability of all servers. Figure 2 on page 12 illustrates the single system availability of the AS/400 system against other platforms. The best availability is achieved from the IBM S/390 Parallel Sysplex model. Sysplex uses a redundant processor, I/O processors, and application resiliency. More details about IBM S/390 Parallel Sysplex can be found at its home page: `http://www.s390.ibm.com/pso/`

The AS/400 system has the highest availability of the single system models. A single system consists of a single processor, main storage, disk, and I/O processors. With cluster technology, the AS/400 system moves forward into another availability level.

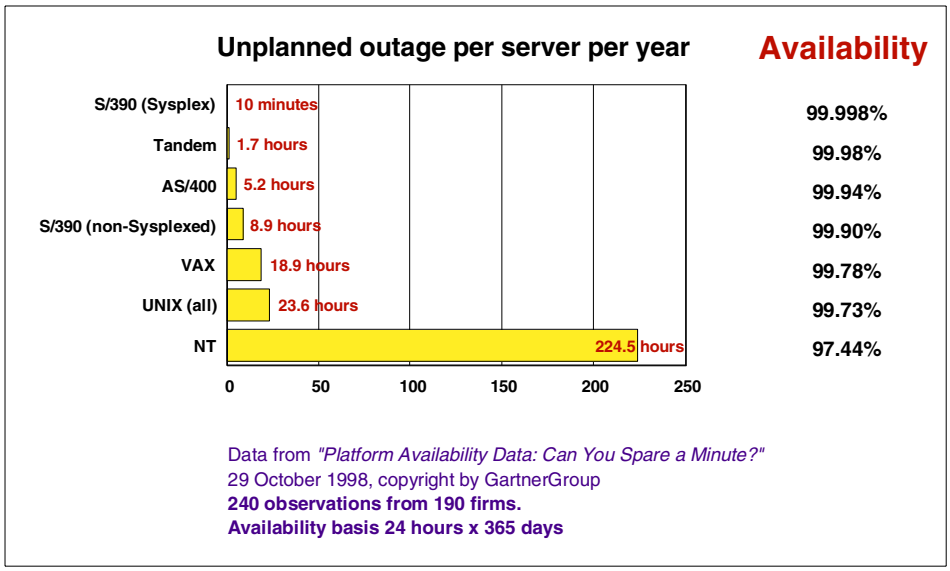


Figure 2. Single system availability

Today the speed of business has greatly increased. Application systems have become more integrated and complex. The cost pendulum has swung the other way, hardware and software costs are diminishing, but personnel costs have skyrocketed. What was once a constraining factor is now an enabler. However, new methodology and processes are needed to take advantage of these empowering technologies. Clusters are one of those empowering technologies.

As more business is conducted via non-traditional methods, such as the Internet and client-server applications, high availability becomes more important. Typically implemented in only the larger companies with huge transactional processing requirements, high availability is moving to mid-tier and small companies at an increasing rate. This has forced many companies to re-evaluate platforms and recovery techniques in order to meet increasing customer demands and satisfaction.

As application integration and complexity increase within companies, it is becoming commonplace to reach outside of the typical business boundaries. A vendor to one company is a customer to another who is a vendor to another and so on. This streamlining increases the critical path for services along the way. Outages at any point can mean lost business. Because of the financial impact, service level agreements are increasingly needed to guarantee delivery of goods and services to all involved.

Figure 3 shows a typical business-to-business environment. Data transfer between businesses is high and critical. Failure of a single processor within the relationship can have a severe impact.

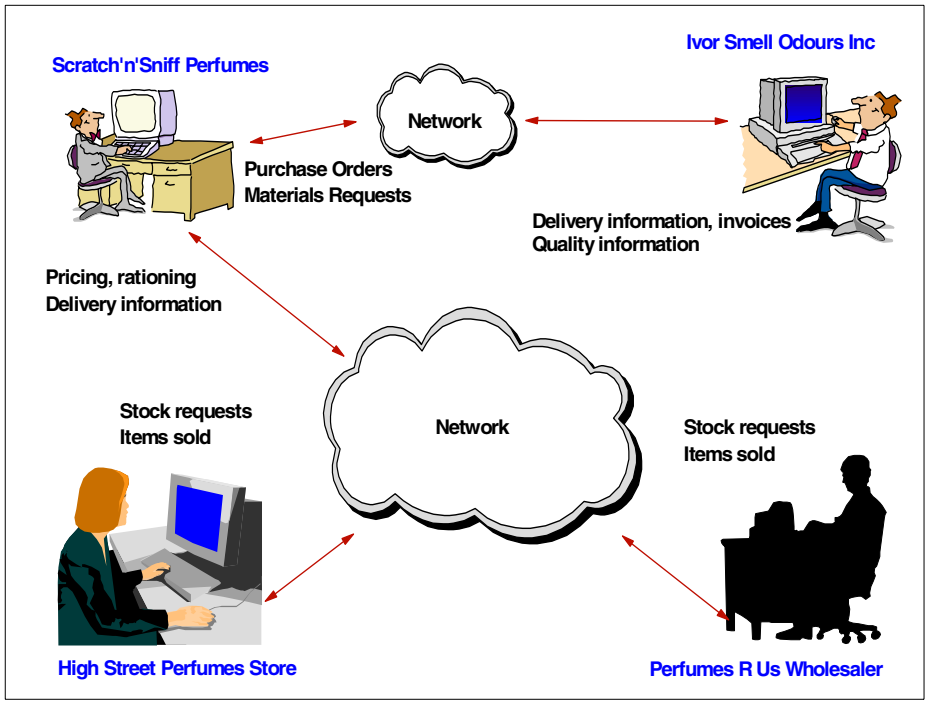


Figure 3. Electronic business-to-business relationship

2.1.2 Financial impact of an outage

The turn of the availability crank becomes even more important as companies drive their business to the Internet. Downtime at any point in this environment can mean the difference between a positive balance sheet or a going out-of-business sign. Figure 4 on page 14 shows the financial impact of outages across different business organizations.

The key to this balancing act of controlling costs on one hand and reaching new customers on the other is to provide existing customers with increasing service level satisfaction. To do this, applications not only have to be available, but must become highly available. The difference is important to many companies' survival as they extend their critical information beyond the traditional walls via Internet or enterprise resource planning (ERP) applications. This new model is forcing businesses to re-look at how to provide this availability. Increasingly companies are looking at linking their

strategic computer resources to each other in an effort to minimize the impact of application outages.

The “cluster” technology helps bring availability and performance to the traditional data center model. Formalized service levels within corporations and between businesses are becoming standard practice. Availability, performance, and reliability are chief among the metrics measured.

Continuous availability on a technical level tends to reduce business planning costs. This would include application availability, batch versus backup contention, decision support systems, use of assets for something other than just disaster recovery and point-in-time reconstruction. These issues lead to cost avoidance, but are very much tangible costs that must be accounted for in the equation.

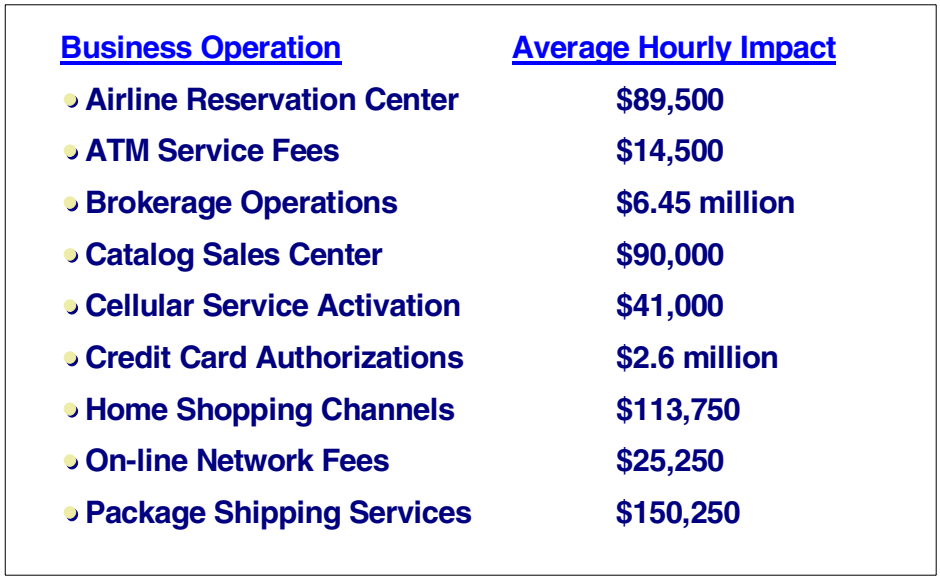


Figure 4. Financial impact of an outage

2.2 Activities that cause downtime

As mentioned earlier, application-level availability is the driving force behind new generation applications. In delivering this availability, the focus should consider both planned and unplanned outages as key costs in evaluating and meeting business requirements.

The following list identifies the more common scheduled and unscheduled outages and an estimate of the percentage of total downtime these represent. The percentage values, 80% and 20%, represent a very rough distribution of outage type over time. For example, 80% of the outage hours are scheduled events.

- Scheduled outages: Controlled actions or activities (80%)
  - New operating system or application software release installs
  - System hardware upgrades, additions, removals, and maintenance
  - System backups or saves
  - Fix installation
  - Recovering storage from deleted files
  - Reorganizing file structures
  - System IPLs
  - Site maintenance
- Unscheduled outages: Uncontrolled or unforeseen events (20%)
  - Power outages
  - System hardware or software errors
  - Site disasters
  - Application malfunction

In cross-business environments with highly integrated applications, the total downtime is added across several architectural levels:

- Server platforms (database server, application server, and client)
- Operating systems (on various systems throughout the network)
- Application software
- Middleware
- Network (local and remote)

These elements of applications availability overall are difficult to maintain within a single corporation or business unit. However, as the application is externalized (e-commerce), the downtime grows exponentially. Clustering, although complex, can improve and enhance application availability.

2.3 Example of business impact analysis

This section presents an example of a business impact summary for a worldwide manufacturing company that has implemented an ERP package across all locations. The name of the company has been removed, but the numbers are real.

This summary takes into account three levels of recovery implementation, the business impact to one business unit and the entire company, and the costs (in dollars) associated with implementing only the physical infrastructure in 1999. The costs of manual processes, employee inefficiencies, lost sales, lost market value of the company, restart of application and synchronization with the manual system is not included. A description of each level follows:

- **Level 1**: No infrastructure is put in place other than agreements to have hardware shipped in a timely manner if a disaster event is declared. This highlights the amount of time to recover all server platforms, operating systems, data, and system re-test and re-sync for the ERP application only. The business impact for Level 1 is three weeks. A loss for this period represents approximately 2.5% of the gross revenue of the company.
- **Level 2**: Minimum infrastructure, major transaction loss, and data restored after the failure. The business impact for Level 2 over 10 days represents approximately 1.7% of gross revenue.
- **Level 3**: Continuous availability, no transaction loss, and little impact to the business. The impact of 30 minutes or less of application outage was rated at minimal. However, the costs to achieve this minimized business risk were not cost prohibitive when compared to Level 2.

Other costs that should be quantified include the following business processes and procedures:

- Possible data integrity problems
- Productivity loss because of inconsistent access to data and applications
- Business loss resulting from lost sales
- Consequences of system outage to the company image

Other server losses include:

- External business consequential loss
- Market value losses

Table 1 shows the three options for disaster recovery. It also shows an estimate of this business's recovery time. This time is translated into the actual dollar cost to the business. This cost is then compared to the approximate cost of implementing the recovery option for that level.

Table 1. ERP disaster recovery options by level

Option	Description	Recovery time	Single business unit lost revenue *	Business impact lost revenue **	Disaster recovery implementation 1999
Level 1	React at the time of disaster	3 weeks	$1,000,000 +	150 million	None
Level 2	Minimum infrastructure build today; data restored after disaster	10 days	$750,000	100 million	$775,000
Level 3	Continuous availability	30 minutes or less	minimum	minimum	$150,000 more than level 2

* Source: Single business unit

** Source: Cumulative Financial Impacts and Exposures. These numbers represent the losses for all global business units.

While the cost to the business of moving to Level 2 appears very high, look at the potential losses of 100 million. Less than one percent of the potential loss is a small price to pay for the business if it wants to survive. To take this a step further, providing a continuously available solution is still a relatively small cost. Customers should try to reach Level 3, rather than staying on Level 2.

The implementation time for Level 2 and Level 3 is different. If a business is serious about availability, Level 2 should be viewed as a tactical business solution to provide protection until the more complex options at Level 3 are implemented.

Some examples of data loss and business impact are:

- 43% of companies experiencing disasters never re-open, and 29% close within two years (McGladrey and Pullen)
- It is estimated that one out of 500 data centers will have a severe disaster each year (McGladrey and Pullen)

A company that experiences a computer outage lasting more than 10 days will never really fully recover financially. 50% will be out of business within five years. For more information, refer to *Disaster Recovery Planning: Managing Risks and Catastrophe in Information Systems* by Jon Toigo.

Table 2 shows the advantages and disadvantages of each level. It is very obvious that the disadvantages and risk are far out weighed by the advantages.

Table 2. Business impact advantages by level

Advantages and cost to implement	Disadvantages and cost to business	Risk protection (insurance coverage)
Level 1 No advantages No investment	Lost revenues may destroy company	None
Level 2 Significant advantages, but with some downtime large investment	Substantial loss of revenue. Annual maintenance and support costs	Low
Level 3 High customer satisfaction Automated processes No downtime Higher cost that Level 2	Highest cost option, but not significantly higher than Level 2.	Extremely high

In conclusion, while the initial outlay for high available systems is viewed as enormous, the resulting savings are far larger.

Chapter 3. Availability technology

This chapter describes the general concepts of availability technology. It also provides an overview of AS/400 cluster support and existing AS/400 availability technologies. There are various combinations of the following terms. They are not mutually exclusive. For example, switch disk can be used to supplement a multiple system availability.

3.1 Single system availability

Most systems in the market place today offer single system availability. These single systems have one processor complex managing memory, DASD and individual I/O adapters/processors. In a single system environment, the processor is the critical object. DASD, memory, and adapters can all have redundancy and failover capability. However, the processor has no failover option.

In this environment, the reliability of the hardware components (chips, circuit boards), the microcode, and the operating system are of the greatest importance. The AS/400 system has maintained dominance in the single system marketplace for some years. This is probably why many customers have been slow in implementing dual systems to get a higher level of availability than they already have with their single AS/400 system. Figure 5 on page 20 shows the components of single system availability in an AS/400 system.

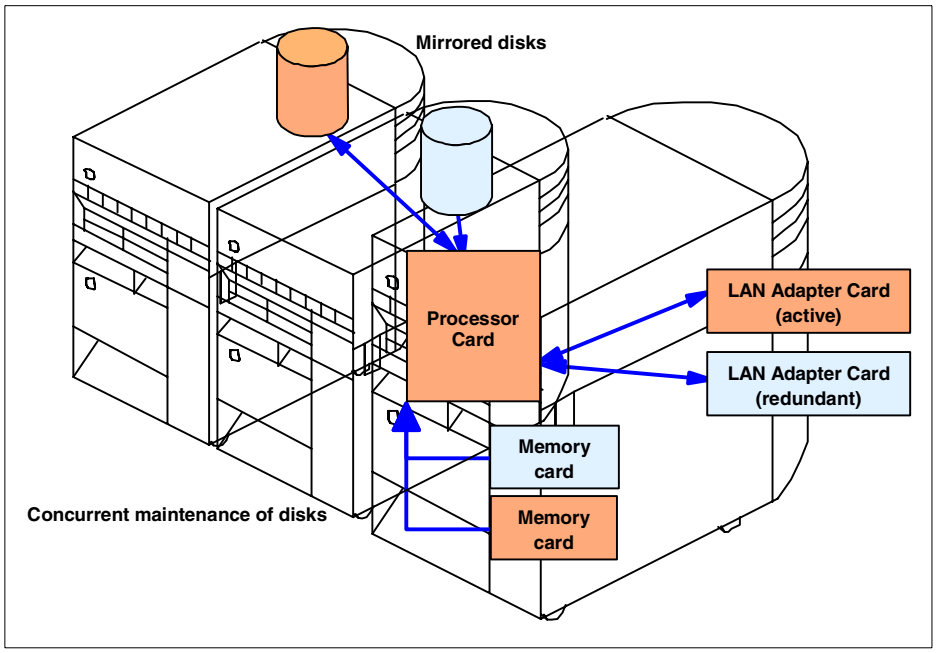


Figure 5. Single system availability

Figure 5 shows the hardware components that make the AS/400 system highly available. There are other OS/400 features that can also improve the application availability, these are journaling and commitment control. Even with a single system, they can improve your recovery.

Let's say a system has failed with a processor malfunction. Once the offending part has been replaced the system can be IPLed. The application transactions can be rolled back to the last complete transaction. Then the application can be restarted and the incomplete transactions can be re-keyed. With this failure, the downtime is the time to replace the part, re-IPL the system, and re-key the incomplete transactions.

To obtain this additional level of protection, you must re-design your application to take advantage of journaling and commitment control.

3.2 The AS/400 system and 99.9+% availability

The AS/400 system has an outstanding single system availability of around 99.9+% as shown in Figure 6. It shows the various components that have

been built into the AS/400 system model over time. These elements are all heading towards the goal of 100% availability and continuous availability.

Excellent reliability is one of the main reasons customers select an AS/400 system to run their mission-critical applications. IBM can to deliver a very reliable system because the IBM development team designs, creates, builds, tests, and services the AS/400 system as one entity. Although reliability is designed throughout the system, it is not sufficient for tomorrow's business-critical needs. One of those critical needs is continuous availability.

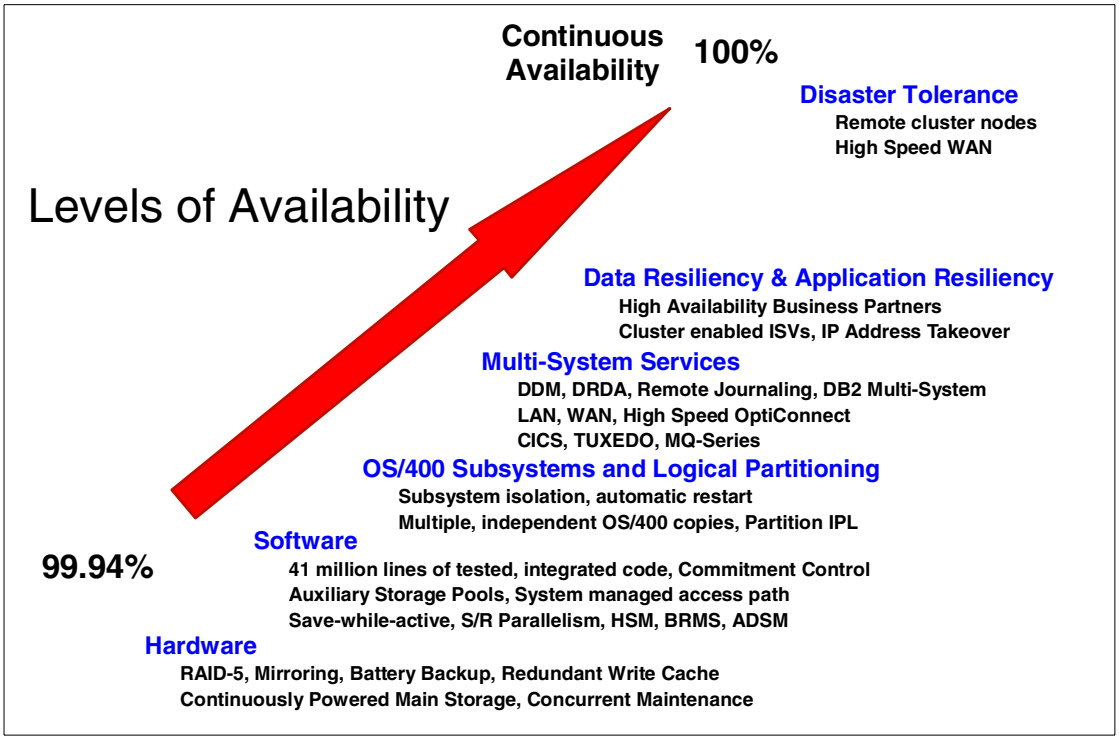


Figure 6. AS/400 levels of high availability

In 1998, the AS/400 system 24-hour, 365-day availability was 99.94%, with only 5.2 hours of unplanned outages per year. However, these measurements do not include certain classes of failures and scheduled outages. For example, scheduled outages are typically the biggest reason for system unavailability. A scheduled outage usually involves either system maintenance, installing a new release, or handling facility repairs of some kind. The limits of what can be achieved in a single system environment are being reached between 99.9% and 99.99% availability. Achieving higher availability (99.999% and above) is only

possible using a multiple system approach. With the AS/400 system's implementation of clustering, it will tend to move closer to the 100% target.

3.3 Standby secondary

The standby secondary mode of backup may be a cold backup or a warm backup. We already mentioned the cold backup arrangement in 2.1, "Basic backup models" on page 9. The backup machine is idle while the primary machine is active. This is the most common of all disaster recovery scenarios. Typically, the standby secondary is a system complex at another site. This secondary system may be configured to match the largest primary system it has to support. I/O processors, processor feature, and disk may match the largest configuration, or the customer may accept a reduction in configuration and, therefore, service. With this reduction, the customer would then choose to support only critical applications that could fit on the reduced configuration.

The advantage of this setup is that it is relatively inexpensive and can support many different systems. However, recovery is nearly as long as a scratch installation of the primary site. Protection from an unplanned outage is this site's main usage, and usually many sites will be supported by one cold site. The primary site is recreated by restoring the primary program and data files from the primary's most recent backup. Following this restore, any work since the last backup must be re-keyed and the business processes must be re-synchronized.

A warm backup is similar, but OS/400 is pre-loaded. The customer must then load their customization file (user profiles, security data, etc.) and then load their applications and data. Again this is very time consuming and does not really even get close to high availability.

In some scenarios where the backups are dedicated, it may be possible to off load some work from the primary site during workload peaks. This would have to be planned some weeks before the event and the time to restore the applications on the backup could be considerable. Another non-trivial task would be re-synchronizing the applications and data back to the production system after the workload peak had passed.

3.4 Active secondary

This configuration is similar to the hot backup described in Figure 1 on page 10. Productive work may be processed on a backup machine while the primary system is active. This is equivalent to a hot backup site. The

productive work running on the backup maybe considered lower priority than the primary unless there is sufficient capacity to run both primary and secondary workloads with degrading production load. If the backup workload is considered a lower priority, it is stopped or degraded until the crisis has pasted.

This is similar to the separate server approach described in 3.8, “Separate server” on page 27. It is more expensive to run than the cold site, but not as expensive as separate servers. In this scenario, the active secondary site may offer support to many primary sites. This configuration is fine unless more than one primary fails.

Recovery can be a long task similar to the standby secondary. It depends on the configuration. Recovery action depends on the currency of the backup database. Any business transactions not completed or included on the backup must be re-keyed to synchronize with the business processes.

In this configuration, it is possible to use either switched disk technology or replication technology to maintain access in the event of a processor failure.

3.5 Replication technology

Replication technology is the database’s ability to make synchronized copies of data and objects from one system to another. On the AS/400 system, this is achieved with journaling and commitment control. These features of OS/400 have been around for many years. Customers who have understood the need for highly available systems have implemented these features on their systems and in their applications. Journaling is the corner stone of the high availability middleware provided by the IBM HABPs.

Journaling allows changes in the database to be recorded and stored. These changes can be transferred to the backup system by a communications method or tape. See B.1, “Journaling” on page 135, for more information on journaling.

Commitment control is implemented at an application level. Basically, it provides transaction boundary points. When a point is reached, the transaction is committed to the database. Any incomplete transactions can be rolled back to the last complete transaction in the event of a failure. This would still involve incomplete transactions to be re-keyed on the backup machine, but it considerably adds to the recoverability of the application. Few application providers have implemented commitment control in their applications. This position will have to change if the application providers

want to deliver continuously available applications. See 2.1.1, “Factors influencing availability” on page 11, for additional information.

3.6 Switched disk

Disk drives can be switched from one system to another (Figure 7). Local access to the data is only available from the owning system.

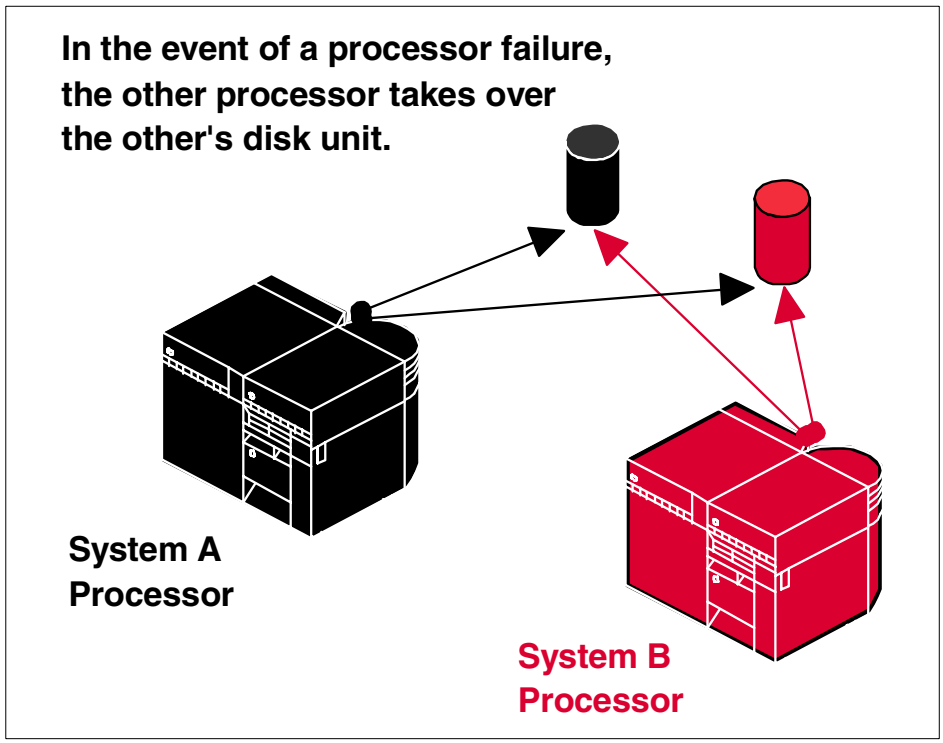


Figure 7. Switched disk cluster

Some operating systems have implemented switched disk technology to improve their reliability from the single system model. Microsoft Cluster Services implements switched disk technology. With switched disk technology, if the processor fails, another processor complex can take over the disk and their associated database. This model is less expensive than dual systems because there is no duplication of DASD and adapters. However, it does not make a significant difference to the single system model unless the hardware is unreliable.

Let's look at what happens to the business transactions in the switched disk model. After a switch, the applications fail and need to be restarted. If a switched system has on-line transaction processing (OLTP) applications and a failure occurs, many users will have partially completed transactions. The database needs to have the ability to roll back the incomplete transactions and restart the application to maintain database integrity. Following roll back, the incomplete transactions need to be re-keyed.

For non-OLTP based applications, whether standalone or server-based, there may be less of a problem, depending when the user last saved their work. For example, a typical word processor or spreadsheet user has less of an impact on the database because their transactions typically only affect their open files. OLTP applications typically have a more pervasive effect and require the database to have additional capabilities.

The AS/400 system has not yet implemented switched disk. However, this function may be made available in a future release.

3.7 Shared disk

Disk drives are attached to multiple systems simultaneously. Local access is available from all systems sharing the disk. Figure 8 on page 26 shows a diagram of a shared disk cluster.

The first attempts at shared disk technology allowed every server to access every disk, which required expensive cabling and switches, plus specialized operating system functions and specialized applications.

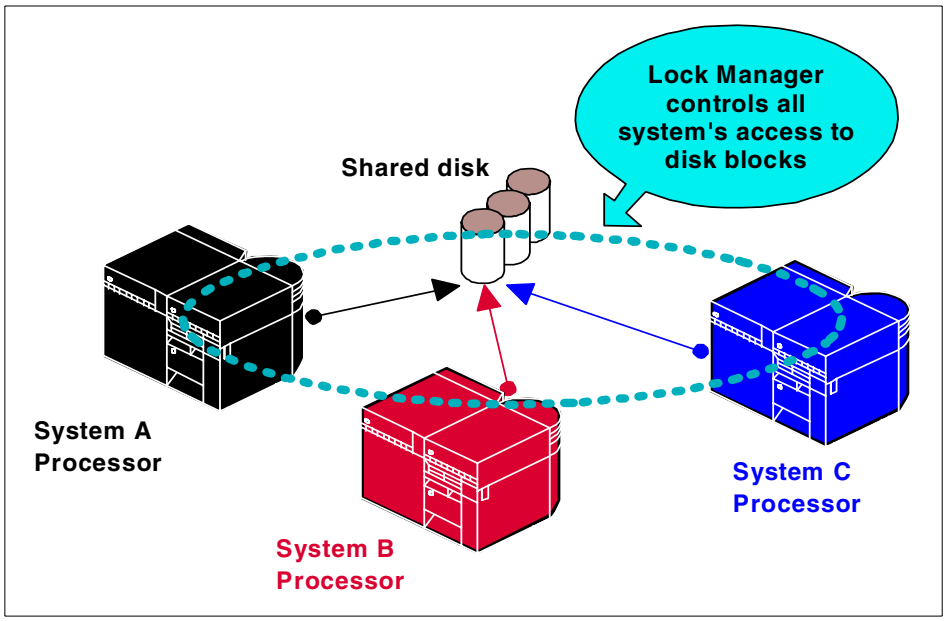


Figure 8. Shared disk

Today's standards, such as small computer systems interface (SCSI), have eliminated the need for expensive cabling and switches. However, shared disk still requires you to buy specially modified applications.

In Figure 8, systems A, B, and C are writing to and reading from the same disk. To manage this, the three systems have some form of DASD block management code. This code controls who has current access to a block of storage. In our example, system A currently has a lock on block 123. Now system B requests block 123 on the shared DASD. The lock manager asks system A to give up block 123. When system A gives up the block, the lock manager changes the ownership of block 123 to system B. System B now has control of the block and can write all over it. At any time, systems C or A can request the block back or could be competing for other blocks. The lock manager can reside on any or all of the three systems.

IBM S/390 Parallel Sysplex successfully uses shared disk technology. This function has developed over time, with a significant investment in the system and applications to manage this function.

The AS/400 system does not implement true shared disk functions. Single level storage and symmetric multi-processing (SMP) have some analogies to shared disk, where multiple applications are running on multiple processors

and in one storage pool. The user does not have to worry about where the data resides. The system takes care of the data management, spreading the data across all the disks. OS/400 also takes care of object lock and task management. These are more examples of the underlying AS/400 functions that provide such high single system availability, but they have largely been taken for granted for many years.

3.8 Separate server

In a separate server cluster or dual system environment, data and objects are replicated from one system to another (Figure 9).

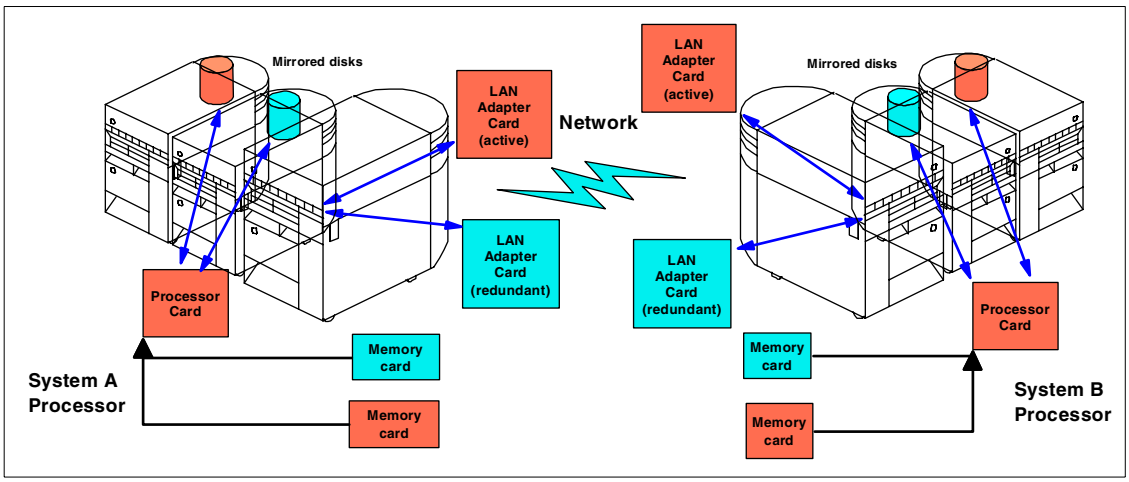


Figure 9. Separate server cluster

Before cluster support was made available in OS/400 V4R4, the AS/400 system provided multi-system services such as distributed data management (DDM), Distributed Relational Database Architecture (DRDA), remote journaling, and DB2 multi-system in terms of software. Local area networks (LANS), wide area networks (WANs), and high-speed OptiConnect were the network connection fabric. This configuration built the multi-system services layer on top of the single system availability provided by the AS/400 system and created a stable model for backup, horizontal growth, and workload balancing.

This could be called a loosely coupled cluster. However, some of the basic cluster functions were not implemented so IBM did not call it a cluster. Now the new cluster support is available with OS/400 V4R4 and is the current AS/400 cluster implementation model.

3.9 The AS/400 system and its availability

The AS/400 system is a highly available system. IBM has tried to balance high-end growth and 24 x 365 usage with continuous availability clusters to improve this availability even further.

Such limits to growth as faster saves and restores, faster normal IPLs and abnormal IPL recovery, improved communications recovery, and avoidance of system crashes and hangs are a high priority to IBM. Scheduled system outages have been reduced by eliminating save-while-active restrictions, providing faster PTF and new releases installations, and faster running commands like the Reclaim Document Library Object (RCLDLO) command. The frequency and duration of scheduled outages has been improved by:

- Increasing the amount of concurrent apply PTFs
- Decreasing the amount of restricted state activities
- Allowing concurrent hardware upgrades and changes

From an unscheduled outage duration and frequency perspective, the AS/400 system provides:

- Efficient database replication
- Transparent switchover in less than five minutes
- Cluster management
- Remote disaster recovery capabilities
- Highly available hardware
- Highly available software
- Fast abnormal IPL times
- Continuously Powered Main (CPM) storage
- Uninterruptable powers supply (UPS) support

3.9.1 AS/400 availability options

A number of availability options existed on the AS/400 system before clusters were introduced in Version 4 Release 4. Those options included:

- **Journal management**: *Journaling* provides additional levels of data integrity and single system recovery. Journaling protects database files by recording changes that occur to those objects. You use a journal to define the files and access paths you want to protect with journal management. This is often referred to as journaling a file or an access path. A journal receiver contains the entries (called journal entries) that the system adds when events occur that are journaled, such as changes to database files, changes to other journaled objects, or security-relevant events.

- **Access path protection**: An *access path* describes the order in which records in a database file are processed. A file can have multiple access paths, if different programs need to see the records in different sequences. If your system ends abnormally when access paths are in use, the system may have to rebuild the access paths before you can use the files again. This is a time-consuming process. To perform an IPL on a large, busy AS/400 system that has ended abnormally can take many hours.

  You can use journal management to keep a record of changes to access paths. This greatly reduces the amount of time it takes the system to perform an IPL after it ends abnormally.
- **Auxiliary storage pools (ASP)**: This is a software definition of a group of disk units on your system. This means that an ASP does not necessarily correspond to the physical arrangement of disks. Conceptually, each ASP on your system is a separate pool of disk units for single-level storage. The system spreads data across the disk units within an ASP. If a disk failure occurs, you need to recover only the data in the ASP that contained the failed unit.
- **Device Parity Protection (RAID-5)**: This is a hardware availability function that protects data from being lost because of a disk unit failure or because of damage to a disk. To protect data, the disk controller or input/output processor (IOP) calculates and saves a parity value for each bit of data. Conceptually, the disk controller or IOP computes the parity value from the data at the same location on each of the other disk units in the device parity set. When a disk failure occurs, the data can be reconstructed by using the parity value and the values of the bits in the same locations on the other disks. The system continues to run while the data is being reconstructed. The overall goal of device parity protection is to provide high availability and to protect data as inexpensively as possible.
- **Mirrored protection**: *Mirrored protection* is a software availability function that protects data from being lost because of failure or because of damage to a disk-related component. Data is protected because the system keeps two copies of data on two separate disk units. When a disk-related component fails, the system may continue to operate without interruption by using the mirrored copy of the data until the failed component is repaired.

While availability options are a complement to a good save strategy, they are not a replacement. Availability options can significantly reduce the time it takes you to recover after a failure. In some cases, you can avoid having to perform a recovery by using availability options. To justify the cost of using availability options, you need to understand:

- The value your system provides
- The cost of a scheduled or unscheduled outage
- Your availability requirements

More information on the above topics can be found in Appendix B, “AS/400 high availability functions” on page 135.

3.9.2 Clusters

With the introduction of *clusters*, the AS/400 system offers a continuous availability solution if your business demands operational systems 24 hours a day, 365 days a year (24 x 365). This solution, called OS/400 Cluster Resource Services, is part of the OS/400 operating system and provides failover and switchover capabilities for your systems that are used as database servers or application servers. If a system outage or a site loss occurs, the functions that are provided on a clustered server system can be switched over to one or more designated backup systems that contain a current copy (replica) of your critical resource. The failover can be automatic if a system failure should happen, or you can control how and when the transfer will take place by manually initiating a switchover.

Figure 10 shows a basic cluster. There are four node systems, A though D. The nodes are connected through a network. Systems A, B, and C are local to each other, and system D is at a remote location. The cluster management tool controls the cluster from anywhere in the network. End users work on servers in the cluster without knowing or caring where their applications are running.

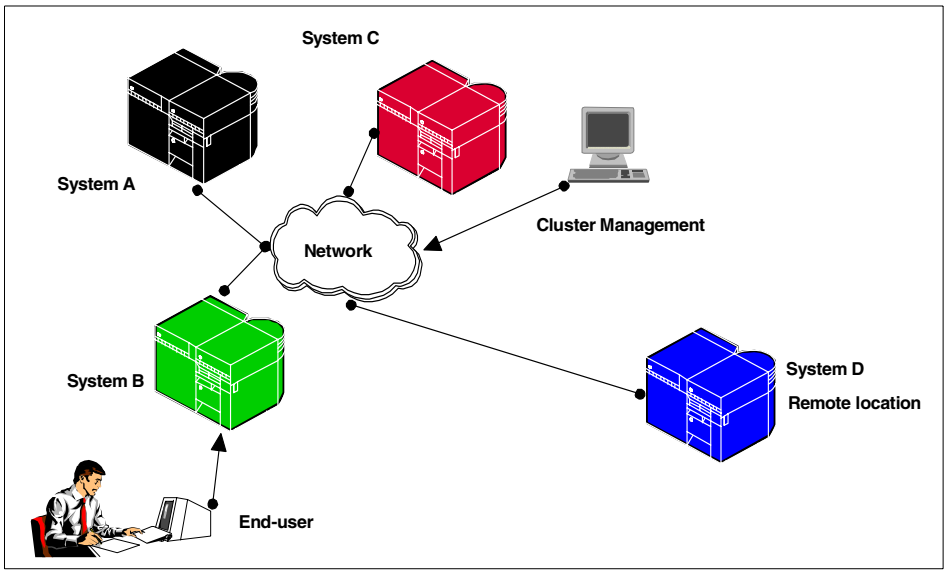


Figure 10. Basic cluster

In the event of a failure, Cluster Resource Services (CRS), which is running on all systems, provides a switchover. This switch causes minimal impact to the end user or applications that are running on a server system. Data requests are automatically rerouted to the new primary system. You can easily maintain multiple data replications of the same data. Clusters contain more than two nodes. You can group a system's resilient data (replicated data) together to allow different systems to act as the backups for each group's resilient data. Multiple backup systems are supported. If a system fails, Cluster Resource Services provides the means to automatically re-introduce or rejoin systems to the cluster, and restore their operational capabilities.

Hardware and software requirements for clusters

Any AS/400 model that can run OS/400 Version 4 Release 4 or later is compatible for implementing clustering. You must have OS/400 Version 4 Release 4 or later installed and Transmission Control Protocol/Internet Protocol (TCP/IP) configured on your AS/400 systems before you can implement clustering. In addition, you can purchase a cluster management package from a High Availability Business Partner (HABP) that will provide the required replication functions and cluster management capabilities.

3.9.3 Logical partitioning

AS/400 logical partitions let you run multiple independent OS/400 instances or partitions (each with its own processors, memory, and disks) in an N-way symmetric multi-processing AS/400e, Model 6xx, Sxx, and 7xx. This also means you can run a cluster environment on a single system image. Up to twelve cluster nodes can exist within one LPAR system. With logical partitioning, you can address multiple system requirements in a single machine to achieve server consolidation, business unit consolidation, and mixed production and test environments.

Logical partitions fall into two categories: primary partitions or secondary partitions. These could be nodes in a cluster. Each logically partitioned system has one primary partition and one or more secondary partitions. All OS/400 V4R4 systems have a primary partition with all resources initially allocated to it. Creating and managing secondary partitions is performed from the primary partition. Movement of processors, memory, and interactive performance between partitions can be achieved with only an IPL of the affected partitions. Movement of IOP resources can be achieved without IPL. Figure 11 shows an example of a cluster created by logical partitioning.

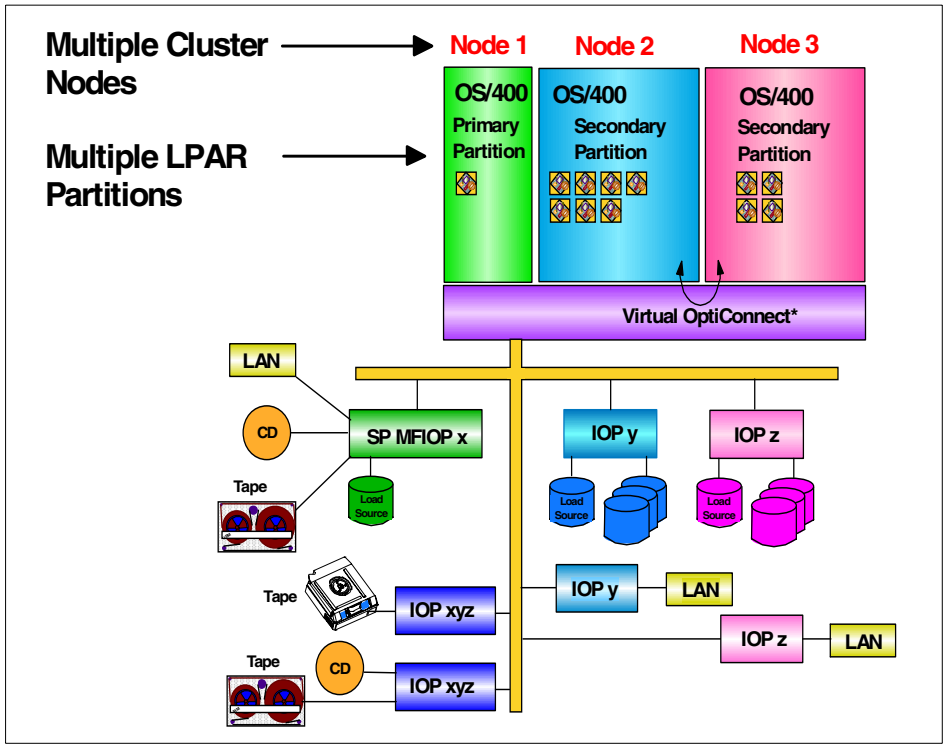


Figure 11. Cluster created by logical partitioning

Each logical partition represents a division of resources in your AS/400e system. Each partition is logical because the division of resources is virtual, not physical. The primary resources in your system are its processors, memory (main storage), I/O buses, and IOPs.

OS/400 is licensed once for the entire system by its normal processor group, regardless of the number of partitions. License management across partitions is not supported in this release. OS/400 V4R4 must be installed on each partition. Previous releases are not supported on a logical partition.

Each logical partition operates as an independent logical system. However, each partition shares a few physical system attributes such as the system serial number, system model, and processor feature code. All other system attributes may vary among partitions. For example, each partition has dedicated hardware such as processors, main storage, and I/O devices.

However, it should be noted that an LPAR solution does not offer a true failover capability for all partitions. If the primary partition fails, all other

partitions also fail. If there are multiple secondary partitions backing each other up, they have the capability to fail over between partitions. These secondary partitions are nodes and are a cluster solution, but not a separate server implementation. LPAR cannot provide the same level of availability as a two or more node cluster solution.

3.10 AS/400 high availability middleware

High availability middleware is the name given to the group of applications that provide the replication and management between AS/400 systems. More recently, they also provide the cluster management middleware.

The High Availability Business Partners (HABPs), shown in Figure 12, provide data resiliency tools. With OS/400 V4R4, they are heading into application resiliency.

You can read more about their solutions at their Web sites. There is a section on the new Cluster Management Utilities in Part 2, “High Availability Business Partners” on page 75, of this redbook.

Figure 12. High Availability Business Partners

You can locate these IBM Business Partners on the Web at the following sites:

- DataMirror: `http://www.datamirror.com`
- LakeView Technology: `http://www.lakeviewtech.com`
- Vision Solutions: `http://www.visionsolutions.com`

Chapter 4. AS/400 clusters explained

Before exploring the underlying AS/400 clustering technology, it is important that you understand AS/400 clustering technology and capabilities.

4.1 What an AS/400 cluster is

Figure 13 shows an example of an AS/400 cluster and introduces some of the unique cluster terminology.

As previously mentioned, a cluster is a collection of complete AS/400 systems that cooperate and interoperate to provide a single, unified computing capability. In Figure 13, the cluster is the group of AS/400 systems shown. The goal of AS/400 clusters is primarily to achieve improved availability (approaching 99.999% and beyond). Each system in the cluster is called a *cluster node*. An AS/400 cluster may have between two and 128 cluster nodes. The set of nodes defined to be in the cluster is referred to as the *cluster membership* list. The cluster nodes must be interconnected via an IP network.

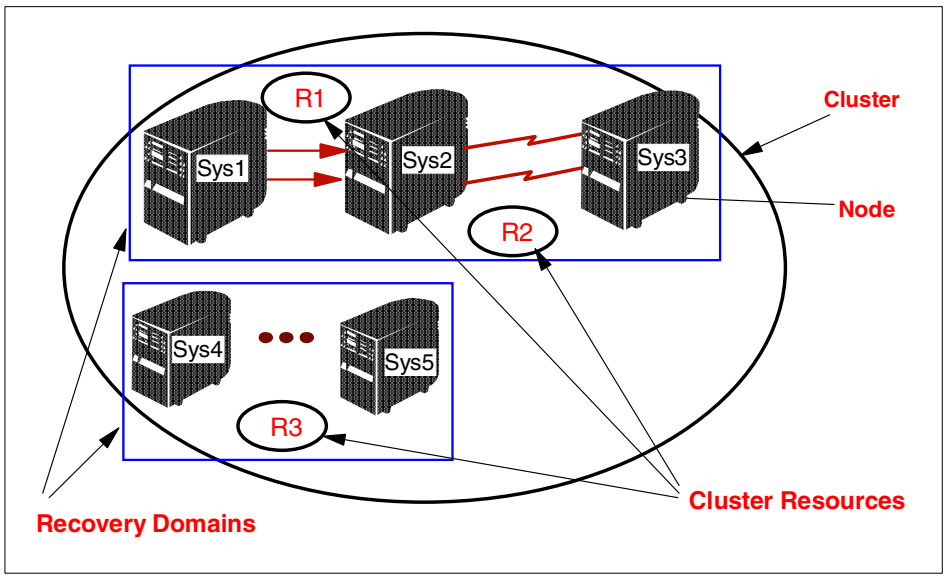


Figure 13. AS/400 cluster terminology

Resources that are available or known across multiple nodes within the cluster are called *cluster resources*. A cluster resource can conceptually be

any physical or logical entity (database, file, application, device, and so forth). Examples of V4R4 cluster resources include AS/400 objects, IP addresses, applications, and physical resources. The objects labeled R1, R2, and R3 in Figure 13 represent cluster resources. When a cluster resource persists across an outage, that is any single point of failure within the cluster, it is known to be a *resilient resource*. As such, the resource is resilient to outages and accessible within the cluster even if an outage occurs to the node currently “hosting” the resource.

The cluster nodes that are grouped together to provide availability for one or more cluster resources is called the recovery domain for that group of cluster resources. A recovery domain can be a subset of the nodes in a cluster and each cluster node may actually participate in multiple recovery domains. Resources that are grouped together for purposes of recovery action or accessibility across a recovery domain are known as a *Cluster Resource Group (CRG)*. The Cluster Resource Group defines the recovery or accessibility characteristics and behavior for that group of resources. Figure 13 on page 37 shows two recovery domains. The first is the top three systems (Sys1, Sys2, and Sys3) and the second is the bottom two systems (Sys4 and Sys5). The recovery domains defined for resource R1 and resource R2 consist of the same set of three systems.

The AS/400 cluster uses the separate server or shared-nothing model. That is, cluster resources are not physically shared between multiple systems. Critical resources are replicated between nodes. Access to the resources may be accomplished through a function shipping protocol. Conceptually, the resource may be viewed as shared since it is accessible from other nodes. However, at any given moment, each resource is owned, or hosted, by a single system.

One additional important cluster concept is the *cluster partition*. The cluster is “partitioned” when a failure occurs that causes a subset of nodes to disconnect from the cluster. The partitioned nodes may communicate among themselves, but they cannot communicate with the rest of the cluster. The cluster cannot determine if the nodes have failed or if they are operational but not reachable.

This partition state is not the same as logical partitions in an LPAR environment (see 3.9.3, “Logical partitioning” on page 32).

4.1.1 How a cluster is used

Although it is possible to use an AS/400 cluster for horizontal growth, the typical intended use is for increased availability. The simplest cluster

configuration is a two-node cluster as shown in Figure 14. There is one primary system for all cluster resources and a second system that is a backup, ready to take over during an outage of the primary system.

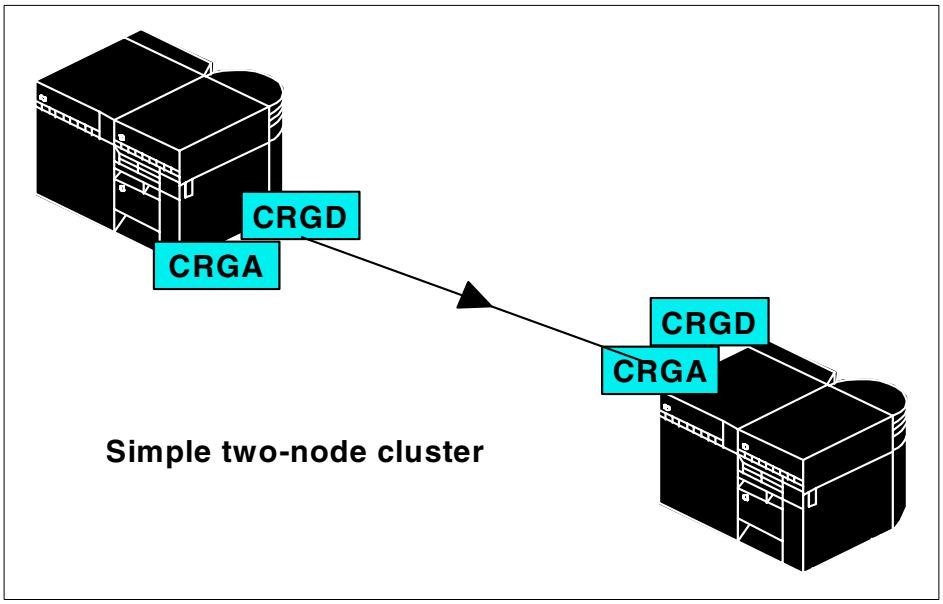


Figure 14. Simple two-node cluster

4.1.2 Four-node mutual takeover cluster

Another typical environment is the mutual takeover environment. Each node in the cluster serves as the primary node for some sets of resources and as the backup node for other sets of resources. With mutual takeover, every system or node is used for production work, and all critical production work is accessible from multiple systems, multiple nodes, or a cluster (Figure 15 on page 40).

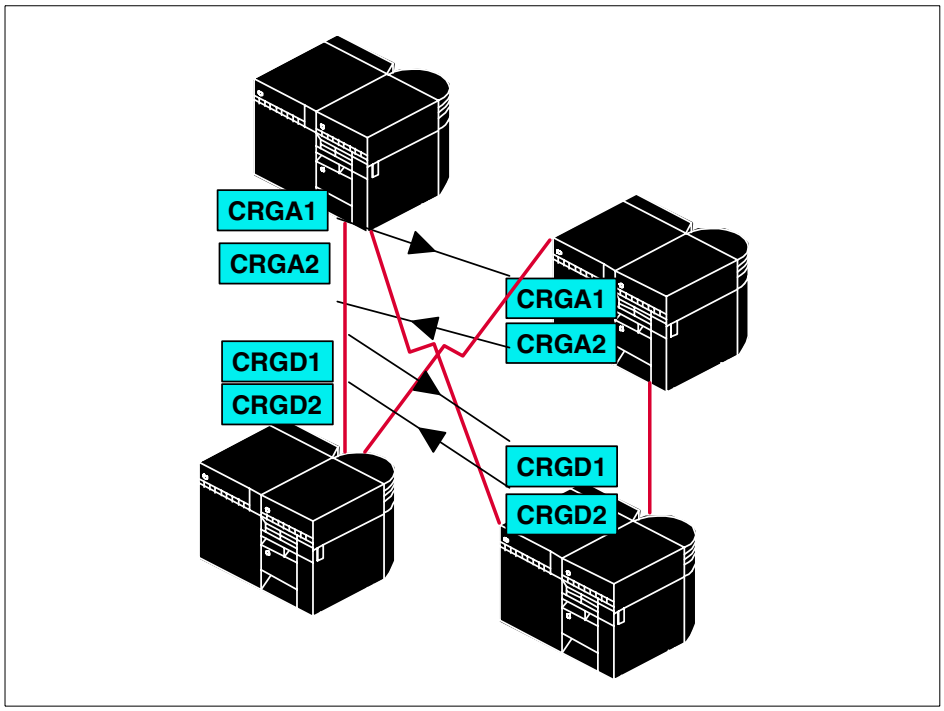


Figure 15. Four-node mutual takeover cluster

The two basic environments shown in Figure 15 can be extended to a cluster of multiple nodes. Combining their capabilities to have multiple ordered backups and to spread the primary point of access for CRGs across different nodes allows significant flexibility in cluster configuration and deployment options. The potential choices are endless.

4.1.3 Application and data resilience

To achieve continuous availability, more than just robust system availability is needed. Critical data and critical applications must also be resilient to outages. Both must be accessible across the cluster even when the normal hosting system for the resource fails. A complete solution is achieved when the critical data and the critical applications are made to be resilient resources and are always available.

Data resilience ensures that a copy of the data is always accessible to end users of the cluster. OS/400 V4R4 provides data resilience through Cluster Middleware Business Partner replication utilities.

Application resilience ensures that the services provided by the application are always accessible to end users of the cluster. Application resilience is provided through IP address takeover and restarting the application on the first backup system.

Later in this chapter, we explore the underlying Cluster Resource Services that enable data and application resilience to be achieved.

4.2 Cluster Resource Services structure

On the AS/400 system, the clustering infrastructure is called Cluster Resource Services. Figure 16 shows the key elements of OS/400 Cluster Resource Services and their relationship.

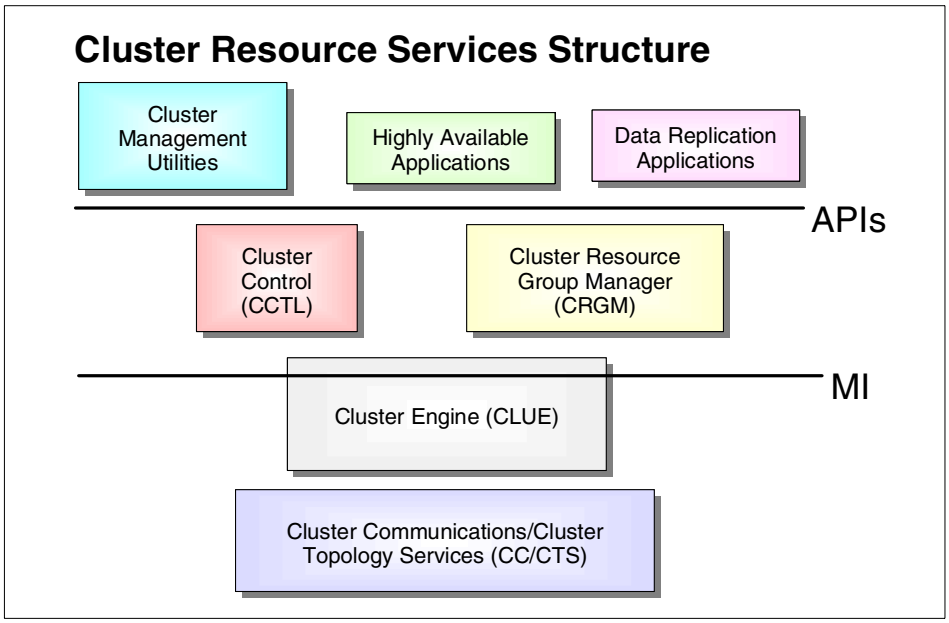


Figure 16. OS/400 Cluster Resource Services

The two boxes labeled Cluster Control and Cluster Resource Group Manager represent OS/400 services that provide APIs. These APIs enable the business partners (BPs), independent software vendors (ISVs), and application providers to deliver a cluster management utility, data resilience through replication, and resilient (highly available) applications. The APIs are documented in the *AS/400 System API Reference Manual (V4R4)*, SC41-5801.

Cluster control provides configuration, activation, and management functions for the cluster and the nodes in the cluster. The cluster definition, configuration, and state information is maintained in a persistent internal object called a cluster *information object*. This object exists on each node in the cluster. Upon request, cluster control starts clustering on a node and coordinates the process of joining that node into the cluster. This process ensures all nodes are equally aware of the action and have the same content in their cluster information object. Cluster control also manages the merging of cluster partitions.

Cluster Resource Group manager provides object management functions, such as creation, deletion, and modification, for the CRG objects. A CRG is a new OS/400 object that defines and controls the behavior for a group of cluster resources across a recovery domain. Conceptually, the CRG is a distributed object. The CRG exists on all nodes in the defined recovery domain. A change on one cluster node is reflected across the recovery domain. Each node in the recovery domain has a defined role of primary, backup, or replicate. The nodes in the recovery domain and their respective roles are defined in the CRG object. When any cluster event occurs that affects that CRG, a user-specified exit program is called on every active node in the recovery domain. A cluster event could be adding a node to the cluster, changing a recovery domain, or a node going offline. The CRG exit program is identified in the *CRG object. Since the exit program provides resource-specific processing for the cluster event, it could be considered the resource manager for the group of resources associated with that CRG. There may be many CRGs on a node, each potentially with a different recovery domain.

The cluster control and Cluster Resource Group manager components use lower-level system services (OS/400 Licensed Internal Code) to ensure:

- The content of all control objects are logically identical across the affected nodes.
- Cluster activity is coordinated across the affected nodes.

The two boxes labeled Cluster Engine and Cluster Communications/Cluster Topology Services in Figure 16 on page 41 provide these system services.

The *Cluster Engine* provides reliable group communications for the distributed processing needed by the other cluster components to achieve coordinated, distributed, and synchronized activity across multiple cluster nodes. The cluster engine services include group membership services and group messaging services. More details on these are provided in 4.2.1, “Underlying technologies” on page 44. Most of the cluster engine is

implemented below the machine interface (MI) to achieve high efficiency, better performance, and better integration with other communication components in the streams stack.

Cluster communications provides low-level internode communications support for the rest of the Cluster Resource Services. It implements the reliable first in, first out (FIFO)-ordered multicast message that takes advantage of the IP multicast support of the underlying network when it is available. This component guarantees that a multicast message is eventually delivered to all its targets, except in the case of failures. When cluster communications fails to deliver a message to a target (after exhausting all retry attempts and alternative paths), it considers the target node unreachable (failed or disconnected). In the case where the local node fails before completing a multicast message, there are no guarantees that all targets receive the multicast message. In addition to multicast messages, cluster communications also supports unreliable unordered messaging, reliable FIFO point-to-point messaging, and unreliable point-to-point messaging. The used components can define many multicast groups, dynamically modify membership of each multicast group, and refer to each multicast group via an identifier (for example, when sending messages). This allows cluster communications to plan message distribution and to maximize parallelism for processing unrelated multicast messages sent in the cluster. Cluster communications is implemented in the streams stack below the MI to achieve high efficiency, better performance, and better integration with other communication components.

Cluster topology services provides a cluster view over existing IP network connectivity. It maintains the knowledge of currently active cluster nodes and cluster nodes known to be partitioned. Two paths may be defined to each node in the cluster. The first path to the node specified on the cluster control API is considered the preferred (primary) path. Cluster topology services continuously checks connectivity of the various network paths and allows a seamless switch to the alternative path when the preferred path is not available. It also allows a seamless switch back to the preferred path, when it becomes available again. In addition, cluster topology services periodically checks connectivity to partitioned nodes to see if connectivity has been re-established. When successful, cluster topology services notifies cluster control and the cluster engine, which then attempt to merge partitions back into the cluster. Part of the continuous check performed by cluster topology services is *heartbeating*, which performs periodic checks on liveness and connectivity of the locally reachable cluster nodes and delivers failure notifications. When a previously connected node becomes unreachable,

cluster topology services notifies the cluster engine. The cluster engine then removes the node from the locally visible cluster or declares a partition.

4.2.1 Underlying technologies

Numerous new and enhanced technologies and architectural features have been added to OS/400 V4R4 to support clustering. The following sections highlight a few of them.

4.2.1.1 Peer cluster nodes

Many cluster implementations follow the paradigm of having a leader for various clustering protocols. A leader may be established as a result of configuration (for example, the primary node is the leader) or it may be determined through some internal algorithm (for example, based on an IP address). AS/400 clustering has chosen to use a leaderless architecture or peer relationship among the cluster nodes. Each active node has all of the information needed to understand the total configuration and operational characteristics of the cluster. As such, a request for a cluster action can be initiated from any node active in the cluster. Furthermore, any node (not necessarily the requesting node) can assume the role as the coordinator for a particular protocol. This helps to ensure that a single outage, or even an outage of several cluster nodes, will rarely constitute a cluster failure.

4.2.1.2 Heartbeating and cluster communications

Heartbeat monitoring determines whether each node is active. When the heartbeat processing for a cluster node fails, the condition is reported so the cluster can automatically fail over resilient resources to a backup node.

A heartbeat failure is more complex than just one missed signal. A heartbeat message is sent every three seconds from every node in the cluster to its upstream neighbor. Each node expects an acknowledgment of this signal in return. In effect, this presents a two-way liveness mechanism. When a heartbeat (or its acknowledgment) is not received, a failure is not immediately reported. Heartbeating continues every three seconds. Heartbeating on remote subnets may be four times that of local heartbeating. If a node misses four consecutive heartbeats, a heartbeat failure is signaled. After this failure is confirmed, the failover process causes access to the cluster resources to be switched over to the designated first backup node.

The heartbeat service within cluster topology services ensures low system overhead during normal operations. Other components of cluster resource service can rely on cluster topology services to determine when a node becomes unreachable. In some circumstances, heartbeat failure may not translate into a node failure, in which case, a failover might not happen.

If the cluster consists of multiple physical networks, the heartbeat processing becomes somewhat more complex. Routers and relay nodes are used to tie the physical networks together as though it were one logical network. A router can be another AS/400 system or a router box that directs communications to another router somewhere else. Every local network is assigned a relay node. This relay node is determined to be the cluster node that has the lowest node ID in the network. For example, if two networks are involved, a logical network containing the two relay nodes is created. The relay nodes can then send heartbeats to each other. By using routers and relay nodes, the cluster nodes in these two networks can monitor each other and signal any node failures. See Figure 17.

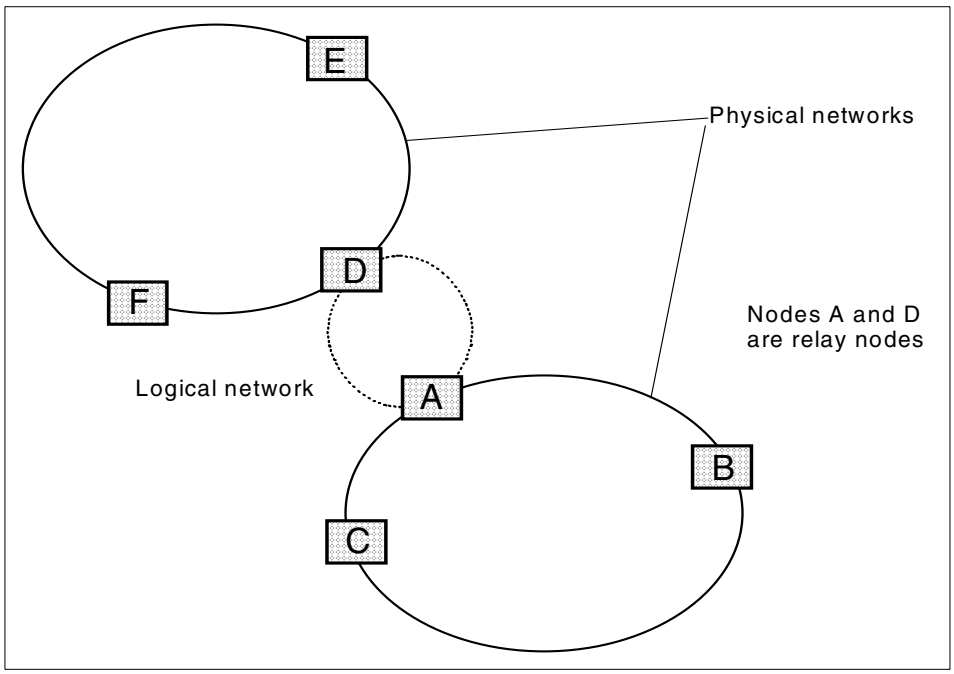


Figure 17. Relay nodes in heartbeat protocols

The AS/400 Cluster Resource Services makes no assumptions about the network throughput, latency, topology, or stability. The heartbeat algorithms are expected to work over any supported network configurations.

4.2.1.3 Distributed activities

Most cluster actions are distributed activities resulting from a user request or system detected event. The synchronization of actions across the nodes of a cluster, or across a subset of the nodes, is accomplished through a

distributed activity. All of the cluster nodes affected by the action need to be involved to ensure that the results are consistently reflected across the cluster. The cluster engine and cluster communications provide the underlying services for building what we refer to as distributed activity groups. The Cluster Engine's Group membership services are used by cluster control and the Cluster Resource Group manager to define *distributed activity groups*. For cluster control, there is a distributed activity group used for the distributed activities associated with defining and administering the cluster. Each node in the cluster is a member in this distributed activity group. There are multiple distributed activity groups associated with the Cluster Resource Group manager. One set, which is one distributed activity group, is defined across the entire cluster and is used to handle the creation of new CRGs on each cluster node in the recovery domain and other similar global activities. It is called the *Cluster Resource Group manager distributed activity group*. There is also a distributed activity group for each CRG defined to handle processing specific to that CRG.

Cluster control and Cluster Resource Group manager can synchronize their services across all affected nodes within the cluster via distributed activities. Any change to internal information or external cluster objects on one cluster node is simultaneously reflected on all nodes in the cluster. Complex protocol flows may be needed to accomplish this processing or to back out changes in the event that an error condition is detected. Because we make no assumptions regarding the guaranteed low latency for the services of the underlying network, we rely on asynchronous distributed agreement solutions.

4.2.1.4 Job structure for Cluster Resource Services

The use of the cluster engine’s group services is also apparent by looking at the Cluster Resource Services job structure. When a cluster is started on a cluster node, a set of system services is started. Each of these services is designed to be highly available (resilient to errors). These services are represented by multi-threaded jobs running in the QSYSWRK subsystem. Anytime a cluster node is active, the following jobs are active in that subsystem:

- A cluster control job called QCSTCTL.
- A Cluster Resource Group manager job called QCSTCRGM.
- Additional jobs are started for handling the Cluster Resource Groups. One job exists for each CRG defined in the cluster. The job name will be the same as the CRG name.

Figure 18 shows an example job structure with just one CRG defined (CRGa). The figure also shows the related jobs. These include the user job that initiated the cluster request (normally in the subsystem for the cluster management processing), the exit program job that is called to handle CRG specific processing, and an application subsystem for a highly available application.

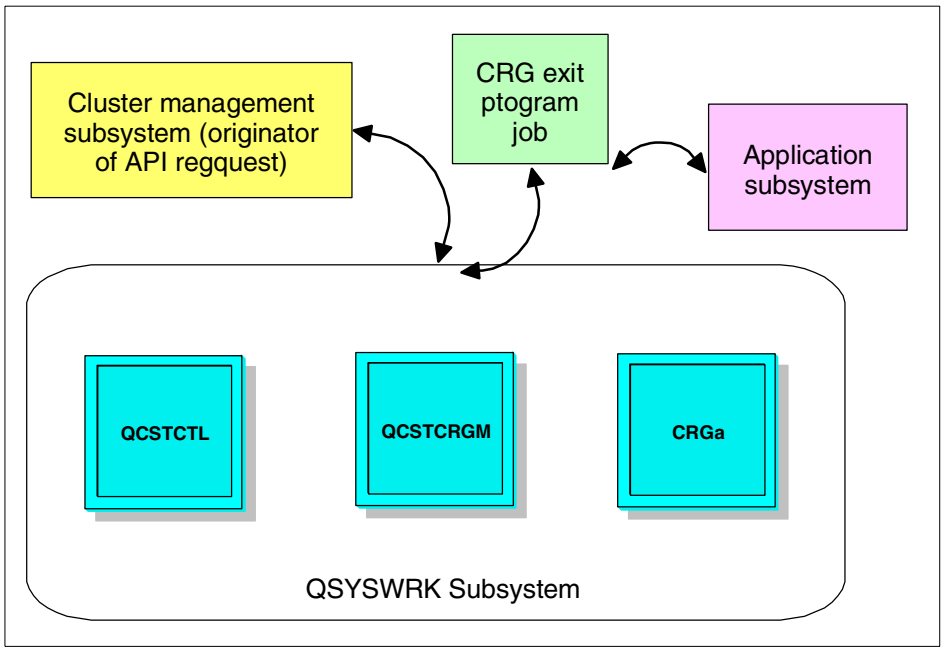


Figure 18. Job structure for Cluster Resource Services

In addition to the system jobs, the user job that originates the request for a cluster service must be considered. The request normally consists of a call to one of the clustering APIs. After validation of the API, the request is passed to the appropriate cluster job in QSYSWRK. The cluster job then handles the distributed processing of the request through the distributed activity group technology. Using the distributed activity group, the request is distributed to other members of the group on the other nodes. There, the appropriate processing of the request takes place and results are returned to the cluster node that initiated the request. Once responses are received from all members that participated in the activity, the results are returned to the results information queue.

Finally, there are CRG exit program jobs associated with CRG activities. They are initiated by Cluster Resource Services on all active nodes in the recovery

domain. These jobs run in a user-specified subsystem, which may be the same as the application subsystem. The exit program jobs are normally transitory, only existing for the duration of the API request. The exception to this is the CRG exit program that is called to start the resilient application, which runs only on the primary system. This job remains active and serves as a daemon job between the application and Cluster Resource Services.

4.2.1.5 Cluster engine services

Cluster engine group membership services provide the ability to define and modify distributed activity group membership definitions. Any changes to the definition or to the state of the members results in notification to live group members of group membership changes. Notification is via a special message called a *membership change message*. The cluster engine ensures that cluster membership changes are handled consistently across affected groups for both administrative changes and changes as the result of a failure. Thus, a consistent view of the membership is guaranteed not only across members of the same distributed activity group, but across related groups as well.

The messaging services that the cluster engine provides to group members include a variety of reliability and ordering guarantees over group messaging:

- **Non reliable, FIFO ordered messaging**: FIFO messaging implies that group messages sent by the same node are received in the same order by all the group members.
- **Reliable, FIFO ordered messaging**: Reliable messaging is a variant of virtually synchronous. Members appearing in two consecutive membership change notifications receive the same set of messages between these notifications.
- **Reliable, totally ordered messaging**: Totally ordered messaging implies that group members who are receiving the same set of messages receive them in the same order.
- **The above guarantees are defined per group**: A cluster engine provides the ability to send non-reliable messages to the group or to a subset of the group.

4.2.2 Partition state

A cluster becomes partitioned when the cluster cannot communicate with one or more nodes and no certain failure has been identified. A cluster partition is not good (unlike a logical partition). The typical case of partitioning occurs when there is a communications link failure and no redundant path has been established as shown in Figure 19.

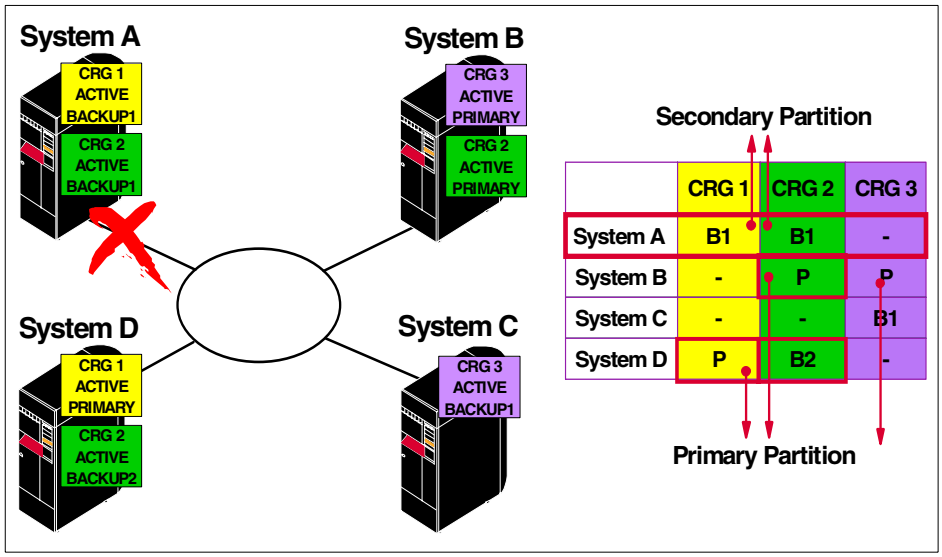


Figure 19. Example of a cluster partition

When the cluster is partitioned, Cluster Resource Services takes overt action to ensure:

- As many operations as possible can continue normally.
- Operations that would cause inconsistencies between partitions are not allowed.

If configuration and operational changes were allowed to be made independently in two or more partitions, there could be no guarantee that merging of the partitions would be successful.

To ensure a partition does not occur, plan for redundant paths within the network and adapters connecting the nodes to the network. See 6.5.1, “Redundancy” on page 71.

4.2.3 Versioning

To ensure that nondisruptive clustering is supported, there is the provision for multiple releases of OS/400 to coexist in a single cluster starting at OS/400 V4R4. This allows individual nodes to be upgraded to the next release without requiring the cluster to be taken down. Therefore, a cluster node must be able to recognize and interoperate with other cluster nodes that are at different release levels.

To support this environment, Cluster Resource Services has implemented levels of versioning beyond what is supported by existing system capabilities. One level of versioning is in the objects used by Cluster Resource Services. Any internal changes to an object cause the version information to be changed. When information is exchanged between nodes, the system services can account for different object versions. The second level of versioning is in the messages passed between nodes and between cluster components. Enhanced messaging, and therefore additional services, can be introduced without hindering the ability to communicate with nodes at the previous release level.

4.2.4 Conclusion

OS/400 V4R4 has introduced new technologies to better support clustering for continuous availability. Several of these technologies have been discussed to highlight the new capabilities in the *System API Reference*, SC41-5801, for V4R4. These technologies provide a firm foundation with low overhead for V4R4 clusters, for example:

- Peer relationships between cluster nodes help ensure no cluster-wide outage.
- Heartbeating and efficient cluster communications provide low overhead internode processing and early detection of potential node outages.
- The distributed activity groups are used to synchronize activities and objects across cluster nodes.
- The cluster engine services provide reliable, ordered messaging and group membership services.
- The job structure of Cluster Resource Services, interjob communications, and internode communications provide a single, consistent view of cluster node and cluster resource status.
- Through cluster partition handling, the system determines the difference between many failure and partition conditions without user intervention.

OS/400 V4R4 enables the delivery of highly available applications and improved data replication utilities as well as coordination between application and data resilience. In addition, the clustering technologies provide obvious foundations for future enhancements. One example shown is internal versioning to enable multiple release levels to interoperate.

Chapter 5. ClusterProven

ClusterProven is an IBM Enterprise Server Group program that covers all four server platforms. This program is designed to assist the independent software vendor (ISV) on their way to continuous availability and recognizes that true availability cannot be achieved without application involvement. The ClusterProven program increases the high availability application portfolio and raises the availability bar to move towards continuous availability. The individual servers have their own definitions for ClusterProven that relate to their specific platform. This chapter describes what is involved in the achievement of ClusterProven for AS/400.

Figure 20. ClusterProven trademark

5.1 Overview of the ClusterProven components

As shown in Figure 21 on page 52, the AS/400 cluster solution is a joint effort between IBM, Cluster Middleware Business Partners, and independent software vendors. The AS/400 system provides the basic cluster infrastructure that works with the Cluster Middleware Business Partner products and the independent software vendor high availability applications.

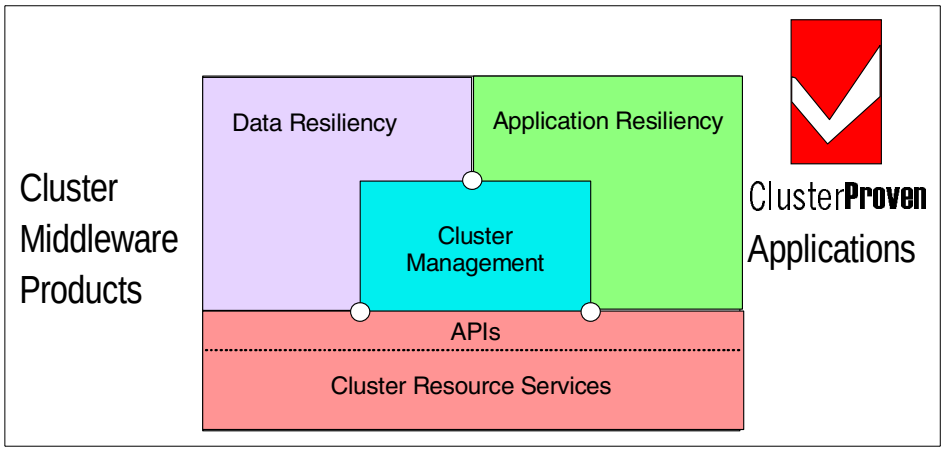


Figure 21. AS/400 cluster partnership

5.1.1 OS/400 Cluster Resource Services

Beginning with V4R4, the OS/400 operating system includes Cluster Resource Services. Cluster Resource Services provides cluster functions and an open set of application programming interfaces (APIs) that applications can use to create and manage a cluster. It establishes the architecture, controlled by the AS/400 system, from which all cluster middleware providers and independent software vendors can build highly availability solutions.

5.1.2 Cluster middleware data resiliency

Data resiliency allows you to maintain one or more copies of the application data on one or more backup systems or logical partitions. DataMirror, Lakeview Technology, and Vision Solutions have provided sophisticated data replication solutions for many years. Taking advantage of the new architecture, each of these business partners now offers new and updated cluster middleware products.

5.1.3 Cluster middleware cluster management

Cluster middleware products provide a user interface to manage the cluster. This interface hides the complexity of creating and managing a highly available solution.

5.1.4 ClusterProven application resiliency

Independent software vendors have the capability of making their applications resilient, allowing them to move users over to an alternate server

if the primary server becomes unavailable. With OS/400 Cluster Resource Services combined with cluster middleware data resiliency, an application can be designed for automatic configuration, activation, and switchover capability that returns the end user to an application screen. Applications that meet the criteria may be approved as ClusterProven for AS/400. There are two levels that applications can be designated as: ClusterProven and Advanced ClusterProven. The criteria for each designation is discussed in the following section.

5.2 ClusterProven for AS/400

ClusterProven for AS/400 means that a purchased application with this designation is enabled as a partner in the cluster solution.

The independent software vendor must follow these steps to obtain the ClusterProven trademark for the AS/400 system:

1. Validate the application against the criteria listed in the following section.
2. Submit the application to PartnerWorld for Developers (PWD).
3. Agree to the terms and conditions that cover the use of the ClusterProven trademark.

5.2.1 Basic ClusterProven for AS/400

Applications must meet the following requirements to be approved for ClusterProven for AS/400:

- The highly available application can switch over to a backup cluster node when the primary node becomes unavailable.
- A cluster management application offers automatic configuration and activation of the application.
- The application provides application resilience by using an application Cluster Resource Group exit program to handle cluster-related events. Using an application Cluster Resource Group exit program takes advantage of OS/400 Cluster Resource Services.
- The application provides restart capabilities that reposition the user to an application menu screen or beyond.

5.2.2 Advanced ClusterProven for AS/400

Applications must meet the following requirements to be approved for Advanced ClusterProven for AS/400:

- The highly available application meets all of the criteria for ClusterProven for AS/400 listed in the previous section.
- The application provides enhanced application resilience through more robust handling of cluster events by the application Cluster Resource Group exit program.
- The application provides a greater level of application restart support:
  - For host-centric applications, the user is repositioned to a transaction boundary using commitment control or checkpointing functions.
  - For client-centric applications, the user experiences a seamless failover with minimal service interruption.

Chapter 6. Planning for AS/400 clusters

This chapter discusses the areas that need to be investigated when planning for an AS/400 cluster solution. It also presents considerations for putting together an implementation plan and the ongoing management of the cluster.

6.1 Cluster planning steps

There are two methods of cluster setup. The first method is very simplistic, and the second method is more detailed.

6.1.1 Simple cluster

In the simple cluster there is an assumption that two systems already exist and a replication environment is in place. This setup does not attempt to meet the stringent availability requirements of the full cluster deployment. It really starts clustering, increasing the ability availability and providing switch over and restart.

1. Produce a cluster development and test environment (6.3.8, “Maintenance” on page 63).
2. Configure and implement resilient hardware and network resources (6.2.3, “Configuration of a cluster” on page 58).
3. Select and deploy a cluster proven application (6.3.2, “Application object inventory” on page 60).
4. Test your cluster under various conditions (6.6, “Cluster testing” on page 72).

6.1.2 Full cluster deployment

These simple cluster planning steps give you a very high level view of your cluster solution. They are the basic tasks that are involved. More details on each task are covered later in this chapter.

1. Calculate your impact to the business costs (6.2.1, “Business impact costs” on page 56).
2. Use the costs to determine the level of availability required to support your business (6.2.2, “Level of availability” on page 57).
3. Develop an availability plan to meet the requirements in Step 2 and your operational procedures (6.4, “Systems management” on page 68).
4. Select and implement High Availability middleware (6.2.4, “Replication environment” on page 59).

5. Produce a cluster development and test environment (6.3.8, “Maintenance” on page 63).
6. Configure and implement resilient hardware and network resources (6.2.3, “Configuration of a cluster” on page 58).
7. Select and deploy a cluster proven application (6.3.2, “Application object inventory” on page 60).
8. Test your cluster under various conditions (6.6, “Cluster testing” on page 72).

6.2 Cluster planning

When embarking on a cluster project, there are a number of detailed tasks that need to be considered during planning: What will it cost? What other non-system related tasks are there? These tasks are reviewed in this section.

6.2.1 Business impact costs

You must first consider an estimate of the cost of downtime. An example of this is in 2.1.2, “Financial impact of an outage” on page 13.

If this project is being started from within the I/T organization, it is imperative that the business has input to the financial case for the cluster solution. This is really the only way the I/T department will justify the costs involved.

Typically, the cost of the hardware configuration should not scare the customer management. Continuous availability is a new paradigm. The implementation costs will look high, even enormous, but in comparison with the potential losses, they are probably small.

Start by investigating what other system outages have occurred over the past couple of years. These outages should include both planned and unplanned activities. Obtain an estimate of the cost of those outages. These costs can be split into two areas: tangible losses and intangible losses:

- Tangible losses
  - Loss to share holdings and profits
  - Losses incurred through product or campaign delays
  - Employee idle time waiting for the system to come back
  - Employee overtime to catch up lost production and transactions
  - Idle time cost of facilities and equipment
  - Consequential loss through penalties from customers or suppliers
  - Goods lost through damage or aging

- Intangible losses
  - Credibility in the marketplace
  - Lost revenue from customers buying elsewhere
  - Market share

Once this data exists, other scenarios can be planned. A rough idea of the cost of each application should be calculated. This helps you to understand the priority of applications. Where applications are integrated, the cost estimate may become complex because of the effects on other parts of the business.

The intangible losses are more difficult to estimate, unless there has been a major disaster within the business in the recent past. These losses should not be ignored since they can be very large, especially if the organization has a very high visibility in the marketplace.

6.2.2 Level of availability

Once the financial case has been established, decide the level of availability that the business can afford. Do not be deterred if the highest level of availability cannot be achieved on the first pass. A tactical solution may need to be considered before the business can move to a more strategic solution.

An example of a tactical solution might be to run with an application that has only the basic ClusterProven status. In this scenario, there would not necessarily be a seamless switchover. If there are 5250 devices, the end users would have to sign on to the backup system. However, when these users open their application session, they would just be positioned back to their last menu. The I/T group would need to establish the integrity if the application and database before users could start performing updates.

This would probably be a far better and more structured solution than what currently exists in most organization. This solution could be implemented relatively quickly and provide high availability while the strategic solution is being developed.

A strategic solution may be to implement Advanced ClusterProven. This solution may take longer to develop, depending on the recoverability of the existing application. The advanced status demands that commitment control for application rollback or a similar function is implemented.

Another alternative is to be very selective over the application or business area that is to be clustered. This would allow the business to tactically implement a cluster that is highly available. It would also allow more time to

plan, implement, and manage the strategic cluster solution with some protection.

When deciding on the level of availability one must also consider the skills needed to manage the cluster. Many organizations have some level of OS/400, development and networking skills. To implement a cluster, the I/T group must also focus on database skills, data management and replication skills (journaling, commitment control, HABP software skills). This is a significant investment that cannot be achieved quickly, so skill development should be considered early in the process.

6.2.3 Configuration of a cluster

The configuration of a cluster is primarily the hardware resiliency. Spending time now to define the configuration will result in a better solution and ultimately give less downtime. Deciding to implement RAID disks now to later find out mirrored disks would have offered more protection could cause a significant impact on your system's availability in the long term. Network changes can also have an effect on the cluster if these changes are significant and made later in the evolution of the cluster. Aim to build the hardware environment early in the project.

Deciding the roles, location, and size of the systems (nodes) in the cluster is also very important. You may already have a number of AS/400 systems in your business. There must be a decision made on how much of the operational environment and application is replicated. You may decide that only the system will be replicated and one key application. This would give the business a simple cluster to work with a develop skills. Then in the future the rest of the hardware and applications could be moved into the cluster environment.

When looking at the systems infrastructure the first task is to document the inventory of all the AS/400 system hardware (processor, storage, disk, IOPs/IOAs), operating system release level, applications, and their release levels. Then, document other related systems that will not be part of the cluster, but affect the same sphere of business operation. These can be other systems (servers or clients) plus their operating systems, databases and applications; network hardware (LAN, ISPs, routers, topology); and peripheral devices (tapes, printer, and displays).

The second task is to select the applications that are to be clustered. Decide where the most appropriate locations are to run these applications. Size the systems at these locations to see if there is spare capacity to implement the availability features and the resilient application under all conditions. Then

decide where any additional capacity is needed. In some cases, the production machine may have plenty of capacity to run additional recovery features and only the backup system needs to have extra capacity added to support the failover or switchover load from production. It is not uncommon to see production systems running 30% utilized with today's price per performance structure.

Remember that the more systems that are in the proposed cluster solution, the more complex the implementation will be. Start simple and build the solution rather than going for the big bang approach. Nodes can always be added to the cluster. Additional Cluster Resource Groups related to other applications and data sets can always be started after the initial cluster is setup.

If you plan to implement multiple applications and multiple nodes in a cluster, select the role that each node will take under various conditions. If it is planned to use the backup system as an end-user query engine and tape backup machine under normal operating conditions, what will happen to the users and tape backups during a role change? Leaving these users and services in an unavailable status, until the backup's role is changed, may be acceptable to those business groups or the database recovery policy. A larger backup system or third node may need consideration. This extra capacity would provide services for the whole load, while the primary is unavailable.

Section 4.1.1, "How a cluster is used" on page 38, discusses the two-node cluster and the four-node mutual takeover cluster. These are just two of the configurations out of many possibilities. How complex the cluster configuration or level of availability is will depend upon the individual company requirements. The demographics of a businesses, the requirements of the service level agreements, and the number of applications to be supported will be key in the decision making.

The most simple design is the two-node cluster. In this design, there is a basic need to replicate the processing complex and database. The backup site could be fairly close to the production machine, within the same location, or the backup could be geographically remote. The machine and applications supported could again be very simple or very complex, for example, multi-application within an LPAR environment.

6.2.4 Replication environment

If there is not a current replication environment in the systems configuration, a High Availability Business Partner must be selected and involved at this stage of the process.

It is possible to develop a replication environment and implement the ClusterProven status. However, this is a complex task and should not be attempted unless the development group has significant skills in OS/400 database, journaling, and commitment control.

The fastest method for deployment is to use one of the three HABPs mentioned in 3.10, “AS/400 high availability middleware” on page 34, and Part 2, “High Availability Business Partners” on page 75.

These HABPs have wide experience in replication setup and hot backup techniques. They have also developed methods of accessing difficult-to-manage objects like user profiles, message queues, and so on. They can provide advice and guidance on the setup and configuration of your cluster as well as application awareness and restart.

6.3 Applications

This section offers guidance on application recoverability, but not in great detail. The characteristics are so different between packages that it would be difficult to be specific.

6.3.1 Application information

If your AS/400 application is provided by an ISV, ask them whether they have ClusterProven status. If they are ClusterProven, their application already has their application information stored in the new cluster data area QCSTHAPPI, and any HABP can easily set up and manage a cluster environment.

If they don’t provide a cluster aware application, ask when they will be moving to this status. It is not a difficult task and there are resources available to help them. If the ISV has no plans to make their application cluster aware, an option is to find another ISV that has such plans.

If your application is developed in house, decide whether you are going to develop it into a ClusterProven application or select a package that is ClusterProven. Again their are resources available to assist with making the application ClusterProven.

6.3.2 Application object inventory

You must first look at the applications that are running throughout the business. This should include both AS/400 applications and applications running on other platforms. Find where continuous availability is required. This could mean changes to applications other than those on the AS/400

system. For AS/400 applications, there are a few objects in the database that will require special handling. Examples of these objects are:

- Temporary files
- Data spaces
- Data queues

Work with your HABP and ISV to investigate whether these types of objects exist as part of your application. They have been dealing with the replication of these types of objects for some time and can offer special handling methods.

6.3.3 Resilient data

Establish which objects on which nodes need to be resilient. This can be done with the help of the inventory completed in the previous planning step. Once these objects are identified, use the HABP middleware to replicate them. This replication involves journaling and replicating the objects between two or more systems in the cluster.

These resilient objects should be entered into the object specifier file associated with QCSTHAPPI data area for the application. This way the HABP Cluster management tool can automatically create data resilience as represented by a data (type-1) CRGs and then setup the replication environment.

6.3.4 Resilient applications

Decide the level of availability for each application. Select an application that is ClusterProven for AS/400 and whether Basic or Advanced ClusterProven is required.

If you have an in-house application, create a development plan for any modifications required to meet the ClusterProven criteria for high availability. Review 5.2, “ClusterProven for AS/400” on page 53, for more information.

Plan the recovery domains and the IP-takeover addresses to be related to application (type-2) CRGs. Decide which nodes will have the applications running and which nodes will be the backups.

Plan the device switchover characteristics. For 5250 devices (displays and printers), a hardware switch needs to be included in the configuration if you are planning to switch or fail over these types of users to a backup machine. The 5250 devices would be switched manually or automatically as a result of

a switch or failover occurring. An automatic switch would also involve some third-party software that may be available as part of the HABP middleware.

For IP devices, such as PCs that are browsing to the node, a simple refresh of the browser reselects the IP takeover address on the new node. The user can then re-access their application.

6.3.5 Switchover

Once you have a ClusterProven application and you have decided which node it will run, you must now consider your business's switchover characteristics. Switchover looks at the following tasks: moving existing users off the application, preventing new users access to the application, stopping the application, completion of apply tasks, and so on.

Some applications may have already implemented these functions, especially in the banking and services sector. These organization tend to have stringent end-of-day (EOD) routines that require all users to be off the system for a short period of time while the EOD tasks run. These applications have methods for removing users from the application.

6.3.6 Failover

A failover is similar to a switchover, but you have less control. In theory, if the failover is completely seamless, it can be used as the switchover (for example, a switchover means press the Off button). However, given the opportunity, it is always safer to switch.

The tasks involved in planning a failover ensure that you can put the users back to the same position that they were in before the failure.

6.3.7 Restart

AS/400 job restart is a complicated task and needs careful planning to achieve the fastest and safest restart of jobs on the backup node. The following sections review the restart characteristics of the jobs running in an AS/400 system.

6.3.7.1 Interactive

Interactive jobs run from a twinax display or in an emulator session. In a switchover situation, these jobs can be normally ended in a controlled fashion, with all transactions completing without loss.

In a failure condition, the jobs end abnormally. Transactions are incomplete and temporary objects are lost. With a good understanding of the application, the losses can be contained and planned.

Since these devices use a 5250 data stream, they cannot be controlled by IP takeover. They must be manually switched. This hardware switch process can be linked to the IP takeover.

6.3.7.2 Batch

Batch jobs are an environment that must be studied in great detail. If you have a long running single thread batch job, you must establish whether this job can have restart points added or must be completely restarted. If you must completely restart the batch job, this can be a long running operation that could seriously effect the overall availability.

Multi-threaded batch jobs are more complex. You may not be able to provide restart points within the job, and it may need to be rolled out and restarted.

6.3.7.3 Client server

Client server is probably the easiest AS/400 environment to cluster. The state of the client is not important to the AS/400 system. Note, you should care about the client under a different part of the cluster or high availability project.

Clients are typically IP-connected devices. IP takeover can handle the movement of these devices from primary to the backup. The client should see very little impact from a failure or switchover.

6.3.8 Maintenance

Application maintenance is another potentially difficult area. If the application only needs changes to the programs running within it, this is an easy task. However, if the application requires changes to the underlying database, maintenance becomes more complicated.

Figure 22 through Figure 25 show how to retain full protection for your cluster while carrying out the application maintenance task.

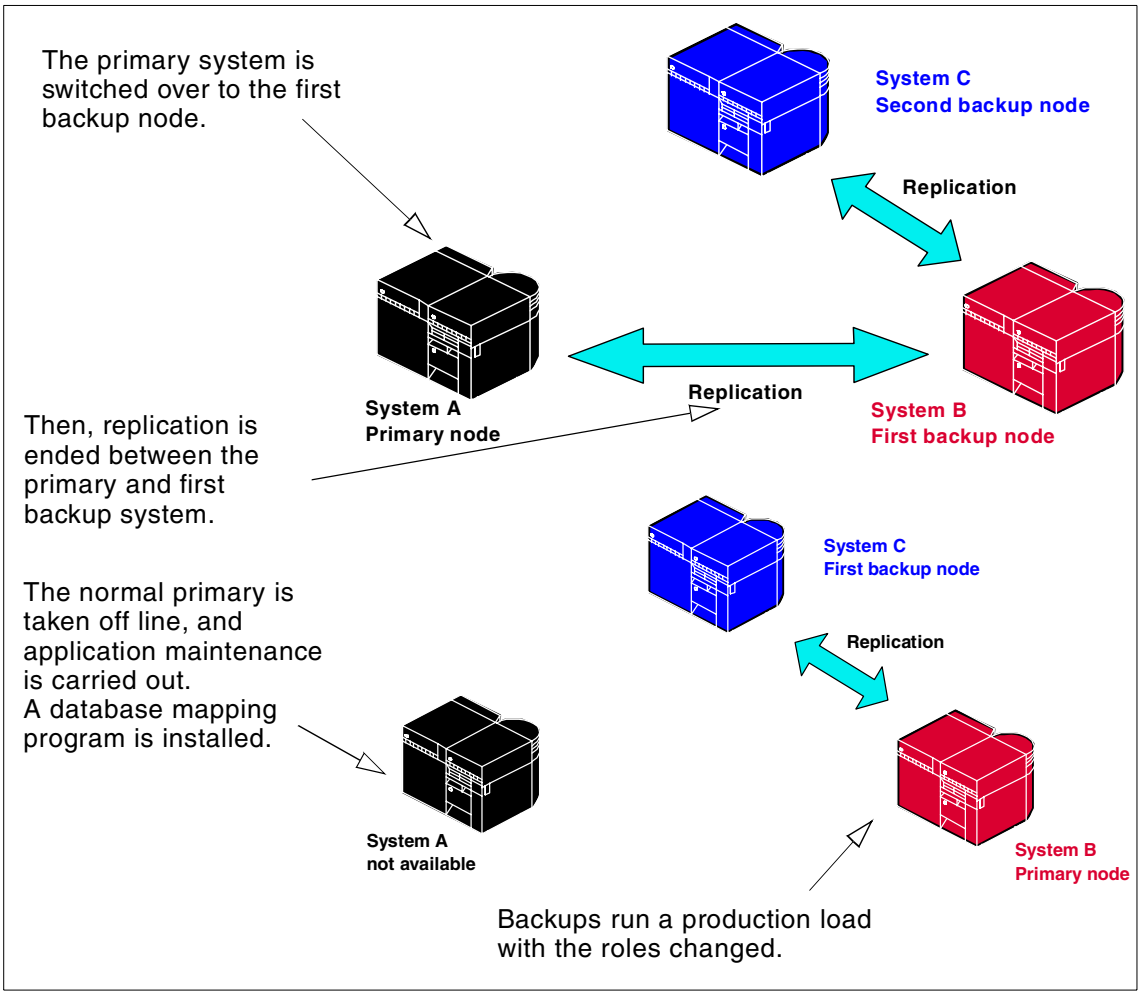


Figure 22. Application maintenance in a cluster (Part 1 of 4)

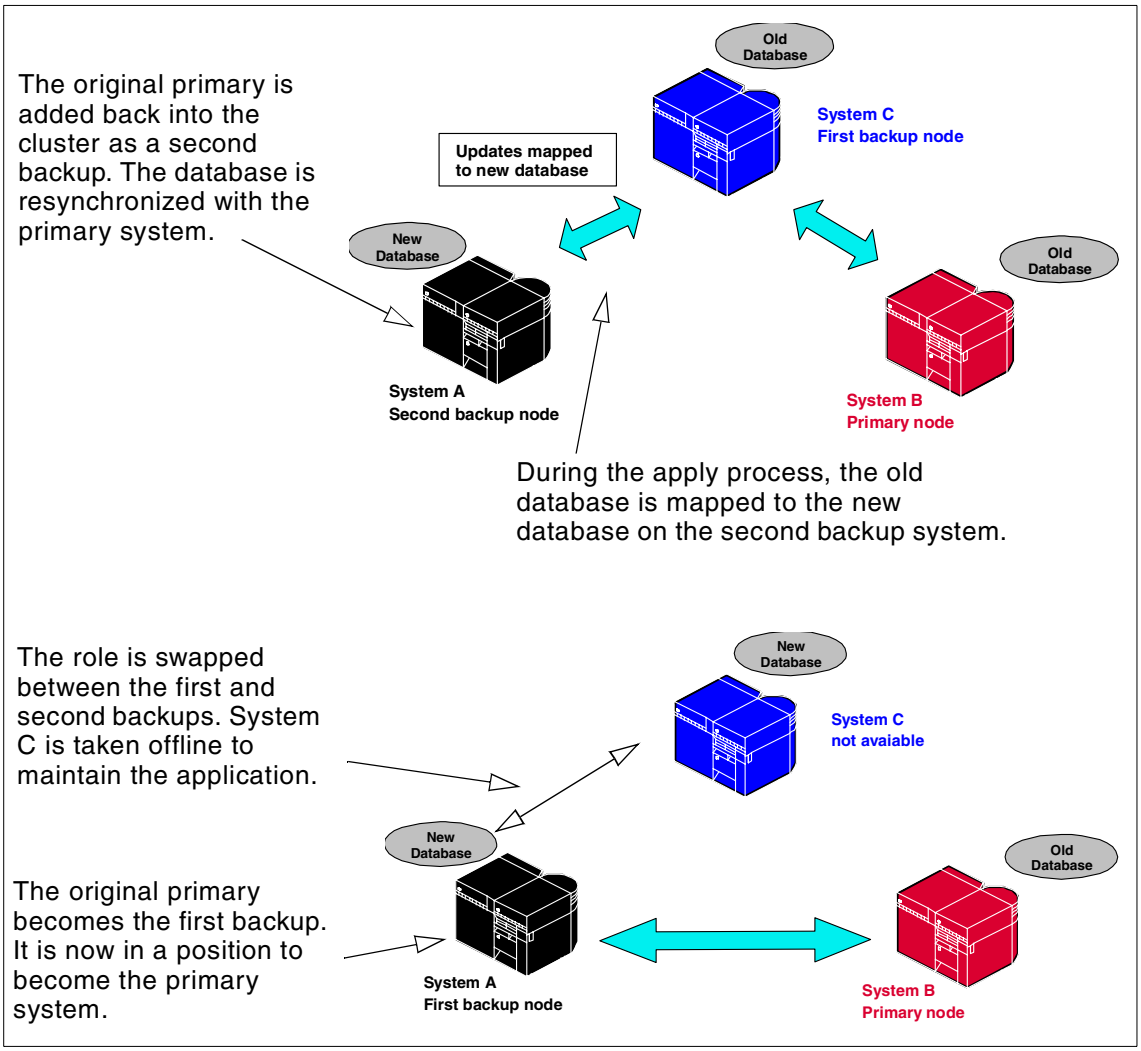


Figure 23. Application maintenance in a cluster (Part 2 of 4)

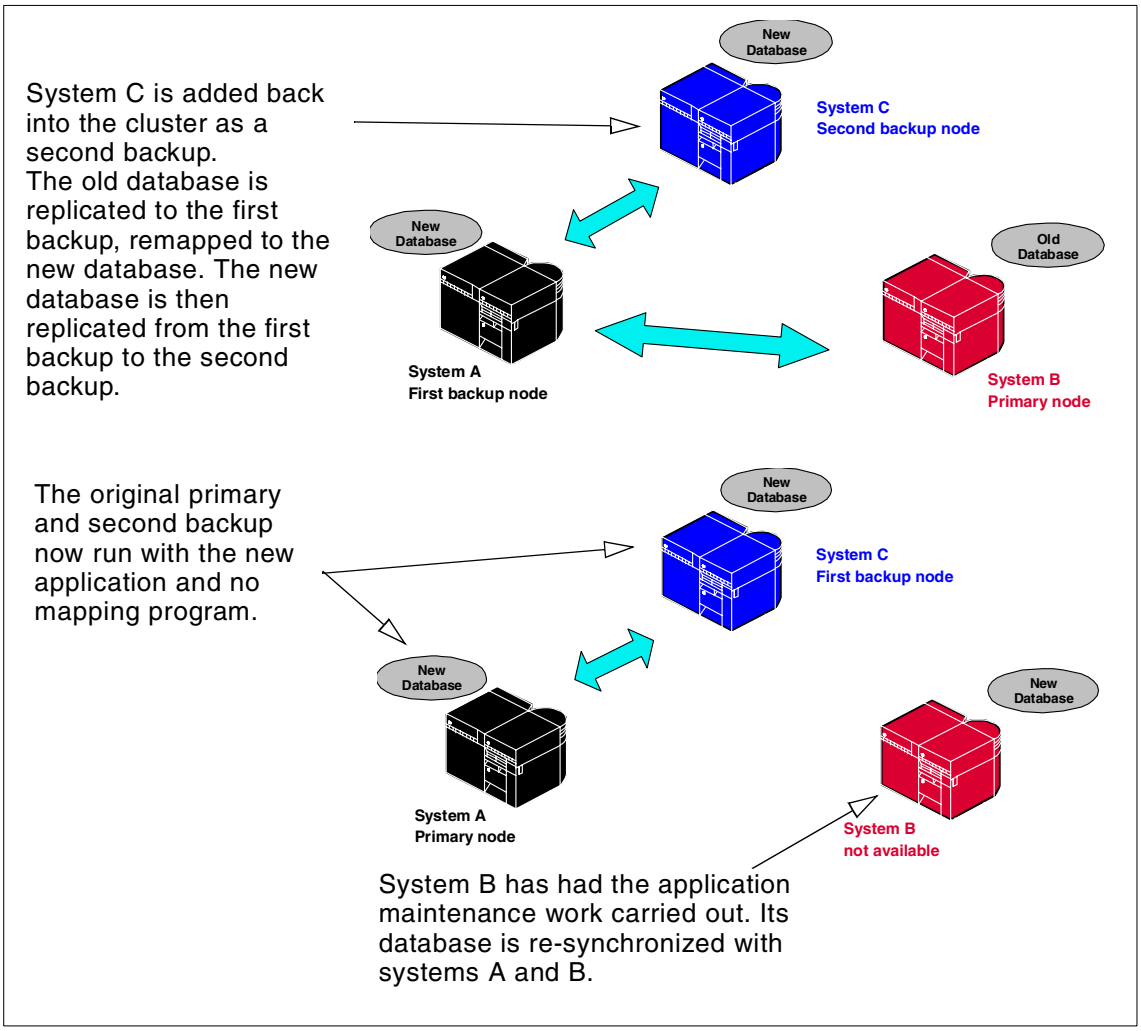


Figure 24. Application maintenance in a cluster (Part 3 of 4)

In the final phase, system B is added back into the cluster as the second backup. Then, its role is switched to the first backup, and return the cluster to its original state.

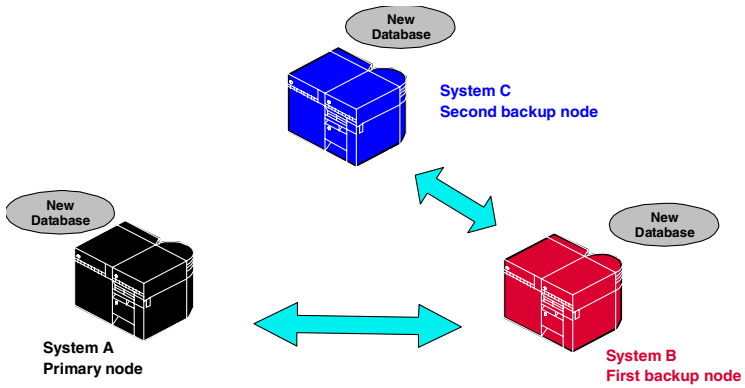


Figure 25. Application maintenance in a cluster (Part 4 of 4)

6.3.9 Database performance

Database performance is not necessarily related to creating a clustered solution. If you are already running replication software, you will understand whether you have a database performance problem and the resolution is no different than if the problem occurred on a single system. Adding journaling and commitment control to a poorly designed database may make things run slower. If this is not normally a journaling or commitment control problem, it's bad database design. Technology improvements in OS/400 and the AS/400 hardware have minimized performance degradation with journaling and commitment control. This is an area that will constantly be improved with later releases.

However, you do not have the significant recovery options made available with journaling and commitment control without some compromise in performance. See *OS/400 Backup and Recovery V4R4*, SC41-5304, for more information on how to implement journaling and commitment control. Contact your ISV for more information on the recovery options implemented in their application.

Many applications are re-developed and customized as a customer's business changes and new data is added or required from the database. Unfortunately few developers go back to the original design to make sure they are not over-normalizing the database. It is better to handle the badly designed database early on in the project and certainly before the strategic solution has been implemented.

6.3.10 HABP selection

There are three High Availability Business Partners to chose from, unless you plan to write you own journaling environment, which we do not recommend. The choice could depend on the relationship between the HABP and your application provider. We do not decide which is the best or do a feature-by-feature comparison. You must review them yourself. Fortunately all the business partners are prepared to provide test versions of their applications or give demonstrations.

6.4 Systems management

System management is another critical area for planning. If you have no SM disciplines in place, make sure you have them before you start a clustering project. You will have great difficulty managing your cluster solution without this backbone of policies.

6.4.1 Service level agreements

When considering service level agreements, start with the business executives. They are your key sponsors for developing the right service level agreements. They are also the source of your financial case to the finance department. Hopefully most of this planning has been done when building the financial case for the availability solution.

6.4.2 Operations management

With a single system environment, the operations staff has had a fairly well-oiled management system. Daily run schedules and overnight batches are well understood, and there is normally a simple but long running recovery process.

In a cluster environment, you may run a more lights-out operation. This does not mean that you will reduce staff, but you will certainly give the existing operations group a great opportunity to investigate new skills. Once you have the cluster planned, you should then overlay the operational environment and fit your resources to this plan.

Once the operational plan is finalized, you need to establish a cutover plan from the existing environment to the new environment. This may mean changing shift patterns and possible taking on extra staff to cover tasks that could run over weekends.

The skills and resources that are required to implement and manage a cluster solution are different from most current I/T departments. Critical skills are

database management and networking. Most customer shops have programming skills and some networking skills. As the mission critical application starts, you must be able to remedy the problem very quickly to meet service level agreements. Having the right skills will help this enormously.

Once you have done all of this, you must update or re-write your operational documentation.

6.4.3 Problem and change management

Establishing an effective problem and change management strategy is critical. Assuming there are problem and change management processes in place, these need to be modified for the new cluster environment. These modifications need to be in tune with the service level agreements.

Reporting and analyzing problems needs to be carried out as quickly as possible. There will no longer be time for waiting for a key person to return a call. The escalation process also needs upgrading to make fast informed decision. For example, a disk reports errors and its impending failure. Waiting for a CE to call is no longer an option. The disk should be pumped, and if the system will be degraded too much, a planned switch should be considered.

Each system throughout of the business should be analyzed and risk assured. The system should be prioritized and failover plans should be documented. This documentation should include decision points with directions on which step to take next.

For example, a processor sends a warning of an overload. One of the applications needs to be switched to another processor that has spare capacity. The applications on the overloaded system should be prioritized. A predefined plan is implemented. A planned switch is initiated for the application using the HABP Cluster management tool. The exit programs cause the CRGs, IP address, and users to be switched to the backup system. This application remains on the backup system until the primary processor has been analyzed and the problems rectified. Then the application can be switched back to the primary system.

Many of the normal problems of maintaining a high available single system will be automated or disappear.

Now, new problems exist and need to be addressed:

- Partition state
- Application maintenance

- Managing a mix of applications with a different resilience characteristics
- Managing batch applications in a cluster

6.4.4 Capacity

How do you size your systems for a cluster? This process is relatively simple. The easiest way to size your cluster is to plan it as a single system and then model the load produced by journaling and the apply process. BEST/1 can be used to model these functions and provide a relatively accurate model of the single node. Cluster functions cause minimal overhead to system resources.

This planning process can become more complex depending on the roles of the nodes and the applications that may be run in different conditions. For example, you may plan to have one system as a primary node for an application in one recovery domain. But, under a failure or switchover, this node is planned to support another application. You must also include this additional application in your planning process. Each node must be reviewed and the worst case scenario should be modelled. You must make sure that the capacity of other components related to the node are also planned. Some examples of these elements are I/O processors, network connections, and backup devices.

Once the cluster is up and running, you must regularly monitor and re-evaluate the capacity of all the nodes. It is very dangerous to allow the resources on the backup nodes to become overstretched. If a failure or switchover was necessary, the additional load on the backup machine could create its own availability problem.

6.4.5 Performance

Performance considerations for clusters are similar to capacity planning for clusters. Measurements must be taken regularly to monitor that service levels are be achieved. Any out-of-line situations should be registered and corrective action taken should be in place to reduce the risk of losing the integrity of the cluster. When switchover occurs for routine maintenance, recordings should be taken from the backup node to ensure that it still meets the performance expectations and that the backup node has not been degraded by growth in its workload.

6.4.6 Security

When clustering a multi-system or multi-company environment, additional security may be required. In a switch or failover to the backup system, there may suddenly be many users on this system. Maintaining the isolation

between discreet groups is important, especially where Internet access is involved.

A thorough review of all the possible user or application grouping should be made during the planning process.

6.5 Hardware

As we already mentioned, hardware is relatively inexpensive when compared to the cost of failure. Do not simply look at the hardware related to the computer system. There are many other hardware components of a complete continuously available solution.

6.5.1 Redundancy

In a continuously available system complex, providing a configuration that includes redundant hardware is also relatively easy. However, this adds complexity to overall system management.

Consider the following items when planning the total solution:

- Processor complex redundancy (includes bus redundancy)
  - Disk redundancy
  - Adapter redundancy
  - Remote site redundancy

  This could be a complete remote location or a remote controller at the main site.
- Site redundancy
  - Machine room
  - Air conditioning
  - Power supply
  - Office space
  - Telephone services
- Network hardware redundancy

  From routers to remote controllers, you need to review all network hardware. If there are critical requirements for remote access, you must provide alternative network paths. This could be as simple as a dial-up link or as complex as a multi-path private network.
- Network connection redundancy

  If you are planning to extend your services to an intranet or the Internet, you are probably looking at one ISP. This gives you another single point of

failure. It is important to have multiple network providers if you want the highest availability.

6.5.2 Network planning

When network planning, you look at capacity and accessibility. The network must be able to maintain the same level of availability as the cluster nodes. Communications providers must give guarantees that they will be available and have sufficient capacity for all possible switch scenarios. There must be alternative network paths to enable the cluster services to manage the cluster resources. These redundant paths should prevent a cluster partition occurring. See C.3, “Recovering from a clustered partition” on page 148, for more information on cluster partitions.

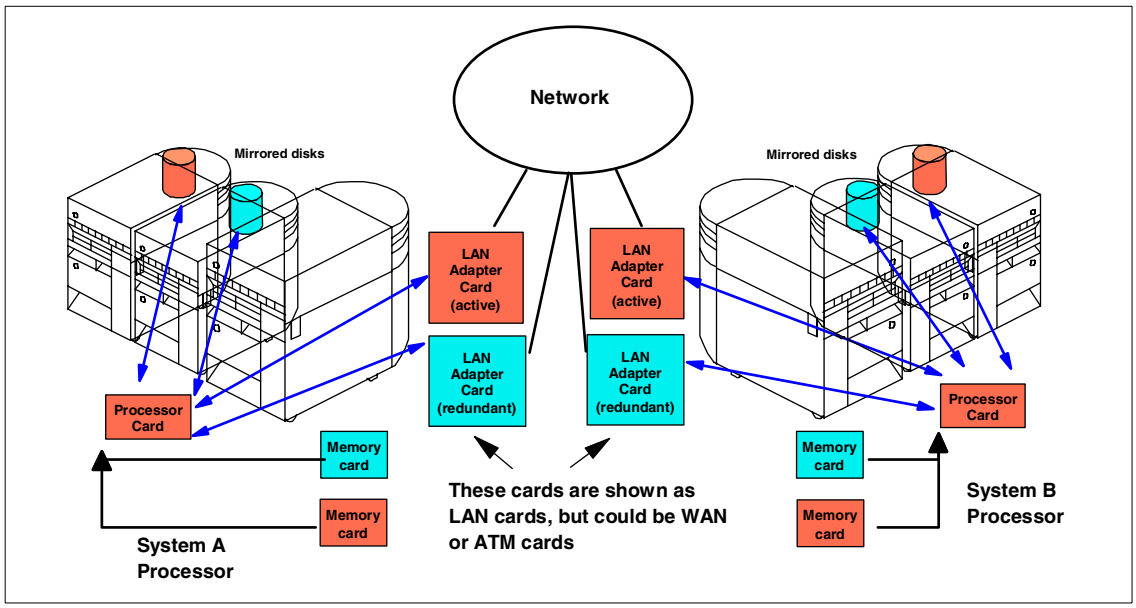


Figure 26. Redundant network connections

6.6 Cluster testing

Testing is a well-established and controlled part of most development shops. Customers often have a second system or separate LPAR partition for their operating system and application test environments. Unfortunately, this testing is not always extended to all facets of the business. Network, hardware, and external link testing are often over looked.

A classic example of this is Web page design. The development shop builds and tests the page on the internal LAN, places it onto the Web, and tests it from their LAN out to the Web and back into the Web site. They typically use very high speed links. Then the application is not tested remotely from, for example, a 14.4 Kbps modem on the other side of the world. When the application is made available, it takes an extended time to load and visitors don't bother to use the site.

Whether you already implemented a highly available solution or you are planning to build a cluster, you must consider your implementation test environment and your ongoing problem or change test environment. Figure 27 shows how a simple customer setup can be changed to produce a more effective test environment.

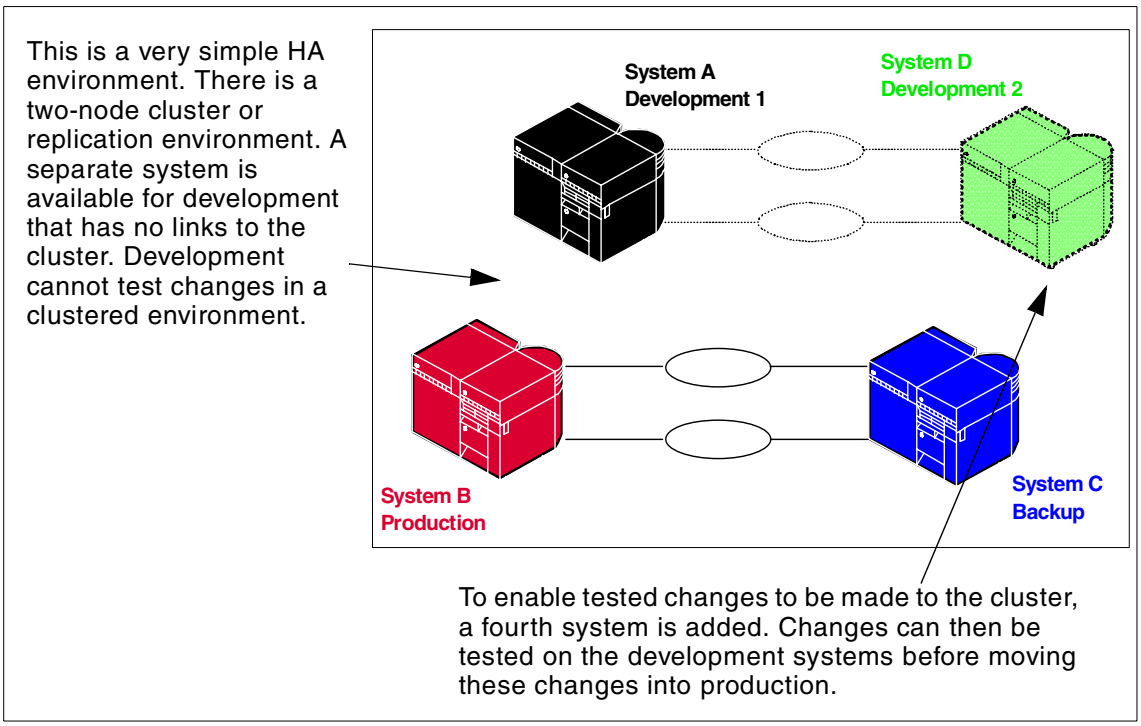


Figure 27. Cluster test scenario

The test environment shown in Figure 27 shows the basic systems. It does not show the routes, LANs, WANs, etc. that need to simulate the real environment. One of the first considerations you must make is whether your development environment is a critical resource. If it is, the development system should be included in the cluster, and you need a totally separate

cluster test configuration. Including development in your simple cluster can produce a configuration similar to the example shown in 6.3.8, “Maintenance” on page 63.

Creating a separate cluster with two small systems would meet most of the needs for testing. The only issues with this arrangement would be the possibility that certain types of hardware and peripherals may not work with very small systems, and it would be difficult to do any accurate volume testing.

6.6.1 General system management-related tests

System management testing is mainly aimed at performance and operations:

- Application process testing (normally part of development testing)
- Application volume testing
- Hardware and peripheral testing (tape, DASD, IOPs, remote devices, clients)
- Interoperability testing
- Network performance testing
- Network hardware testing

When performing volume-related testing, it is important to have a well-documented script for producing the test. If the capacity is not available on the local test machines, you may have to consider an external testing source, for example, one of the IBM Benchmark Centers (`http://www.partnerworld.ibm.com`). Many large businesses regularly use the Benchmark Center for pre-production testing of their applications.

6.6.2 Cluster management-related tests

Some of the scenarios that should be tested before moving your cluster into production are shown in the following list. These scenarios should also be re-tested after a system upgrade or major change to any of the cluster components:

- Planned switch
- Failover
- Rejoin
- Adding a cluster node

In summary, testing is as critical as any application on the system. If your cluster is running for six months, you make a change, and then at a critical moment find, it will not fail over, you will have a large problem. Since the cost of the test systems is trivial, do not try to save money in this area.

Part 2. High Availability Business Partners

This part showcases each cluster management utility provided by the three IBM High Availability Business Partners:

- DataMirror
- Lakeview Technology
- Vision Solutions

Chapter 7. DataMirror iCluster

DataMirror's iCluster product provides a set of easy-to-use interfaces to set up and manage an IBM AS/400 cluster for high availability. iCluster also provides replication support for data Cluster Resource Groups (CRGs) and resilient applications using DataMirror HA Suite. DataMirror HA Suite is a proven high performance real-time replication utility for AS/400 data and objects.

iCluster provides three interfaces for AS/400 cluster management. All three can be used interchangeably and provide a consistent view of the cluster. The three iCluster interfaces are:

- An AS/400 green-screen menu interface
- A Java graphical user interface (GUI) client running on a PC or workstation
- A full set of AS/400 commands for cluster setup and management

In addition to cluster management and data and object replication, iCluster allows you to:

- Check whether your objects and data are synchronized across two systems
- Monitor replication processes
- Stop and start replication apply processes while continuing the replication journal scrape processes
- Define synchronization points in a replication process, with optionally specified user exits to be executed when a synchronization point is reached
- Define user exits to be executed automatically before or after a group switchover or failover (failure of a group's primary node)
- Define message queues where messages will be placed by iCluster in the event of a failure of a group's primary node

The basic steps that have to be performed to set up and run a cluster using DataMirror iCluster are explained in the following sections. Each step is illustrated with an example from one of the interfaces.

7.1 Getting started with iCluster

Once your system administrator installs DataMirror iCluster on the nodes that will form to your cluster and, optionally, the iCluster GUI interface (called iCluster Administrator) on your PC or workstation, you can set up a cluster.

If you are using the AS/400 menu interface, on any AS/400 command line, type:

```
GO <iCluster library name>/DMCLUSTER
```

Press Enter. You see the iCluster main menu as shown in Figure 28.

```
                        DataMirror iCluster Main Menu
                                                          System:   S100A94R
Select one of the following:

  System
     1. Work with nodes
     2. Work with groups
     3. Work with resilient applications
     4. Display event log
     5. Clear event log
     6. Set cluster-wide system values

  Operations
    11. Start clustering
    12. End clustering
    13. Start cluster operations at a node
    14. End cluster operations at a node
    15. Start cluster operations for a group
                                                                    More...
Selection or command
===> _____________________________________________________________________

F3=Exit   F4=Prompt   F9=Retrieve   F12=Cancel   F22=DM Cmds
```

Figure 28. iCluster main menu

If you are using the iCluster Administrator on your PC or workstation, you are asked to log in with your AS/400 user ID and password. Then you are presented with the iCluster Administrator main window, which is shown in Figure 29.

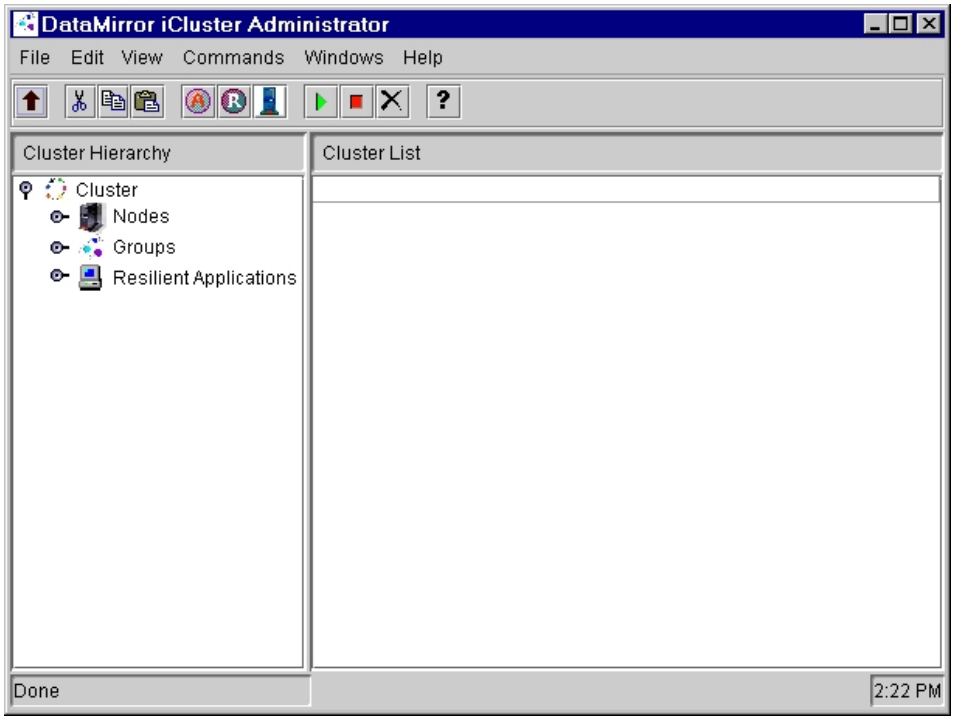


Figure 29. iCluster Administrator main window

7.2 Creating a cluster

After you define the first node in the cluster, you have created a cluster. iCluster automatically activates (start) each node as it is added to the cluster. Nodes can be de-activated (ended) and re-activated (re-started) at any time.

The cluster's first node must be defined as the AS/400 system that you are currently using. Other nodes in the cluster must be defined from a system that is already an active node in the cluster. If you define a node from a system that is not a node in the cluster, you create a new cluster with that system as its first node.

The first node defined in the cluster becomes its *master node*. The master node is responsible for maintaining the information that iCluster needs for data and object replication. This information has to be maintained on all the nodes in the cluster. That way, in the event of a failure or removal of the master node, any other node can automatically assume the role of the master node. For this reason, the master node, or any node that can potentially

become the master node, must be directly accessible to all the nodes of the cluster via the TCP/IP interface given when each node is defined.

7.2.1 Adding a node to the cluster

A node can be defined using the Add option on the iCluster Work with Nodes screen, the DataMirror (DM) iCluster Add Node (DMADDNODE) command (Figure 30), or the iCluster Administrator Add Node input dialog (Figure 31).

```
                      DM iCluster Add node (DMADDNODE)

Type choices, press Enter.

Node . . . . . .   S100A94R      Name
IP Address . . . . . .   126.1.1.31
                                                                    +
                                                                    +
                                                                    +
Alternate IP Address . . . . . .                                    +
                                                                    +
                                                                    +
                                                                    +
Port . . . . . .   4545          1-65535
DM iCluster product library  . .   DMCLUSTER     Character value
Description  . . . . . .                                            +
                                                                    +
Job description  . . . . . .   *CLUSTER      Name, *CLUSTER
  Library  . . . . . .                       Name
User profile status  . . . . . .   *CLUSTER      *CLUSTER, *PRIMARY, *NOCHG...
                                                                  More...
F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
F24=More keys
```

Figure 30. The iCluster Add node input screen on the AS/400 system

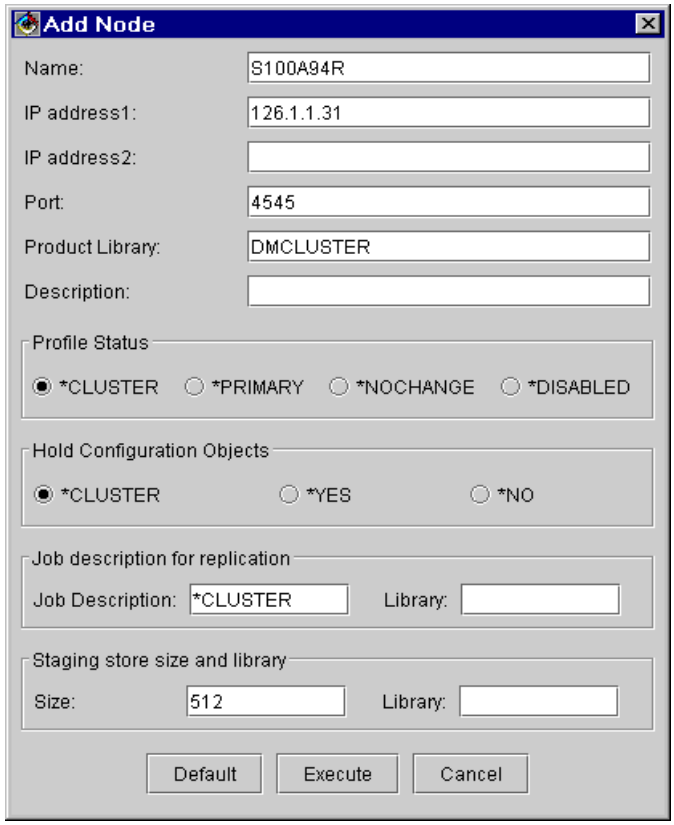


Figure 31. The iCluster Administrator Add Node input dialog

The complete list of nodes in the cluster and their current status can be easily viewed by expanding the node list in the iCluster Administrator main window. You can also view them with the Work with nodes option from the iCluster main menu on the AS/400 system (Figure 32 on page 82).

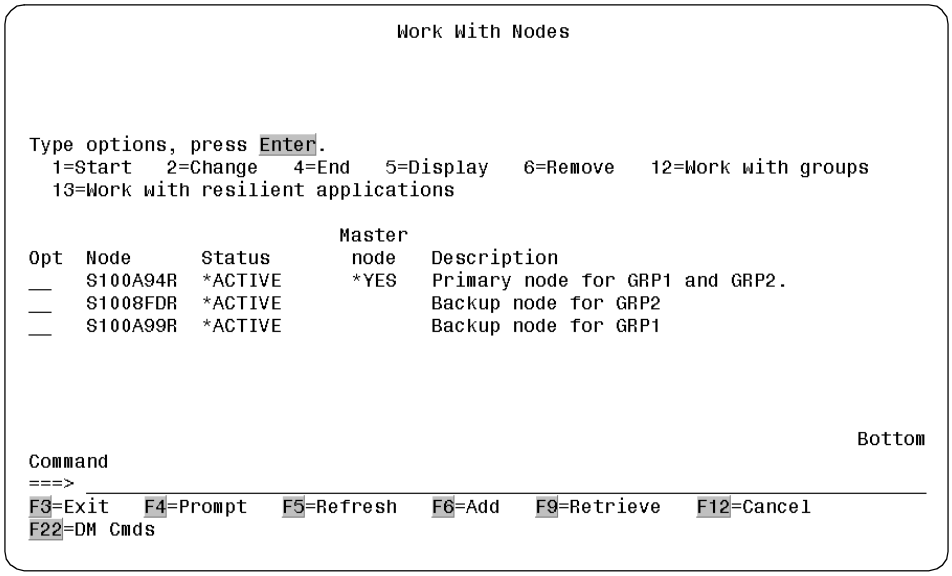


Figure 32. The iCluster Work With Nodes screen

7.2.2 Activating and de-activating nodes in the cluster

Nodes can be activated by selecting option 1 (Start) on the Work With Nodes screen. When the node becomes active, its status is shown as *ACTIVE. Alternatively, you can activate a node with the DM iCluster Start Node (DMSTRNODE) AS/400 command, for example:

```
DMSTRNODE NODE(S100A99R)
```

Nodes can be de-activated by selecting option 4 (End) on the Work with Nodes screen. When the node becomes inactive, its status is shown as *INACTIVE. Alternatively, you can de-activate a node with the DM iCluster End Node (DMENDNODE) AS/400 command, for example:

```
DMENDNODE NODE(S100A99R)
```

Here, *S100A99R* is the name of a node in the cluster.

If you want to de-activate all the nodes in the cluster, use the `DMENDCLSTR` command with the ENDNODES parameter set to *YES, for example:

```
DMENDCLSTR ENDNODES(*YES)
```

7.3 Creating and using Cluster Resource Groups (CRGs)

In V4R4, there are two types of Cluster Resource Groups:

- Data CRGs
- Application CRGs

The initial release of DataMirror iCluster with V4R4 of OS/400 allows you to create both data CRGs and application CRGs that have either one node (the primary) or two nodes (a primary and a backup) in their recovery domain.

Using two node groups, you can create more complex cluster scenarios. For example, you can set up a cluster consisting of a single primary node with two or more backup nodes. Simply create as many data CRGs as there are backup nodes, all with the same primary node, and select the same object specifiers to all the CRGs (Figure 33).

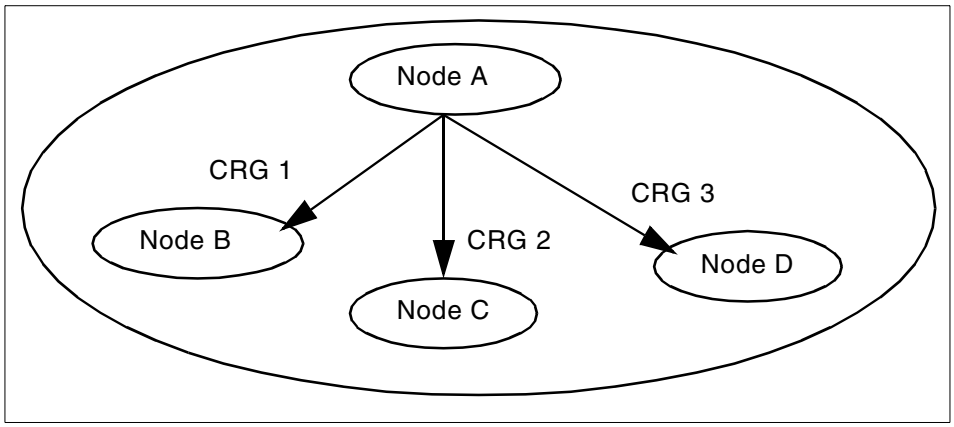


Figure 33. Cluster with a single primary node and three backup nodes

7.3.1 Creating data CRGs

You can create data CRGs in two ways:

- Use the DM iCluster Add Group (DMADDGRP) command, the Add option from the iCluster Work With Groups screen, or the iCluster Administrator Add Group input dialog. This creates a data CRG in the cluster.
- Use the DM iCluster Add Group (DMADDAPP) command, the Add option from the iCluster Work With Resilient Applications screen, or the Add option from the iCluster Administrator Resilient Applications window. This sets up a resilient application that contains one or more data CRGs.

Use the first approach when you have specific, known high-availability requirements for data and objects. This approach allows you to directly select the objects that you require for high availability.

The second approach is primarily intended for setting up a ClusterProven resilient application on your cluster. ClusterProven resilient applications are provided by application vendors who have enabled their applications for clustering. See 7.4, “Using ClusterProven applications” on page 91.

The remainder of this section deals with data CRGs created using the first approach, that is, as individual groups not associated with any resilient application.

Figure 34 shows the input required to create a data CRG with the Add option from the Work With Groups screen on the AS/400 system.

```
                    DM iCluster Add group (DMADDGRP)

Type choices, press Enter.

Group  . . . . . . . . . . . . . > GRP1          Name
Recovery domain source . . . . .   *LIST         Character value, *LIST
Primary node . . . . . . . . . .   S100A94R      Name
Backup nodes . . . . . . . . . .   S100A99R      Name
               + for more values   _____
Replicate nodes  . . . . . . . .   *NONE         Name, *NONE
               + for more values   _____
Failover message queue name  . .   *NONE         Name, *NONE
  Library  . . . . . . . . . . .     __________  Name
Do role switch at failover . . .   *YES          *YES, *NO
User exit before role switch . .   *NONE         Name, *NONE
  Library  . . . . . . . . . . .     __________  Name
User exit after role switch  . .   *NONE         Name, *NONE
  Library  . . . . . . . . . . .     __________  Name


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

Figure 34. The iCluster Add group input screen

You can view the groups in the cluster and their current status with either the iCluster Work With Groups screen (Figure 35) or the iCluster Administrator Groups window.

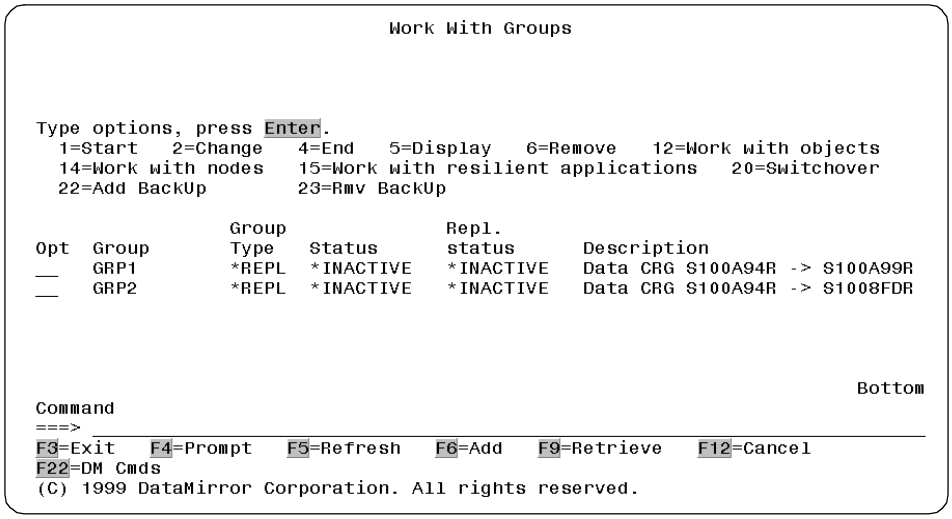


Figure 35. The iCluster Work With Groups screen

Note that CRGs are not automatically activated (started) when they are created. A group will remain in *INACTIVE status until you activate it. Note also that data CRGs (group type *REPL) have a second status value: the *replication status*. The replication status indicates whether replication processes are active for the group. Normally, if a data CRG is in *ACTIVE status, its replication status should also be *ACTIVE.

7.3.2 Selecting objects for a data CRG for high availability

After defining a data CRG, you can select the objects required for high availability with either the DM iCluster Select Object (DMSELOBJ) AS/400 command, the Select option on the iCluster Work With Object Specifiers By Group screen from the iCluster Work With Groups screen, or the iCluster Administrator Select Objects input dialog. The DM iCluster Select Object screen is shown in Figure 36 on page 86.

```
                    DM iCluster Select Object (DMSELOBJ)

Type choices, press Enter.

Group  . . . . . . . . . . . . . > GRP1          Name
Object . . . . . . . . . . . .   *ALL          Name, generic*, *ALL
  Library  . . . . . . . . . .     IVCS1        Name
Object type  . . . . . . . . .   *FILE         *ALL, Supported type
Object extended attribute  . . .  *ALL         Name, *ALL, CMNF38, DDMF...
Description  . . . . . . . . .   ____________________________________________
_____
Include or exclude . . . . . .   *INCLUDE      *INCLUDE, *EXCLUDE
Object polling interval  . . .   *GROUP        000010-235959, *GROUP...









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

Figure 36. The iCluster Select Object input screen

The objects that match the object specifiers selected to the group are replicated from the group's primary node to its backup node when you activate (start) the group. You can easily see which object specifiers have been selected to a particular group by using the Show Selected function (F16) on the Work with Object Specifiers By Group screen (Figure 37).

```
                    Work With Object Specifiers By Group

Group . . . :   GRP1
Description :   Data CRG S100A94R -> S100A99R

Type options, press Enter.
  2=Change Selection   4=De-select   5=Display

Opt  Group       Inc  Object      Type        Library     Attribute
 _   GRP1        INC  *ALL        *DTAARA     IVCS1       *ALL
 _   GRP1        INC  *ALL        *DTAQ       IVCS1       *ALL
 _   GRP1        INC  *ALL        *FILE       IVCS1       *ALL









                                                                       Bottom
Command
===> ____________________________________________________________________________
F3=Exit   F4=Prompt   F5=Refresh   F6=Select   F9=Retrieve   F12=Cancel
F16=Show Selected     F22=DM Cmds
(C) 1999 DataMirror Corporation. All rights reserved.
```

Figure 37. iCluster Work With Object Specifiers By Group screen

Note that the objects to be replicated do not have to exist when the object specifier is selected to the group. iCluster's real-time auto-registration technology can detect when an object that matches the specifier is created and will begin replicating the object as soon as it is created.

You can remove (de-select) object specifiers from a data CRG with either the DMDSELOBJ AS/400 command, the De-select option on the iCluster Work With Object Specifiers By Group screen, or the iCluster Administrator Deselect Object input dialog.

Note: You cannot directly select or de-select object specifiers from CRGs that are part of a resilient application. See 7.3, "Creating and using Cluster Resource Groups (CRGs)" on page 83, for more information.

7.3.3 Creating application CRGs

Application CRGs are created according to the specification in the QCSTHAAPPI automated installation data area architected for application resiliency by IBM. See 7.3, "Creating and using Cluster Resource Groups (CRGs)" on page 83, for details.

7.3.4 Changing a CRG recovery domain

With DataMirror iCluster, you can initially define a CRG with either one node (the primary) or two nodes (primary and backup) in its recovery domain. If you define the CRG with a primary node, you can add a backup node later before activating the group. You can add and remove backup nodes as necessary when the group is inactive.

You can view a group's current recovery domain by selecting option 5 (Display) on the iCluster Work With Groups screen (Figure 35 on page 85) or selecting Groups and Display details in the DataMirror iCluster Administrator window (Figure 29 on page 79). Figure 38 on page 88 shows the DM iCluster Display Group screen.

```
                    DM iCluster Display Group (DMDSPGROUP)

Type choices, press Enter.

Group  . . . . . . . . . . . . > GRP1
Description  . . . . . . . . .   'Data CRG S100A94R -> S100A99R
  '
Primary node . . . . . . . . .   S100A94R
Backup node  . . . . . . . . .   S100A99R
Failover message queue name  . .   *NONE
  Library  . . . . . . . . . . .
Do role switch at failover . . .   *YES
User exit before role switch . .   *NONE
  Library  . . . . . . . . . . .
User exit after role switch  . .   *NONE
  Library  . . . . . . . . . . .
Group polling interval . . . . .   000500
Max. objects per savefile  . . .   *MAX
Save active objects  . . . . . .   *NO
Max. wait for spool files  . . .   000015
                                                                    More...
F3=Exit   F4=Prompt   F5=Refresh   F12=Cancel   F13=How to use this display
F24=More keys
```

Figure 38. The iCluster Display Group output screen for a data CRG

If a CRG has only one node in its recovery domain (by definition the primary node), you can add a backup node with the DM iCluster Add backup Node (DMADDBACK) AS/400 command, the Add Backup option on the Work With Groups screen, or the iCluster Administrator Add Backup input dialog. Figure 39 shows the DM iCluster Add Backup Node screen.

```
                    DM iCluster Add Backup Node (DMADDBACK)

Type choices, press Enter.

Group or resilient application   > GRP1           Name
Node . . . . . . . . . . . . . .   S1008FDR       Name
Backup role  . . . . . . . . . .   *LAST          0-127, *PRIMARY, *FIRST...









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

Figure 39. The iCluster Add Backup Node input screen

If the CRG has two nodes in its recovery, you can change the backup node by removing it and adding another node as the backup. To remove the existing

backup, use either the DMRMVBACK AS/400 command, the Remove Backup option on the iCluster Work With Groups screen, or the iCluster Administrator's Remove Backup dialog. Figure 40 shows the DM iCluster Remove Backup Node screen.

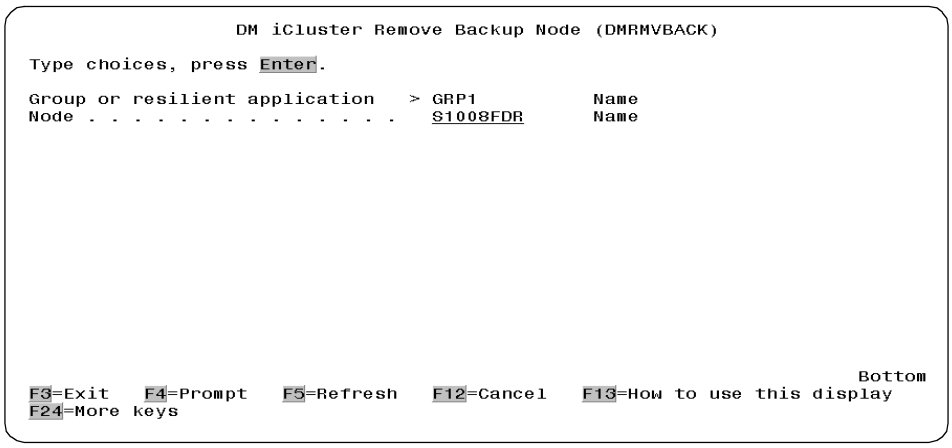


Figure 40. The iCluster Remove Backup Node input screen

Note that you cannot directly change a CRG's primary node. To change the primary node, you can perform a switchover on the group so that the current backup node becomes the primary node (see 7.3.7, "Switching over a data CRG" on page 90). Or, you can re-define the group with a different primary node.

7.3.5 Activating or starting a data CRG

You can activate (start) a data CRG that is not part of a resilient application with the DM iCluster Start Group (DMSTRGRP) AS/400 command, the Start option on the iCluster Work With Groups screen, or the iCluster Administrator Start Group input dialog.

Note: You can activate application CRGs and data CRGs that are part of a resilient application by activating the resilient application with which they are associated. See 7.4.5, "Activating or starting a resilient application" on page 95.

Once the CRG is activated, its status changes to *ACTIVE. If the CRG has objects or data selected to it, replication of the objects or data from the primary node to the backup node begins and the group's replication status changes to *ACTIVE. Replication activation typically takes longer than CRG activation due to the number of jobs that have to be started.

7.3.6 De-activating or ending a data CRG

You can de-activate or end a data CRG that is not part of a resilient application by using either the DM iCluster End Group (DMENDGRP) AS/400 command, the End option on the iCluster Work With Groups screen, or the iCluster Administrator End Group input dialog.

Note: You can de-activate application CRGs and data CRGs that are part of a resilient application by de-activating the resilient application with which they are associated. See 7.4.6, “De-activating or ending a resilient application” on page 95, for more information.

7.3.7 Switching over a data CRG

Switchover is the process of interchanging the primary and backup roles of a CRG’s recovery domain and changing the direction of object and data replication in a data CRG. You can switch over an active data CRG that is not part of a resilient application with the DM iCluster Start Switch Over (DMSTRSWO) AS/400 command, the Start Switch Over option on the iCluster Work With Groups screen, or the iCluster Administrator Switch Over Group input dialog.

Switching over a group may not happen instantaneously, particularly if large amounts of objects and data are being replicated by the group. Other factors that can increase the time required for switchover to complete are:

- Latency in the apply processes on the backup node
- Switchover user exit processing
- Starting journalling of database files on the new primary node, particularly if many files need to be journaled
- Setting up trigger programs and enabling database file constraints on the new primary node

While a group is undergoing the switchover process, you see the group’s status displayed as *SWO_PENDING. When switchover is complete and the group starts replicating in the opposite direction, the group’s status reverts to *ACTIVE.

Note: You can switch over application CRGs and data CRGs that are part of a resilient application by switching over the resilient application with which they are associated. See 7.4.7, “Switching over a resilient application” on page 95.

7.4 Using ClusterProven applications

AS/400 independent software vendors (ISVs) can enable their applications for clustering on the AS/400 system. Applications that have been enabled for clustering and have passed the IBM cluster-enablement certification test are called *ClusterProven applications*. iCluster provides a simple interface for setting up ClusterProven applications in your cluster.

The result of setting up a ClusterProven application for clustering is called a *resilient application*. This is basically a set of application CRGs, data CRGs, object specifiers selected to the data CRGs, and a takeover IP address that can be reassigned to the new primary node for the application in the event of a switchover or failure of the application's primary node. Note that all the CRGs in the resilient application have the same recovery domain.

You can set up, start, end, switch over, change, update, and remove resilient applications with DataMirror iCluster. Note that you cannot add or remove CRGs from a resilient application once the application has been set up. Nor can you perform any cluster operation on the groups that comprise the resilient application on their own, but only as part of the resilient application.

7.4.1 Setting up a resilient application

You can set up a resilient application with either one node (the primary) or two nodes (a primary and a backup) in its recovery domain with DataMirror iCluster. The other items you need to know to set up a resilient application are:

- The takeover IP address of the resilient application.
- The name of the ClusterProven application's installation library on your systems. This library must exist on all nodes of the application's recovery domain and must contain a data area named QCSTHAAPPI, which defines what CRGs are to be created for the resilient application and what object specifiers are to be selected to the application's data CRGs. This data area is provided by the application vendor.

You set up a resilient application with the DM iCluster Add Application (DMADDAPP) AS/400 command, the Add option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Add Resilient Application input dialog. Figure 41 on page 92 shows the DM iCluster Add Application screen.

```
                    DM iCluster Add Application (DMADDAPP)

 Type choices, press Enter.

 Application name . . . . . > APP1          Name
 Application data library . . . . > APP1LIB     Name
 Takeover IP address  . . . . . > '126.1.1.55'
 Recovery domain source . . . . .   *LIST        Character value, *LIST
 Primary node . . . . . .  .  .  .  S100A99R     Name
 Backup nodes . . . . . .  .  .  .  S100A94R     Name
                + for more values
 Replicate nodes  . . . . . .  .  .  *NONE       Name, *NONE
                + for more values
 Description  . . . . . .  .  .  .  Resilient app for APP1
_______


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

Figure 41. The iCluster Add Resilient Application input screen

After a resilient application is created, you can see it on the iCluster Work With Resilient Applications screen (Figure 42) or the iCluster Administrator Resilient Applications window.

```
                        Work With Resilient Applications



 Type options, press Enter.
   1=Start   2=Change   3=Update        4=End   6=Remove   12=Work With Groups
   14=Work with nodes   20=Switchover   22=Add BackUp      23=Rmv BackUp

                                      Takeover
      Application  Primary   First     IP
 Opt     name      node      backup    address          Description
 __   APP1         S100A99R  S100A94R  126.1.1.55       Resilient app for APP1




                                                                        Bottom
 Command
 ===> ____________________________________________________________________________
 F3=Exit   F4=Prompt   F5=Refresh   F6=Add   F9=Retrieve   F12=Cancel
 F22=DM Cmds
```

Figure 42. iCluster Work With Resilient Applications screen

You can view the list of groups that are associated with an application by choosing the Work With Groups option on the iCluster Work With Resilient Applications screen. This displays the Work With Groups By Resilient Application screen (Figure 43).

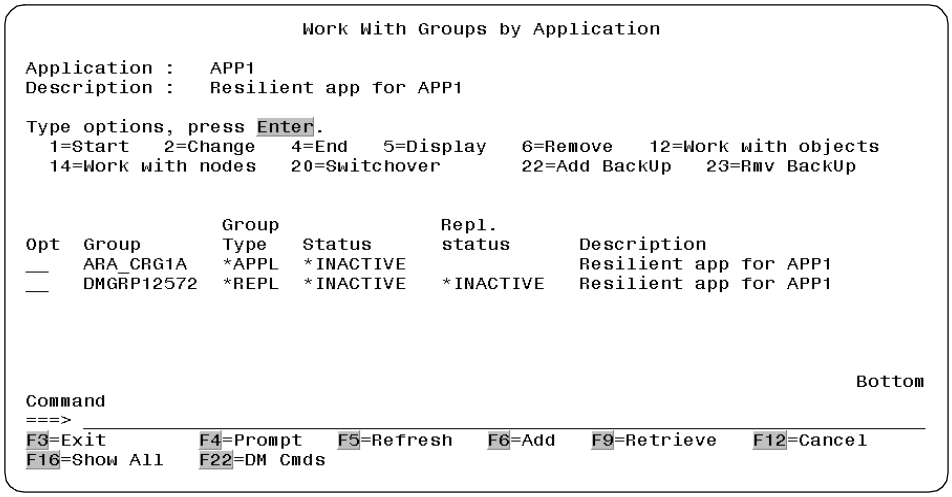


Figure 43. iCluster Work With Groups by Application screen

This screen also displays the status of the groups associated with the resilient application. Note that the replication status field of group ARA_CRG1A is blank. This indicates that ARA_CRG1A is an application CRG (type *APPL), not a data CRG (type *REPL).

7.4.2 Selecting objects to a resilient application

You do not have to select object specifiers to a resilient application or its associated groups. The object specifiers required for a resilient application are listed in a file that is named in the QCSTHAAPPI data area for the application. iCluster reads this file when defining the resilient application and automatically selects the object specifiers to the appropriate data CRGs that are associated with the resilient application.

Similarly, you do not have to de-select object specifiers from a resilient application or a group that is associated with a resilient application. This is done automatically by iCluster when the application is updated or removed.

7.4.3 Changing or updating a resilient application

You can change a resilient application's takeover IP address and its description directly with the DM iCluster Change Application (DMCHGAPP) AS/400 command, the Change option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Change Resilient Application input dialog. You can also change a resilient application's

recovery domain (see 7.4.4, “Changing a resilient application’s recovery domain” on page 94).

However, no other parts of a resilient application’s definition can be changed directly. To change any other aspect of a resilient application’s definition (for example, the object specifiers selected for replication or the number of groups associated with the application), you must update the application. The update process removes the groups currently associated with the application and reads the application’s QCSTHAAPPI data area to re-define the groups and re-select the object specifiers required for the application.

You can update a resilient application with the DM iCluster Update Application (DMUPDAPP) AS/400 command, the Update option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Update Resilient Application input dialog.

If you upgrade your ClusterProven application with a new version supplied by the application vendor, the upgrade may also include some changes to the resilient application’s definition. Your application vendor will provide you with a new QCSTHAAPPI data area to take account of these changes. In this situation, you should update the resilient application using the method described in the previous paragraph.

7.4.4 Changing a resilient application’s recovery domain

Using iCluster, you can define a resilient application with either one (the primary) or two nodes (a primary and a backup) in its recovery domain. Backup nodes can be added and removed as necessary when the resilient application is inactive.

If a resilient application has only a primary node in its recovery domain, you can add a backup node with either the DMADDBACK AS/400 command, the Add Backup option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Add Backup Node input dialog.

If a resilient application has two nodes in its recovery domain, you can remove the backup node with either the DMRMVBACK AS/400 command, the Remove Backup option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Remove Backup Node input dialog.

Note that you cannot directly change a resilient application’s primary node. To change the primary node, you can perform a switchover on the resilient application so that the current backup node becomes the primary node (see 7.4.7, “Switching over a resilient application” on page 95). Or you can re-define the resilient application with a different primary node.

7.4.5 Activating or starting a resilient application

You can activate or start a resilient application with the DM iCluster Start Application (DMSTRAPP) AS/400 command, the Start option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Start Resilient Application input dialog.

If the resilient application has data CRGs with objects selected to them, replication will also be activated for those CRGs.

7.4.6 De-activating or ending a resilient application

You can de-activate or end a resilient application with the DM iCluster End Application (DMENDAPP) AS/400 command, the End option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator End Resilient Application input dialog.

7.4.7 Switching over a resilient application

Switchover is the process of interchanging the primary and backup roles of a resilient application's recovery domain and changing the direction of object and data replication in the data CRGs associated with the application. You can switch over an active resilient application with either the DM iCluster Switch Over Application (DMSWOAPP) AS/400 command, the Start Switch over option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Switchover Resilient Application input dialog.

Switching over a resilient application may not happen instantaneously, particularly if it has many associated CRGs or large amounts of objects and data are replicated by the associated CRGs. Other factors that can increase the amount of time spent doing a switchover are:

- The need for the application to back up to a recovery point if a switchover took place in the middle of a transaction
- Latency in the apply processes on the backup node, for example, if a switchover takes place in the middle of a batch run on the backup node
- Starting journalling of database files, particularly if many files need to be journaled, on the new primary node
- Setting up trigger programs and enabling database file constraints on the new primary node

While the CRGs associated with the application are undergoing the switchover process, their status is displayed as *SWO_PENDING. When the switchover is complete, their status reverts to *ACTIVE.

7.5 Removing the cluster and its components

This section provides details on how to handle more of the on-going cluster management tasks. Once the cluster is up and running, some of its components or even the cluster itself need to be removed. This is how to carry out these tasks.

7.5.1 Removing a resilient application

You can remove a resilient application with either the DM iCluster Remove Application (DMRMVAPP) command, the Remove option on the iCluster Work With Resilient Applications screen, or the iCluster Administrator Remove Resilient Application input dialog.

When a resilient application is removed, the object specifiers selected to the application are also removed. However, the objects that were replicated by the application's CRGs are not affected on either the primary or backup node of the application.

7.5.2 Removing a data CRG

You can remove a data CRG that is not associated with a resilient application with either the DM iCluster Remove Group (DMRMVGRP) command, the Remove option on the iCluster Work With Groups screen, or the iCluster Administrator Remove Group input dialog.

When a data CRG is removed, the object specifiers that were selected to the group are also removed, but the objects that were replicated by the group are not affected on either the primary or backup node of the group.

7.5.3 Removing a node from the cluster

You can remove a node from the cluster at any time. We recommend that you remove a node when it is active to ensure a complete cleanup of cluster data from the node. If any cluster data is left over after the node is removed from the cluster, it may lead to difficulties if you try to add the node to a new cluster or the same cluster at a later time.

You can remove a node from the cluster with the DM iCluster Remove Node (DMRMVNODE) command, the Remove option on the iCluster Work With Nodes screen, or the iCluster Administrator Remove Node input dialog.

7.5.4 Removing the entire cluster

You can remove all resilient applications, data CRGs and nodes from the cluster with the DM iCluster Delete Cluster (DMDLTCLSTR) command. This command can be invoked on the command line or as an option in the DM iCluster Commands (DMCMDS) menu, which is accessible from any iCluster menu or screen on the AS/400 system. This command is also accessible from the iCluster Administrator.

Note that removing the entire cluster only means that the cluster is de-activated, and nodes and other cluster components are removed. The objects that were replicated by the cluster's data CRGs are not affected. The iCluster product itself is still accessible as it was before the cluster was created.

If the cluster is not partitioned and all the nodes in the cluster are active, you only need to call the DM iCluster Delete Cluster (DMDLTCLSTR) command on one node to remove the entire cluster.

However, if the cluster is partitioned, you must call the DMDLTCLSTR command once in the primary partition and once on each node in the secondary partitions. Similarly, if any nodes in the cluster are inactive, you must call this command on each inactive node of the cluster and in the active part of the cluster.

The DMDLTCLSTR command can be used to delete any cluster.

7.6 Using iCluster commands to access Cluster Services operations

Most iCluster commands correspond directly to an OS/400 Cluster Services operation or API. You can use the iCluster commands to access the OS/400 Cluster Services operations, for example, when recovering from a partition or node failure. Appendix C, "Problem determination" on page 147, describes the cluster operations that are allowed in a cluster partition situation and shows how to recover from a node failure with Cluster Services operations. Recovery from a cluster partition or node failure can be performed with the iCluster commands that map to the Cluster Services operations.

Table 3 on page 98 lists the mapping between the Cluster Services APIs and the DataMirror iCluster commands.

Table 3. Mapping Cluster Services operations to iCluster commands

Cluster Services operation	iCluster command
Add a node to the cluster	DMADDNODE
Change a cluster node	DMCHGNODE
Remove a node from the cluster	DMRMVNODE
Start a cluster node	DMSTRNODE
End a cluster node	DMENDNODE
Delete the cluster	DMDLTCLSTR

Table 4 lists the mapping between Cluster Resource Group operations and iCluster commands.

Table 4. Mapping Cluster Resource Group operations to iCluster commands

Cluster Resource Group operation	iCluster commands
Create a Cluster Resource Group	DMADDGRP, DMADDAPP
Change a Cluster Resource Group	DMCHGGRP, DMCHGAPP
Delete a Cluster Resource Group	DMRMVGRP, DMRMVAPP
Add node to recovery domain	DMADDBACK
Remove node from recovery domain	DMRMVBACK
Start a Cluster Resource Group	DMSTRGRP, DMSTRAPP
End a Cluster Resource Group	DMENDGRP, DMENDAPP
Initiate switchover	DMSTRSWO, DMSWOAPP

Chapter 8. Lakeview Technology availability solutions

Lakeview Technology is a High Availability Business Partner specializing in availability management for the IBM AS/400 with its MIMIX solutions suite. MIMIX ClusterServer and MIMIX FastPath components, combined with the IBM AS/400 system, provides a robust clustering environment for data and application resiliency. This section describes the high availability solution suite offered by Lakeview Technology.

8.1 MIMIX ClusterServer

The MIMIX ClusterServer, from Lakeview Technology, offers a new and completely integrated clustering solution for the availability of applications and data, centralized cluster management, and a worldwide single-point-of-contact 24 x 365 support.

Figure 44. MIMIX ClusterServer logo

8.1.1 The need for availability

The issue of high availability and continuous operations has never been greater among IT managers. To meet the demands of their users, IT managers are adopting clustering technology to reduce planned and unplanned downtime. The solution requires that applications must be written to specific OS/400 APIs. In addition, certified cluster middleware for replication services is required, and a cluster manager is necessary to deploy a complete clustering solution. The open architecture allows for each component to be provided by separate vendors. The integration of the solution and the complexity of the support in a multi-vendor environment can be more than challenging. MIMIX ClusterServer provides the solution to answer the challenge.

8.1.2 MIMIX ClusterServer for AS/400 solution

MIMIX ClusterServer provides a Java-based GUI Cluster Manager, high performance MIMIX Replication Services, MIMIX Application Cluster Templates, and MIMIX Cluster Optimizer. The result is an environment that delivers applications that leverage robust high availability coordinated with planned application switchovers or unplanned application failovers.

8.1.2.1 MIMIX-ACE for AS/400

MIMIX Application Cluster Enabler (ACE) is a highly integrated combination of software and services for analyzing and meeting ISV application cluster requirements and creating the required program elements.

8.1.2.2 MIMIX-ACT for AS/400

The MIMIX Application Cluster Template (ACT) is software that provides unique templates for cluster-enabled applications.

8.1.2.3 MIMIX Cluster Manager

MIMIX Cluster Manager is a Java-based GUI or green-screen cluster administrator that manages cluster resources and facilities.

8.1.2.4 MIMIX Cluster Optimizer for AS/400

This tool is customized for the total customer environment. It includes ClusterProven and non-ClusterProven applications, along with MIMIX Cluster Manager for an automated and highly integrated cluster environment.

8.1.2.5 MIMIX Replicator for AS/400

This tool is the standard in AS/400 data and object replication.

8.1.2.6 Open architecture

MIMIX ClusterServer has two deployment options: the open architecture model using MIMIX ClusterServer and MIMIX-ACT in combination with any HABP replication services and any cluster manager. When the flexible, open architecture model is used, any cluster manager or any high availability business partner software can exploit the AS/400 clustering technology.

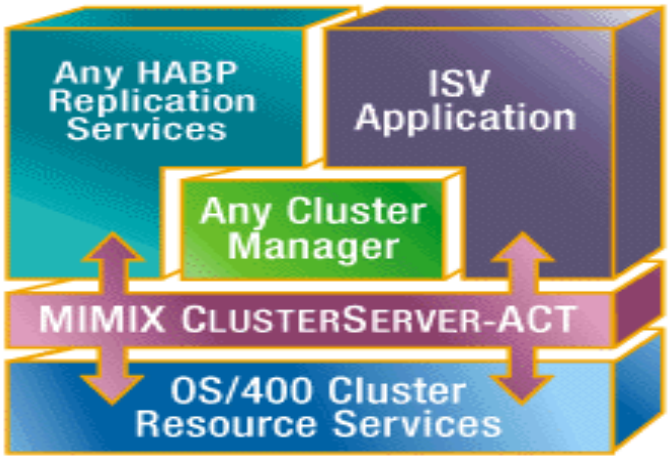


Figure 45. AS/400 cluster implementation open architecture

8.1.2.7 Optimal architecture

Optimal cluster implementation uses all of the MIMIX ClusterServer components including MIMIX Replicator. Under the optimal implementation, the solution is tightly integrated and highly optimized and includes consulting services with a worldwide single point-of-contact for support.

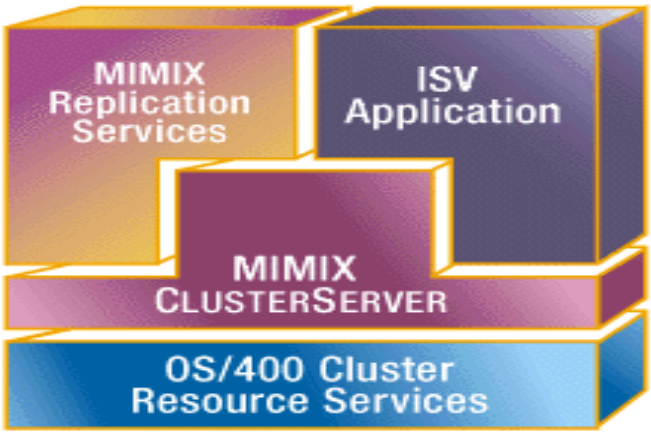


Figure 46. Optimal AS/400 cluster implementation MIMIX ClusterServer

8.1.2.8 Cluster consulting services

Lakeview Technology and its network of worldwide partners offer a full range of cluster services including planning, implementation, customization, skills transfer, and application integration. Lakeview's *MIMIX FastPath* also

provides application cluster enablement to ISVs and custom-written applications.

8.1.2.9 Technical support

In a multi-vendor cluster environment, it is often difficult to isolate the problem and call in the appropriate vendor.

With MIMIX ClusterServer, “Follow the Sun” Support is provided for your entire cluster through a worldwide single point-of-contact by Lakeview Technology and its worldwide network of partners.

8.2 MIMIX FastPath

MIMIX FastPath for AS/400 is an exclusive Solution Services offering consisting of tools, services, support, and clustering expertise designed to fully enable applications for AS/400 ISVs and custom developed applications to work in an AS/400 clustering environment with little or no modification of the application code.

8.2.1 The need for ClusterReady applications

In the era of e-commerce, distributed users, and supply-chain management, the business demand for high availability and continuous operation of applications continue to concern AS/400 customers. To meet these demands, IBM has introduced AS/400 clustering in OS/400 V4R4. Applications that are written to a specific set of OS/400 APIs can take advantage of the Cluster Resource Services and deliver availability and continuous operations to its users.

Note

ClusterReady is a trademark of Lakeview Technology and is not part of the IBM ClusterProven program.

8.2.2 Why MIMIX FastPath

Accelerated time to market! And time to market means competitive advantage and profitability.

Business managers are facing increasing pressure every day to grow their business and maintain competitive advantage. These managers rely on core systems and business-critical applications that can have no tolerance for

downtime. To meet the needs, managers need clustering solutions that are robust, easy to mange, and rich in functionality.

To take advantage of clustering, three elements must be in place:

- OS/400 V4R4
- High availability replication services
- Applications that have been written or modified to work with the specific cluster APIs in the operating system

The critical issue for the application providers is the time, resource investment, and high availability expertise required when modifying applications to take advantage of the Cluster Resource Services as well as the ongoing support of these modifications. Many ISVs and customers may find these challenges too great to readily adopt any clustering technology.

MIMIX FastPath provides the optimal solution by providing services that enable applications with a minimal amount of effort and investment.

8.2.3 The MIMIX FastPath process

Lakeview Technology Solution Services Consultants begin the MIMIX FastPath process by determining the level of availability required, assessing the application, and using MIMIX ACE to create the program elements necessary to achieve ClusterReady status. The result is MIMIX ACT, which enables the application to interface with the OS/400 cluster APIs, replication services, and a cluster manager. The process is completed with extensive testing, skills transfer, and certification for ClusterReady.

8.2.3.1 MIMIX ACE for AS/400

The MIMIX Application Cluster Enabler (ACE) is a highly integrated software tool. It is used by Solution Services to analyze an application and create the required cluster elements for an application to work in an AS/400 clustered environment without the need to modify the actual application code.

8.2.3.2 MIMIX ACT for AS/400

The MIMIX Application Cluster Template (ACT) is an application-specific software module. It enables the application to interface with the OS/400 cluster resource APIs, replication services, and the cluster manager.

8.2.3.3 MIMIX FastPath services

Professional services include application assessment, cluster planning, creation of the ACT, skills transfer, testing, and documentation necessary for achieving ClusterReady status. Also included is an analysis and written

documentation to detail any additional work that may be required coupled with on-site services, support, and expertise to achieve IBM ClusterProven branding.

8.2.3.4 MIMIX FastPath support

Maintaining the clustering modifications throughout the life of the application is a significant challenge. MIMIX FastPath provides a single point of worldwide support for ongoing modifications, updates, and upgrades to the MIMIX FastPath work through the life of an agreement. This assures compliance with new releases and fixes of the operating system, application version changes, and other required updates, while freeing critical ISV resources to focus on the development of core product functionality.

Chapter 9. Vision Solutions

With OMS/400 Cluster Manager, recovery from unplanned failovers or planned switchovers can now be both seamless and rapid. Building upon the Vision Suite of middleware High Availability software products, OMS/400 Cluster Manager extends your ability to create highly available and resilient data, application, and user environments.

9.1 Vision Solutions OMS/400 Cluster Manager

In addition to non-clustering related product features, such as mirroring database files, data areas, and data queues in real-time using IBM journalling abilities, OMS/400 Cluster Manager now provides object mirroring support for data CRGs and ClusterProven applications.

OMS/400 Cluster Manager provides these capabilities through new menus and screens presented to the user in two ways:

- The traditional AS/400 green-screen interface
- A client-server Java application with a graphical user interface running on a PC or workstation

Both of these interfaces are fully integrated with OMS/400. They allow you to define sets of objects and, using bi-directional communication paths, and create and maintain one or more additional sets of synchronized data.

9.1.1 Implementation goals

In addition to supporting clustered environments, Vision Solutions objectives in implementing OMS/400 Cluster Manager include:

- Building upon the OMS/400 high level of data integrity to increase data resiliency on all clustered systems
- Working with ISVs to build highly resilient application environments
- Assisting ISVs in the process of obtaining IBM ClusterProven status

9.2 Getting started with OMS/400 Cluster Manager

Before installing the client, the AS/400 systems must be setup for clustering. This includes ensuring that all managed systems are on the same operating system level and enabled for TCP/IP and clustering. In addition, OMS/400 R6.3 (or higher) must be installed.

9.2.1 Installing the client

In our Windows implementation of the OMS/400 Cluster Manager, there are five installation panels. The first one is shown in Figure 47.

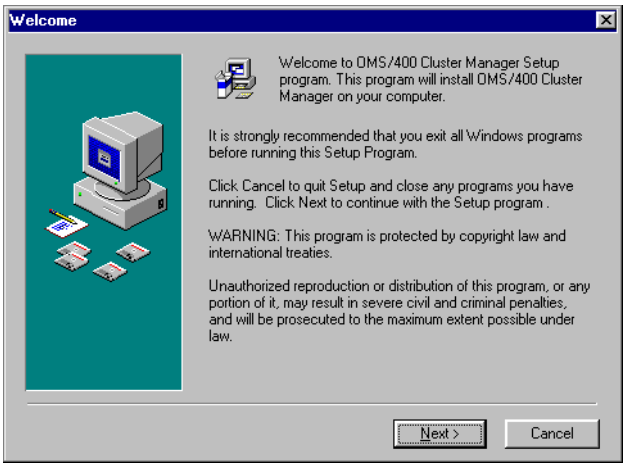


Figure 47. OMS Welcome page

9.2.2 Starting the product

The first time you start using the GUI, OMS/400 Cluster Manager asks you the hostname of the cluster node to which you want to initially connect. To login and begin managing your clusters, you need to know either the node's hostname or TCP/IP address of at least one cluster-enabled node.

9.2.3 Defining host systems

If you are not sure which AS/400 systems in your network are currently cluster-enabled, you can use the Client Server Configuration Wizard built into OMS/400 Cluster Manager to automatically detect and report on the cluster-enabled status and operating system level of all nodes reachable from your client computer.

Once you configure at least one cluster-enabled node to OMS/400 Cluster Manager, you can begin managing your clustered environment. If you want to work with pre-defined clusters (for example, clusters built with the green screen version of OMS/400), you can simply send a request for cluster information to any node in an existing cluster. Even if that request finds that the node is inactive, OMS/400 Cluster Manager will attempt to forward that request to other nodes.

If you are configuring new clusters using the GUI interface, log on to any AS/400 system. Then, using OMS/400 Cluster Manager's built-in wizards, begin defining your clusters, CRGs, and recovery domains.

9.2.4 Auto-detection of clustered nodes

As you send requests to various systems, those AS/400 systems are automatically added to the list of configured host machines. The next time you use OMS/400 Cluster Manager, you can forward requests directly to the additional cluster nodes.

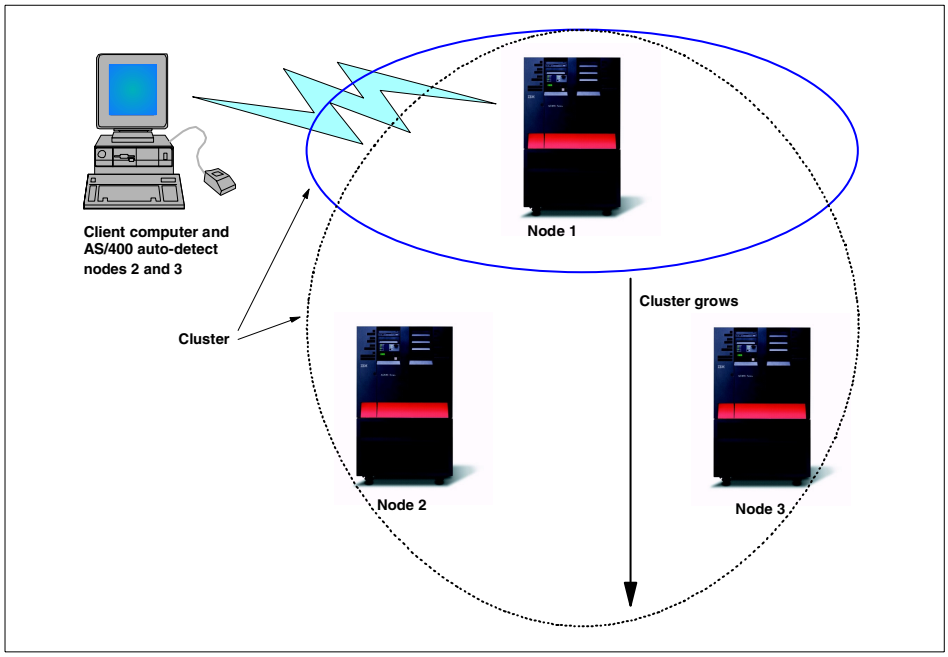


Figure 48. Auto-detecting nodes

In Figure 48, the client has only node one currently defined in its list of host systems. When you request cluster information from node one, the response tells the client that there are actually three nodes in the cluster, and stores in memory the additional nodes' host configuration information. When you close the client application, that information is stored on your client computer for retrieval next time the application is started.

Similarly, if new nodes have been added to an existing cluster since last using OMS/400 Cluster Manager, the client recognizes those hosts as "new" and adds them to the list of currently defined host systems. This ability to

auto-forward requests to any AS/400 system reachable via TCP/IP allows organizations to rapidly configure and implement clustering environments.

9.2.5 IP interface selection

An additional feature is IP interface selection. As you go through your systems, adding and configuring them for clusters, you can view all IP interfaces through which you may interconnect the nodes in a clustered environment. This features allows organizations to define specific routing paths for their cluster-enabled networks and reduce IP traffic on other, non-clustered networks.

9.2.6 Working with ClusterProven applications

As part of Vision Solutions' continuing support for application integrity, OMS/400 Cluster Manager works with mixed data and application CRG environments for seamless switchovers. To prevent the loss of in-flight data transactions, OMS/400 Cluster Manager, working with ClusterProven applications, waits until the application completes its activities before notifying the data CRG that switchover or failover can commence.

The interaction between the application CRG and the data CRG varies depending on the specific resiliency requirements of the application. For example, OMS/400 ensures the data associated with a ClusterProven application is in sync. The term "in-sync" in this example means the recovery domain and switchover or failover information is the same (such as the current roles of the primary node and first and second backups are the same for both application and data CRGs). If the ISV uses a commitment control scheme to increase application resilience, OMS/400 Cluster Manager only begins a data switch over when the following conditions are met:

- All users are removed and disconnected from the application, ensuring no more transactions are created.
- Transactions not committed are rolled back.
- No more transactions are coming into the journals.

Only then will the application CRG notify the data CRG that a data switchover can begin.

Similarly, when the switchover or failover is completed (assuming the former primary node is now a backup node, and the node that was the first backup is now the primary node), the application can restart. This allows users to log back into the application on the new primary system and begin working again.

The data CRG then notifies OMS/400 to begin picking up new transactions and send them to the new backup system.

9.3 OMS/400 Cluster Manager sample displays

The following illustrations show you how to perform various clustering tasks using the GUI version of OMS/400 Cluster Manager.

9.3.1 Working with clusters and CRGs

Figure 49 shows the OMS Cluster Manager window and contains selection buttons to create a cluster and gather cluster information. From the File pull-down menu, a range of other cluster related activities is available.

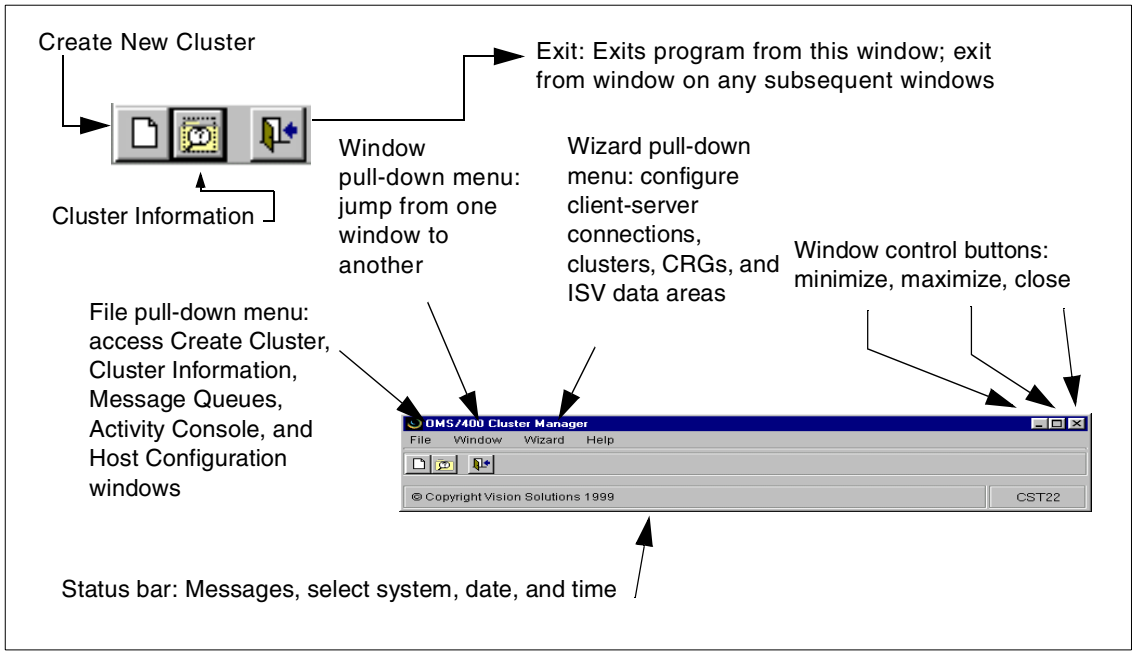


Figure 49. OMS Cluster Manager

9.3.2 Creating new clusters

When creating a new cluster, the window shown in Figure 50 on page 110 appears.

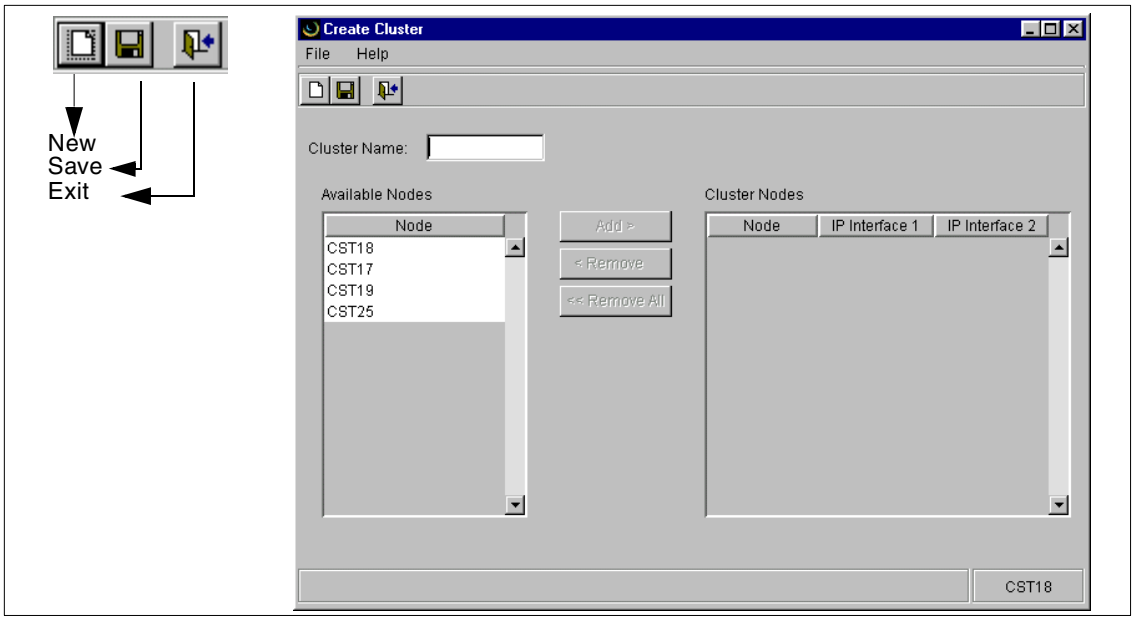


Figure 50. Creating a cluster window

9.3.3 Viewing cluster information

Once you create the cluster, you can view the cluster information and resources from the cluster information window (Figure 51).

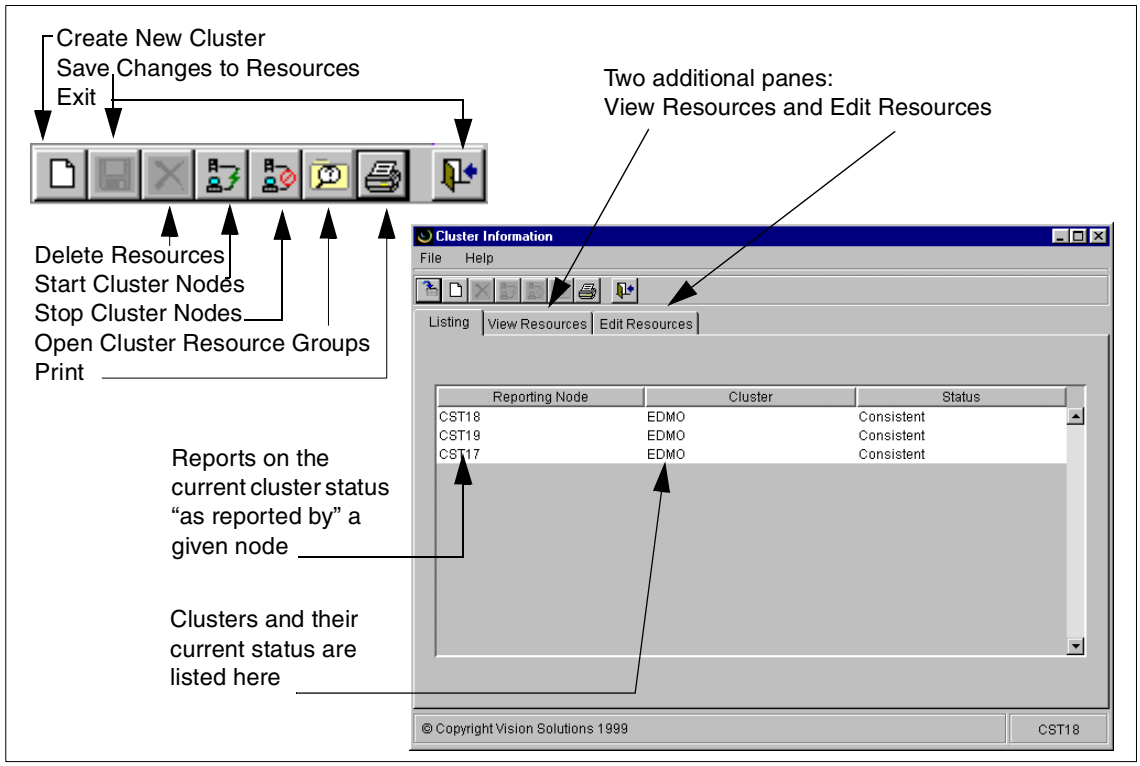


Figure 51. Cluster information windows

9.3.4 Adding a node to the cluster

OMS Cluster Manager allows you to add node to the cluster by selecting them from a standard window list (Figure 52 on page 112).

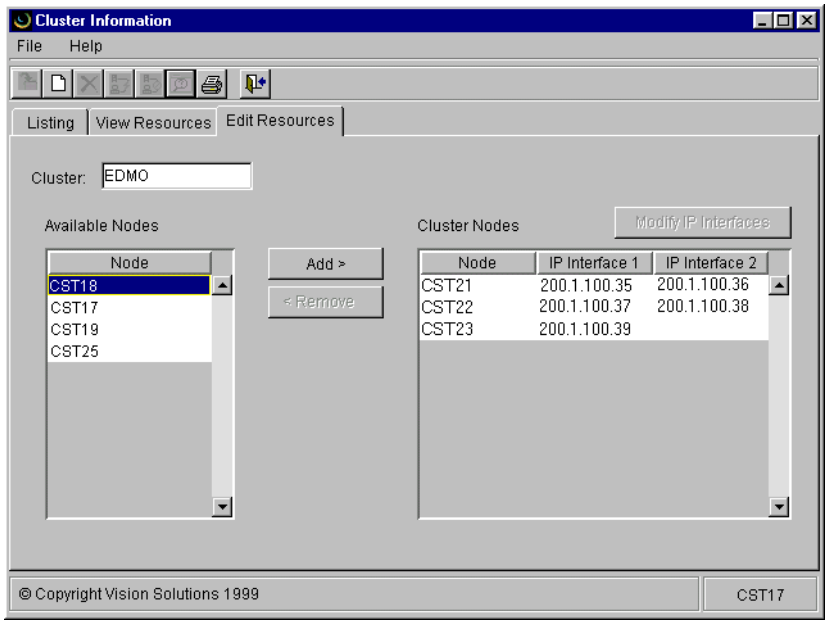


Figure 52. Adding a node

9.3.5 Activating and de-activating nodes in the cluster

Once the initial setup tasks are completed, the cluster can be activated and then when it's running, it can be de-activated. The nodes shown can be selected and processed by pressing the End or Start buttons (Figure 53).

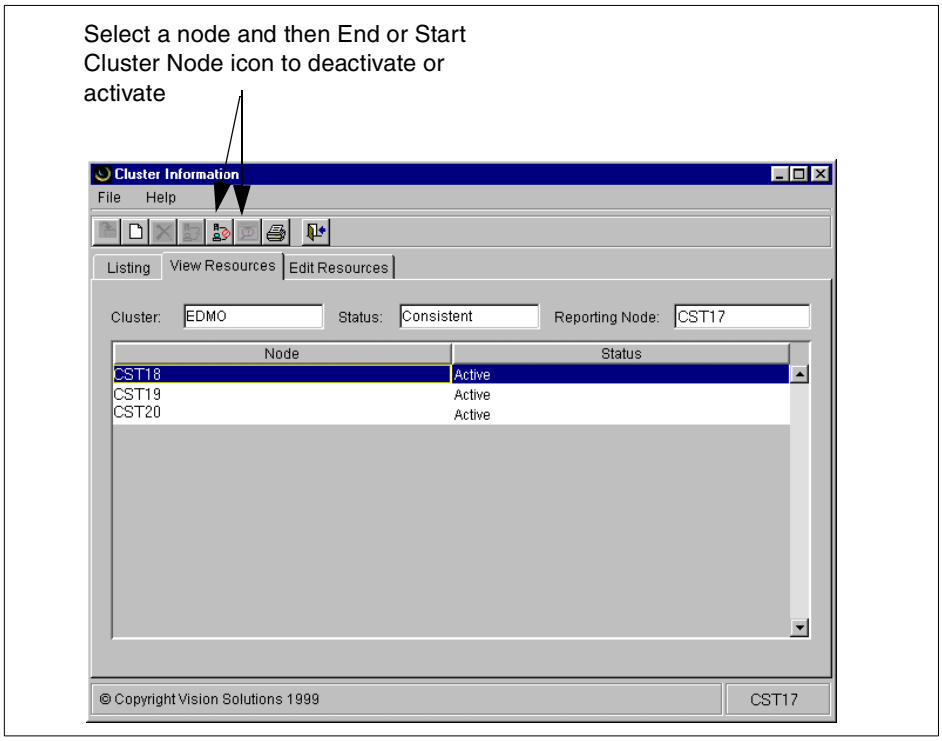


Figure 53. Cluster activation

9.3.6 Creating and using Cluster Resource Groups (CRGs)

To create resilient objects in the Cluster Resource Groups window, select the Edit CRG Configuration tab. This panel allows the creation of both data and application CRGs. Depending on the type of CRG, not all input boxes are required.

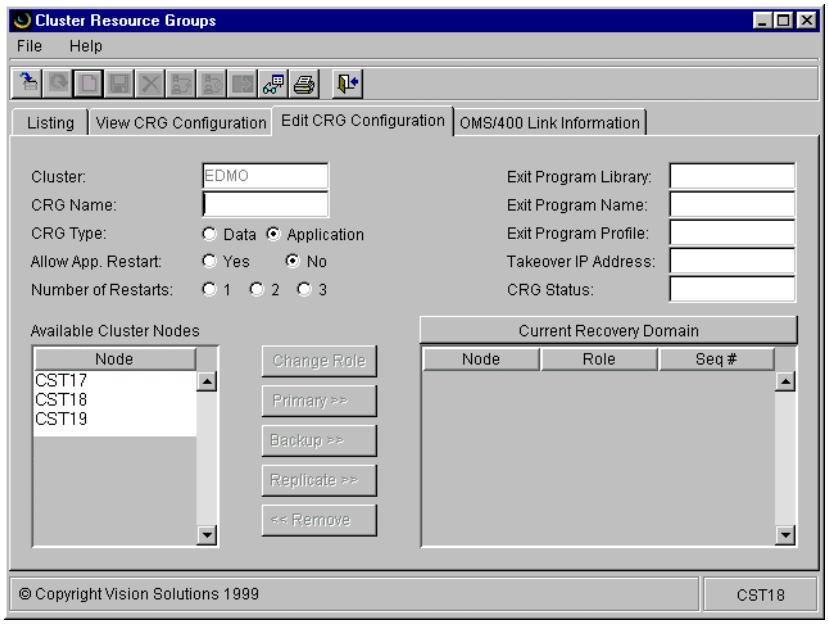


Figure 54. Creating CRGs with iCluster

9.3.7 Changing a CRG recovery domain

Changing the recovery domain and altering the role of a node is one of the tasks performed by the operations group when managing the cluster (Figure 55).

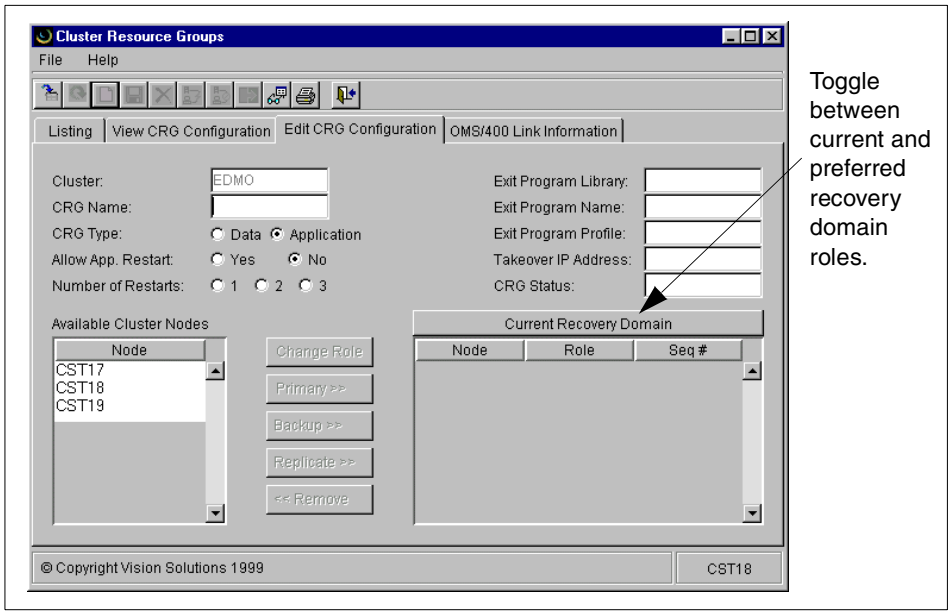


Figure 55. Changing a recovery domain

9.3.8 Activating or starting a data or application CRG

Once you create the application or data CRGs, you can select them for activation (Figure 56 on page 116).

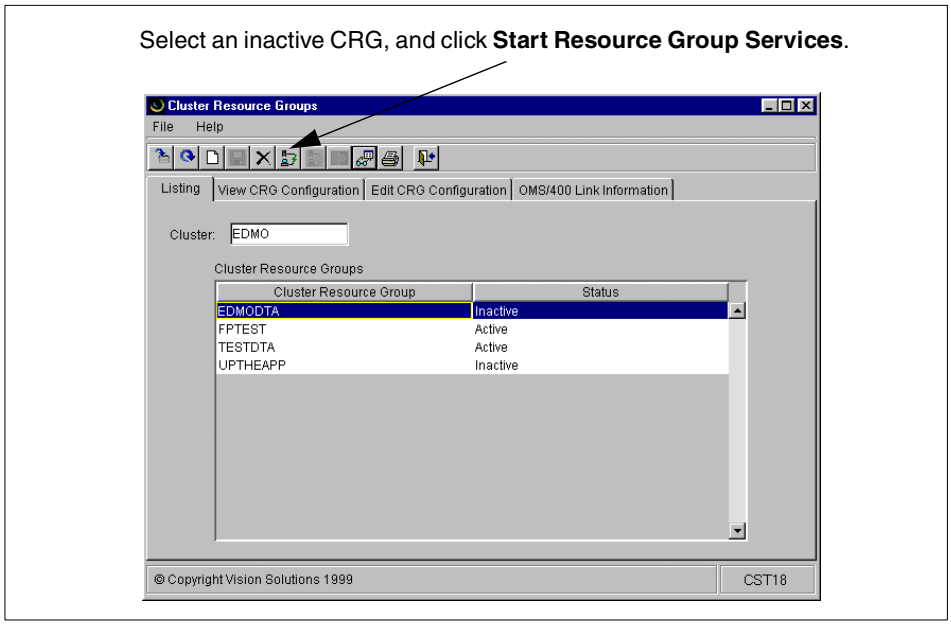


Figure 56. CRG activation

9.3.9 De-activating or ending a data or application CRG

To end a data or application CRG, first highlight the CRG. Then, click the **Stop Resource Group Services** button (Figure 57).

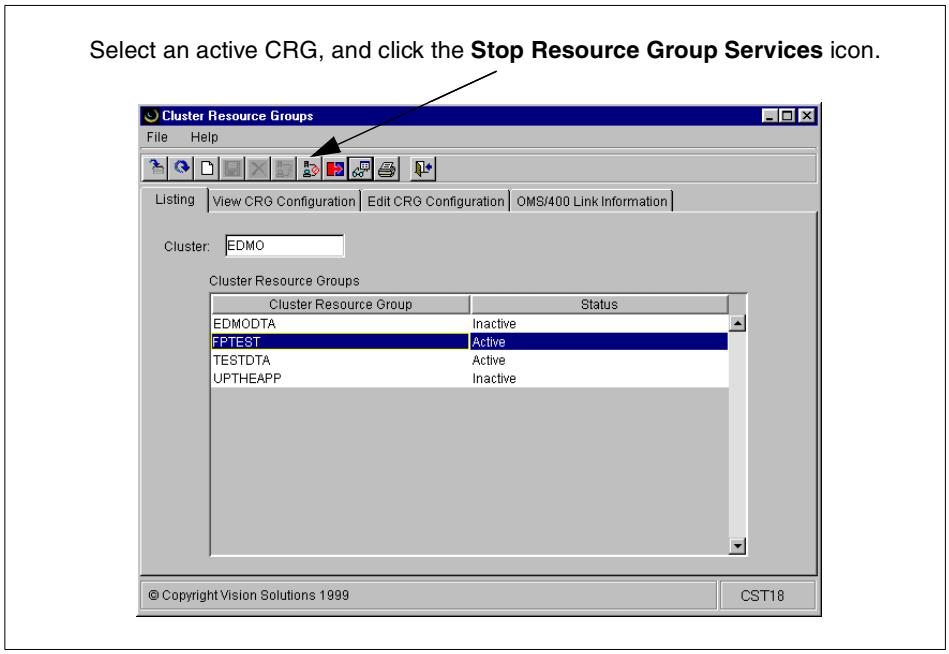


Figure 57. Stop Resource Group Services

You can perform a data switchover or application CRG in the Cluster Resource Groups Listing view (Figure 58 on page 118).

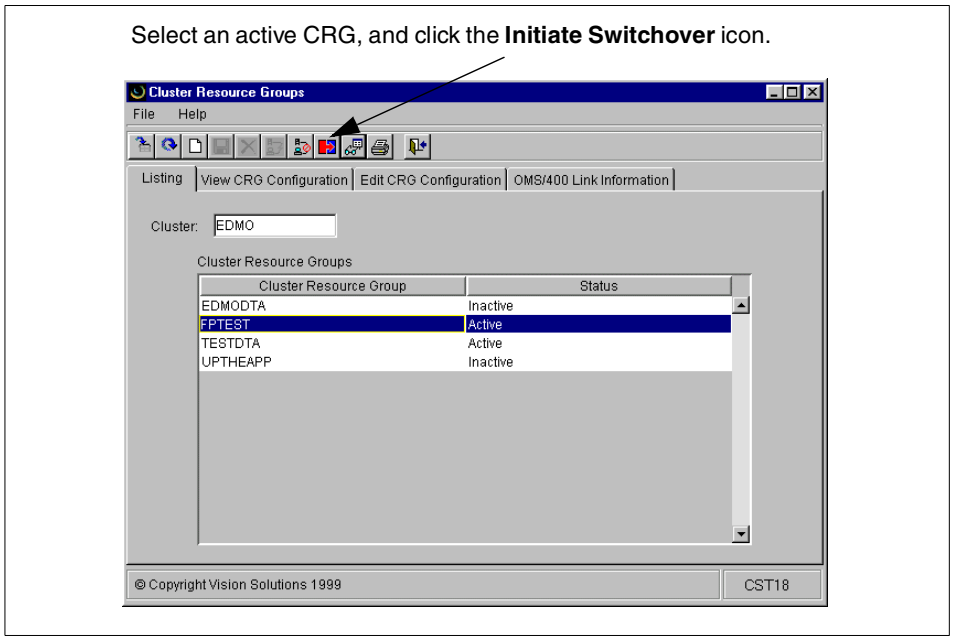


Figure 58. Switching over CRG

9.3.10 Creating an application CRG recovery domain

When you create an application CRG recovery domain, you must specify the takeover IP address as shown in Figure 59.

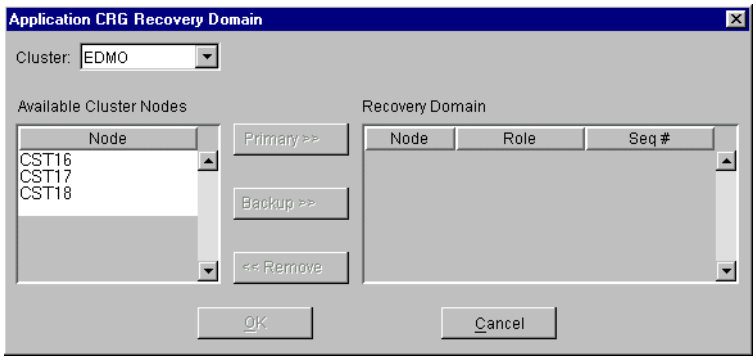


Figure 59. Creating an application CRG

The takeover IP address must not be active on any of the nodes (Figure 60).

Figure 60. Takeover IP address

9.3.11 Removing a data or application CRG

Removing a data or application CRG is a standard management function. Figure 61 shows an example of this function.

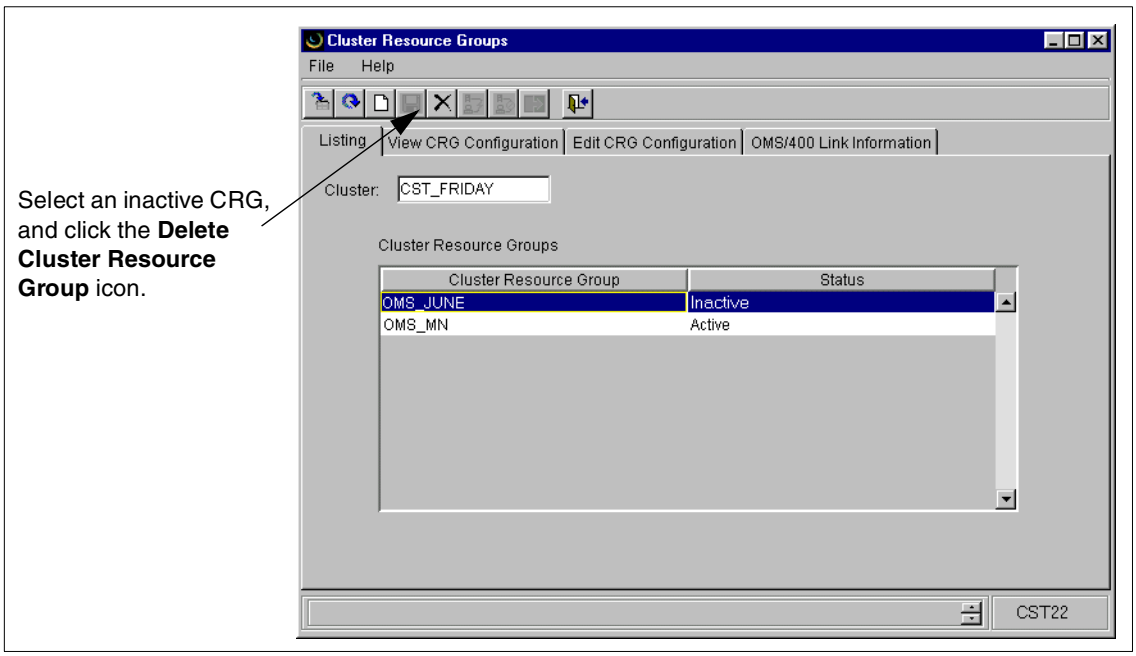


Figure 61. Removing CRGs

9.3.12 Removing a node from the cluster

Back at the Cluster information window, select the **Edit Resources** tab. The panel that appears enables you to select a node and remove it from the cluster.

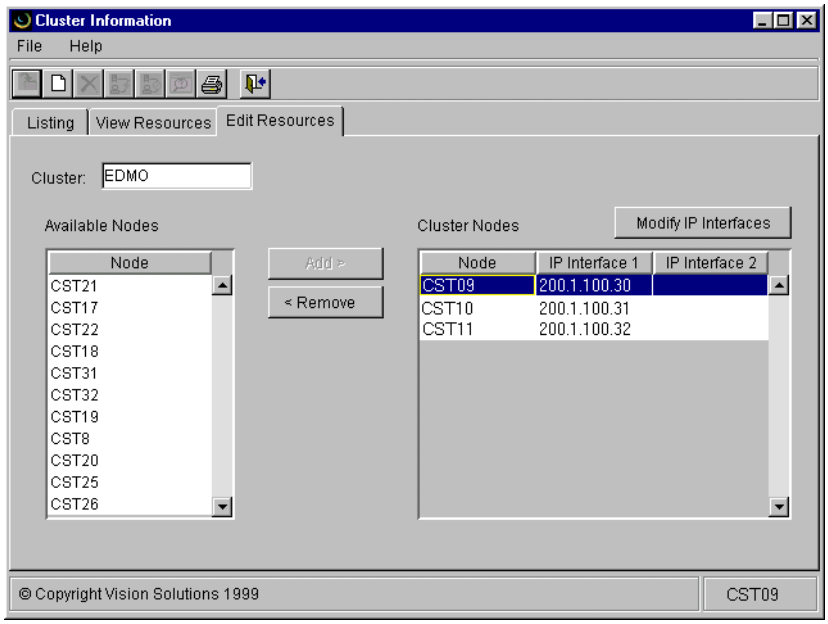


Figure 62. Removing a node from a cluster

9.3.13 Removing the entire cluster

In certain cases, you need to remove the entire cluster. At the Cluster Information window, select the **Listing** tab. In the view, select the cluster and click the **Delete** button (Figure 63).

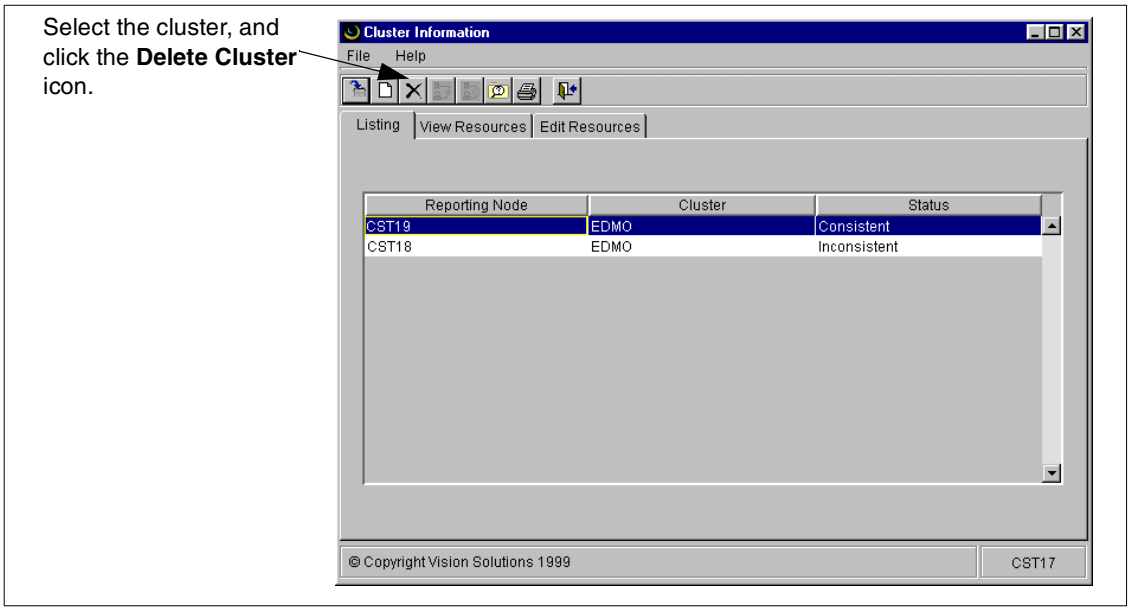


Figure 63. Removing the cluster

9.4 Working with applications

If the application is to be cluster aware, you must edit the ISV data area QCSTHAPPI. This data area will be changed if the application is developed by an ISV. If the application is developed in-house, change this data area to make your application cluster aware.

9.4.1 ISV data area contents

The ISV Data Area Management window (Figure 64 on page 122) allows you to modify the QCSTHAPPI data area.

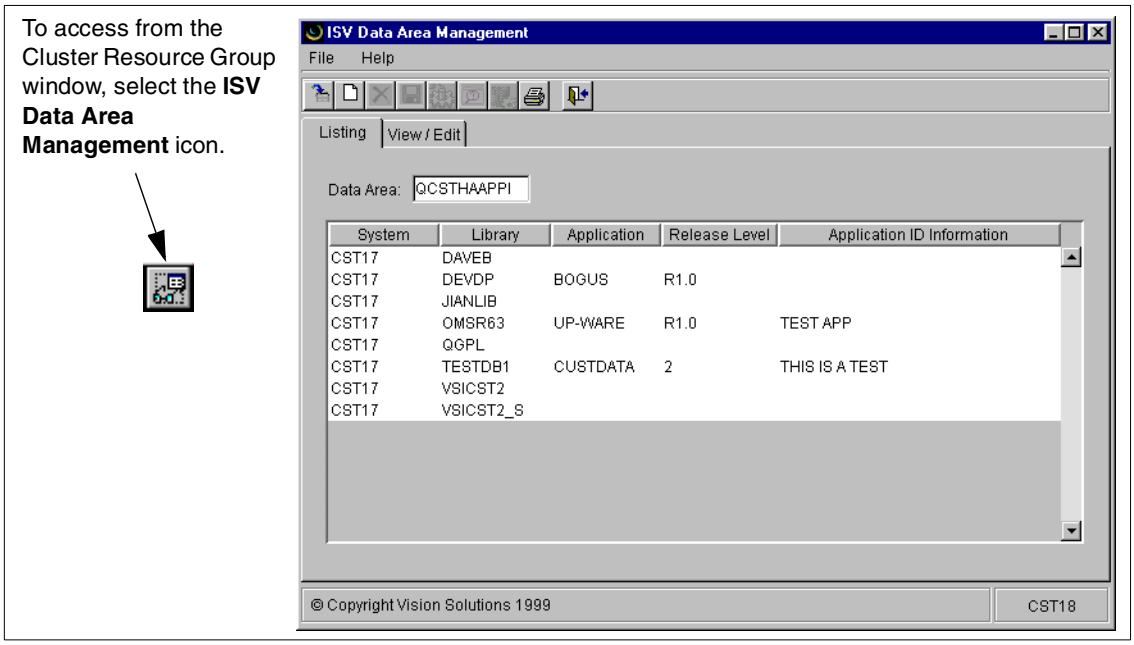


Figure 64. QCSTHAPPI contents

9.4.2 Creating ISV data areas for application CRGs

When creating ISV data areas for application CRGs, select the **View/Edit** tab for the data area input fields (Figure 65).

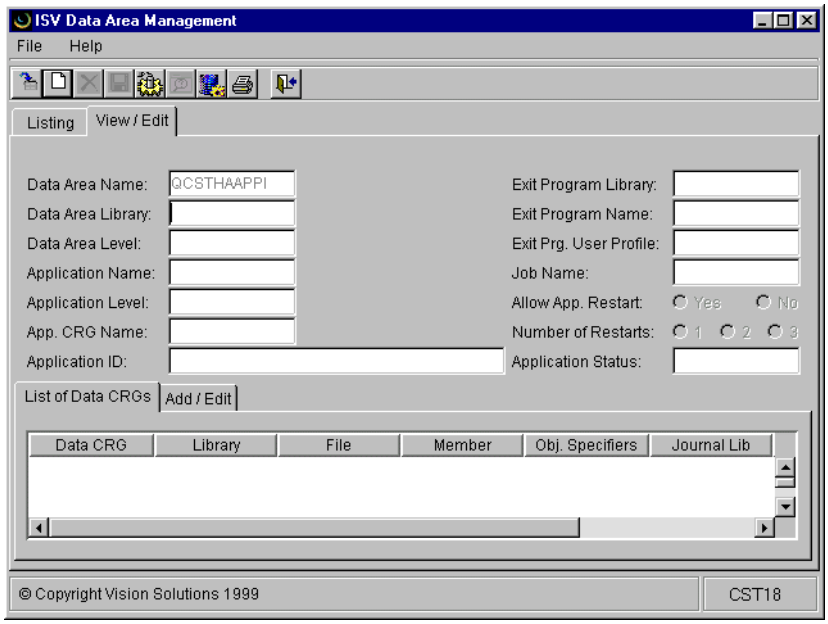


Figure 65. Creating QCSTHAPPI

9.4.3 Changing or updating data areas

To change or update a data area, select the **View/Edit** tab from the ISV Data Area Management window (Figure 66 on page 124). Then, select the CRG to be changed in the List of Data CRGs panel.

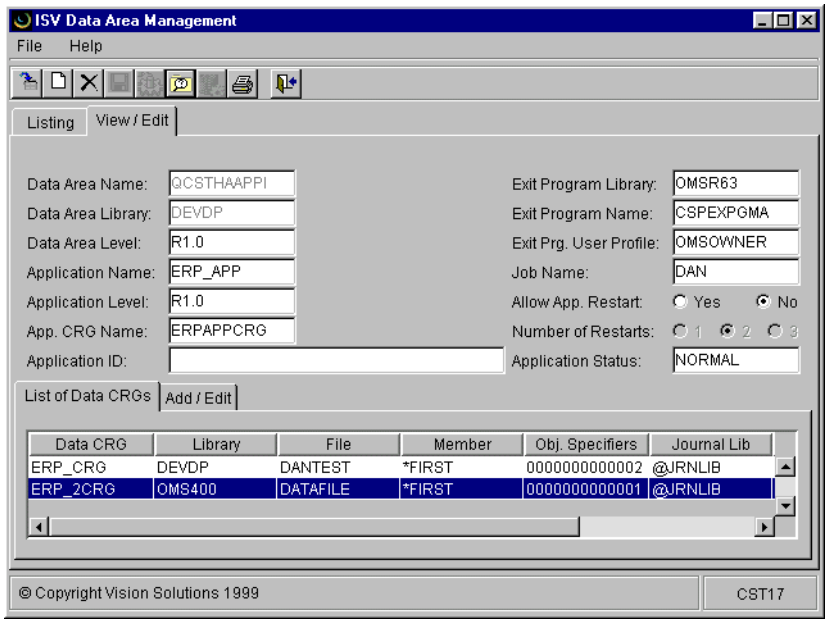


Figure 66. Changing QCSTHAPPI

9.4.4 Changing a resilient application's data area contents

The data area contents are displayed and are available for update in the Add/Edit panel (Figure 67).

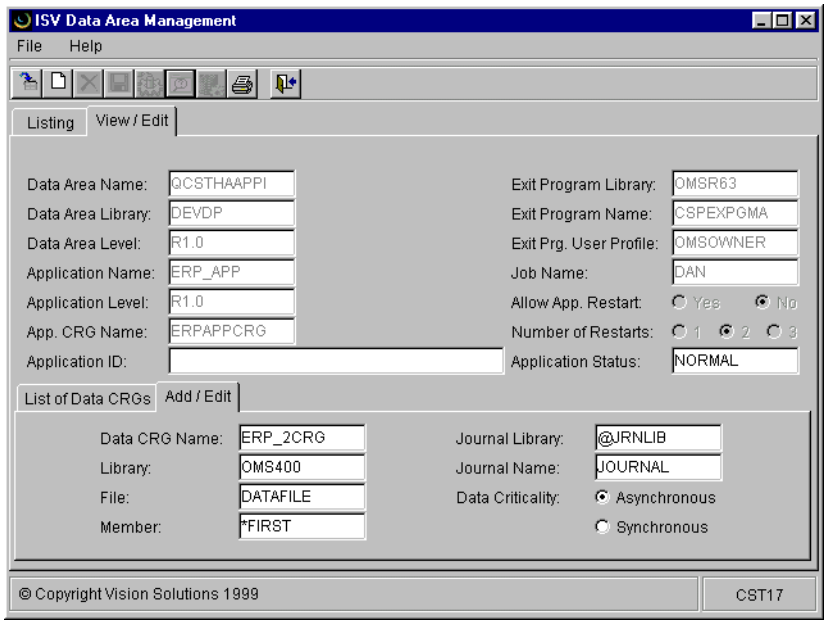


Figure 67. Updating QCSTHAPPI contents

9.4.5 Working with object specifiers

Object specifiers are the files that contain the resilient information associated with a particular application CRG. The Object Specifier Management window (Figure 68 on page 126) allows the management of these object specifier files.

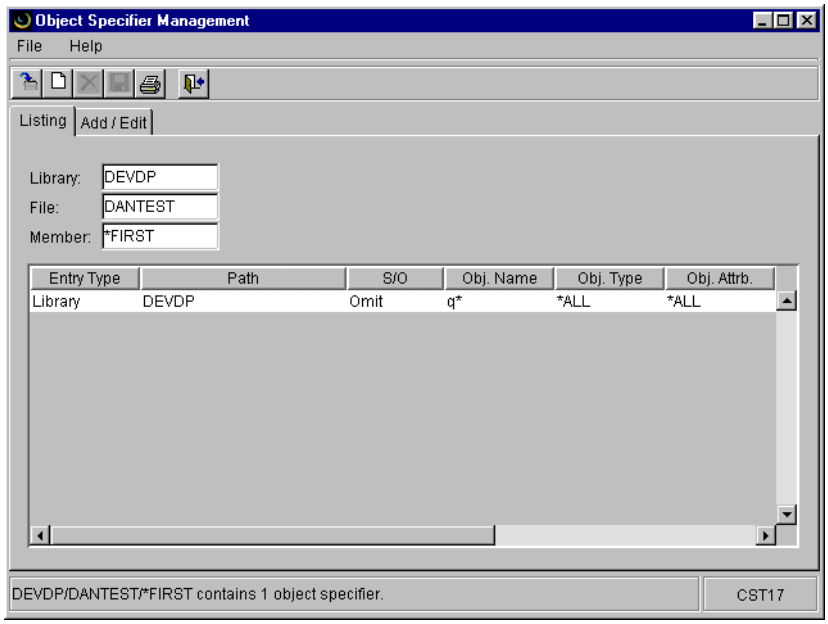


Figure 68. Object specifier list

To work with object specifiers, select the **Add/Edit** tab. The object specifier details are then displayed (Figure 69).

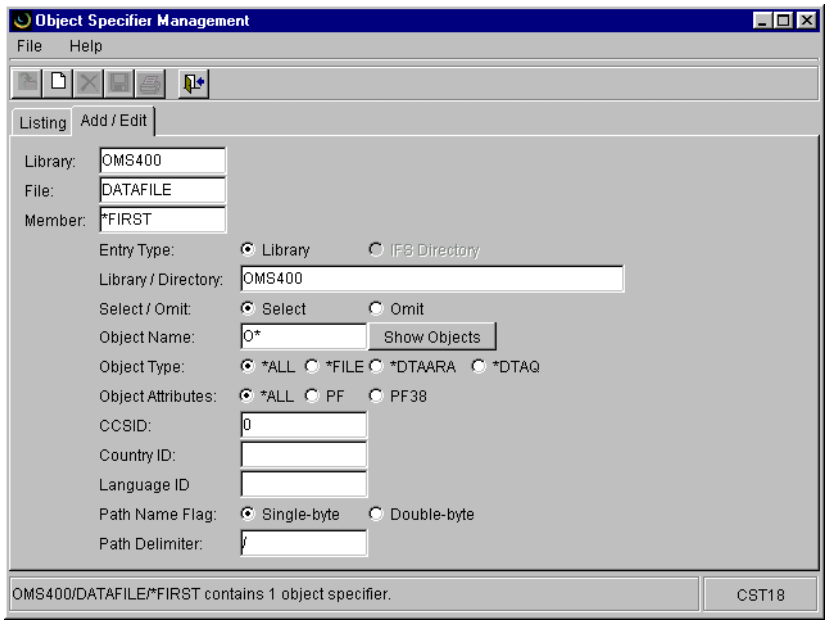


Figure 69. Working with object specifiers

9.4.6 Object Selection Results

The Object Specifier Results panel displays objects that are found within the library or directory that is selected.

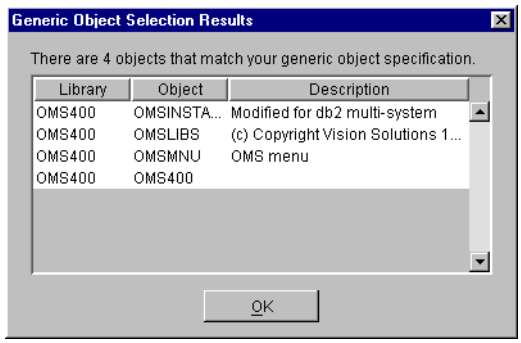


Figure 70. Object selection

9.4.7 Creating a list of objects for high availability

The wizard for ISV Data Management enables easy selection of objects for resiliency (Figure 71 on page 128).

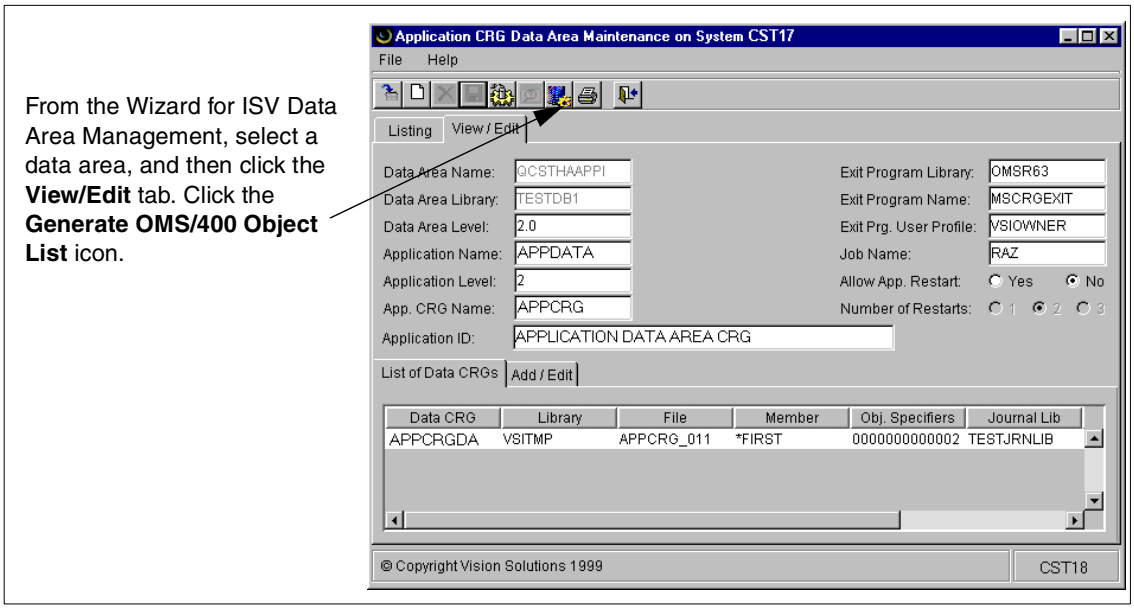


Figure 71. Creating a resilient object list

9.4.8 Viewing OMS/400 links and statistics

From the Cluster Resource Groups main window, select the **OMS/400 Link Information** tab. On this panel, the resilient resources status is displayed (Figure 72).

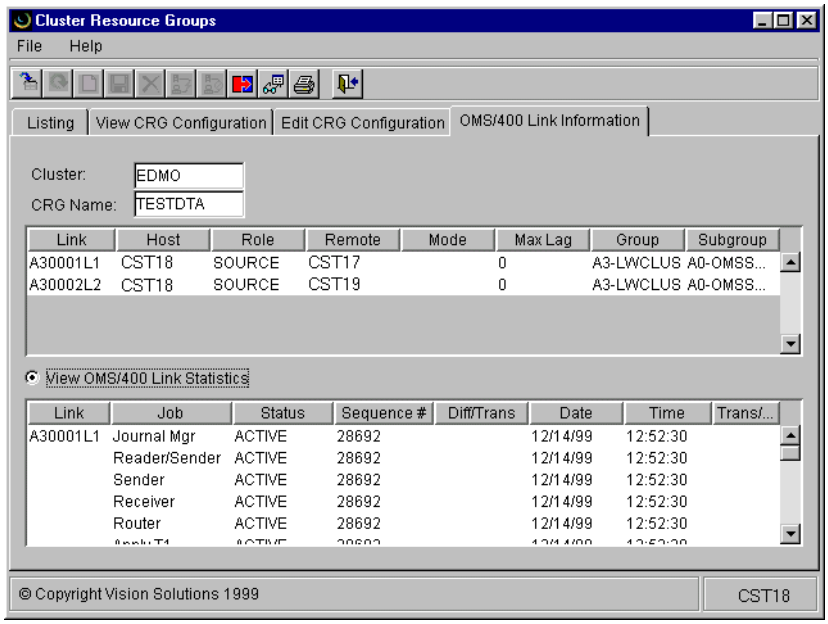


Figure 72. Viewing OMS/400 links and statistics

Part 3. Appendices

This part includes the appendices, which complement the material presented in this redbook. They contain information on the AS/400 software and hardware functions that are not specific to cluster support, but vital for highly available solutions.

Appendix A. AS/400 cluster resources

For customers, independent software vendors, and Business Partners who want to investigate AS/400 Highly Available solutions and clusters, refer to the following resources and contacts:

- **Technical Solutions Center**

  Send e-mail to rchtsc@us.ibm.com

- **PartnerWorld for Developers**

  http://www.developer.ibm.com/

- **Information Center Web site**

  http://www.as400.ibm.com/infocenter

  AS/400e Information Center CD-ROM (English version: SK3T-2027)

- **AS/400 home page**

  - http://www.as400.ibm.com/ha/ha2_99.htm
  - http://www.ibm.com/servers/clusters

- **IBM Marketing Representative**

  Call 1 800 IBM 4YOU (1-800-426-4968). This is for the US only. For other geographies, contact your local marketing representative.

- **Rochester AS/400 Support Center**

  Call 1-800-237-5511. Select option **2**, BPLUS team.

Appendix B. AS/400 high availability functions

This appendix discusses the basic AS/400 hardware and OS/400 software availability options. Hardware and software recovery has been available with the AS/400 system for several years. Some customers are well aware of these functions and use them to improve the resilience of their business systems. To others, these may be new concepts.

B.1 Journaling

Journaling recovers the changes to database files or other objects that have occurred since your last complete save. You use a journal to define what files and access paths you want to protect with journal management. This is often referred to as journaling a file or an access path. A journal receiver contains the entries (called *journal entries*) that the system adds when events occur that are journaled, such as changes to database files, changes to other journaled objects, or security-relevant events.

You can use the remote journal function to set up journals and journal receivers on a remote AS/400 system. These journals and journal receivers are associated with journals and journal receivers on the source system. The remote journal function allows you to replicate journal entries from the source system to the remote system.

The main purpose of journal management is to assist in recovery. You can also use the information that is stored in journal receivers for other purposes, such as:

- An audit trail of activity that occurs for database files or other objects on the system.
- Assistance in testing application programs. You can use journal entries to see the changes that were made by a particular program.

Journaling provides the following benefits:

- Reduces the frequency and amount of data saved.
- Improves the ability and speed of recovery from a known point to the failure point.
- Provides file synchronization if the system ends abnormally.

The disadvantages of journal management are:

- Increases auxiliary storage requirements.
- May have an impact on performance because of an increase in the activity of your disks and processing unit.
- Requires file and application knowledge for recovery.

B.2 Access path protection

An *access path* describes the order in which records in a database file are processed. A file can have multiple access paths, if different programs need to see the records in different sequences. If your system ends abnormally when access paths are in use, the system may have to rebuild the access paths before you can use the files again. This is a time-consuming process. Performing an IPL on a large, busy AS/400 system that has ended abnormally can take many hours.

You can use journal management to record changes to the access paths. This greatly reduces the amount of time it takes the system to perform an IPL after it ends abnormally.

Two methods of access-path protection are available:

- System-managed access-path protection (SMAPP)
- Explicit journaling of access paths

Access-path protection provides the following benefits:

- Avoids rebuilding access paths after most abnormal system ends
- Manages the required environment and makes adjustments as the system changes if SMAPP is active
- Successful even if main storage cannot be copied to storage unit 1 of the system ASP during an abnormal system end
- Generally faster and more dependable than forcing access paths to auxiliary storage for the files (FRCACCPTH parameter)

The disadvantages of access-path protection include:

- Increases auxiliary storage requirements.
- May have an impact on performance because of an increase in the activity of your disks and processing unit.
- Requires file and application knowledge for recovery. There is a small additional processor overhead if *RMVINTENT is specified for the

RCVSIZOPT parameter for user-created journals. However, the increase in storage requirements for access path journaling is reduced by using *RMVINTENT.

- Normally requires a significant increase in the storage requirements for journaling files. The increase with SMAPP is less than when access paths are explicitly journalled.

B.3 Auxiliary storage pools

An *auxiliary storage pool (ASP)* is a software definition of a group of disk units on your system. This means that an ASP does not necessarily correspond to the physical arrangement of disks. Conceptually, each ASP on your system is a separate pool of disk units for single-level storage. The system spreads data across the disk units within an ASP. If a disk failure occurs, you need to recover only the data in the ASP that contained the failed unit.

There are two types of ASPs:

- System auxiliary storage pool
- User auxiliary storage pool

Your system may have many disk units attached to it for auxiliary storage of your data. To your system, they look like a single unit of storage. The system spreads data across all disk units. You can use auxiliary storage pools to separate your disk units into logical subsets.

When you assign the disk units on your system to more than one ASP, each ASP can have different strategies for availability, backup and recovery, and performance.

ASPs provide a recovery advantage if the system experiences a disk unit failure resulting in data loss. If this occurs, recovery is only required for the objects in the ASP that contained the failed disk unit. System objects and user objects in other ASPs are protected from the disk failure. There are also additional benefits as well as certain costs and limitations that are inherent in using ASPs.

Placing objects in user ASPs can provide several advantages:

- Additional data protection. By separating libraries, documents, or other objects in a user ASP, you protect them from data loss when a disk unit in the system ASP or other user ASPs fails. For example, if you have a disk unit failure, and data contained on the system ASP is lost, objects

contained in user ASPs are not affected and can be used to recover objects in the system ASP. Conversely, if a failure causes data that is contained in a user ASP to be lost, data in the system ASP is not affected.

- Improved system performance. Using ASPs can also improve system performance. This is because the system dedicates the disk units that are associated with an ASP to the objects in that ASP. For example, suppose you are working in an extensive journaling environment. Placing libraries and objects in a user ASP can reduce contention between the journal receivers and files if they are in different ASPs, which improves journaling performance. However, placing many active journal receivers in the same user ASP is not productive. The resulting contention between writing to more than one receiver in the ASP can slow system performance. For maximum performance, place each active journal receiver in a separate user ASP.
- Separation of objects with different availability and recovery requirements. You can use different disk protection techniques for different ASPs. You can also specify different target times for recovering access paths. You can assign critical or highly used objects to protected, high-performance disk units. You might assign large, low-usage files, like history files, to unprotected, low-performance disk units.

There are specific limitations that you may encounter when using ASPs:

- The system cannot directly recover lost data from a disk unit media failure. This situation requires you to perform recovery operations.
- Using ASPs can require additional disk devices.
- Using ASPs requires you to manage the amount of data in an ASP and avoid an overflowed ASP.
- You need to perform special recovery steps if an ASP overflows.
- Using ASPs requires you to manage related objects. Some related objects, such as journals and journaled files, must be in the same ASP.

B.4 Device parity protection

Device parity protection (RAID-5) is a hardware availability function that protects data from being lost because of a disk unit failure or because of damage to a disk. To protect data, the disk controller or input/output processor (IOP) calculates and saves a parity value for each bit of data. Conceptually, the disk controller or IOP computes the parity value from the data at the same location on each of the other disk units in the device parity set. When a disk failure occurs, the data can be reconstructed by using the

parity value and the values of the bits in the same locations on the other disks. The system continues to run while the data is being reconstructed. The overall goal of device parity protection is to provide high availability and to protect data as inexpensively as possible.

Device parity protection is built into the 2726, 2740, 2741, 6502, 6512, 6532, 6533, 6751, and 6754 input/output processors (IOPs). It can be activated for disk units that are attached to those IOPs. It is also built into the high-availability models of the 9337 Disk Array Subsystem.

If possible, you should protect all the disk units on your system with either device parity protection or mirrored protection. This prevents the loss of information when a disk failure occurs. In many cases, you can also keep your system operational while a disk unit is being repaired or replaced.

Remember

Device parity protection is not a substitute for a backup and recovery strategy. Device parity protection can prevent your system from stopping when certain types of failures occur. It can speed up your recovery process for certain types of failures. However, device parity protection does not protect you from many types of failures, such as a site disaster or an operator or programmer error. It does not protect against system outages that are caused by failures in other disk-related hardware (such as disk controllers, disk I/O processors, or a system bus).

Device parity protection provides the following benefits:

- Lost data is automatically reconstructed by the disk controller after a disk failure.
- The system continues to run after a single disk failure.
- A failed disk unit can be replaced without stopping the system.
- Device parity protection reduces the number of objects that are damaged when a disk fails.

There are also costs and limitations involved with using device parity protection:

- Device parity protection can require additional disk units to prevent slower performance.
- Restore operations can take longer when you use device parity protection.

B.5 Mirrored protection

Mirrored protection is a software availability function that protects data from being lost because of failure or because of damage to a disk-related component. Data is protected because the system keeps two copies of data on two separate disk units. When a disk-related component fails, the system may continue to operate without interruption by using the mirrored copy of the data until the failed component is repaired.

When you start mirrored protection or add disk units to an ASP that has mirrored protection, the system creates mirrored pairs using disk units that have identical capacities.

The overall goal is to protect as many disk-related components as possible. To provide maximum hardware redundancy and protection, the system attempts to pair disk units that are attached to different controllers, IOPs, and buses.

If a disk failure occurs, mirrored protection is intended to prevent data from being lost. Mirrored protection is a software function that uses duplicates of disk-related hardware components to keep your system available if one of the components fails. It can be used on any model of the AS/400 system and is a part of the Licensed Internal Code.

Different levels of mirrored protection are possible, depending on the hardware that is duplicated. You can duplicate:

- Disk units
- Disk controllers
- Disk I/O processors
- A bus

The system remains available during the failure if a failing component and the hardware components that are attached to it are duplicated.

In deciding whether to use mirrored protection on your system, you must evaluate the cost of potential downtime against the cost of additional hardware, over the life of the system. The additional cost in performance or system complexity is usually negligible. You should also consider other availability and recovery alternatives, such as device parity protection. Mirrored protection normally requires twice as many storage units. For concurrent maintenance and higher availability on systems with mirrored protection, other disk-related hardware may be required.

Remote mirroring support allows you to have one mirrored unit within a mirrored pair at the local site, and the second mirrored unit at a remote site. For some systems, standard DASD mirroring will remain the best choice. For others, remote DASD mirroring provides important additional capabilities. You must evaluate the uses and needs of your system, consider the advantages and disadvantages of each type of mirroring support, and decide which is best for you.

B.6 Separate servers

Separate servers provide a solution to businesses that require very high availability. Some or all data is maintained on two systems. The secondary system can take over critical application programs if the primary system fails.

The most common method for maintaining data on the secondary system is through the use of journaling. Journal entries from the primary system are transmitted to the secondary system. A user-written program on the secondary system receives the journal entries and uses them to update files and other journalled objects.

In this method, the journal entries are transmitted at the application layer by using the Receive Journal Entry (RCVJRNE) command or the Retrieve Journal Entries (QjoRetrieveJournalEntries) API. Using the Remote Journal function improves this method. This function allows the primary system to transmit the journal entries to a duplicate journal receiver on the secondary system at the Licensed Internal Code layer.

A third method is to copy journal receivers to tape regularly. The journal receivers are then restored to the secondary system. A user-written program uses the journal entries to update the files on the secondary system.

Several software packages are available from independent software vendors to support dual systems on the AS/400 system.

B.7 Clusters

Clusters are a configuration or a group of independent servers that appear on a network as a single machine. Or, as shown in Figure 73 on page 142, a cluster is a collection of complete systems that work together to provide a single, unified computing resource.

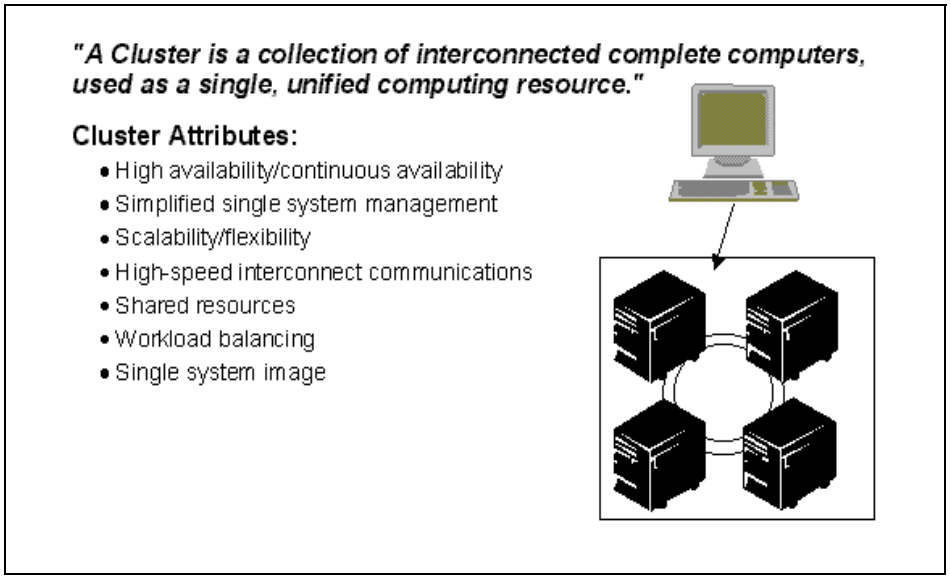


Figure 73. What is a cluster?

The cluster group shown in Figure 73 is managed as a single system or operating entity and is designed specifically to tolerate component failures and to support the addition or subtraction of components in a way that is transparent to users. Clusters let you efficiently group systems together to set up an environment that provides availability that approaches 100% for critical applications and critical data. Resources can be accessed without regard to location. A client interacts with a cluster as if it were a single system.

The major benefits that clusters offer your business are:

- Continuous availability of your systems, data, and applications
- Simplified administration of servers by allowing you to manage a group of systems as a single system or single database
- Increased scalability by allowing you to seamlessly add new components as your business growth requires

As described earlier, the AS/400 system is highly available. By replicating data and applications, system outages do not affect the ability for a business to keep on running. There is a single point of management and control and the various cluster systems are continuously communicating with each other. New systems can be added non-disruptively, incremental, and in a variety of

configuration options. Each added system has input/output capabilities that can increase overall cluster bandwidth.

The ultimate cluster in Figure 74 would be a single system image that, along with these benefits, would also handle workload balancing, allow the sharing of resources, high-speed interconnect communications, and so on. Although the V4R4 AS/400 system is not there yet, the AS/400 development team continues to work towards these goals.

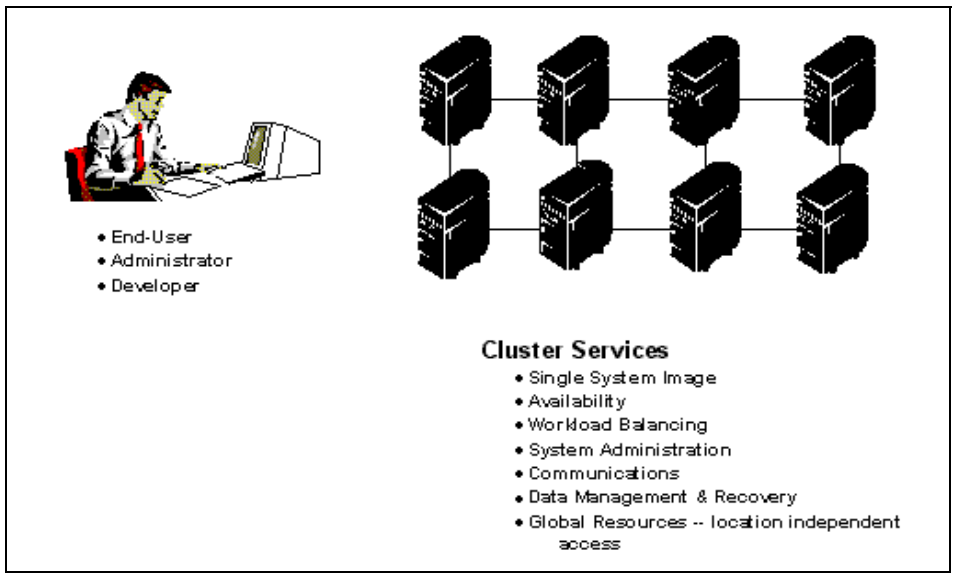


Figure 74. Ultimate cluster

B.8 Logical partitioning

AS/400 logical partitions let you run multiple independent OS/400 instances or partitions (each with its own processors, memory, and disks) in an N-way symmetric multiprocessing AS/400e, Model 6xx, Sxx, and 7xx. You can now address multiple system requirements in a single machine to achieve server consolidation, business unit consolidation, mixed production and test environments, and integrated clusters.

Logical partitions fall into two categories: primary partitions or secondary partitions. Each logically partitioned system has one primary partition and one or more secondary partitions. All V4R4 systems have a primary partition with all resources initially allocated to it. Creating and managing secondary partitions is performed from the primary partition. The movement of

processors, memory, and interactive performance between partitions can be achieved with only an IPL of the affected partitions. The movement of IOP resources can be achieved without an IPL.

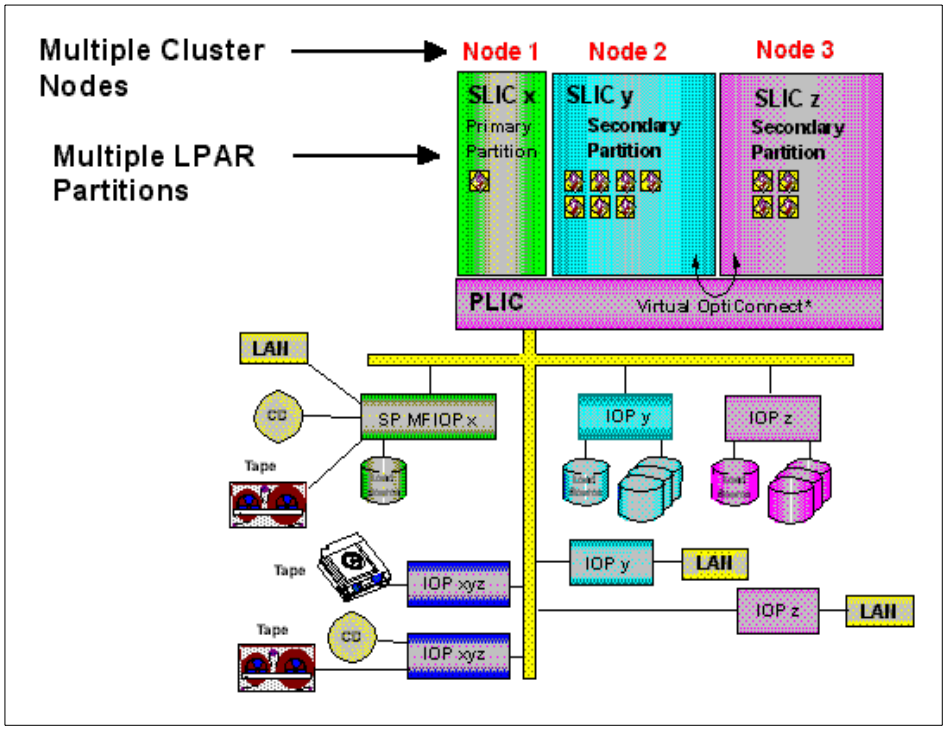


Figure 75. Logical partition cluster

Each logical partition represents a division of resources in your AS/400e system. Each partition is logical because the division of resources is virtual, not physical. The primary resources in your system are its processors, memory (main storage), I/O buses, and IOPs.

OS/400 is licensed once for the entire system by its normal processor group, regardless of the number of partitions. License management across partitions is not supported in this release. OS/400 V4R4 must be installed on each partition. Previous releases are not supported on a logical partition.

Each logical partition operates as an independent logical system. However, each partition shares a few physical system attributes such as the system serial number, system model, and processor feature code. All other system attributes may vary among partitions. For example, each partition has dedicated hardware such as processors, main storage, and I/O devices.

Logical partitions on a single AS/400 system can prove beneficial in the following scenarios.

- **Maintaining independent systems:** Dedicating a portion of the resources (disk storage unit, processors, memory, and I/O devices) to a partition achieves logical isolation of software. Logical partitions also have some hardware fault tolerance if configured properly. Interactive and batch workloads that may not run well together on a single machine can be isolated and run efficiently in separate partitions.
- **Consolidation:** A logically partitioned system can reduce the number of AS/400 systems that are needed within an enterprise. You can consolidate several systems into a single logically partitioned system. This eliminates the need for, and expense of, additional equipment. You can shift resources from one logical partition to another as needs change.
- **Mixed production and test environment:** You can create a combination production and test environment. You can create a single production partition in the primary partition. For multiple production partitions, see creating a multiple production partition environment in the following point. You can use a logical partition as a test partition or a production partition. A production partition runs your main business applications. A failure in a production partition could significantly hinder business operations and cost the customer time and money. A test partition tests software. This may include Year 2000 (Y2K) testing or compiling and running new software. A failure in a test partition, while not necessarily planned, will not disrupt normal business operations.
- **Multiple production partition environment:** You should create multiple production partitions only in your secondary partitions. In this situation, you dedicate the primary partition to partition management.
- **Hot backup:** When a secondary partition replicates another logical partition within the same system, switching to the backup during partition failure would cause minimal inconvenience. This configuration also minimizes the effect of long save windows. You can take the backup partition off line and save, while the other logical partition continues to perform production work. You need special software to use this hot backup strategy.
- **Integrated cluster:** Using OptiConnect/400 and high availability application software, your partitioned system can run as an integrated cluster. You can use an integrated cluster to protect your system from most unscheduled failures within a secondary partition. Figure 75 shows a conceptual view of multiple LPAR partitions in a clustered environment.

B.9 Additional information

For further information, visit the AS/400 Information Center at:
`http://www.as400.ibm.com/infocenter`

You can also refer to *OS/400 Backup and Recovery V4R4*, SC41-5304.

Appendix C. Problem determination

This chapter covers some of the more common problems you may run into when you set up and manage your cluster. First, you need to know where to look for error messages that deal specifically with clustering. You also need to know how to fix those errors once you find them.

C.1 Monitoring for problems

There are several places that you need to monitor for cluster error messages. You can find alertable messages in both the QHST history log or in the QSYSOPR message queue. Use the Display Log (DSPLOG) command to display the history log, or use the Display Message (DSPMSG) command to see what is in QSYSOPR. If you find that you are unable to correct the problem and need to call service, it is helpful to note the message ID and message text. For example, message ID CPFBB05 contains the help text: `Cluster node xx cannot be started.`

You can run the Work with Active Job (WRKACTJOB) command to display messages about your cluster resource service jobs. Cluster control messages go to the QCSTCTL job and Cluster Resource Group manager messages go to the QCSTCRGM job. Both of these job logs are found in the QSYSWRK subsystem.

You can also use the WRKACTJOB command to display messages about your Cluster Resource Group jobs. Under the QSYSWRK subsystem, look for the name of the Cluster Resource Group job that you created. For example, if you named your Cluster Resource Group CRG1, look for the job named CRG1 in QSYSWRK. To find an application Cluster Resource Group job, you need to know in which subsystem the application job will run. You can determine this by looking at the Cluster Resource Group. It contains the name of the user profile that is used when the application Cluster Resource Group job is submitted. Every user profile is associated with a job description and every job description is associated with a subsystem.

C.2 Common cluster questions

This section addresses commonly asked questions when working with clusters.

C.2.1 Why won't my cluster start?

Use the WRKACTJOB command to make sure that TCP/IP is active. Under the QSYSWRK subsystem, look for a job named QTCPIP. If this job exists, TCP/IP is running on your system. You can also use Netstat and select option 1 to see if TCP/IP is active. If it is not active, run the Start TCP/IP (STRTCP) command from your AS/400 command line. You also need to be sure that the *INETD server is started. Under the QSYSWRK subsystem, look for a job named QTOGINTD. If this job exists, the *INETD server is started. You can also use Netstat and select option 3 to see if the *INETD server is started. If it is not started, you can run the STRTCPSVR *INETD command from your AS/400 command line.

C.2.2 Why is my CRG hung up?

The Cluster Resource Group (CRG) may have submitted an exit program to a job queue in a subsystem that already has a job running. The CRG is waiting for the exit program to return a completion message. If this exit program does not complete, the CRG appears to be in a hung state. You need to be sure that the subsystem that your Cluster Resource Group exit program is running in allows more than one job to run at a time. Use the Change Subsystem Description (CHGSBSD) command and specify *NOMAX for the maximum jobs parameter. If it is not possible to change maximum jobs to *NOMAX, consider creating a separate subsystem.

C.2.3 Is my cluster up and running?

To determine if Cluster Resource Services is active on your system, run the WRKACTJOB command from an AS/400 command line. Under the QSYSWRK subsystem, look for two jobs named QCSTCTL and QCSTCRGM. If these jobs exist, Cluster Resource Services is active and your cluster is up and running.

C.2.4 Why do I have two clusters after fixing my cluster partition?

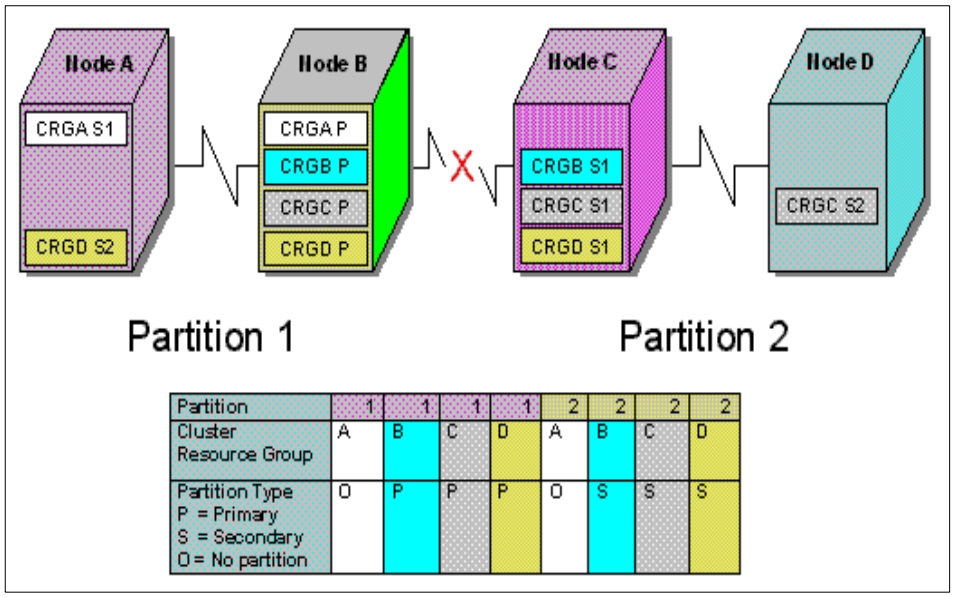
The most common reason for this happening that is you most likely ran the Start Cluster Node (QcstStartClusterNode) API on the inactive node. You need to run this on an active node in your cluster to start Cluster Resources Services on the inactive node.

C.3 Recovering from a clustered partition

A cluster partition happens if you lose contact between one or more nodes in the cluster and a failure of the lost nodes cannot be confirmed. If you receive

error message CPFBB20, you are in a partitioned cluster situation and you need to know how to recover. You will find this error message in the QHST history log and in the QCSTCTL job log in the QSYSWRK subsystem.

The example in Figure 76 shows a partition condition that involves a cluster made up of four nodes A, B, C, and D. The example shows a loss of communication between cluster nodes B and C has occurred, which results in the cluster dividing into two cluster partitions. Before the cluster partition occurred, there were four Cluster Resource Groups called CRGA, CRGB, CRGC, and CRGD. The example shows the recovery domain of each Cluster Resource Group.



Partition	1	1	1	1	2	2	2	2
Cluster Resource Group	A	B	C	D	A	B	C	D
Partition Type P = Primary S = Secondary O = No partition	O	P	P	P	O	S	S	S

Figure 76. Cluster partition

To determine the types of Cluster Resource Group actions that you can take within a cluster partition, you need to know whether the partition is a primary or a secondary cluster partition. The cluster partition that contains the current primary node in the recovery domain of a Cluster Resource Group is considered the primary partition of the Cluster Resource Group. All other partitions are secondary partitions. The primary partitions may not be the

same for all Cluster Resource Groups. The restrictions for each Cluster Resource Group API are shown in Table 5.

Table 5. Cluster Resource Group API partition restrictions

Cluster Resource Group API	Partition allowed In
Add Node to Recovery Domain	Allowed only in a primary partition
Change Cluster Resource Group	Allowed only in a primary partition
Create Cluster Resource Group	Not allowed in any partition
Delete Cluster Resource Group	Allowed in any partition, but only affects the partition running the API
End Cluster Resource Group	Allowed only in a primary partition
Initiate Switch Over	Allowed only in a primary partition
List Cluster Resource Groups	Allowed in any partition
List Cluster Resource Group Information	Allowed in any partition
Remove Node from Recovery Domain	Allowed only in a primary partition
Start Cluster Resource Group	Allowed only in a primary partition

By applying these restrictions, Cluster Resource Groups can be resynchronized when the cluster is no longer partitioned. As nodes rejoin the cluster from a partitioned status, the version of the Cluster Resource Group in the primary partition is copied to nodes from a secondary partition.

When a partition is detected, neither the Add Cluster Node Entry or the Create Cluster API can be run in any of the partitions. All of the other Cluster Control APIs may be run in any partition. However, the action performed by the API takes affect only in the partition running the API.

Once you correct the partitioned cluster situation, you receive the message CPFBB21. This message lets you know that you have recovered from the cluster partition. You can find this message in the QHST history log and in the QCSTCTL job log in the QSYSWRK subsystem.

C.3.1 Cluster partition tips

The following list offers some cluster partition tips:

- The rules for restricting operations within a partition are designed to make merging the partitions feasible. Without these restrictions, reconstructing the cluster would require extensive work by you.

- If the nodes in the primary partition have been destroyed, special processing may be necessary in a secondary partition. The most common scenario that causes this condition would be the loss of the site that made up the primary partition. Use the example shown in Figure 76 on page 149 and assume that Partition 1 was destroyed. In this case, the primary node for Cluster Resource Groups B, C, and D must be located in Partition 2. To do this, perform these operations:

  a. Delete Cluster Resource Groups B, C, and D in Partition 2.

  b. Remove Nodes A and B from the cluster in Partition 2. Partition 2 is now the cluster.

  c. Create Cluster Resource Groups B, C, and D in Partition 2 specifying Nodes C and D as the recovery domain.

  d. Establish any replication environments needed in the new cluster.

  Since nodes have been removed from the cluster definition in Partition 2, an attempt to merge Partition 1 and Partition 2 will fail. To correct the mismatch in cluster definitions, run the Delete Cluster API on each node in Partition 1. Then add the nodes from Partition 1 to the cluster, and reestablish all the Cluster Resource Group definitions, recovery domains, and replication. This requires a great deal of work and is also prone to errors. It is important that you do this procedure only in a site loss situation.

- Processing a start node operation depends on the status of the node that is being started:
  - The node either failed or an End Node operation ended the node:
    - Cluster Resource Services is started on the node that is being started.
    - A cluster definition is copied from an active node in the cluster to the node that is being started.
    - Any Cluster Resource Group that has the node being started in the recovery domain is copied from an active node in the cluster to the node being started. No Cluster Resource Groups are copied from the node that is being started to an active node in the cluster.
  - The node is a partitioned node:
    - The cluster definition of an active node is compared to the cluster definition of the node that is being started. If the definitions are the same, the start continues as a merge operation. If the definitions do not match, the merge stops, and the user needs to intervene.

- If the merge continues, the node that is being started is set to an active status.
- Any Cluster Resource Group that has the node being started in the recovery domain is copied from the primary partition of the Cluster Resource Group to the secondary partition of the Cluster Resource Group. Cluster Resource Groups may be copied from the node that is being started to nodes that are already active in the cluster.

C.3.2 Merging a cluster partition example

A merge operation is similar to a rejoin operation except that it occurs when a cluster has become partitioned. The partition may be a true partition in that Cluster Resource Services is still active on all nodes. However, some nodes can't communicate with other nodes due to a communication line failure. Or, the problem may be that a node actually failed, but was not detected as a failure.

In the first case, the partitions are merged back together automatically once the communication problem is fixed. This happens when both partitions periodically try to communicate with the partitioned nodes and eventually re-establish contact with each other. In the second case, Cluster Resource Services must be restarted on the failed node. CRS must be restarted by calling the Start Cluster Node API from one of the nodes that is active in the cluster. If you call the Start Cluster Node API on the failed node, it becomes a one node cluster and will not merge back into the rest of the cluster.

As shown in Figure 77, a merge operation can occur with one of the configurations that are present.

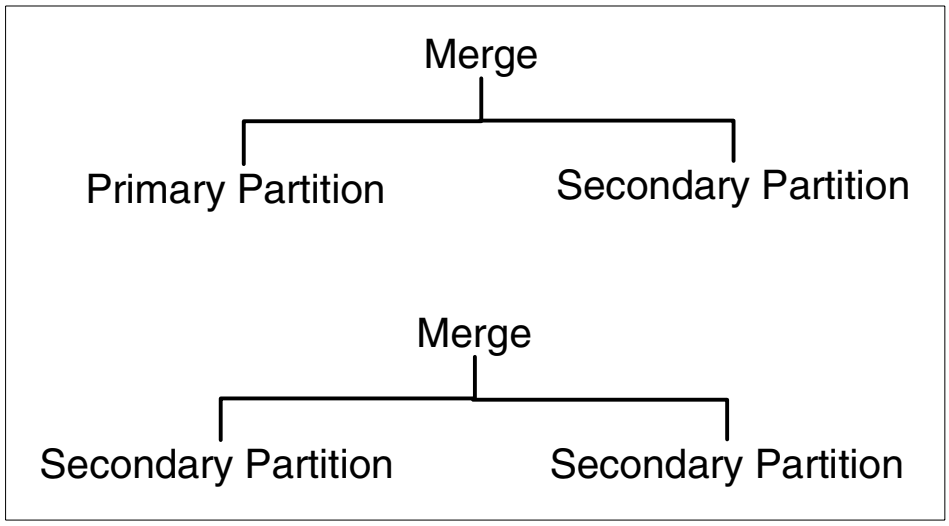


Figure 77. Possible merge operations

Primary and secondary partitions are unique to Cluster Resource Groups. For a CRG, a primary partition is defined as a partition that has the CRG's primary node active in it. A secondary partition is defined as a partition that does not have the primary node active in it. For example, a cluster has two nodes, A and B, and two CRGs, CRG1 and CRG2. Node A is the primary node for CRG1, and node B is the backup node. Node B is the primary node for CRG2, and node A is the backup node. If a partition occurs, node A is the primary partition for CRG1 and the secondary partition for CRG2. Node B is the primary partition for CRG2 and the secondary partition for CRG1. See Figure 78 on page 154.

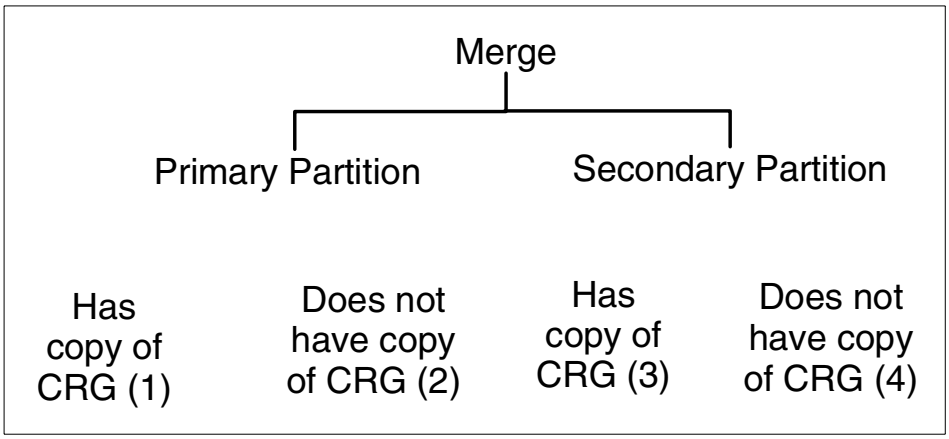


Figure 78. Primary-secondary merge operation

During a primary secondary merge as shown in Figure 78, the following situations are possible:

- 1 and 3
- 1 and 4
- 2 and 3: Cannot happen since a primary partition has the primary node active and must have a copy of the CRG
- 2 and 4: Cannot happen since a primary partition has the primary node active and must have a copy of the CRG

C.3.2.1 Primary-secondary merge situations

A copy of the CRG object is sent to all nodes in the secondary partition. The following actions can result on the nodes in the secondary partition:

- No action since the secondary node is not in the CRG's recovery domain.
- A secondary node's copy of the CRG is updated with the data from the primary partition.
- The CRG object is deleted from a secondary node since the secondary node is no longer in the CRG's recovery domain.
- The CRG object is created on the secondary node since the object does not exist. However, the node is in the recovery domain of the CRG copy that is sent from the primary partition.

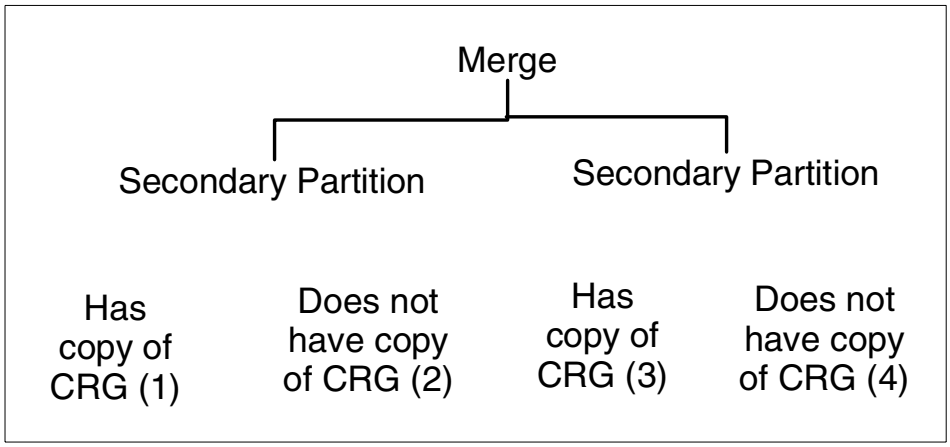


Figure 79. Secondary-secondary merge operation

During a secondary-secondary merge as shown in Figure 79, the following situations are possible:

- 1 and 3
- 1 and 4
- 2 and 3
- 2 and 4

C.3.2.2 Secondary-secondary merge situation 1

The node with the most recent change to the CRG is selected to send a copy of the CRG object to all nodes in the other partition. If multiple nodes are selected because they all appear to have the most recent change, the recovery domain order is used to select the node. The actions that can occur on the receiving partition nodes are:

- No action since the node is not the CRG's recovery domain.
- The CRG is created on the node since the node is in the recovery domain of the copy of the CRG object it receives.
- The CRG is deleted from the node since the node is not in the recovery domain of the copy of the CRG object is receives.

C.3.2.3 Secondary-secondary merge situations 2 and 3

A node from the partition that has a copy of the CRG object is selected to send the object data to all nodes in the other partition. The CRG object may be created on nodes in the receiving partition if the node is in the CRG's recovery domain.

C.3.2.4 Secondary-secondary merge situation 4

Internal data is exchanged to ensure consistency throughout the cluster.

A primary partition can subsequently be partitioned into a primary and secondary partition. If the primary node fails, CRS detects it as a node failure. The primary partition becomes a secondary partition. The same result would occur if you ended the primary node that uses the End Cluster Node API. A secondary partition can become a primary partition if the primary node becomes active in the partition either through a rejoin or merge operation.

For a merge operation, the exit program is called on all nodes in the CRG's recovery domain regardless of the partition in which the node is located. The same action code as rejoin is used. No roles are changed as a result of the merge, but the status of the nodes in the CRG's recovery domain is changed from partition to active. Once all partitions merge together, the partition condition is cleared, and all CRG APIs can be used.

Appendix D. Special notices

This publication is intended to help customers, IBMers, and Business Partners to understand the AS/400 cluster environment, so that high levels of system and application availability can be achieved. See Appendix E, “Related publications” on page 161, for currently available information.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter 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 the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594 USA.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

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

AIX	AS/400
AS/400e	AT
C/400	ClusterProven
CT	DB2
Distributed Relational Database Architecture	DRDA
IBM ®	Manage. Anything. Anywhere.
Netfinity	OS/400
Parallel Sysplex	PartnerWorld
RS/6000	S/390
SP	System/38
System/390	Wizard
400	Lotus
Tivoli	TME
NetView	Cross-Site
Tivoli Ready	Tivoli Certified

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere.,TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

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

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

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

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

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

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

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

Appendix E. Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

E.1 IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 163.

- *DB2/400 Advanced Database Functions*, SG24-4249
- *AS/400 Remote Journal Function for High Availability and Data Replication*, SG24-5189
- *Slicing the AS/400 with Logical Partitioning: A how to Guide*, SG24-5439
- *Database Parallelism on the AS/400*, SG24-4826

  This publication can only be access online in softcopy format from the redbooks Web site at: `http://www.redbooks.ibm.com/`

  At the site, click **Redbooks Online** and enter the title or publication number in the search field that appears. Then, click on the appropriate book title.

E.2 IBM Redbooks collections

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at `http://www.redbooks.ibm.com/` for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	Collection Kit Number
System/390 Redbooks Collection	SK2T-2177
Networking and Systems Management Redbooks Collection	SK2T-6022
Transaction Processing and Data Management Redbooks Collection	SK2T-8038
Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
AS/400 Redbooks Collection	SK2T-2849
Netfinity Hardware and Software Redbooks Collection	SK2T-8046
RS/6000 Redbooks Collection (BkMgr)	SK2T-8040
RS/6000 Redbooks Collection (PDF Format)	SK2T-8043
Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

E.3 Other resources

These publications are also relevant as further information sources:

- *TCP/IP Tutorial and Technical Overview*, GG24-3442
- *System API Reference V4R4*, SC41-5801
- *OS/400 Backup and Recovery V4R4*, SC41-5304
- Gartner Group. *Platform Availability Data: Can You Spare a Minute?* October 1998.
- Toigo, Jon. *Disaster Recovery Planning: Managing Risks and Catastrophe in Information Systems*. Yourden Press Computing Services, 1989 (ISBN 0132149419).

E.4 Referenced Web sites

These Web sites are also relevant as further information sources:

- More details of IBM S/390 Parallel Sysplex can be found at its home page at: `http://www.s390.ibm.com/pso/`
- DataMirror, an IBM Business Partner, can be accessed online at: `http://www.datamirror.com`
- LakeView Technology, an IBM Business Partner, can be accessed online at: `http://www.lakeviewtech.com`
- Vision Solutions, an IBM Business Partner, can be accessed online at: `http://www.visionsolutions.com`
- Access the Benchmark Center for pre-production testing of your applications. The center can be found online at: `http://www.partnerworld.ibm.com`
- Visit the AS/400 Information Center at: `http://www.as400.ibm.com/infocenter`
- Visit the IBM AS/400 home page at: `http://www.as400.ibm.com/ha/ha2_99.htm`
- Visit the IBM ClusterProven home page at: `http://www.ibm.com/servers/clusters`
- Visit the IBM PartnerWorld for Developers home page at: `http://www.developer.ibm.com`

How to get IBM Redbooks

This section explains how both customers and IBM employees can find out about IBM Redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** http://www.redbooks.ibm.com/

  Search for, view, download, or order hardcopy/CD-ROM Redbooks from the Redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

  Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders by e-mail including information from the IBM Redbooks fax order form to:

	e-mail address
In United States	usib6fpl@ibmmail.com
Outside North America	Contact information is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the Redbooks Web site.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and Redbooks by accessing the IBM Intranet Web site at http://w3.itso.ibm.com/ and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access MyNews at http://w3.ibm.com/ for redbook, residency, and workshop announcements.

IBM Redbooks fax order form

Please send me the following:

Title	Order Number	Quantity

First name	Last name	
Company		
Address		
City	Postal code	Country
Telephone number	Telefax number	VAT number

☐ Invoice to customer number _____

☐ Credit card number _____

Credit card expiration date	Card issued to	Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Index

Numerics

24x365 30
24x365 usage 28
5250 datastream 63
99.9+% availability 20

A

access path 29, 136
access path protection 29
activating and de-activating nodes 82
  in the cluster 112
activating or starting a data CRG 89
activating or starting a data or application CRG 115
activating or starting a resilient application 95
active secondary 22
adapter redundancy 71
adding a node to a cluster 74, 80, 111
Advanced ClusterProven 53
API 41, 43, 47, 52, 97, 99, 103
application CRG 83, 90
application error 5
application maintenance 69
application object inventory 60
application resilience 40, 41, 54
application resiliency 52, 99
application servers 30
application-level availability 14
apply process 90
AS/400 cluster 37
AS/400 clustering technology 37
ASP (auxiliary storage pool) 29, 136, 137
ASP overflow 5
auto-detection of clustered nodes 107
auxiliary storage pool (ASP) 29, 137
availability technology 5, 19

B

backup 15, 27, 42, 87, 94
backup model 9
backup system 69
basic cluster 30
Basic ClusterProven 53
batch 63
batch applications in a cluster 70
BEST/1 70
business boundaries 12
business critical 102
business impact costs 56
business unit consolidation 32

C

capacity 70
changing a CRG's recovery domain 114
changing or updating a resilient application 93
changing or updating data areas 123
client server 63
client-centric 54
cluster 3, 30, 141
  hardware requirements 31
  software requirements 31
cluster communications 43
cluster control 41, 42, 46
cluster engine 42, 48, 50
cluster management 28, 52, 77
cluster management middleware 34
cluster management related tests 74
cluster management solution 7
cluster management tool 30, 69
cluster management utility 34, 41
cluster membership 48
cluster middleware 51, 52, 99
cluster node 37
cluster partition 48, 97
Cluster Resource Group (CRG) 38, 59, 98
Cluster Resource Group manager 41, 42, 46
Cluster Resource Services (CRS) 6, 30, 31, 41, 50, 52
cluster resources 37, 39
Cluster Services 97
cluster testing 72
cluster topology services 43
ClusterProven 6, 84, 91, 100, 102, 105, 108
ClusterProven applications 91
cold backup 9, 22
commitment control 10, 20, 23
communication bandwidth 11
components of single system availability 19
concurrent apply PTF 28
configuration of a cluster 58
configure 55, 56
consolidation 145
continuosly power main storage 28
continuous availability 4, 16, 21, 28, 30, 40

continuous operations 4
continuously available 3
CPM 28
creating a cluster 79
creating an application CRG recovery domain 118
creating and using Cluster Resource Groups 83, 113
creating application CRGs 87
creating data CRGs 83
creating ISV data areas for application CRGs 122
creating new clusters 109
CRG 38, 40, 42, 46, 69, 77, 108, 113
CRG (Cluster Resource Group)) 38
critical applications 142
critical data 40
critical resource 30
CRS (Cluster Resource Services) 31

D

data CRG 83, 90, 93, 95, 97, 108
data integrity 16
data queue 61
data resilience 40, 41
data resiliency 52
data space 61
database servers 30
DataMirror 7, 52, 77
DataMirror HA Suite 7
DB2 multi-system 27
DDM (distributed data management) 27
de-activating or ending
  a data CRG 90
  a data or application CRG 116
  a resilient application 95
device parity protection 29, 138
disaster 3, 9
disaster recovery 4, 16, 22, 28
disk 11
disk redundancy 71
distributed activities 45
distributed activity group membership 48
distributed activity groups 46, 50
distributed data management (DDM) 27
Distributed Relational Database Architecture (DRDA) 27
distributed systems 3
downtime 14
downtime issues 4, 9
DRDA (Distributed Relational Database Architecture) 27

E

e-commerce 5, 15
Emergency Power Off (EPO) 4
emulator session 62
enterprise resource planning 13
environmental system 5
example of business impact 15

F

factors influencing availability 11
failover 30, 44, 59, 69, 74, 77
failure 9
financial impact 13
fix 15
fix applies 5
four-node mutual takeover 39
four-node mutual takeover cluster 59

G

general system managment related tests 74
group membership 48
group membership services 42, 46, 48
group messaging services 42

H

HABP 23, 68, 69, 100
HABP (High Availability Business Partner) 34
hardware 5, 71
hardware upgrade 5, 15, 28
heartbeating 43, 44, 50
High Availability 3
high availability 3, 4, 51, 135
High Availability Business Partner (HABP) 34
High Availability Business Partners (HABP) 7
High Availability middleware 34
horizontal growth 3, 9, 27, 38
host-centric 54
hot backup 10, 22, 145
hot or cold sites 9

I

I/T infrastructure 6
IBM Benchmark Center 74
IBM S/390 Parallel Sysplex 11, 26

IBM Server Group 51
iCluster 7, 77
iCluster Administrator 78
impact to the business cost 55
independent software vendor 41, 51, 91
information object 42
input/output processor 29, 138
interactive job 62
inventry of all the system hardware 58
IOP 29, 32, 138, 144
IP address 44, 69
IP address takeover 41
IP multicast 43
IP network 37
IP takeover 63
IPL 15, 28, 136, 144
ISV 41, 51, 91, 105
ISV data area 121

J

job structure 46
journal entries 135
journal management 28
journaling 10, 20, 23, 28, 67, 90, 95, 135

L

Lakeview Technology 7, 52, 99
LAN (local area network) 27, 58
level of availability 55, 57
local area network (LAN) 27
logical partition 33
logical partitioning (LPAR) 32, 143
loosely coupled cluster 27
LPAR 32, 59, 72
LPAR (logical partitioning) 32

M

machine interface 43
maintenance 63
master node 79
membership change message 48
membership list 37
memory 11
messaging services 48
MI 43
Microsoft Cluster Services 24
MIMIX 7, 99
MIMIX Cluster Manager 100
MIMIX Cluster Optimizer 100
MIMIX ClusterServer 7, 99
MIMIX FastPath 7, 99
MIMIX Replicator 100
MIMIX-ACE 100
MIMIX-ACT 100
mirrored disks 58
mirrored protection 29, 140
mission critical 21
mixed production and test environment 32
multiple nodes 40
multi-system services 27

N

network connection redundancy 71
network failure 5
network hardware redundancy 71
network planning 72
network resources 55, 56
node 30, 31, 50, 59, 70, 79

O

object replication 77
object specifier 87, 94, 125
OLTP 25
OMS/400 Cluster Manager 7, 105
operations management 68
OptiConnect 27, 145
OS/400 49
OS/400 V4R4 9, 27, 33, 34, 40, 50, 102
OS/400 Version 4 Release 4 6
outage 4, 38, 39, 50

P

partition 38
partition state 48, 69
PartnerWord for Development AS/400 53
peer cluster nodes 44
peer relationships 50
performance 67, 70, 138
periferal devices 58
planned switch 74
planning for AS/400 clusters 6, 55
planning steps 55
power outages 15
primary 39, 42, 87, 94

primary node 77, 90
primary partitions 32
primary server 53
primary system 31, 141
problem and change management 69
problem determination 8
processors 11
PTF 28

Q

QCSTCRGM 46
QCSTCTL 46
QCSTHAAPPI 91, 93, 94
QSYSWRK 46, 47

R

RAID 58
RAID-5 29, 138
RCLDLO (Reclaim Document Library Object) 28
Reclaim Document Library Object (RCLDLO) 28
recovering storage 15
recovery domain 38, 42, 87, 90, 91
redundancy 71
redundant computer centers 9
rejoin 74
release installations 15
reliability 21
remote journaling 27
remote site redundancy 71
removing a data CRG 96
removing a data or application CRG 119
removing a node from the cluster 96, 119
removing a resilient application 96
removing the cluster 96
removing the entire cluster 97, 120
reorganizing file 15
replicate 42
replication 23, 28, 41, 77, 79, 100
replication technology 23
replication utility 7
resilience 70
resilient 77
resilient application 48, 58, 77, 84, 87, 89, 90, 91, 95, 97, 105
  data area contents 124
  recovery domain 94
resilient data 31, 105
resilient hardware 55, 56
resilient processes 3
resilient resource 38, 44
restart 20, 54, 62
restricted state 28
roll back 20
routers 45, 58

S

scalability 3
scheduled downtimes 5
scheduled outage 15, 21, 28
secondary partitions 32
secondary system 141
security 70
selecting objects to a data CRG 85
selecting objects to a resilient application 93
separate server 23, 27, 38, 141
server consolidation 32
server platforms 15
service level agreements 12, 68
setting up a resilient application 91
shared disk 25, 26
shared-nothing 38
single system 11
single system availability 5, 9, 19, 27
single system environment 68
single system image 32
site disasters 15
site loss 30
site maintenance 15
site redundancy 71
SMAPP 136
SMP (symmetric multi-processing) 26
software upgrades 5
standby secondary 22, 23
strategic solution 57
switch over 28, 31, 59, 62, 77
switched disk 23, 24
switching over a data CRG 90
switching over a data or application CRG 117
switching over a resilient application 95
symmetric multi-processing (SMP) 9, 26, 32
system auxiliary storage pool 137
system outage 30
system-managed access-path (SMAP) 136
systems management 6, 68

T
tactical solution 57
takeover IP address 91
temporary files 61
test environment 55, 56, 73, 145
twinax display 62
two-node cluster 39, 59
type of disaster 4

U
uninterruptable power supply (UPS) 28
unscheduled outage 15, 28
UPS (uninterruptable power supply) 28
user auxiliary storage pool 137

V
versioning 49
Vision Solutions 7, 52, 105
Vision Suite 7, 105

W
WAN (wide area network) 27
warm backup 22
Web site crashes 3
wide area network (WAN) 27
working with clusters and CRGs 109
workload balancing 9, 27
workload peaks 22

IBM Redbooks review

Your feedback is valued by the Redbook authors. In particular we are interested in situations where a Redbook "made the difference" in a task or problem you encountered. Using one of the following methods, **please review the Redbook, addressing value, subject matter, structure, depth and quality as appropriate.**

- Use the online **Contact us** review redbook form found at **ibm.com**/redbooks
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Document Number **Redbook Title**	SG24-5194-00 AS/400 Clusters: A Guide to Achieving Higher Availability
Review	
What other subjects would you like to see IBM Redbooks address?	
Please rate your overall satisfaction:	O Very Good O Good O Average O Poor
Please identify yourself as belonging to one of the following groups:	O Customer O Business Partner O Solution Developer O IBM, Lotus or Tivoli Employee O None of the above
Your email address: The data you provide here may be used to provide you with information from IBM or our business partners about our products, services or activities.	O Please do not use the information collected here for future marketing or promotional contacts or other communications beyond the scope of this transaction.
Questions about IBM's privacy policy?	The following link explains how we protect your personal information. **ibm.com**/privacy/yourprivacy/

AS/400 Clusters: A Guide to Achieving Higher Availability

(0.2”spine)
0.17”<->0.5”
90<->249 pages

AS/400 Clusters

A Guide to Achieving Higher Availability

Explore and understand your AS/400 high availability options

Find out how AS/400 clusters can improve business uptime

Preview solutions by AS/400 High Availability Business Partners

Gain a broad understanding of the new cluster architecture available with OS/400 Version 4 Release 4. In this era of e-commerce, availability is of the utmost importance for business survival. This new cluster architecture provides support for customers who want to make their businesses continuously available.

This redbook presents an overview of a generic cluster and the basic terminology surrounding clusters. It also examines the AS/400 cluster and its implementation. It introduces you to the new brand initiative ClusterProven for AS/400 and explains how it applies to AS/400 customers and independent software vendors.

This redbook targets IBM customers, technical representatives, and Business Partners who are planning business solutions and systems that are continuously available.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-5194-00 ISBN 0738417297