IBM SAN Volume Controller Stretched Cluster with PowerVM and PowerHA

- Learn about the IBM Stretched Cluster solution and architecture
- See how to implement it with Power Systems
- Learn about diagnosis and recovery guidelines

Shivendra Ashish
Kameswara Bhaskarabhatla
Huang Chen
William Genovese
Jon Tate

ibm.com/redbooks

International Technical Support Organization

IBM SAN Volume Controller Stretched Cluster with PowerVM and PowerHA

July 2013

SG24-8142-00

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

First Edition (July 2013)

IBM SAN Volume Controller code level was 6.4.1.2 (build 75.0.1211301000). The IBM Storwize V7000 back-end storage was 6.4.1.2 (build 75.0.1211301000) as was the quorum storage.

© Copyright International Business Machines Corporation 2013. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices ix
Trademarks x

Preface xi
Authors xi
Now you can become a published author, too! xiii
Comments welcome xiv
Stay connected to IBM Redbooks xiv

Chapter 1. Introduction 1
1.1 High Availability concepts, definitions, and building blocks 3
1.1.1 Starting points and key definitions: Reliability, Availability, Serviceability 4
1.1.2 Key industry definitions 5
1.1.3 Infrastructure availability continuum 6
1.1.4 High Availability building blocks: IBM Power Systems best practices 10
1.1.5 Tiers of DR 29
1.1.6 Blending tiers into an optimized solution 31
1.2 Storwize V7000 33
1.2.1 Overview and concepts 33
1.2.2 Architecture and solution context 36
1.3 SAN Volume Controller 38
1.3.1 Overview and Concepts 38
1.3.2 Architecture and solution context 42
1.4 SAN Volume Controller Stretched Cluster 46
1.4.1 Overview and concepts 46
1.4.2 Architecture and solution context 49
1.5 Fibre Channel SAN 50
1.5.1 Overview and concepts 50
1.5.2 Architecture and solution context 51
1.6 Power 750 Express and Hardware Management Console 53
1.6.1 IBM Power 750 (8233-E8B) Express Server 53
1.6.2 IBM Hardware Management Console 55
1.7 Architecture and solution context 61
1.7.1 Power 750 system architecture 61
1.7.2 Power 750 and HMC 62
1.8 PowerVM technology 66
1.9 PowerVM 66
1.9.1 Architecture and solution context 68
1.9.2 Virtual I/O Server 71
1.9.3 Power Systems network virtualization 72
1.10 Live Partition Mobility 78
1.10.1 Overview and concepts 78
1.10.2 LPM prerequisites and requirements 78
1.10.3 Architecture and solution context 79
1.11 N_Port ID Virtualization 82
1.11.1 Overview and concepts 82
1.11.2 Architecture and solution context 85
1.11.3 SDDPCM 87
1.11.4 Legacy failure categories 88

© Copyright IBM Corp. 2013. All rights reserved.

1.11.5 Emerging failure categories 88
1.11.6 Multi-path I/O 91
1.12 PowerHA 95
1.12.1 Overview and concepts 95
1.13 Architecture and solution context 99
1.13.1 Reliable Scalable Cluster Technology 99
1.13.2 SAN Volume Controller Stretched Cluster solution context: PowerHA clusters and workload 101
1.14 SAN Volume Controller, Layer-2 IP Network, Storage Networking infrastructure and LPM integration 104
1.14.1 Overview and concepts 104
1.14.2 Application mobility over distance 105
1.14.3 IBM, PowerVM, and Layer 2 IP Network solution 105
1.15 Architecture and solution context 106
1.15.1 IP and Layer 2 network architecture 106
1.16 Open data center Interoperable Network 110
1.16.1 Overview and concepts 110
1.16.2 Architecture and solution context 111

Chapter 2. Hardware and Software Description 115
2.1 IBM System Storage SAN Volume Controller 117
2.2 SAN directors and switches 117
2.2.1 SAN384B-2 and SAN768B-2 directors 117
2.2.2 SAN24B-5, SAN48B-5, and SAN80B-4 switches 118
2.3 FCIP routers 120
2.3.1 8 Gbps Extension Blade 120
2.3.2 SAN06B-R extension switch 121
2.4 Ethernet switches and routers 121
2.4.1 IBM system networking switches 121
2.4.2 Brocade IP routers and Layer 4 - 7 application delivery controllers 123
2.5 P750 and Hardware Management Console 127
2.5.1 P750 127
2.5.2 HMC 129
2.6 Software 130
2.6.1 IBM Fabric OS 130
2.6.2 IBM Network Advisor 131
2.6.3 IBM 3759 Fabric Ethernet Network Networking Operating System 133
2.6.4 IBM Storwize V7000 134
2.6.5 IBM SAN Volume Controller 134
2.7 IBM Power Systems operating system 135
2.7.1 IBM server firmware 139
2.8 IBM PowerVM: VIOS, NPIV, and SDDPCM/MPIO 140
2.8.1 PowerVM Editions 141
2.8.2 IBM Virtual I/O Server 143
2.8.3 IBM LPM 143
2.8.4 N_Port ID Virtualization 143
2.8.5 Subsystem Device Driver Path Control Module 144
2.8.6 MPIO 144
2.9 PowerHA 145
2.9.1 PowerHA SystemMirror 7.1 deprecated features 145
2.9.2 PowerHA SystemMirror 7.1 new features 146
2.10 IBM DB2 147
2.11 IBM WebSphere Application Server 147

Chapter 3. SAN Volume Controller Implementation 149
3.1 Test environment 150
3.2 ADX: Application Delivery Controller 151
3.2.1 VIP and real server configuration 152
3.2.2 GSLB configuration 153
3.3 IP networking configuration 155
3.4 IBM System Networking Rack Switch G8264CS 156
3.4.1 Layer 2 switch configuration 159
3.4.2 IP Core (MLXe) configuration 161
3.4.3 Data Center Interconnect configuration 165
3.5 IBM FC SAN 169
3.5.1 Creating the logical switches 170
3.5.2 Creating FCIP tunnels 179
3.6 IBM SAN Volume Controller using Stretched Cluster 182
3.7 SAN Volume Controller volume mirroring 183
3.7.1 Volume mirroring prerequisites 184
3.8 SAN Volume Controller I/O operations on mirrored volumes 185
3.8.1 Preferred node 185
3.8.2 Primary copy 185
3.9 Read operations 185
3.10 Write operations 186
3.10.1 `-mirrorwritepriority` settings 187
3.10.2 Best Practices 188
3.11 SAN Volume Controller quorum disk 188
3.12 Quorum disk requirements and placement 189
3.13 Automatic SAN Volume Controller quorum disk selection 190
3.13.1 SAN Volume Controller cluster state and voting 194
3.14 Failure scenarios in SAN Volume Controller Split Cluster configuration 195
3.15 SAN Volume Controller Split Cluster configurations 198
3.15.1 No ISL configuration 198
3.15.2 ISL configuration 201
3.15.3 FCIP configuration 204
3.16 Fibre Channel settings for distance 206
3.17 Back-end storage allocation to the SAN Volume Controller Cluster 208
3.18 Volume allocation 209

Chapter 4. PowerVM and PowerHA Implementation 213
4.1 PowerVM overview 215
4.2 PowerVM considerations 216
4.2.1 NPIV planning 218
4.3 PowerVM initial configuration 219
4.3.1 Power Systems management console 219
4.4 Installing VIOS 220
4.4.1 Creating a Virtual I/O Server 220
4.4.2 Installation of Virtual I/O Server 224
4.4.3 Configuring VIOS for client network 225
4.4.4 Configuring VIOS for client storage 232
4.5 Installing AIX and creating VIOS clients 237
4.5.1 AIX client partition installation 239
4.5.2 AIX HOST path configuration 243
4.6 NPIV initial configuration 245
4.7 Live Partition Mobility implementation 248
4.7.1 LPM components 248

4.7.2 Other components affecting LPM 250
4.8 LPM operations 251
4.8.1 Inactive migration 251
4.8.2 Active migration 253
4.9 Combining LPM with other features 258
4.9.1 High availability clusters 258
4.10 LPM configuration 259
4.10.1 LPM prerequisites 259
4.10.2 System level preparation 260
4.10.3 VIOS preparation 261
4.10.4 Client configuration 261
4.11 LPM 262
4.11.1 Source system validation 262
4.11.2 Migration 266
4.12 Error identification and resolution 267
4.13 LAM considerations 268
4.14 PowerHA SystemMirror initial configuration and other considerations 269
4.14.1 Considerations before PowerHA 7.1.1 is used 269
4.14.2 PowerHA installation and configuration 270
4.14.3 High-level PowerHA configuration steps 271
4.14.4 PowerHA SystemMirror configuration with DB2 273
4.14.5 PowerHA SystemMirror configuration with WebSphere Application Server . . . 274
4.14.6 PowerHA SystemMirror configuration examples 275
4.15 Advanced configuration 284
4.15.1 VIOS partition and system redundancy 284
4.15.2 Virtual storage redundancy 285
4.15.3 Multipathing in Virtual I/O Server 287
4.15.4 Multipathing in client partition 287
4.15.5 LVM mirroring considerations 287
4.15.6 Virtual Ethernet redundancy 288
4.15.7 IBM PowerHA SystemMirror for AIX virtual I/O client 289

Chapter 5. Stretched Cluster Diagnostic and Recovery Guidelines 293
5.1 SAN Volume Controller Stretched Cluster solution testing 294
5.1.1 Test environment 294
5.1.2 Test scenarios and results 300
5.1.3 Test conclusions 302
5.2 Solution recovery planning 305
5.3 SAN Volume Controller recovery planning 305
5.3.1 Planning for SAN Volume Controller recovery first steps 305
5.4 PowerVM recovery planning 312
5.5 PowerHA SystemMirror recovery planning 312
5.6 SAN Volume Controller Diagnosis and recovery guidelines 312
5.6.1 Critical event scenarios and complete domain failure 312
5.6.2 SAN Volume Controller diagnosis guidelines 312
5.6.3 SAN Volume Controller recovery guidelines 328
5.7 PowerVM diagnosis and recovery guidelines 340
5.7.1 PowerVM diagnosis guidelines 340
5.7.2 PowerVM recovery guidelines 340
5.8 PowerHA SystemMirror diagnosis and recovery guidelines 340
5.8.1 PowerHA SystemMirror diagnosis guidelines 340
5.8.2 PowerHA SystemMirror recovery guidelines 340

Related publications 341
IBM Redbooks 341
VMware online resource 342
Other publications 342
Web sites 342
Help from IBM 343

Notices

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

© Copyright IBM Corp. 2013. All rights reserved.

Trademarks

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

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

Active Memory™
AIX 5L™
AIX®
BladeCenter®
DB2 Universal Database™
DB2®
developerWorks®
Domino®
DS4000®
DS6000™
DS8000®
Easy Tier®
Electronic Service Agent™
EnergyScale™
Enterprise Storage Server®
eServer™
FICON®
FileNet®
FlashCopy®
GDPS®
Geographically Dispersed Parallel Sysplex™
Global Technology Services®
HACMP™
IBM Flex System™
IBM Systems Director Active Energy Manager™
IBM®
InfoSphere®
Lotus®
Micro-Partitioning®
Optim™
Parallel Sysplex®
Power Architecture®
POWER Hypervisor™
Power Series®
Power Systems™
Power Systems Software™
POWER6+™
POWER6®
POWER7+™
POWER7®
PowerHA®
PowerVM®
POWER®
pureScale®
pureXML®
RackSwitch™
Rational®
Real-time Compression™
Redbooks®
Redpaper™
Redbooks (logo) ®
RS/6000®
Storwize®
System i5®
System i®
System p5®
System p®
System Storage®
System x®
System z®
SystemMirror®
Systems Director VMControl™
Tivoli®
WebSphere®
XIV®
z/OS®
zSeries®

The following terms are trademarks of other companies:

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

ITIL is a registered trademark, and a registered community trademark of The Minister for the Cabinet Office, and is registered in the U.S. Patent and Trademark Office.

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

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

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

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

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

Preface

This IBM® Redbooks® publication describes the IBM Storage Area Network and IBM SAN Volume Controller Stretched Cluster solution when combined with PowerVM® and PowerHA®. We describe guidelines, settings, and the implementation steps that are necessary to achieve a successful implementation.

This book is for administrators who are familiar with the SAN, IBM SAN Volume Controller, and IBM PowerVM and PowerHA Systems.

Authors

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

Shivendra Ashish is a software developer in PowerHA Development and holds Master's degree in Computer Science and Applications. He also worked across various Power hardware and software, AIX®, PowerVM technologies. He has over five years of experience working on these technologies. He also has extensive experience in configuring and managing IBM SAN Storage subsystems and worked on various customer use cases and environment by using his expertise. He has a developerWorks® article on LPM, developerWorks wikis and blogs on Federated security in PowerHA, a disclosure rated filed on virtualization and he is an IBM certified administrator and technical support specialist in AIX System p® and virtualization.

Kameswara Bhaskarabhatla is an IT Specialist at IBM Global Technology Services®, Delhi, India. He joined IBM Czech Republic in 2007 as a Level 3 IT Specialist. Kamesh provides daily and ongoing support to SAN and Disk systems, and works with SAN designs and solutions for customers. Kamesh also is involved in the development of Storage Automation tools, such as TPCHC and PDT.

© Copyright IBM Corp. 2013. All rights reserved.

Huang Chen is a Storage Solution support specialist at IBM Systems & Technology Group, Client Enablement & Systems Assurance, Research Triangle Park, North Carolina, US. He has worked in the IBM Storage Area Network and Storage Solution support center since 2002. Before he joined Storage Solution support, he worked in the development of zSeries® I/O Channel, embedded OS and Open system firmware and device drivers. Huang has 29 years of experience in I/O development, service, and support. Huang is a Certified Brocade SAN Designer and an IBM Open Systems Storage Solutions Specialist. He also is a member of the Storage Networking Industry Association.

William Genovese is an Engagement Leader, Senior Enterprise IT Architect, and Senior Managing Technology Consultant in the IBM Systems & Technology Group (STG), Lab Services Executive Consulting group - High Availability Center of Competency (HACoC). Bill has been with IBM for nine years (STG, GTS, and GBS), and has over 20 years experience working with Financial Services customers. He is an IBM Certified IT Architect and an IBM Certified Consultant with specializations in Technical Architecture, Systems Engineering and Application Architecture. Bill also is a patent inventor with an IBM team for dynamic generation of user interfaces and automated mapping of input data for service-oriented, architecture-based system management applications.

Jon Tate is a Project Manager for IBM System Storage® SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 27 years of experience in storage software and management, services, and support, and is an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He also is the UK Chairman of the Storage Networking Industry Association.

There are many people that contributed to this book. In particular, we thank the development and PFE teams in IBM Hursley, UK.

We also would like to thank the following people for their contributions:

- Chris Canto
- Dave Carr
- Robin Findlay
- Carlos Fuente
- Geoff Lane
- Andrew Martin
- Paul Merrison
- Bill Scales
- Matt Smith
- Barry Whyte

IBM Hursley, UK

- Thomas Vogel
- Torsten Rothenwaldt

  IBM Germany

- Duane Bolland
- John Lindley
- Mayur Shetty

  IBM US

- Mary Connell
- Chris Saul
- Bill Wiegand

  IBM US

Special thanks to the following Brocade Communications Systems staff in San Jose, California, for their unparalleled support of this residency in terms of equipment and support in many areas:

- Steven Tong
- Silviano Gaona
- Brian Steffler
- Marcus Thordal
- Brian Larsen
- Jim Baldyga

  Brocade Communications Systems

Now you can become a published author, too!

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

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

http://www.ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  http:///www.ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

1

Introduction

Business continuity and continuous application availability are among the top requirements for many organizations. In today's increasing business demands, critical applications must be available all of the time. The system also must be fault tolerant to provide continuous service availability. Disruption in service to business-critical applications can lead to customer dissatisfaction, lost productivity, and lost revenue, which drives businesses to require stringent service level agreements (SLAs).

These fault tolerant systems often come with a heavy cost. Technological advancements in virtualization, storage, and networking made enhanced high availability and business continuity possible for critical systems. Information technology solutions can now be designed to manage planned and unplanned outages and provide other advantages in flexibility and cost efficiencies that are available from virtualization and entry into cloud-computing models. In addition, these cost reduction investments made in virtualization technology and automation provide cloud-computing technology entry points. These entry points, combined with development and maturity of service and systems management processes, can provide higher levels of availability and business continuity for customers.

In today's complex environments, providing continuous service for applications is a key component of a successful IT implementation. Best practices in high availability in system design that supports critical services and application provide for protection through each architectural layer to minimize business impact should there be a single component or a larger systemic failure of multiple components that support the applications and overall service. A high availability solution can ensure that the failure of any component does not cause the application and its data to be unavailable to the user community. This is achieved through the elimination (or masking) of planned and unplanned downtime by eliminating single points of failure.

High availability is one of the components that contribute to providing continuous service for the application clients by masking or eliminating planned and unplanned outages. A high availability solution ensures that the failure of any system component (hardware, software, or system management) does not cause the application and its data to become permanently unavailable to the user.

© Copyright IBM Corp. 2013. All rights reserved.

High availability solutions must eliminate single points of failure through appropriate design, planning, selection of hardware, configuration of software, control of applications, a carefully controlled environment, and service management discipline.

IBM worked to realize an integrated solution that offers significant functionality for maintaining high availability and business continuity in an IBM Power Series® (pSeries - AIX- UNIX Operating System) Virtual Machine (PowerVM) environment including Power High Availability (PowerHA) and Power LPM. This technology, which is deployed across data centers, relies on the industry-proven PowerVM Live Partition Mobility (LPM), storage area network (SAN) Volume Controller Stretched Cluster solution, IBM Storwize® V7000 disk system, and a Brocade Layer 2 IP Network and storage networking infrastructure, with high performance traffic management and DC interconnect. IBM PowerHA SystemMirror® for AIX and SAN Volume Controller can be combined with the IBM Storwize V7000 disk system to address high availability and fault tolerance requirements for today's business critical applications. The IBM Storwize V7000 disk system targets medium-sized business to large enterprise users who seek enterprise-class storage efficiency and ease of management, with support for moving volumes between storage pools without causing disruption to running applications.

This book describes the technology solution, its architecture, and the solution context and provides the implementation details.

This chapter includes the following topics:

- High Availability concepts, definitions, and building blocks
- Storwize V7000
- SAN Volume Controller
- SAN Volume Controller Stretched Cluster
- Fibre Channel SAN
- Power 750 Express and Hardware Management Console
- Architecture and solution context
- PowerVM technology
- PowerVM
- Live Partition Mobility
- N_Port ID Virtualization
- PowerHA
- Architecture and solution context
- SAN Volume Controller, Layer-2 IP Network, Storage Networking infrastructure and LPM integration
- Architecture and solution context
- Open data center Interoperable Network

1.1 High Availability concepts, definitions, and building blocks

Availability is a concept that is becoming more established in today's IT environments. As more customer-facing applications are developed, the demand for high availability is increasing. As more revenue producing applications are developed, the cost of downtime continues to rise. Tremendous growth in system capabilities, with business demands for around the clock operations, has put increased demands on efforts to provide the highest levels of system availability. Companies are beginning to measure service availability from the user's perspective, and there is a growing trend to measuring availability on a 24 x 7 basis, such that planned and unplanned outages are included.

Many customers in multiple industries struggle with a specific definition of high availability for their systems and services as the overall trend in customer requirements is moving from two or three nines of availability, to four and even five nines (99.999%) of availability with near zero Recovery Time Objectives (RTO) and Recovery Point Objectives (RPO). With these requirements, some companies are moving towards a three-site data center solution that blurs the line between HA and Disaster Recovery (DR).

IBM Power Systems™ servers are especially designed to help achieve high availability; however, the need for planned downtime is required for periodic maintenance (for hardware and software) and cannot be completely eliminated. Although both types of outages affect the overall availability of a server, we must understand the following distinctions between planned and unplanned downtime in today's business environments:

- Planned downtime is scheduled and often is a result of a maintenance action to the hardware, operating system, or an application. Scheduled downtime is used to ensure that the server can operate optimally and reliably in the future. Because this type of event can be planned for in advance, it can be scheduled at a time that least affects system or application availability.
- Unplanned downtime is downtime that occurs as a result of a type of physical event or failure, along with human error, and cannot be planned in advance.

"Despite the potential consequences of unplanned downtime, less than 10% of all downtime can be attributed to unplanned events, and only a fraction of that is due to a site disaster. The other 90+%—the kind that companies face on a regular basis—are those caused by system maintenance tasks."[1]

The following typical system maintenance tasks are included in planned downtime:

- Data backups (nightly, weekly, and monthly)
- Reorganization of files to reclaim disk space and improve performance
- Vendor software upgrades and data conversions
- IBM software release upgrades and patches, such as program temporary fixes (PTFs)
- New application software installations
- Hardware upgrades
- System migrations

The number of unplanned outages continues to shrink quickly as hardware and software technology becomes more resilient. Each generation of IBM hardware and storage technology products provides enhancements and improvements in terms of Reliability, Availability, and Serviceability (RAS). Even as reliability continues to improve, human exception handling skills are decreasing in terms of overall service availability for critical systems in many customer environments.

[1] Vision Solutions, Inc. white paper, *An introduction to System i High Availability 2010*

Although unplanned outages must still be eliminated, in the desire to achieve 24 x 365 availability, planned downtime has now become a primary focus, as enabled through system automation and less human intervention. The technology must be managed, and many companies are moving towards an Information Technology Infrastructure Library (ITIL)[2] based model. However, as many companies move towards ITIL, they are taking a silo-based approach to improve their IT processes, which have long been built on manual interventions at the component levels. It also is driven by planned outages and not the end-to-end service or system level. Higher levels of availability require more automation, and as such, there are tremendous opportunities to improve IT management systems and solutions to provide this automation and better service (including high availability) at lower costs. Through virtualization technology at the Power Systems and SAN controller levels, customers can move further up the high availability continuum while reducing costs as provided through virtualization and entry level automation.

1.1.1 Starting points and key definitions: Reliability, Availability, Serviceability

Servers must be designed to help avoid every possible outage, focusing on applications availability. For almost two decades now, the IBM Power Systems design and development teams have worked to integrate industry-leading IBM System z® mainframe reliability features and capabilities into the IBM Power Systems servers line. These RAS capabilities together with the IBM Power Systems virtualization features help implement fully virtualized and highly available environments.

RAS is defined the following ways[3],[4]:

- Reliability

  Reliability refers to how infrequently a defect or fault is seen in a server. Freedom from operational failure. The probability that an item performs the intended function for a specified interval under stated conditions. A measure of how long an item or IT Service can perform its agreed upon function without interruption.

- Availability

  Availability refers to how infrequently the functionality of a system or application is affected by a fault or defect. The ability of a component or service to perform its required function at a stated instant or over a stated period of time. It is usually expressed as the availability ratio; that is, the proportion of time that the service is actually available for use by the customers within the agreed service hours ((Agreed-Unavailable/Agreed)*100%. The availability of the infrastructure is one part of overall IT availability, and IT availability is, in turn, a subset of the overall organization or business availability. IBM uses the term *resilience* to address these broader capabilities. Resilience is then subdivided into business resilience and IT resilience, which are defined in the next section.

---

[2] The Information Technology Infrastructure Library (ITIL) is a set of practices for IT Service Management (ITSM) that focuses on aligning IT services with the needs of business. In its current form (known as ITIL 2011 edition), ITIL is published in a series of five core publications, each of which covers an ITSM lifecycle stage. ITIL underpins ISO/IEC 20000 (previously BS15000), the International Service Management Standard for IT Service Management, although differences between the two frameworks do exist. ITIL describes processes, procedures, tasks and checklists that are not organization-specific and are used by an organization for establishing integration with the organization's strategy, deliver value and maintain a minimum level of competency. It allows the organization to establish a baseline from which it can plan, implement, and measure. It is used to demonstrate compliance and to measure improvement.

[3] D. Henderson, J.Mitchell, G. Ahrens. “POWER7® System RAS Key Aspects of Power Systems Reliability, Availability, and Serviceability” POW03056.doc, November 2010

[4] IBM Systems Journal, Vo. 47 No. 4, October 2008 Morrill, Beard, and Clitherow, “Achieving Continuous Availability of IBM Systems Infrastructures”, page 495.

- Serviceability

  How well faults and their impacts are communicated to users and services, and how efficiently and non-disruptively they are repaired.

Reliability in hardware is all about how often a hardware fault requires a system to be serviced (the less frequent the failures, the greater the reliability). Availability is how infrequently such a failure affects the operation of the system or application. For high levels of availability, correct system operation must not be adversely affected by hardware faults. A highly available system design ensures that most hardware failures do not result in application outages. Serviceability relates to identifying what fails and ensuring an efficient repair of that component, firmware, or software)

IBM POWER7 is designed for RAS by including technologies to detect and isolate component faults the first time that they appear without the need to re-create the situation or perform further tests. This technology helps to minimize the risk of repeating the same error and causing similar or even larger problems.

Downtime can start with a small problem and end up costing an organization millions of dollars a day in lost revenue. While many companies spend large amounts of IT budgets to ensure a resilient data center, it only takes one event to throw a thriving business into chaos. The latest white paper from IDC, *IBM Power Systems and Their Support for Business Resiliency in Challenging IT Environments* outlines the best practices for companies in their search for bullet-proof business continuity. The white paper is available for download at this website:

http://www-03.ibm.com/systems/power/advantages/assets/resiliency.html?lnk=wf&cmp=101AR&ct=101AR12W&cr=sme_tipsheet&csr=power&cm=h&ccy=us&cpb=stg&cd=20130122&S_TACT=101AR12W&sf8864043=1

It is important that the starting point for the solution in this book provides some further explanation and understanding of IBM's RAS strategy principles in terms of how reliable servers are built. These RAS strategies are part of the DNA of the core design principles for Power Servers, along with equally as important business drivers such as efficiency and performance gains through each generation of IBM servers. However, this is not the end of the journey to high availability in terms of the end-to-end service from the user's perspective. A highly dependable, highly reliable single server does not provide or enable continuous operations.

In the next section, we describe the various levels of service availability, the key industry definitions for these, and example solution building blocks for service availability on the journey or continuum to continuous availability and operations. We also describe the positioning of the solution context in this book against the service availability building blocks as to where it fits to enable high availability for customers.

1.1.2 Key industry definitions

Terminology surrounding availability first came into the lexicon of IT in the 1980s, from the mainframe user community and from the joint committee on fundamental concepts and terminology of the IEEE Computer Society and the International Federation for Information Processing (IFIP)[5].

[5] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic Concepts and Taxonomy of Dependable and Secure Computing," IEEE Transactions on Dependable and Secure Computing 1, No. 1, 11–33 (2004).

Today, there is a common body of terms that pervades the online and other literature on IT availability. However, contradictions and discrepancies in terminology can still occur in the discussion of availability. Most often when there is a debate, the question is one of scope. For example, when the claim is made that a system infrastructure provides continuous availability, it is necessary to determine whether the availability capability describes the hardware alone or also extends to the operating system and data.

In recent years, the claim of continuous availability for some platforms described hardware capabilities. For System z, it described the hardware, operating system, and selected middleware configured in a parallel sysplex, but did not pertain to the data storage subsystem. Thus, readers must be cautious concerning terminology. You must understand what is meant by the terms that are used and what is the scope of each statement.

ITIL has a comprehensive, vendor-neutral glossary that includes many terms related to availability[6] and is one source for the definitions that follow. Terms are also found in various IBM technical publications where IT availability is discussed.[7]

The availability of the infrastructure is one part of overall IT availability, and IT availability is, in turn, a subset of the overall organization or business availability.

Business resilience is the ability of the business to rapidly adapt and respond to opportunities, regulations, and risks, to maintain secure and continuous business operations, be a more trusted partner, and enable growth. Business resilience spans business strategy, organizational structure, business and IT processes, IT infrastructure, applications and data, and facilities. The term *dependability* is used in a similar vein as a global concept that subsumes the attributes of reliability, availability, safety, integrity, maintainability, and so on.

IT resilience refers to the ability for IT to rapidly adapt and respond to opportunities, demands, disruptions, or threats, and enable continued business operations without significant impact. Achieving IT resilience encompasses a broad range of disciplines, including infrastructure, applications, and processes.

IT infrastructure availability is a necessary starting point, but not a sufficient component of IT resilience. It can be described and measured as shown in the next section.

1.1.3 Infrastructure availability continuum

The trend and end goal for many customers is to achieve continuous availability. The definitions that are presented in this section are a subset of terms in the availability discussion that most closely relate to the infrastructure, and are important in the discussion of the roadmap to continuous availability for customers. If we consider reliability as the basic level of infrastructure availability, there is a continuum of increasingly stringent capabilities starting with reliability and culminating in continuous availability and disaster recovery. It is important to note that the infrastructure availability continuum is not unique to a specific platform. This can be applied to any platform, including z/OS®, UNIX, Intel/x86, or Linux. In addition, it can be applied to network and SAN infrastructure.

[6] OGC Common Glossary, ITIL Version 3 (May 2007), http://www.best-management-practice.com/officialsite.asp?FO¼1230366&action¼confirmation&tdi¼575004.

[7] F. Kyne, C. Matthys, U. Bengtsson, A. Clifton, S. Cox, G. Hines, G. McGeoch, D. Lawson, G. Nicholls, and D. Raften, Achieving the Highest levels of Parallel Sysplex® Availability, IBM Corporation Redbooks (December 2004), http://www.redbooks.ibm.com/redbooks/pdfs/ sg246061.pdf.

There are various configurations that match the levels of availability on the continuum from reliability to CA and DR, as show in Figure 1-1.

Figure 1-1 Infrastructure availability continuum

Reliable resources are the starting point for IT infrastructure availability. As a starting point in the continuum, IBM systems currently have inherent reliability that is far higher than what was possible 20 - 30 years ago. For many systems and uses of IT, this level of availability is sufficient to meet business requirements. However, the key concern becomes not whether systems are reliable enough, but whether or customers are fully using the capabilities of the current systems. IBM provides a systems assurance process to guide the installation of various systems and assist in verifying whether an installation is properly configured and deployed to optimize availability.

A second aspect of reliability is the extent to which installations unduly depend on the reliability of an IBM system. Regardless of how reliable it might be, a single system is still a single point of failure (SPOF). Cases like that of the insurance company that stores its mission-critical data on a single disk subsystem because it has not seen a disk subsystem outage in three years are all too common. The solution in this book uses recommended multisystem and multisite configurations in its end-to-end architecture, beyond just based on a single Power Systems server, single storage, or single SAN Volume Controller installation.

High Availability

As a next step in the continuum, *High Availability* (HA) is the ability of a system to provide service during defined periods at acceptable or agreed-upon levels, and to mask unplanned outages from users. HA is achieved by the use of fault tolerance, automated failure detection, recovery, bypass reconfiguration, testing, and problem and change management. HA begins with the identification of SPOFs and using redundancy. System design must also encompass the ability to failover to and transparently recover from running on the alternative resources without affecting service. This requires monitoring, data replication, and workload balancing.

Many installations consider an availability of 99.7 percent as an accepted minimum level for HA systems. The fact that the service is available only during scheduled periods implies that there is time available to apply changes and maintenance outside these periods.

Given the reliability of hardware, failover capability is important for software. It is a recognized phenomenon that software errors are responsible for a significant share of system unplanned outages. A large number of these errors occur because of an unrepeatable sequence of events, which is known as a *Heisenbug*. Failover to a node where work can run again can resolve these transient errors.[8]

It has been debated whether adding components increases availability or actually introduces more points of failure and diminishes the availability of the overall system. In serial configurations, all components fail if any single component fails; in parallel configurations, the system fails only if all of its components fail.[9]

Continuous operations

The term *continuous operations* (CO) indicates the ability of a system to continuously operate and mask planned outages from users. This is achieved by the use of nondisruptive hardware and software change technology by using nondisruptive configuration upgrades and software that is designed so that new levels coexist with current levels, which allows changes to be rolled through the suite of software, one component at a time. That a system provides continuous operation does not imply that it is highly available. A system might have the potential to run 24 x 7, but might be available only 95% of the time because of intermittent system failures.

COs center on whether hardware, software, and microcode upgrades can be made without affecting service. It requires communication and automation among redundant systems and the capability for different levels of technology to coexist in the system infrastructure. Design issues in this area include how tightly coupled the systems in the cluster need to be for transparency and compatibility. For more information, see *'Clustering Solutions Overview*,' which outlines various clustering alternatives and their key attributes.[10]

Continuous availability

Continuous availability (CA) is the ability of a system to deliver non-disrupted service to the user 24 x 7, with no planned or unplanned outages)

Consideration of CA extends the discussion to multiple sites and revisits the questions about HA and continuous operations for sites that are geographically distant. Questions about location, number of sites, extent of automation, bandwidth, and latency all come into play. CA includes the capability to operate during, and recover from, disasters. CA typically requires more system capacity in the infrastructure to accommodate the loss of any one component. This allows the workload requiring CA to be run in parallel on two or more systems, or to be switchable (in a nondisruptive manner) from a primary to an alternative system.

Disaster recovery or disaster tolerance

The term d*isaster recovery* or *disaster tolerance* indicates recovery after a disaster, such as a fire, that destroys or otherwise disables a system. DR techniques typically involve restoring data to a second (recovery) system, and then using the recovery system in place of the destroyed or disabled application system.

DR might not require or provide CA. DR encompasses the ability of the system to transfer data and workload offsite and to restart the workload at a new site. The transfer often involves a system and service outage, but can be made completely seamless to important users.

---

[8] G. F. Pfister, In Search of Clusters: The Ongoing Battle in Lowly Parallel Computing, Prentice Hall PTR, Upper Saddle River, NJ (1995).

[9] K. S. Trivedi, Probability and Statistics with Reliability, Queuing and Computer Science Applications, John Wiley & Sons, Hoboken, NJ (2002).

[10] K. S. Trivedi, Probability and Statistics with Reliability, Queuing and Computer Science Applications, John Wiley & Sons, Hoboken, NJ (2002).

There are organizations today for which one multisite architecture provides HA, CA, and DR. *Combining high availability and disaster recovery solutions for highly critical IT environments*, by David Clitherow, et. al.[11], describes various system configuration and data replication techniques for HA and DR solutions. It offers examples of combining HA and DR environments to deliver the highest levels of availability through planned and unplanned outages, including events where DR must be started.

The following measurements are unique to DR:

- Business recovery objective (RO): The wanted time within which business processes should be recovered, along with the minimum staff, assets, and services that are required to perform this recovery.
- Network recovery objective (NRO): The maximum acceptable time that is required to transfer network connectivity from the failed site to the backup site.
- Recovery time objective (RTO): How long the customer can afford to function without access to its applications. This determines how much time is available following a disaster to restore all applications to functionality.
- Recovery point objective (RPO): The amount of data that the customer can afford to re-create or lose as a result of a disaster. If the RPO is set at zero data loss, a synchronous remote copy solution is generally the only option. If the RPO is set at a few minutes and there is a significant distance between sites, asynchronous remote copy might be an appropriate solution.

It is important to note and describe the starting points of reliability, availability, and serviceability in context to the SAN Volume Controller Stretched Cluster, PowerHA and LPM solution in this book. Customers are starting in a solid position with IBM servers and storage in terms of reliable infrastructure. The next steps in the infrastructure availability continuum include the deployment of redundant virtualized resources.

In this book, we describe PowerVM virtualization technology at a high level as a necessary prerequisite to the next step of Clustered Resources in the Infrastructure Availability Continuum. Because PowerVM is not the primary focus of this book, applicable links to other PowerVM books and other reference information is provided. The solution in this book fits between Clustered Resources and Cross-site Clusters in the Infrastructure Availability Continuum. The SAN Volume Controller Stretched Cluster, PowerHA, and LPM technology solution is ideally positioned as an active-active cross site clustered solution for data access protection, cluster management, and a solution to reduce planned outage windows for IBM Power Systems server maintenance for critical workloads.

In the next section, we describe the infrastructure building blocks and continuum as applied to IBM Power Systems to achieve high availability, continuous operations, continuous availability and mature, comprehensive disaster recovery.

[11] For more information, see this website:
http://ieeexplore.ieee.org/xpl/login.jsp?reload=true&tp=&arnumber=5386509&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5386509

1.1.4 High Availability building blocks: IBM Power Systems best practices

In the previous section, the Infrastructure Availability Continuum was described. Remember that this continuum in terms of the building blocks can be applied against many of the infrastructure and hosted system solutions, including Storage, SAN, Network, Servers, Middleware, Databases, and Information Management. Supporting these building blocks, it is important to note the following best practices framework for IBM Power Systems that aligns to the high availability building blocks and should be considered as part of the journey on the continuum to continuous availability:

- Quantify the cost benefits of availability.

  Remember that technology alone is not enough. There is not one solution that meets all requirements for HA. HA costs time, money, and people. Fully identify the set of availability problems you are attempting to address and their costs for each partition or server, operating system, and application.

- Reduce planned outage windows by using IBM strategic tools for migration, build, and maintenance.

  Usually, operations staff must shut down all application tiers for hardware, firmware, and software maintenance. For planned maintenance, Live Partion Mobility (LPM) can be used to relocate workload. Deploy Workload Partitions (WPAR) for mobility for static workloads. In addition, consider Live Application Mobility (LAM) as an important component of workload partitioning to provide for increased availability. LAM provides automatic, policy-based relocation of workload between systems that use WPARs.

  Implement local provisioning automation for Power infrastructure builds. Products such as the IBM STG Power Systems Provisioning Tool Kit might be introduced safely and at lower cost to further reduce lower deployment time for server builds.

- Use Power Systems availability features, which are provided through power Virtual Machine (VM).

  Deploy dual Virtual I/O Servers with Logical Volume Manager (LVM) for Power Systems. Implementing a dual virtual I/O configuration provides redundancy and allows for protection for both planned and unplanned outages. LVM provides protection against data loss at the Power Systems host level.

  Implement redundant network infrastructure components. Implementing a dual integrated virtual Ethernet configuration provides a redundant logical network infrastructure for logical partitions (LPARs) and VIO Servers.

  Use improved options for SAN integration, improved load balancing, and increased resiliency. N_Port ID Virtualization (NPIV) is a Fibre Channel industry standard method for virtualizing a physical Fibre Channel port. Use NPIV for critical system environment by using enterprise storage to ease storage administration, zoning provisioning, and to improve load balancing for clustered solutions.

  Manage capacity proactively for HA to prevent outages. Understand and manage the capacity requirements of the target systems. In doing so, clearly define the application priorities (tiers). Use PowerVM features such as shared processor pools that dynamically allocate the processor resources when needed. Use dynamic LPARs to reconfigure memory, processor, and adapters on the target LPAR; further automate through PowerHA.

  Eliminate the requirement for an immediate fallback. Lack of capacity on the backup server requires a quick fallback to the primary server, which is a second outage.

  Consider the following power-on demand offerings:

  - Capacity Upgrade on Demand: Permanent Upgrade
  - On/Off Capacity on Demand: HACMP™ Integration
  - Utility on Demand: Automated through Shared Processor Pools

Use Active Memory™ Sharing for maximizing LPAR usage. This ensures the maximum efficiency of LPAR usage and helps to eliminate outages because of potential spikes in memory usage by moving memory from one partition to another.

Consider Active Memory Expansion for specific applications. This provides memory expansion beyond the physical limits of the LPAR when needed by giving more memory capacity to the partition by using compression and decompression of the contents in true memory.

- Use clustering technologies and replicated data for critical workloads and applications.

  Deploy PowerHA for AIX (formerly HACMP). PowerHA software automates recovery actions upon detection of failure. It provides seamless integration with AIX and offers numerous new features for easier management of clusters; for example, web graphical user interface (GUI), nondisruptive cluster maintenance, and cluster verification tools. Configure middleware integration for added control and protection via SmartAssists for DB2®, Oracle, and WebSphere®. Use PowerHA as a tool during planned maintenance actions to minimize disruption to users and enable rolling upgrades.

  The goal is to reduce failover time, but the duration is often dependent on application characteristics. Enhanced concurrent VGs, fast disk takeover, and fast failure detection help speed failover. HA clustering is an important part of an overall availability strategy, but it cannot provide continuous availability on its own.

  For more information about clustering best practices for Power Systems, see the following resources:

  – *Virtualization and Clustering Best Practices using IBM System p Servers*, SG24-7349:

    http://www.redbooks.ibm.com/abstracts/sg247349.html

  – Power Systems HA and DR Best Practices - IBM High Availability Center of Competency (Page 56-59):

    https://w3-connections.ibm.com/wikis/form/anonymous/api/wiki/b7b38f1c-91b5-43ac-88d8-18f55aeb8abc/page/9b72d7a9-a33f-4607-b554-e2fac21b5cc3/attachment/5d46d62d-af79-4709-9021-a2cec7e72df0/media/HACoC%20POWER_HA%20and%20DR%20Best%20Practices.pdf

- Consider cross platform disaster recovery.

  An active-active architecture where all applications are clustered can provide continuous availability, scalability, and disaster recovery

- Adopt an availability driven culture.

  High levels of availability cannot be achieved without continual focus. Develop a culture that is focused on the following availability practices and standards:

  – Identify acceptable levels for the frequency, duration, and scope of outages.

  – Track, report, and enforce Service Level Agreements (SLA); monitor and automate responses.

  – Provide incentives that reward improvements.

  – Assign an Availability Manager; team with Executive Support.

- Proactively work to prevent incidents and reduce problems through preventive maintenance through the adoption of IBM standard methodologies and tooling.

  For more information, see these resources:

  – http://www-304.ibm.com/support/customercare/sas/f/best/home.htmlIBMPowerSystemsFirmware_Best_Practices_v6.pdf

  – http://www-03.ibm.com/systems/power/software/aix/whitepapers/

  - Chapter 2, “Hardware and Software Description” on page 115.

- Control changes and test each one in advance.

  Changes increase causes outage risk; testing mitigates that risk. Do not allow testing in production environments (except regularly testing PowerHA Cluster failover and failback). No changes should be allowed in the Production environment unless approved by the Change Control Board and the changes must be verified through testing first.

  The test environment must closely mimic production. However, keep the test and production environments separate. Testing for Integrated Infrastructure Power Systems environments should always included storage microcode testing and SAN Volume Controller software.

- Maintain an energy efficient and well-managed data center, as shown in the following examples:

  - Enhanced cooling (water) through rear door heat exchangers.
  - Separate systems through fireproof and floodproof walls.
  - Control and track access.
  - Implement dual power cords.
  - Implement dual UPS (Uninterruptible Power Supplies).
  - Implement dual power grids.
  - Implement POWER7 processor-based server Thermal Power Management.

- Centralize and simplify configuration management.

  “Sick but not dead” symptoms are often caused by unrecognized interdependencies. Understand cross infrastructure resource dependencies with an enterprise configuration management database (CMDB). De-complex the environment by monitoring components end-to-end and build alerts to show when redundant components failed.

It is vitally important that these best practices are followed on the journey to continuous service availability through the IBM Power Systems High Availability building blocks. Next, we describe these building blocks with the representative technology building blocks for each step towards continuous service availability.

Figure 1-2 on page 13 shows the IBM Power Systems HA building blocks. This can be compared to the Infrastructure Availability Continuum that is described in 1.1.3, “Infrastructure availability continuum” on page 6. It also can be seen where the IBM Power Systems Technology aligns to the Infrastructure Availability Continuum, specifically with the solution components in this book: PowerVM, LPM, NPIV, PowerHA, and SAN Volume Controller.

Figure 1-2 IBM Power Systems High Availability Building Blocks

It is important to note that, as described in section 1.1.3, “Infrastructure availability continuum” on page 6, there are levels associated with the building blocks from HA through Service Availability, as shown in Figure 1-2. Next, we describe the specific building block categories and the IBM Power Systems technologies that are fit into each of these categories.

Reliable systems by design

The starting point HA building block for IBM Power Systems begins with highly resilient servers that contain robust capabilities, such as hot swappable component and concurrent upgrades. Each generation of IBM Power Systems improves in terms of RAS capabilities as described in 1.1.1, “Starting points and key definitions: Reliability, Availability, Serviceability” on page 4.

So, what does it take to build a reliable server by design? We provide the solid starting point by instilling the following the strategy, design and build lifecycle:

1. Build to a RAS Architecture:
   a. Reliability follows an ordered, engineered strategy.
   b. Continuous improvements are clear, concise, and well-structured.
   c. RAS is a key design aim as is performance and cost.
2. Ensure reliability by minimizing outage opportunities. Minimize total components and use higher-grade components.

3. Instill autonomic computing principles to keep running despite problems (which result in higher availability):
   a. Engineer with self-diagnostics capabilities; work around faulty components, or better yet, “self-heal”, which automatically repair with internal redundancy.
   b. Reduce frequency and duration of outages.
4. Make it easy to service; minimize outages and downtime (which result in improved serviceability):
   a. Avoid repeat failures with accurate diagnostics.
   b. Use concurrent repair on higher failure rate items.
   c. Provide self-healing where possible; guided repair otherwise.

If we look closer at our IBM strategy for detecting failures in IBM Power Systems, it is important to note that runtime monitoring is a critical component of First Failure Data Capture (FFDC). The objective of FFDC is to capture as much information about the current state of the system when an error occurs. We design and engineer extensive “checkers” throughout the hardware that provides the following tasks:

- Monitor error conditions:
  - Might occur during processing
  - For power supply and fan status
  - For environmental conditions
- Report non-critical errors out of band (without system performance impact)

We also include an analytical service processor that correlates and processes error information that uses thresholding and other techniques to determine when action must be taken. The service processor offloads detection and analysis from system operation, which allows the server to continue performing and processing workload, as shown in the following examples:

- Requesting a part to be replaced
- Dynamically starting built in redundancy to repair the failing component
- Dynamically deallocating failing components so that system availability is maintained

In IBM Power Systems, we provide extensive error detection and self-healing, which is available during run time, at Initial Program Load time, and after a critical fault (for example, checkstop).

Figure 1-3 on page 15 shows the POWER7 System Architecture for detecting failures. As a point of reference, a Power 795 server has approximately 598,000 error checkers captured in 96,000 Fault Isolation Register (FIR) bits.

Figure 1-3 Power System Architecture for Error Detection and Analysis

In summary, IBM Power Systems provide failure data analysis before a service call and reliably identify potential failing components, thereby reducing costly downtime.

For more information, see the white paper, *POWER7 System RAS: Key Aspects of Power Systems Reliability, Availability, and Serviceability*, POW03056-USEN, which is available at this website:

http://www-07.ibm.com/tw/imc/seminar/download/2010/POWER7_RAS_Whitepaper.pdf

Redundant and virtualized systems

The next step for the Power Systems building blocks for HA is redundant and virtualized systems. Here we are starting to build out capabilities for HA by introducing more than one server to protect critical workloads. IBM Power Virtualization Technology (PowerVM) PowerVM is a combination of hardware and software that enable the virtualization platform for AIX, Linux, and IBM i environments for IBM Power Systems. By implementing PowerVM, you can perform the following functions to increase high availability:

- Easily and quickly deploy new partitions.
- Run isolated workloads for production, development, and test systems.
- Reduce downtime for relocation workloads for hosted applications in LPARs or WPARs for planned maintenance without interruption to running applications through LPM and LAM. Increase your overall availability by making workloads independent of the physical hardware and by adding the capabilities to move those workloads to another server without disruption, thus eliminating planned downtime.

- Optimize the usage of I/O adapters and provide more redundancy through dual Virtual I/O Servers.
- Optimize the usage of network adapters, increase redundancy, and simplify network connections through VLAN trunking through Shared Ethernet Adapters (SEA) and Integrated Virtual Ethernet (IVE).
- Reduce storage administration, improve load balancing, and simplify HBA Card connectivity through NPIV.
- Reduce the complexity and management of the entire environment.
- Optimize resource usage and enable HA by effectively allocating resources to those workloads that need them and when workloads are relocated through Dynamic LPARs and shared resource pool functionality.

Key benefit of PowerVM: You also can reduce costs by consolidating AIX, IBM i, and Linux workloads into high-end IBM Power Systems through the use of PowerVM. Because enterprise-wide infrastructure cost reduction and workload and server consolidation initiatives are a high priority, customers can reduce costs through PowerVM while increasing HA. Thus, this is an example of making HA pay for itself, where cost reduction through implementing PowerVM is also helping to pay for increased HA.

- Add extra resiliency through LVM for the IBM Power environments and critical systems to reduce the effect of data loss at the disk subsystem level. LVM provides for synchronous replication of data between primary and failover volumes and requires no storage administrator intervention in the case of a failover. Remember that the end goal is to sustain failure of any single infrastructure component in the Power Systems integrated environment. For more information about LVM, see the following IBM Redbooks:
  - *AIX Logical Volume Manager from A to Z: Introduction and Concepts*, SG24-5432:

    http://www.redbooks.ibm.com/abstracts/sg245432.html?Open
  - *Quick Reference: AIX Logical Volume Manager and Veritas Volume Manager*, REDP-0107:

    http://www.redbooks.ibm.com/abstracts/redp0107.html?Open

In this chapter, we describe each of the key concepts, architecture, and the solution context for the solution in this book for each component area. We describe the solution hardware and software in Chapter 2, “Hardware and Software Description” on page 115, and the implementation of these technologies in Chapter 3, “SAN Volume Controller Implementation” on page 149, and Chapter 4, “PowerVM and PowerHA Implementation” on page 213, with testing results in Chapter 5, “Stretched Cluster Diagnostic and Recovery Guidelines” on page 293.

Figure 1-4 on page 17 shows a PowerVM enabled environment, with redundancy and virtualization that was designed and implemented throughout the solution architecture from the dual VIOS Servers, VIOS Clients, NPIV, dual SAN directors, and dual storage frames-IBM DS8K.

Figure 1-4 POWER7 Reference Architecture: End-to-end resiliency

Figure 1-4 includes the following components:

- Four HBA ports (two per VIOS) for tape traffic to be allocated (one port per VIOS). Each client partition receives two virtual feeds for tape traffic (one from each VIOS).
- For data traffic, eight ports (four per VIOS) are allocated. The partition ID (Odd/Even) is used as criteria for distributing data traffic between the eight ports.
- Zone LUNs, tape drives for Virtual Fibre Channel (VFC) Client adapter WWPN pairs.
- Out example uses the practice of zoning two storage ports per WWPN. Consider zoning both the WWPN pairs of Virtual Fibre Channel (VFC) Client adapters for two storage ports because each VFC adapter has two WWPN numbers.
- Standardize RootVG size while SAN storage is used. Review application requirements and standardize. (For example, 70 GB for non-DB servers, 100 GB for DB servers.)
- Recommend the use of SAN Boot for all VIO Clients.
- Data center environment at the end state continues to have two SAN switches or directors to support full and balanced workload.

Clustered systems and replicated data

As requirements for continuous operations and continuous availability further develop (for example, the amount of uptime that is required increases and RTO and RPO requirements numbers decrease closer to 0), it is necessary to invest in clustered systems (which might encompass hardware, middleware, and databases, depending on the requirements) with replicated data at the database or storage levels to provide an active-active architecture. There are many types of clusters, such as load balancing and hardware failover (which is provided through PowerHA) application clusters (WebSphere/Middleware) that can improve business service availability and support an active-active architecture. Also called *dual active*, active-active is a phrase that is used to describe a network of independent processing nodes where each node has access to a replicated database, which gives each node access and usage of a single application.

In an active-active system, all requests are load-balanced across all available processing capacity. Where a failure occurs on a node, another node in the network takes its place. An active-active architecture can provide resiliency in a short-distance, campus-like two data center design, or involve Local HA protection within a single data center on different floors or across data center room partitions. Data is replicated synchronously across the single data center or across the campus data centers.

PowerHA SystemMirror is one of the building blocks for IBM Power Systems for HA. It can be used for Local HA in one data center, across two campus-type data centers that are relatively close together (under 5 KM), or even at extended distances. For the solution in this book, we used PowerHA SystemMirror Standard Edition version 7.1.1.2 + FixPack APAR:IV25311 to ensure interoperability and compatibility with the software stack and hardware. It also provides a real-world environment for testing distances from 0 KM (Local HA), 5 KM (Campus), and various metro-mirror like distances from 15 KM, 75 KM, 150 KM, and 300 KM. For more information about testing, results, conclusions, see Chapter 5, “Stretched Cluster Diagnostic and Recovery Guidelines” on page 293.

Figure 1-5 on page 18 shows a high-level PowerHA design that is used in a single data center with LVM Mirroring and synchronous data replication for Local HA, which targets meeting RTO and RPO requirements of 0.

Figure 1-5 Single Datacenter: PowerHA with LVM Mirroring

Based on HA requirements and existing data center locations, a Metro Mirror-type solution with synchronous data replication can be designed and implemented across a metropolitan city to provide HA for applications that are hosted on Power Systems. Peer-to-Peer Remote Copy (PPRC) (which is also referred to as Metro Mirror when compared to Global Mirror) is the IBM named protocol to replicate a storage volume to another control unit in a remote site. Synchronous PPRC causes each write to the primary volume to be performed to the secondary as well, and the I/O is only considered complete when the update to the primary and secondary volume is complete. Asynchronous PPRC flags tracks on the primary volume to be duplicated to the secondary when time permits. PPRC can be used to provide fast data recovery because of failure of the primary site. In the SAN Volume Controller, PPRC is used to mirror a virtualized storage volume to remote (or the same) cluster. PPRC is designed to support replication at “Metropolitan” distances of (normally) less than 300 km.

For more information, see this website:

http://publib.boulder.ibm.com/infocenter/dsichelp/ds8000ic/index.jsp?topic=/com.ibm.storage.ssic.help.doc/f2c_pprc_1qbw8w.html

Figure 1-6 on page 19 shows a high-level PowerHA active-active design that is used across two data centers with PPRC synchronous data replication, which targets at meeting RPO requirements of 0 and RTO <1 HR.

Figure 1-6 PowerHA with PPRC, SAN Volume Controller: Cross data centers

More information about PowerHA is presented in the key concepts, architecture, and solution context for the solution in this chapter. We cover the solution hardware and software description in later chapters of this book.

The SAN Volume Controller is another key that supports the HA building block component for Power Systems that are hosting clustered systems with replicated data. In computer data storage, the SAN Volume Controller is a block storage virtualization appliance that belongs to the IBM System Storage product family. SAN Volume Controller implements an indirection, or virtualization layer in a Fibre Channel storage area network (FC SAN).

The SAN Volume Controller is an inline virtualization (or gateway) device. It logically sits between hosts and storage arrays, which presenting itself to hosts as the storage provider (target) and storage arrays as a single, large host (initiator). SAN Volume Controller is physically attached to any available port in one or several SAN fabrics.

The virtualization approach allows for nondisruptive replacements of any part in the storage infrastructure, including the SAN Volume Controller devices. It also aims at simplifying compatibility requirements in strongly heterogeneous server and storage landscapes. Therefore, all advanced functions are implemented in the virtualization layer, which allows switching storage array vendors without much impact to the environment. Finally, spreading an SAN Volume Controller installation across two or more sites (stretched clustering) enables basic disaster protection paired with continuous availability.

Because it hides the physical characteristics of storage from host systems, SAN Volume Controller helps applications to continue to run without disruption while you make changes to your storage infrastructure, which can help your business increase its availability to customers.

Figure 1-7 shows a high-level view of the SAN Volume Controller.

Figure 1-7 SAN Volume Controller

SAN Volume Controller is a vitally important building block for HA for Power Systems because it helps to preserve data access to clustered systems that are hosted in Power environments in the event of a storage failure.

For more information about the SAN Volume Controller architecture, see 1.3, “SAN Volume Controller” on page 38. For more information about the SAN Volume Controller implementation for this solution, see Chapter 3, “SAN Volume Controller Implementation” on page 149.

Automated cross-site DR and remote DR

Per the best practices for Power Systems environments and depending on your requirements as you move to continuous availability, you should consider cross platform and automated cross-site disaster recovery and remote disaster recovery if the DR data center is out of the

region from the production data center at an extended distance (typically greater than 300 KM). This is not an obstacle to having continuous availability because IBM has a PowerHA extended distance solution available.

Extended distance capabilities through PowerHA automated protection was introduced in v5.5, and further enhancements were added to the latest version that is available today through v7.1.2 Enterprise Edition. The XD feature of PowerHA SystemMirror provides HA across site-wide failures or large-scale disasters by mirroring business-critical data across sites while enabling failover to the remote sites. In PowerHA SystemMirror terms, a site is where nodes of the cluster are. PowerHA SystemMirror supports a maximum of two sites.

PowerHA System Mirror/XD uses the following mirroring or replication approaches to achieve its purpose:

- Data replication or mirroring over TCP/IP
- Data replication by using storage PPRC

PowerHA 7.1.2 Enterprise Edition offers host replication (synchronous and asynchronous) supported mirroring technologies through Geographical Logical Volume Manager (GLVM).

The GLVM is an AIX software-based technology that is built on AIX Logical Volume Manager to create a mirror copy of data at a geographically distant location. GLVM relies fully upon LVM for mirroring and is based on the concept of Remote Physical Volume (RP)). GLVM is not in the scope of the solution in this book because it is for distances greater than 300 KM, which is the upper limit for the SAN Volume Controller Stretched Cluster solution in this book. It is important to note, because it is one of the next building block technologies where there is a remote DR data center in the customer environment.

For more information about GLVM, see this website:

https://www.ibm.com/developerworks/mydeveloperworks/blogs/PowerHA/entry/geographic_logical_volume_manager_glvm1?lang=en

Storage array replication (synchronous and asynchronous) includes support for the following products:

- IBM Series DS8000®
- IBM Storwize
- IBM SAN Volume Controller
- IBM XIV®

Planned support: Support for EMC SRDF, Hitachi Universal Replicator and Truecopy, and HP Continuous Access is planned for 2013.

Another component in the Automated Cross-Site and Remote DR Building Block category is Global Mirror. Global Mirror is an IBM technology that provides data replication over extended distances between two sites for business continuity and disaster recovery. If adequate bandwidth exists, Global Mirror provides a recovery point objective (RPO) of as low as 3 - 5 seconds[12]between the two sites at extended distances with no performance impact on the application at the primary site. It replicates the data asynchronously and forms a consistency group at a regular interval that allows a clean recovery of the application. The two sites can be on separate continents or on different utility grids.

[12] For more information, see this website:
http://www-03.ibm.com/support/techdocs/atsmastr.nsf/5cb5ed706d254a8186256c71006d2e0a/5d86c53d414d5aaa862570a2005c2b74/$FILE/Global%20Mirror%20whitepaper%20v2.pdf

Global Mirror is based on existing IBM Copy Services functions: Global Copy and FlashCopy®. Global Mirror periodically pauses updates of the primary volumes and swaps change recording bitmaps. It then uses the previous bitmap to drain updates from the primary volumes to the secondaries. After all primary updates are drained, the secondary volumes are used as the source for a FlashCopy to tertiary volumes at the recovery site. This ensures that the tertiary copy of the volumes has point-in-time consistency. By grouping many volumes into one Global Mirror session, multiple volumes can be copied to the recovery site simultaneously while maintaining point-in-time consistency across those volumes.[13]

Global Mirror can be combined with a wide area clustering product like Geographically Dispersed Parallel Sysplex™ (GDPS®), HACMP/XD, or IBM TotalStorage Continuous Availability for Windows (CAW) to provide for automated failover between sites. This combined solution provides lower recovery time objective (RTO), because it allows most applications to automatically resume productive operation in 30 - 600 seconds.

The Global Mirror function is available on the IBM enterprise DS8K storage family, the Enterprise midrange storage servers DS4000® family, and the IBM SAN Volume Controller.

For more information about the comparison of Metro Mirror to Global Mirror, see the IBM Storage Information Center at this website:

http://publib.boulder.ibm.com/infocenter/dsichelp/ds8000ic/index.jsp?topic=%2Fcom.ibm.storage.ssic.help.doc%2Ff2c_pprcxsyncview_2mm5lj.html

Figure 1-8 shows a high-level PowerHA active-active design that is used across two data centers with PPRC synchronous data replication. It targets at meeting RPO requirements of seconds, and RTO <1 HR.

[13] For more information, see this website:
http://www.redbooks.ibm.com/redpieces/pdfs/sg246374.pdf

Figure 1-8 PowerHA with GLVM/PPRC and SAN Volume Controller: Cross data centers remote DR

So what does it all look like in a multi-cross platform, cross data center site architecture design? Figure 1-9 on page 24 shows an active-active cross-site architecture where clustered systems provide continuous availability, scalability, and disaster recovery.

Figure 1-9 Automated cross-site and cross platform active-active HA architecture and DR

Cross-system workload integration

In the Infrastructure Availability Continuum, the highest level is DR. In the IBM Power Systems HA building blocks diagram in Figure 1-2 on page 13, the highest level is Service Availability from a continuum perspective. Once again, looking at the definition of Availability from the IBM definitions for RAS.

Availability is how infrequently the functionality of a system or application is impacted by a fault or defect. It is the ability of a component or service to perform its required function at a stated instant or over a stated period of time. It is usually expressed as the availability ratio, for example, the proportion of time that the service is actually available for use by the customers within the agreed service hours (Agreed-Unavailable/Agreed)*100%. The availability of the infrastructure is one part of overall IT availability, and IT availability is, in turn, a subset of the overall organization or business availability.

The challenges and complexities that are involved in delivering and maintaining end-to-end, cross-site, and cross-system workload that is integrated move beyond the capabilities of Power Systems in isolation. They depend on other building blocks and complimentary technologies as seen in the diagram in Figure 1-2 on page 13, including Automation for Cross-Site DR and Remote DR. Many of these technologies also begin where disaster recovery starts, and the lines between an HA solution versus disaster recovery can become blurred because the complete solution can evolve to cross-site and cross-platform integration that continuous service availability requires.

There are a number of supporting technologies that can provide cross-platform or cross-system automation that compliment Power Systems. Remember that if automation and cluster management is required for middleware and databases that are hosted on Power Systems, the preferred and recommended solution is still PowerHA with SmartAssists to manage the application (middleware and database) clusters. If there are other components of the end-to-end service that is hosted on multiple platforms in addition to Power Systems (that is, Intel x86, Mainframe), there are other technologies that can work with Power Systems to provide higher levels of service availability.

Tivoli® System Automation Application Manager can work with IBM Power Systems to provide high levels of end-to-end service availability. These solutions are beyond the intended scope of this book.

IBM Tivoli System Automation Application Manager helps coordinate and manage application availability across cluster technologies on multiple platforms, so you can better control your enterprise business services. It helps you to establish one operations and automation team that is responsible for Linux, AIX, Windows, and Solaris applications. It also can simplify problem determination and resolution.

Figure 1-10 shows a high-level view of IBM Tivoli System Automation Application Manager.

Figure 1-10 Tivoli System Automation Application Manager

IBM Tivoli System Automation Application Manager includes the following features:

- Provides High Availability and Disaster Recovery management of composite business applications (Cross-Cluster + Cross-System + Cross-Site), and automation to decrease end-to-end system and service level RTO.
- Automates startup, shutdown, and resynchronization of composite business applications that are running on different platforms.
- Integrates IBM Tivoli Monitoring and IBM Tivoli Composite Application Manager managed endpoints for a consistent management of end-to-end composite business applications.
- Uses one web-based console for operations of system resource groups. All resources look the same and can be controlled the same way within one cluster for Tivoli System Automation Application Manager.
- Integrates various High-Availability Clusters (for example, Tivoli System Automation for Multiplatforms, SA z/OS, PowerHA, MSCS, Veritas Cluster Server). Automatic cross cluster recovery can be triggered by Tivoli System Automation Application Manager.
- Integrates application components with replicated storage devices to provide a multi-site integrated disaster recovery solution.

For more information, see these websites:

- http://www-01.ibm.com/software/tivoli/products/sys-auto-app-mgr/
- http://www.ibm.com/developerworks/wikis/display/tivoli/Tivoli+System+Automation+Application+Manager+Overview

DB2 high availability options

The following complimentary technologies are focused solely on making DB2 more highly available-hosted in a multiplatform environment in various options:

- High Availability Disaster Recovery

  The High Availability Disaster Recovery (HADR) feature provides an HA solution for partial and complete site failures. In a HADR environment, log data is shipped continuously from a primary database to one or more standby databases and reapplied to the standby databases. When the primary database fails, applications are redirected to a standby database that automatically takes over the role of the primary database.

- DB2 HA feature

  The DB2 HA feature enables integration between DB2 and cluster managing software. IBM Tivoli System Automation for Multiplatforms is bundled with DB2 10.1 on AIX and Linux as part of the DB2 HA Feature, and integrated with the DB2 installer. On Windows operating systems, Tivoli System Automation for Multiplatforms also is bundled with DB2 10.1, but it is not integrated with the DB2 database product installer.

- Operating system-dependent solutions

  Most operating systems provide tools or products to create clusters. The techniques and concepts are similar, and the intention always is to present to the world a single virtual server, although the services can be spread among the nodes or members of the cluster. In the event of a service or resource failure, cluster processes try to restart the resource in the affected node. If the restart is not possible, resources on another node take their place.

  DB2 10.1 supports the following operating system-dependent cluster managing software:

  - IBM PowerHA SystemMirror for AIX (formerly known as HACMP). This is the solution implemented in this book.
  - Cluster Multi-Processing for AIX or HACMP
  - Microsoft Cluster Server for Windows
  - Multi-Computer/ServiceGuard for Hewlett-Packard
  - Sun Cluster for Solaris

- Operating system-independent solutions

  Operating system-independent solutions (for example, Tivoli System Automation Application Manager, IBM Tivoli System Automation for Multiplatforms, Veritas Cluster Server) have characteristics similar to the operating system-dependent clustering software. This type of clustering software can work with various operating systems. They usually provide more sophisticated features to manage and control the clusters. Some allow frameworks that manage and interchange resources among groups of clusters.

  You can use the operating system independent clustering software under the following circumstances:

  - When you have a heterogeneous environment with different hardware providers.
  - If platform change is a possibility.
  - If your operating system does not offer a cluster solution or has restrictions that can be avoided with an operating system-independent clustering software.

- If you require to centrally manage a complex set of existing base-level clusters on differing platforms, which individually comprise only part of a larger overall application, then a multi-tiered cluster management solution such as Tivoli.
- Tivoli System Automation for Multiplatforms can plug in to existing base-level clusters without any requirement to swap architecture or vendors.

► HA through disk mirroring

Mirroring is the process of writing data to two separate hard disks at the same time. One copy of the data is called a mirror of the other. You can use disk mirroring to maintain a secondary copy of your primary database. You can use DB2 10.1 suspended I/O functionality to split the primary and secondary mirrored copies of the database without taking the database offline. After the primary and secondary databases copies are split, the secondary database can take over operations if the primary database fails. For more information about disk mirroring, see this website:

http://pic.dhe.ibm.com/infocenter/db2luw/v10r1/index.jsp?topic=%2Fcom.ibm.db2.luw.admin.ha.doc%2Fdoc%2Fc0006356.html

► HA through log shipping

Log shipping is the process of copying log files from a primary to a secondary database copy from an archive device or through a user exit program that is running against the primary database.

When the production machine fails, a failover occurs and the following processes must take place:

- The remaining logs are transferred over to the standby machine.
- The standby database rolls forward to the end of the logs.
- The clients reconnect to the standby database and resume operations.

When this approach is used, the primary database is restored to the standby machine by using the DB2 10.1 restore utility or the split mirror function.

► Automatic Client Reroute

Automatic Client Reroute (ACR) is a DB2 10.1 feature that you can use in an HA database solution to redirect client applications from a failed server to an alternative server so the applications can continue their work with minimal interruption. ACR can be accomplished only if an alternative server was specified before the failure.

For more information about automatic client reroute, see this website:

http://pic.dhe.ibm.com/infocenter/db2luw/v10r1/index.jsp?topic=%2Fcom.ibm.db2.luw.admin.ha.doc%2Fdoc%2Fr0023392.html

► DB2 pureScale®

You can use the IBM DB2 pureScale feature to scale a database for a set of servers in an active-active approach. Traffic that is intended for a failed node is passed on to an existing node or load balanced for the remaining nodes. This technology is similar to the proven data sharing architecture that is found in DB2 for IBM z/OS. This feature provides the benefits of transparency, continuous availability, and scalability at a lower operational cost. The DB2 pureScale technology provides an active-active configuration that transparently delivers high throughput, continuous availability, and scalability.

Figure 1-11 shows the main components of a DB2 pureScale cluster.

Figure 1-11 DB2 pureScale Cluster

As shown in Figure 1-11, a DB2 pureScale cluster is composed of the following components:

- A DB2 member (also referred to as a *member*) is the core component that handles and processes database transactions (application requests) within the clustered system. DB2 members are the database engine processes that manage and retrieve the data from the database. Adding a DB2 member refers to adding a processing node within the cluster for data sharing.
- The cluster caching facility (CF) is responsible for instance-wide lock management and global caching. This component relieves the members from having to handle locking, caching, and communication with each other to maintain vital locking and data consistency information.
- The shared storage is a storage unit that is connected to all DB2 members and CFs in the cluster to allow for all the data to be shared and used simultaneously. The DB2 pureScale Feature supports all SANs that work as a storage switch and fibre attached shared block storage, which is referenced as a logical unit number (LUN).

– The cluster interconnect refers to the high speed, low latency networking that allows the DB2 members and CFs to communicate with each other and transfer data or logs to process transactions. The cluster interconnect uses a 10 Gigabit Ethernet (GbE) or InfiniBand connection to communicate. These features are used to take advantage of Remote Direct Memory Access (RDMA), which helps use centralized locking and buffer caching.

– The clients (application) refer to clients and the applications that are running transactions against the DB2 pureScale Feature and associated databases.

For more information, see the following publications:

– *Delivering Continuity and Extreme Capacity with the IBM DB2 pureScale Feature*, SG24-8018-00, which is available at this website:

  http://www.redbooks.ibm.com/redbooks/pdfs/sg248018.pdf

– *High Availability and Disaster Recovery Options for DB2 for Linux, UNIX, and Windows*, SG24-7363-02, which is available at website:

  http://www.redbooks.ibm.com/abstracts/sg247363.html

In the next section, we describe the tiers (or building blocks) for DR, where the Infrastructure Availability Continuum and IBM Power Systems building blocks leave off, and continue with DR. It is important to note that these are different solution paths, and the solution in this book is positioned as an HA solution building block. However, where the solution fits or aligns along the DR tiering continuum also is described, based on what other prerequisite DR building block solutions already are implemented in your environment.

1.1.5 Tiers of DR

From an IT infrastructure standpoint, there are various valid DR products available. The fundamental challenge is to select the optimum blend of all these DR products and technologies, to provide an all encompassing solution, that will pick up where HA coverage ends, and to ideally be used in the event of a site level disaster event. The common problem in the past was a tendency to view the DR solution as individual product technologies and piece parts. Instead, DR solutions must be viewed as a whole integrated multiproduct solution.

To combine and properly select among multiple products, disciplines, and skills to affect a successful IT DR solution, we first observe that we can categorize all valid DR IT technologies into the following component domains:

- Servers
- Storage
- Software and automation
- Networking and physical infrastructure
- Skills and services that are required to implement and operate the other components

All IT infrastructure that is necessary to support the DR solution can be inserted into one of these components, as shown in Figure 1-12 on page 30.

Figure 1-12 Five components of DR

These five categories provide a framework to organize the various component evaluation skills that are needed. Gathering the proper mix of evaluation skills facilitates an effective comparison, contrast, and blending of all five product component areas to arrive at an optimum solution.

The concept of tiers is a common method used in today's best practices for DR solution design. The concept of tiers is powerful and central to our selection methodology because the tiers concept recognizes that for a client RTO, all DR products and technologies can be sorted into an RTO solution subset that addresses that particular RTO range.

By categorizing DR technology into the various tiers, we can more easily match our wanted RTO time with the optimum set of technologies. The reason for multiple tiers is that as the RTO time decreases, the optimum DR technologies for RTO must change. For any given RTO, there are always a particular set of optimum price and performance DR technologies.

The tiers concept is flexible. As products and functions change and improve over time, the tiers chart only needs to be updated by the adding that new technology into the appropriate tier and RTO.

The DR tiers chart that is shown in Figure 1-13 on page 31 gives a generalized view of some of today's IBM DR technologies by tier. As the recovery time becomes shorter, more aggressive DR technologies must be applied to achieve that RTO (carrying with them their associated increase in value and capital cost).

Figure 1-13 Tiers of DR

The concept and shape of the tiers chart continues to apply even as the scale of the application or applications changes. Large-scale applications tend to move the curve to the right, and small scale applications tend to move the curve to the left. But in both cases, the general relationship of the various tiers and DR technologies to each other remains the same. Finally, although some DR technologies fit into multiple tiers, clearly there is not one DR technology that can be optimized for all the tiers.

Of course, your technical staff can and should, when appropriate, create a specific version of the tiers chart for your particular environment. After the staff agrees on what tier or tiers and corresponding RTO a solution delivers for your enterprise, then DR technical evaluation and comparisons are much easier, and the technology alternatives can be tracked and organized in relation to each other. Although the technology within the tiers has changed through time, the concept continues to be as valid today as when it was first described by the US SHARE user group in 1988.

1.1.6 Blending tiers into an optimized solution

Best practices today in designing a DR solution is to further use the tiers concept to derive a blended DR solution for the entire enterprise. The most common result, from an enterprise standpoint, is a strategic architecture of three tiers in a blended DR solution. Three tiers generally appear as an optimum number because at the enterprise level, two tiers generally are insufficiently optimized (in other words, overkill at some point and not enough at others), and four tiers are more complex, but generally do not provide enough additional strategic benefit.

To use the tiers to derive a blended, optimized enterprise DR architecture, we suggest the following steps:

1. Categorize the business' entire set of applications into the following bands:
   - Low tolerance to outage
   - Somewhat tolerant to outage
   - Very tolerant to outage

   Of course, although some applications are not critical, they do feed the critical applications. Therefore, those applications must be included in the higher tier.
2. Within each band, there are tiers. The individual tiers represent the major DR technology choices for that band. It is not necessary to use all the tiers, and of course, it is not necessary to use all the technologies.
3. After we segment the applications (as best we can) into the three bands, we usually select one best strategic DR methodology for that band. The contents of the tiers are the candidate technologies from which the strategic methodology is chosen.

A blended architecture optimizes and maps the varying application recovery time requirements with an appropriate technology at an optimized cost. The net resulting DR architecture, based on the blended solution tiers, provides the best possible application coverage for the minimum cost.

As mentioned previously, the SAN Volume Controller Split Cluster, PowerHA, and LPM technology solution is ideally positioned as an active-active cross site clustered solution for data access protection, cluster management, and a solution to reduce planned outage windows for IBM Power Systems server maintenance for critical workloads.

In context to the DR Tiers, this solution is ideally positioned for Tier 1 critical applications, which have a low tolerance to outages, whether the outages are unplanned or planned. It is important to note that the best practice is a blending of the solution tiers to maximize application coverage at the lowest possible cost. One size, one technology, or one methodology does not fit all applications.

In addition, the higher levels of availability and DR protection are reached through a building block approach. In other words, prerequisite solutions should be implemented on the journey to HA, continuous operations, continuous availability, and mature, comprehensive disaster recovery. Customers should use their investments through these building blocks for HA, and tiering through DR to reduce their RTO and RPO times.

The solution in this publication is complimentary to the investments that are made in DR technology. Also, because it should be used to support applications with low tolerance to outages, it is important that the customer environment should be at least Tier 5 in terms of prerequisite DR technology to gain the full benefits of this SAN Volume Controller Split Cluster, PowerVM, PowerHA and LPM solution.

1.2 Storwize V7000

This chapter provides an overview of the IBM Storwize V7000 architecture as it applies to the SAN Volume Controller Split Cluster configuration and solution context in this book. It assumes a base understanding of storage architecture in general. It also includes a brief explanation of storage virtualization.

The Storwize V7000 is a solid storage offering from IBM that is ideal for mid-size environments and as an entry point to storage virtualization. It is not in the same product class as the IBM DS8K Enterprise Storage product family.

At the time of writing, it cannot be implemented as a Split Cluster solution. In this book, we use it as back-end storage for our solution only.

1.2.1 Overview and concepts

In this section, we provide an overview of storage virtualization and concepts.

Storage virtualization

Storage virtualization, like server virtualization, is one of the foundations of building a flexible and reliable infrastructure solution that allows companies to better align their business and IT needs.

Storage virtualization allows an organization to implement pools of storage across physically separate disk systems (which might be from different vendors). Storage can then be deployed from these pools and migrated between pools without any outage of the attached host systems. Storage virtualization provides a single set of tools for advanced functions, such as instant copy and remote mirroring solutions, which means that deploying storage can be performed by using a single tool, regardless of the underlying storage hardware.

Figure 1-14 on page 34 shows a storage virtualization scenario.

Figure 1-14 Storage virtualization

For more information about storage virtualization, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

IBM Storwize V7000

The IBM Storwize V7000 solution provides a modular storage system that includes the capability to virtualize external SAN-attached storage and its own internal storage. The IBM Storwize V7000 solution is built upon the IBM SAN Volume Controller technology base and uses technology from the IBM System Storage DS8000 family. Storwize V7000 is an entry to midrange virtualizing RAID storage subsystem. It incorporates all of the functions of the SAN Volume Controller virtualization appliance and provides the following features and benefits:

- Exceptional ease of deployment and use; a specialized storage administrator often is not required.
- Scalable and extensible RAID storage, which allows you to grow the system incrementally as storage capacity and performance needs change.
- Full range of disk-drive technology, including Solid-State drives (SSD), which allows for implementation of multiple storage tiers in a single system with non-disruptive migration between them.
- Strong ecosystem support and ISV integration, which provides easier deployment into your existing server environment.
- Included with SAN Volume Controller and the IBM Storwize family storage systems is a simple and easy-to-use graphical user interface (GUI), which is designed to allow storage to be deployed quickly and efficiently. The GUI runs through a web browser window on the SAN Volume Controller, Storwize V7000, and Storwize V3700 systems, and through the

IBM Flex System™ Manager on Flex System V7000, so there is no need for a separate console.

An IBM Storwize V7000 system provides various configuration options that are aimed at simplifying the implementation process. It also provides automated wizards, called *Directed Maintenance Procedures* (DMP), to assist in resolving any events that might occur. An IBM Storwize V7000 system is a clustered, scalable, and midrange storage system and an external virtualization device.

Figure 1-15 shows a high-level overview of the IBM Storwize V7000.

Figure 1-15 IBM Storwize V7000 overview

The GUI runs on the IBM Storwize V7000 system, so there is no need for a separate console. The management GUI contains a series of preestablished configuration options called *presets* that feature commonly used settings to quickly configure objects on the system. Presets are available for creating volumes and IBM FlashCopy mappings and for setting up a RAID configuration.

The IBM Storwize V7000 solution provides a choice of up to 240 x 3.5-inch or 480 x 2.5-inch Serial Attached SCSI (SAS) drives for the internal storage and uses SAS cables and connectors to attach to the optional expansion enclosures.

When external storage arrays are virtualized, an IBM Storwize V7000 system can provide up to 32 PB of usable capacity. An IBM Storwize V7000 system supports various external disk systems similar to what the SAN Volume Controller supports today.

The IBM Storwize V7000 solution consists of a one or two control enclosures and, optionally, up to 18 expansion enclosures (and supports the intermixing of the different expansion

enclosures). Within each enclosure are two canisters. Control enclosures contain two node canisters, and expansion enclosures contain two expansion canisters.

When external storage arrays are virtualized, SAN Volume Controller and the Storwize family of storage systems support a wide range of external disk systems from leading storage vendors. The Storwize V7000, Flex System V7000, and Storwize V3700 systems consist of a pair of redundant controllers, called *nodes*, and expansion enclosures that can support solid-state disk, serial-attached SCSI (SAS), and nearline SAS drive types. The SAN Volume Controller does not support expansion enclosures or internal disk storage, but virtualizes external disk storage only. The pair of redundant nodes combines to form an I/O group and clustering allows multiple I/O groups to be managed as a single storage system. There can be up to four I/O groups in an SAN Volume Controller, Storwize V7000, or Flex System V7000 cluster, while only a single I/O group is allowed in a Storwize V3700 cluster.

Figure 1-16 shows the Storwize V7000 disk system.

Figure 1-16 IBM Storwize V7000 Disk System

The SAN Volume Controller and Storwize family storage systems include flexible host connectivity options with support for 8 Gb Fibre Channel, 1 Gb iSCSI, and 10 Gb iSCSI connections.

The following full array of advanced software features are available:

- Seamless data migration
- Thin provisioning
- Volume mirroring
- Global Mirror and Metro Mirror replication
- FlashCopy: 256 targets, cascaded, incremental, space efficient (thin-provisioned)
- Integration with IBM Tivoli Productivity Center for end-to-end SAN management
- IBM System Storage Easy Tier®, which provides a mechanism to seamlessly migrate hot spots to a higher-performing storage pool

1.2.2 Architecture and solution context

This section describes the architecture a solution context.

Deployment architecture

A Storwize V7000 subsystem consists of a set of drive enclosures. Control Enclosures contain disk drives and two SAN Volume Controller nodes (an I/O group), which are attached to the SAN fabric. Expansion Enclosures contain drives and are attached to control enclosures.

The simplest use of Storwize V7000 is as a traditional RAID subsystem, as show in Figure 1-17 on page 37. The internal drives are configured into RAID arrays and virtual disks are created from those arrays.

Figure 1-17 Storwize V7000 as a RAID subsystem

Storwize V7000 also can be used to virtualize other storage controllers, as shown in Figure 1-18. Multiple Storwize V7000 control enclosures can be clustered to scale out the system. Storwize V7000 control enclosures and SAN Volume Controller appliance nodes cannot be part of the same cluster.

Storwize V7000 supports regular and solid-state drives (SSDs) and IBM System Storage Easy Tier to automatically place volume hot-spots on better-performing storage.

Figure 1-18 Storwize V7000 as a virtualization solution

For the solution in this book Storwize V7000 6.3 (as back-end storage) is implemented with SAN Volume Controller V6.4.

1.3 SAN Volume Controller

This section provides an overview of the SAN Volume Controller, the topology and architecture, and the solution context for SAN Volume Controller in this solution.

1.3.1 Overview and Concepts

SAN Volume Controller is a storage virtualization system that enables a single point of control for storage resources to help support improved business application availability and greater resource usage. The objective is to manage storage resources in your IT infrastructure and to make sure they are used to the advantage of your business and do it quickly, efficiently, and in real time, while avoiding increases in administrative costs.

SAN Volume Controller combines software and hardware into a comprehensive, modular appliance that uses symmetric virtualization. Symmetric virtualization is achieved by creating a pool of managed disks (MDisks) from the attached storage systems. Those storage systems are then mapped to a set of volumes for use by attached host systems. System administrators can view and access a common pool of storage on the SAN. This functionality helps administrators to use storage resources more efficiently and provides a common base for advanced functions.

SAN Volume Controller configuration flexibility means that your SAN Volume Controller implementation can start small and then grow with your business to manage very large storage environments.

SAN Volume Controller combines hardware and software into an integrated, modular solution that is highly scalable. An I/O Group is formed by combining a redundant pair of storage engines based on IBM System x® server technology. Highly available I/O Groups are the basic configuration element of a SAN Volume Controller cluster.

SAN Volume Controller software performs the following functions for the host systems that attach to SAN Volume Controller:

- Creates a single pool of storage
- Provides logical unit virtualization
- Manages logical volumes
- Mirrors logical volumes

The SAN Volume Controller system also provides the following functions:

- Large scalable cache
- Copy Services:
  - IBM FlashCopy (point-in-time copy) function, including thin-provisioned FlashCopy to make multiple targets affordable
  - Metro Mirror (synchronous copy)
  - Global Mirror (asynchronous copy)
  - Data migration
- Space management:
  - IBM System Storage Easy Tier to migrate the most frequently used data to higher performing storage
  - Metering of service quality when combined with IBM Tivoli Storage Productivity Center
  - Thin-provisioned logical volumes

– Compressed volumes to consolidate storage

A SAN is a high-speed Fibre Channel (FC) network that connects host systems and storage devices. In a SAN, a host system can be connected to a storage device across the network. The connections are made through units such as routers and switches. The area of the network that contains these units is known as the fabric of the network.

Figure 1-19 shows hosts, SAN Volume Controller nodes, and RAID storage systems that are connected to a SAN fabric. The redundant SAN fabric comprises a fault-tolerant arrangement of two or more counterpart SANs that provide alternative paths for each SAN-attached device.

Figure 1-19 SAN Volume Controller in a Fabric

A system of SAN Volume Controller nodes presents volumes to the hosts. Most of the advanced functions that SAN Volume Controller provides are defined on volumes. These volumes are created from MDisks that are presented by the RAID storage systems. All data transfer occurs through the SAN Volume Controller nodes, a process that is described as symmetric virtualization.

Figure 1-20 shows the data flow across the fabric.

Figure 1-20 Data flow in SAN Volume Controller

SAN Volume Controller supports attachment to servers by using FC protocols and iSCSI protocols over IP networks at 1 Gbps or 10 Gbps speeds, which can help reduce costs and simplify server configuration. SAN Volume Controller also supports FCoE protocol. I/O operations between hosts and SAN Volume Controller nodes and between SAN Volume Controller nodes and RAID storage systems are performed by using the SCSI standard. The SAN Volume Controller nodes communicate with each other by using private SCSI commands.

Fibre Channel over Ethernet (FCoE) connectivity is supported on SAN Volume Controller node model 2145-CG8 only, after the system software is upgraded to version 6.4.

Table 1-1 shows the fabric type that can be used for communicating between hosts, nodes, and RAID storage systems. These fabric types can be used at the same time.

Table 1-1 SAN Volume Controller Communications Types

Communications Type	Host to SAN Volume Controller	Host to SAN Volume Controller	SAN Volume Controller to SAN Volume Controller
Fibre Channel SAN	Yes	Yes	Yes
iSCSI (1 Gbps Ethernet or 10 Gbps Ethernet)	Yes	No	No
Fibre Channel over Ethernet SAN (10 Gbps Ethernet)	Yes	Yes	Yes

SAN Volume Controller supports SSDs enabling scale-out high performance SSD support. These internal SSDs can be used to create RAID MDisks that in turn can be used to create volumes. SSDs provide host servers with a pool of high-performance storage for critical applications. The scalable architecture of this solution and the tight integration of SSDs enable businesses to take advantage of the high throughput capabilities of the SSDs. The scalable architecture is designed to deliver outstanding performance with SSDs for critical applications.

SAN Volume Controller also includes the System Storage Easy Tier function, which is designed to help improve performance at lower cost through more efficient use of SSDs. The Easy Tier function automatically identifies highly active data within volumes and moves only the active data to an SSD. In this way, the Easy Tier function targets the use of an SSD to the data that benefits the most, which helps deliver the maximum benefit even from small amounts of SSD capacity. SAN Volume Controller helps move critical data to and from SSDs as needed without application disruption.

SAN Volume Controller helps increase the amount of storage capacity that is available to host applications by pooling the capacity from multiple disk systems within the SAN.

In addition, SAN Volume Controller combines various IBM technologies including thin provisioning, automated tiering, storage virtualization, Real-time Compression™, clustering, replication, multiprotocol support, and a next-generation GUI. Together, these technologies enable SAN Volume Controller to deliver extraordinary levels of storage efficiency.

SAN Volume Controller includes a dynamic data migration function that moves data from one storage system to another while maintaining access to the data.

The SAN Volume Controller Volume Mirroring function also stores two copies of a volume on different storage systems. This function helps improve application availability in the event of failure or disruptive maintenance to an array or disk system. SAN Volume Controller is automatically uses whichever copy of the data remains available.

SAN Volume Controller enables administrators to apply a single set of advanced network-based replication services that operate in a consistent manner, regardless of the type of storage that is used. The Metro Mirror and Global Mirror functions operate between SAN Volume Controller systems at different locations to help create copies of data for use in the event of a catastrophic event at a data center. For even greater flexibility, Metro Mirror and Global Mirror also support replication between SAN Volume Controller systems and IBM Storwize V7000 Unified systems.

The IBM FlashCopy function creates an almost instant copy of active data that can be used for backup purposes or for parallel processing activities. This capability enables disk backup copies to be used to recover almost instantly from corrupted data, which significantly speeds application recovery.

Figure 1-21 shows the SAN Volume Controller structure and its components.

Figure 1-21 SAN Volume Controller structure

1.3.2 Architecture and solution context

The IBM System Storage SAN Volume Controller is a SAN block aggregation virtualization appliance that can attach to various host computer systems.

The following major approaches can be considered for the implementation of block-level aggregation and virtualization:

- Symmetric: In-band appliance

  The device is a SAN appliance that sits in the data path, and all I/O flows through the device. This implementation is also referred to as *symmetric virtualization* or *in-band*. The device is target and initiator. It is the target of I/O requests from the host perspective, and the initiator of I/O requests from the storage perspective. The redirection is performed by issuing new I/O requests to the storage. The SAN Volume Controller uses symmetric virtualization. For the solution in this publication, we are using the SAN Volume Controller symmetrically as an In-band appliance.

- Asymmetric: Out-of-band or controller-based

  The device is usually a storage controller that provides an internal switch for external storage attachment. In this approach, the storage controller intercepts and redirects I/O requests to the external storage as it does for internal storage. The actual I/O requests are themselves redirected. This implementation is also referred to as *asymmetric virtualization* or *out-of-band*.

Figure 1-22 shows variations of the two virtualization approaches.

Figure 1-22 Overview of block-level virtualization architectures

SAN-based storage is managed by the SAN Volume Controller in one or more pairs of SAN Volume Controller hardware nodes,. This configuration is referred to as a *clustered system* or *system*. These nodes are attached to the SAN fabric with RAID controllers and host systems. The SAN fabric is zoned to allow the SAN Volume Controller to “see” the RAID controllers, and for the hosts to “see” the SAN Volume Controller. The hosts are not allowed to “see” or operate on the same physical storage (LUN) from the RAID controller that was assigned to the SAN Volume Controller. Storage controllers can be shared between the SAN Volume Controller and direct host access if the same LUNs are not shared. The zoning capabilities of the SAN switch must be used to create distinct zones to ensure that this rule is enforced.

Figure 1-23 on page 44 shows a conceptual diagram of a storage system that uses the SAN Volume Controller. It shows a number of hosts that are connected to a SAN fabric or LAN. In practical implementations that have HA requirements (the majority of the target clients for SAN Volume Controller), the SAN fabric cloud represents a redundant SAN. A *redundant SAN* consists of a fault-tolerant arrangement of two or more counterpart SANs, which provide alternative paths for each SAN-attached device.

Both scenarios (which are using a single network and two physically separate networks) are supported for iSCSI-based and LAN-based access networks to the SAN Volume Controller. Redundant paths to volumes can be provided in both scenarios.

For simplicity, Figure 1-23 shows only one SAN fabric and two zones, namely host and storage. In a real environment, it is a preferred practice to use two redundant SAN fabrics. We implemented two redundant fabrics for the solution in this book. The SAN Volume Controller can be connected to up to four fabrics.

Figure 1-23 SAN Volume Controller conceptual and topology overview

A clustered system of SAN Volume Controller nodes that are connected to the same fabric presents logical disks or volumes to the hosts. These volumes are created from managed LUNs or MDisks that are presented by the RAID disk subsystems. The following distinct zones are shown in the fabric:

- A host zone, in which the hosts can see and address the SAN Volume Controller nodes.
- A storage zone, in which the SAN Volume Controller nodes can see and address the MDisks or logical unit numbers (LUNs) that are presented by the RAID subsystems.

Hosts are not permitted to operate on the RAID LUNs directly, and all data transfer happens through the SAN Volume Controller nodes. This design is commonly described as *symmetric virtualization*.

For iSCSI-based access, the use of two networks and separating iSCSI traffic within the networks by using a dedicated virtual local area network (VLAN) path for storage traffic prevents any IP interface, switch, or target port failure from compromising the host servers access to the volumes LUNs.

In terms of the specific solution context for the solution in this book, we implemented the SAN Volume Controller Split Cluster solution. For the sake of simplicity, we implemented a single node/single I/O group architecture with the single node and I/O split over the two data center locations. This can, in fact, also be a real-world recommend implementation for a critical system in which the I/O activity characteristics are well-known and monitored.

It is not advisable to group and implement all I/O activity in a single I/O group or node configuration. It is best to have these balanced and redundant in at least a two-node and I/O group configuration. However, a single node/single I/O group configuration for SAN Volume Controller split cluster is a highly available solution for a critical system that is hosted in its own single I/O group.

Figure 1-24 shows a well-balanced and highly available SAN Volume Controller configuration with two redundant fabrics. For illustrative purposes, the storage in Figure 1-24 is DS4000; a similar design can be implemented with IBM Storwize V7000, XIV, and DS8K.

Figure 1-24 SAN Volume Controller recommended practice reference architecture

The solution context and role of SAN Volume Controller in the Split Cluster solution in this book is described in 1.4, “SAN Volume Controller Stretched Cluster” on page 46.

For more information about the SAN Volume Controller, see the following resources:

- *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933-01, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/SG247933.html?Open

- IBM SAN Volume Controller Information Center:

  http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp

1.4 SAN Volume Controller Stretched Cluster

This section focuses on the SAN Volume Controller architecture as it applies to the SAN Volume Controller Stretched Cluster configuration, and solution context in this book. It assumes a base understanding of the general SAN Volume Controller architecture.

1.4.1 Overview and concepts

In a standard SAN Volume Controller configuration, all nodes are within the same rack. When SAN Volume Controller was first introduced, the maximum supported distance between nodes within an I/O Group was 100 m (328.08 ft.).

The SAN Volume Controller Stretched Cluster configuration provides a continuous availability platform whereby host access is maintained in the event of the loss of any single failure domain. This is accomplished through the inherent active-active architecture of SAN Volume Controller along with the use of volume mirroring. In the event of a failure, the SAN Volume Controller node that the associated mirrored copy of the data is on remains online and is available to service all host I/O.

With SAN Volume Controller Stretched Cluster, the two nodes in an I/O Group are split or separated by distance between two locations and a copy of the volume is stored at each location. This means that you can lose the SAN or power at one location and still access the disks that remain available at the alternative location. The use of this feature requires clustering software at the application and server layer to failover to a server at the alternative location and resume access to the disks. Because the SAN Volume Controller keeps both copies of the storage in synchronization and the SAN Volume Controller cache is mirrored between both nodes, the loss of one location causes no disruption to the alternative location.

As with any clustering solution, avoiding a possible situation called split-brain (where nodes cannot communicate with each other) a tie break is needed. SAN Volume Controller is no exception. The SAN Volume Controller uses a tie-break mechanism that is facilitated through the implementation of a quorum disk. The SAN Volume Controller selects three quorum disks from the MDisks that are attached to the cluster that is used for this purpose. For the most part, the management of the quorum disks is not apparent to the SAN Volume Controller users. However, in an SAN Volume Controller Stretched Cluster configuration, the location of the quorum disks must be assigned manually to ensure that the active quorum disk is in a third location. This must be done to ensure the survival of one location in the event a failure were to occur at another location.

The quorum disk fulfills the following functions for cluster reliability:

- Acts as tiebreaker in split brain scenarios
- Saves critical configuration metadata

The SAN Volume Controller quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. There are three quorum disk candidates. At most, one of these candidates acts as the active quorum disk. The other two are reserved to become active if the current active quorum disk fails. All three quorum disks are used to store configuration metadata, but only the active quorum disk acts as tiebreaker and is used in T3 recovery.

Important: Each quorum disk should be placed in one site (failure domain). The quorum disk in the third site (quorum site) should be set as active quorum disk.

For more information about quorum disk requirements and placement, see 3.11, “SAN Volume Controller quorum disk” on page 188.

> **Critical:** A quorum disk must be placed in each of the three failure domains. The quorum disk in the third failure domain should be set as the active quorum disk. The loss of all quorum disks results in a loss of data.

In SAN Volume Controller version 5.1, SAN Volume Controller introduced support for the Stretched Cluster configuration where nodes within an I/O Group can be separated by a distance of up to 10 km (6.21 miles). This capability allows for nodes to be placed in separate failure domains, which provide protection against failures effecting a single failure domain. The initial support for Stretched Cluster delivered in SAN Volume Controller 5.1 included the restriction that all communication between SAN Volume Controller node ports cannot traverse through inter-switch links (ISLs), which limited the maximum supported distance between failure domains.

Starting with SAN Volume Controller 6.3, the ISL restriction was removed, which allowed the distance between failure domains to be extended to 300 km (186.41 miles). Additionally, in SAN Volume Controller 6.3, the maximum supported distance for non-ISL configurations was extended to 40 km (24.85 miles).

The links between fabrics at either site have certain requirements that must be validated.

For more information about Stretched Cluster prerequisites, see this website:

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102134

SAN Volume Controller is a flexible solution, especially when you consider that you can use the storage controller of your choice at any of the three locations and with SAN Volume Controller, they can be from different vendors. Also, this is all possible by using the base SAN Volume Controller virtualization license with no additional charge.

SAN Volume Controller fully makes use of two major I/O functions that were introduced beginning with SAN Volume Controller software release 4.3: thin provisioning and volume mirroring. The volume mirroring function is a mechanism by which a single volume has two physical copies of the data on two independent MDisk Groups (storage pools and storage controllers). This feature provides the following capabilities:

- Changes the extent size of a volume.
- Migrates between storage controllers, or split off a copy of a volume for development or test purposes.
- Increases redundancy and reliability of lower-cost storage controllers and provides a temporary mechanism to add a second copy to a set of volumes to enable disruptive maintenance to be performed to a storage controller without any loss of access to servers and applications.

Another capability provided by VDM is the ability to split the cluster while maintaining access to clustered servers and applications.

Imagine that you have two servers acting as a cluster for an application. These two servers are in different machine rooms and power domains and are attached to different fabrics. You also have two storage controllers, one in each computer room. You want to mirror data between the controllers, and at the same time provide access to users when you lose power or access to disks within one of the machine rooms. This can now be done through the implementation of the SAN Volume Controller Stretched Cluster configuration.

The solution we explore in this book focuses on SAN Volume Controller and the PowerVM environment. However, the SAN Volume Controller Stretched Cluster configuration can be applied to any other operating system and environment, such as native Microsoft Cluster, Intel with VMWare, or Linux Cluster.

All of the SAN Volume Controller Stretched Cluster benefits and protection criteria apply and use data protection and business continuity requirements regardless as to the operating system your application is using.

For more information about interoperability of the SAN Volume Controller Stretched Cluster configuration, see this website:

http://www-03.ibm.com/systems/storage/software/virtualization/svc/interop.html

Failure domains

In an SAN Volume Controller Stretched Cluster configuration, the term *failure domain* is used to identify components of the SAN Volume Controller cluster that are contained within a boundary such that any failure that can occur within that boundary (such as power failure, fire, and flood) is contained within that boundary and cannot propagate or affect components that are outside of that boundary. The components that comprise an SAN Volume Controller Stretched Cluster configuration must span three independent failure domains. Two failure domains contain SAN Volume Controller nodes and the storage controllers containing customer data. A third failure domain contains a storage controller where the active quorum disk is located.

Failure domains are typically areas or rooms in the data center, buildings on the same campus, or even buildings in different towns. Different kinds of failure domains protect against different types of failure conditions, as shown in the following examples:

- If each failure domain is an area with a separate electrical power source within the same data center, the SAN Volume Controller can maintain availability if a failure of any single power source were to occur.
- If each site is a different building, the SAN Volume Controller can maintain availability in the event that a loss of any single building were to occur (for example, power failure or fire).

Ideally, each of the three failure domains that are used for the SAN Volume Controller Stretched Cluster configuration are in separate buildings and powered by separate power sources. While this configuration offers the highest level of protection against all possible failure and disaster conditions, it is not always possible and some compromise is often required.

In cases where a third building is unavailable, it is possible to place the failure domain that contains the active quorum disk in the same building as one of the other two failure domains. When this is done, the following rules apply:

- Each failure domain must be powered by an independent power supply or uninterruptible power supply (UPS).
- The storage controller that is used for the quorum disk must be separate from the storage controller that is used for the customer data.
- Each failure domain should be on independent and isolated SANs (separate cabling and switches).
- All cabling (power, SAN, and IP) from one failure domain must not be physically routed through another failure domain.
- Each failure domain should be placed in separate fire compartments.

Important: The key prerequisite for failure domains as they pertain to the SAN Volume Controller Stretched Cluster is that each node from an I/O Group must be placed in separate failure domains. Each I/O Group within the SAN Volume Controller cluster must adhere to this rule.

A diagram of the typical failure domains for the SAN Volume Controller Stretched Cluster solution with SAN connectivity is shown in 1.4.2, “Architecture and solution context” on page 49.

1.4.2 Architecture and solution context

Figure 1-25 shows the components that are used for the SAN Volume Controller Stretched Cluster and the SAN connectivity in our solution.

Figure 1-25 SAN Volume Controller Stretched Cluster architecture with SAN connectivity

This book does not describe the physical installation of the SAN Volume Controller. It is intended to supplement the *SAN Volume Controller V6.3.0 Configuration Guidelines for Extended Distance Split-System Configurations for IBM System Storage SAN Volume Controller*, S7003701, which is available at this website:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

We assume that the reader is familiar with the major concepts of SAN Volume Controller clusters such as nodes, I/O groups, MDisks, and quorum disks. If you are not, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

For more information, see this website:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp

Also, for more information about the implementation and configuration with the failure domain scenarios for the specific SAN Volume Controller Stretched Cluster solution in this book, see Chapter 3, “SAN Volume Controller Implementation” on page 149.

1.5 Fibre Channel SAN

This section provides an overview of the Fibre Channel (FC) SAN, along with the specific architecture and solution context for the solution in this book.

1.5.1 Overview and concepts

We assume that the reader is familiar with general FC SAN design and technologies. Typical best practice SAN design features servers and storage that are connecting into dual, redundant fabrics. In our lab configuration, we implement a redundant fabric design by using two IBM SAN768B-2 chassis at each data center site. Each IBM SAN768B-2 is equipped with the following components:

- IBM FC 8 Gbps FCIP Extension blade in Slot 1
- IBM FC 16 Gbps 48-port blade in Slot 8

IBM SAN Volume Controller Stretched Cluster requires the following types of SAN:

- Public SAN: Dedicated for host attachment, storage system attachment, IBM SAN Volume Controller, and links for metro or Global Mirror.
- Private SAN: Dedicated for IBM SAN Volume Controller node-to-node communication.

Each IBM SAN768B-2 is split into two logical chassis to implement segregated Private and Public SANs on the same chassis, as seen in Figure 1-26.

Figure 1-26 Public and Private SAN connections at each data center site

SAN Volume Controller split I/O group clusters support the following approaches for node-to-node intra-cluster communication between production sites:

- Attach each SAN Volume Controller node to the FC switches in the local production site and in the remote production site directly. Thus, all node-to-node traffic can be done without passing ISLs. This configuration is referred to as split I/O group configuration without ISLs between SAN Volume Controller nodes.
- Attach each SAN Volume Controller node only to local FC switches and configure ISLs between production sites for SAN Volume Controller node-to-node traffic. This configuration is referred to as split I/O group configuration with ISLs between SAN Volume Controller nodes. For longer distances, the ISL can imply the connection through FCIP router or Active/Passive WDM. Before v6.3, no ISLs between nodes of same I/O group are supported. SAN Volume Controller support for FCIP was introduced in SAN Volume Controller 6.4.

In our implementation of AIX LPM, the second approach with FCIP is used (the distance implies latency).

The use of ISLs for SAN Volume Controller split I/O group clusters requires the private SAN and public SAN to ensure the node-to-node communication. In a private SAN, the SAN is dedicated for SAN Volume Controller node-to-node communication. In a public SAN, the SAN is dedicated for host attachment, storage system attachment, and SAN Volume Controller Metro Mirror or Global Mirror.

Important: The following conditions require the use of private SAN:

- The connection between SAN Volume Controller nodes must ensure a minimum bandwidth of 4 Gbps or 2 times the peek host write I/O workload, whichever is higher. To avoid performance bottlenecks, each SAN Volume Controller node requires two ports that are operating at full speed (for example, 2 x 8 Gbps) worth of bandwidth to other nodes within the same cluster.
- Fabric topology changes can interrupt cluster communication for too long. In large fabrics, it is possible to take longer time for completion of topology changes. Private and dedicated (small) fabric with one hop is the requirement because the time for completion of topology changes is short.
- ISL congestion can affect cluster communication. One non-shared ISL per I/O group per fabric is needed to prevent congestion.

1.5.2 Architecture and solution context

The SAN architecture and solution is shown in Figure 1-27 on page 52. In this configuration, FC IP links are used between failure domains.

Advantage

The advantage of this architecture and solution is that it uses existing IP networks for extended distance connectivity.

Requirements

The architecture and solution include the following requirements:

- At least two FCIP tunnels between failure domains.
- Private SAN and public SAN.
- Third failure domain for quorum disk placement.

- Failure domain 3 (quorum disk) must be FC or FCIP attached. If FCIP attached is used, the response time to the quorum disk cannot exceed 80 ms.
- Storage controllers that contain quorum disks must be FC or FCIP attached.
- A guaranteed minimum bandwidth of 2 MBps is required for node-to-quorum traffic.
- No more than one ISL hop is supported for connectivity between failure domains.

Zoning requirements

The SAN Volume Controller Stretched Cluster FCIP configuration requires private and public SANs. The two SANs must be configured according to the following rules:

- Two ports of each SAN Volume Controller node are attached to the public SANs.
- Two ports of each SAN Volume Controller node are attached to the private SANs.
- A single trunk between switches is required for the private SAN.
- Hosts and storage systems are attached to fabrics of the public SANs. Links that are used for SAN Volume Controller Metro Mirror or Global Mirror must be attached to the public SANs.
- Failure Domain 3 (quorum disk) must be attached to the public SAN.
- ISLs belonging to the private SANs must not be shared with other traffic and must not be over-subscribed.

Figure 1-27 SAN architecture and solution

For more information, see this website:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

Important: Depending on the SAN Volume Controller SAN approach (with or without ISLs for node-to-node communication), other buffer-to-buffer credits should be assigned to all long distance F_Ports (node-to-node communication without ISLs) or to all long distance E_Ports (node-to-node communication with ISLs). For more information about the configuration of the Buffer to Buffer credits, see Chapter 3, “SAN Volume Controller Implementation” on page 149.

1.6 Power 750 Express and Hardware Management Console

This section provides an overview of the IBM Power Systems Power 750 Express servers and Hardware Management Console (HMC) along with the specific architecture and solution context for the solution in this publication. The IBM SAN Volume Controller Split Cluster solution in this book is on Live Partition Mobility (LPM) and uses two Power 750 Express servers (8233-E8B), one server for each data center site.

1.6.1 IBM Power 750 (8233-E8B) Express Server

The IBM Power 750 server delivers the outstanding performance of the POWER7 processor. The performance, capacity, energy efficiency, and virtualization capabilities of the Power 750 Express makes it an ideal consolidation, database, or multi-application server. It is “fit for purpose” for consolidation of UNIX, IBM i, and x86 Linux workloads and virtualized application servers and for medium to large database server hosting.

As a consolidation or highly virtualized multi-application server, the Power 750 Express offers tremendous configuration flexibility to meet the most demanding capacity and growth requirements. Realize the full capability of the system by using industrial-strength PowerVM virtualization for AIX, IBM i, and Linux. PowerVM offers the capability to dynamically adjust system resources based on workload demands so that each partition gets the resources it needs. Active Memory Expansion is a new POWER7 technology that enables the effective maximum memory capacity to be much larger than the true physical memory. Innovative compression and decompression of memory content can enable memory expansion up to 100%. This can enable a partition to do significantly more work or enable a server to run more partitions with the same physical amount of memory.

For database servers, the performance of the POWER7 processor makes it possible for applications to run faster with fewer processors, which result in lower per core software licensing costs. The POWER7 processor contains innovative technologies that maximize performance based on client workloads and computing needs. Intelligent Threads technology enables workload optimization by selecting the most suitable threading mode: Single Thread (per core) or Simultaneous Multi-thread-2 or 4 modes. Consequently, Intelligent Threads technology can provide improved application performance. The Power 750 server can deliver up to 128 simultaneous compute threads. In addition, POWER7 processors can maximize cache access to cores, which improves performance by using Intelligent Cache technology. These capabilities are designed to satisfy even the most demanding processing environments and can deliver business advantages and higher client satisfaction.

Figure 1-28 shows the Power 750 Express.

Figure 1-28 Power 750 Express (8233-E8B)

The Power 750 Express is a one- to four-socket server that supports up to 32 cores with outstanding energy efficiency and diagnostic features in a 4 U (EIA Units) package. The Power 750 is an ENERGY STAR qualified server that includes features to help clients become more energy efficient. IBM Systems Director Active Energy Manager™ makes use of EnergyScale™ technology, which enables Intelligent Energy management features to dramatically and dynamically conserve power. These Intelligent Energy features enable the POWER7 processor to operate at a higher frequency if environmental conditions permit for increased performance and performance per watt. Alternatively, they can operate at a reduced frequency if user settings permit for significant energy savings. The Power 750 also implements Light Path diagnostics, which provide an obvious and intuitive means to positively identify failing components.

Figure 1-29 on page 55 provides a summary of the high- level Power 750 configuration options.

Power 750 Express at a glance	
Configuration options	
POWER7 processor modules – one per processor card	4-core 3.7 GHz or 6-core 3.7 GHz or 8-core 3.2 GHz or 8-core 3.6 GHz
Sockets	1 to 4
Level 2 (L2) cache	256 KB per core
Level 3 (L3) cache	4 MB per core
Memory	8 GB[1] to 512 GB of RDIMM DDR3 Active Memory Expansion
Solid State Drives (SSD)	Up to eight SFF drives
Disk drives	Up to eight SFF SAS drives
Disk capacity	Up to 2.4 TB
Media bays	Slimline for DVD-RAM Half height for tape drive or removable disk
PCI Adapter slots	Two PCI-X 2.0; Three PCI Express 8x

Figure 1-29 Power 750 Express configuration options

For more information, see the following resources:

- The IBM Power 750 Express server website:

  http://www-03.ibm.com/systems/power/hardware/750/index.html

- *IBM Redbooks Power 750 and 755 Technical Overview and Introduction*, REDP-4638-00, which is available at this website:

  http://www.redbooks.ibm.com/redpapers/pdfs/redp4638.pdf

1.6.2 IBM Hardware Management Console

The HMC is a dedicated workstation that provides a GUI for configuring, operating, and performing basic system tasks for the POWER5, POWER5+, POWER6®, POWER6+™, and POWER7 processor-based systems that function in non-partitioned, partitioned, or clustered environments. The HMC is part of the underlying infrastructure for Power Systems Management. In addition, the HMC is used to configure and manage partitions. One HMC can control multiple POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor-based systems. At the time of this writing, one HMC supports up to 1000 LPARs by using the HMC machine code Version 7 Release 710. It can also support up to 48 Power 750 or 755 servers. For updates of the machine code and HMC functions and hardware prerequisites, see this website:

https://www14.software.ibm.com/webapp/set2/sas/f/hmc/home.htm

With the HMC, a system administrator can perform logical partitioning functions, service functions, and various system management functions by using the browser-based user interface or the command-line interface (CLI). The HMC uses its connection to one or more systems (referred to as *managed systems*) to perform the functions that are described in the next section.

HMC functions

The HMC provides three groups of functions: server, virtualization, and HMC management.

Server management

The first group contains the following functions that are related to the management of the physical servers that are under the control of the HMC:

- System password
- Status bar
- Power On and Off
- Capacity on Demand
- Error management:
  - System Indicators
  - Error and event collection reporting
  - Dump collection reporting
  - Call Home
  - Customer notification
  - Hardware replacement (Guided Repair)
  - SNMP events
- Concurrent add and repair
- Redundant service processor
- Firmware updates
- Support for Power 750 and 755 Server

Virtualization management

The second group includes the following functions that are related to virtualization features, such as the partitions configuration or dynamic reconfiguration of resources:

- System plans
- System profiles
- Partitions (create, activate, shut down)
- Profiles
- Partition mobility
- Dynamic LAPR (DLPAR), including processors, memory, and I/O
- Custom groups
- Remove limit of 128 PowerVM Active Memory Sharing Partitions
- Increased limit of partitions that are managed by an HMC to 1024
- Active memory expansion

HMC management

The last group relates to the management of the HMC, its maintenance, security, or configuration, including the following examples:

- Set up wizard
- User management:
  - User IDs
  - Authorization levels
  - Customizable authorization
- Disconnect and reconnect

- Network security:
  - Remote operation enable and disable
  - User-definable SSL certificates
- Console logging
- HMC redundancy
- Scheduled operations
- Back up and restore
- Updates and upgrades
- Customizable message of the day
- Increase Capacity on Demand billing capacity
- Ongoing HMC performance improvements

The HMC should be set up in a redundant configuration. Ideally, there should be two HMCs per managed server. Figure 1-30 shows a redundant HMC design and configuration with an HMC for remote operations, a Local HMC, and a second HMC for redundancy.

Figure 1-30 Local, remote, and redundant HMC design

Figure 1-31 shows the IBM 7042 C06/CR4 HMC.

Figure 1-31 IBM 7042 Hardware Management Console

HMC Version 7 uses a browser-based user interface. This interface uses a tree-style navigation model that provides hierarchical views of system resources and tasks by using drill down and launch-in-context techniques to enable direct access to hardware resources and task management capabilities. It provides views of system resources and tasks for system administration.

Figure 1-32 on page 59 shows the HMC GUI (Welcome: Main menu window) that we deployed for our solution in this book.

Figure 1-32 HMC Welcome window

Figure 1-33 shows one of the Power 750 servers and LPARs for this solution. This screen is accessed by clicking **Systems Management** → **Servers** → **Server-8233-E8B-SN0617** from the Systems Management menu of the HMC.

Figure 1-33 HMC Systems Management Servers window

For our solution, we implemented and provisioned four LPARs (WebSphere and DB2 workloads) and two VIOS servers per server, which is an IBM best practice for resiliency and redundancy.

For more information, see the following resources:

- The developerWorks public wikis:

  http://www.ibm.com/developerworks/wikis/display/virtualization/HMC
- *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491-01:

  http://www.redbooks.ibm.com/abstracts/sg247491.html
- IBM Information Center for Power Systems:

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp
- *IBM Hardware Management Console Best Practices*:

  http://www-304.ibm.com/webapp/set2/sas/f/best/hmc_best_practices.pdf

1.7 Architecture and solution context

This section describes the architecture and solution context.

1.7.1 Power 750 system architecture

The Power 750 is one of the IBM pSeries systems to feature the latest processor technology (POWER7) and designed for virtualized consolidation. This system uses the latest in multi-core design, intelligent multi-threading design, and large bandwidths for memory and I/O interconnects. POWER7 can deliver three times the performance per watt while using 20% less energy. Increased consolidation capability drives higher server usage and more energy savings. The overall solution design includes continuous availability with PowerVM Live Partition Mobility for planned system downtime without application disruption.

For more information about the overall Power 750 system architecture, see *IBM Power 750 and 755 (8233-E8B, 8236-E8C) Technical Overview and Introduction*, REDP-4638, which is available at this website:

http://www.redbooks.ibm.com/abstracts/redp4638.html?Open&pdfbookmark

This publication expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power 750 (and Power 755) systems. It is strongly recommended as foundational reference material for the Power 750 server architecture and technology in this Redbooks solution.

The goal of the I*BM Power 750 and 755 Technical Overview and Introduction* Redpaper™ is to introduce the following major innovative Power 750 and 755 offerings and their prominent functions:

- The POWER7 processor that is available at frequencies of 3.0 GHz, 3.3 GHz, and 3.55 GHz.
- The specialized POWER7 Level 3 cache that provides greater bandwidth, capacity, and reliability.
- The 1 Gb or 10 Gb Integrated Virtual Ethernet adapter that is included with each server configuration and provides native hardware virtualization.
- PowerVM virtualization including PowerVM Live Partition Mobility and PowerVM Active Memory Sharing.
- Active Memory Expansion that provides more usable memory than what is installed on the system.
- EnergyScale technology that provides features such as power trending, power-saving, capping of power, and thermal measurement.

This IBM Redpaper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the 750 and 755 systems. It is intended as another source of information that, with existing sources, can be used to enhance your knowledge of IBM server solutions.

1.7.2 Power 750 and HMC

POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor technology-based servers that are managed by an HMC require Ethernet connectivity between the HMC and the server Service processor. In addition, if dynamic LPAR, Live Partition Mobility, or PowerVM Active Memory Sharing operations are required on the managed partitions, Ethernet connectivity is needed between these partitions and the HMC. A minimum of two Ethernet ports are needed on the HMC to provide such connectivity. The rack-mounted 7042-CR5 HMC default configuration provides four Ethernet ports. The deskside 7042-C07 HMC standard configuration offers only one Ethernet port. Be sure to order an optional PCIe adapter to provide more Ethernet ports.

For any logical partition in a server, it is possible to use a Shared Ethernet Adapter set in Virtual I/O Server or Logical Ports of the Integrated Virtual Ethernet card for a unique or fewer connections from the HMC to partitions. Therefore, a partition does not require its own physical adapter to communicate with an HMC.

A good practice is to connect the HMC to the first HMC port on the server, which is labeled as HMC Port 1, although other network configurations are possible. You can attach a second HMC to HMC Port 2 of the server for redundancy (or vice versa). Figure 1-34 shows a simple network configuration to enable the connection from HMC to server and to enable Dynamic LPAR operations.

For more information about HMC and the possible network connections, see *Power Systems HMC Implementation and Usage Guide*, SG24-7491.

Figure 1-34 HMC to service processor and LPARs network connection

The default mechanism for allocation of the IP addresses for the service processor HMC ports is dynamic. The HMC can be configured as a DHCP server, which provides the IP address at the time the managed server is powered on. In this case, the service processor is allocated an IP address from a set of address ranges predefined in the HMC software. These predefined ranges are identical for version 710 of the HMC code and for previous versions.

If the service processor of the managed server does not receive DHCP reply before timeout, predefined IP addresses are set up on both ports. Static IP address allocation also is an option. You also can configure the IP address of the service processor ports with a static IP address by using the Advanced System Management Interface (ASMI) menus.

Important: The service processor is used to monitor and manage the system hardware resources and devices. The service processor offers two Ethernet 10/100 Mbps ports as connections. Note the following information:

- Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI options from a client web browser by using the HTTP server integrated into the service processor internal operating system.
- When not configured otherwise (DHCP or from a previous ASMI setting), both Ethernet ports of the first service processor have the following predefined IP addresses:
  - Service processor Eth0 or HMC1 port is configured as `169.254.2.147` with netmask `255.255.255.0`
  - Service processor Eth1 or HMC2 port is configured as `169.254.3.147` with netmask `255.255.255.0`

HA by using the HMC

The HMC is an important hardware component. When in use, POWER7 processor-based servers and their hosted partitions can continue to operate when no HMC is available. However, in such conditions, some operations cannot be performed, such as a DLPAR reconfiguration, a partition migration that uses PowerVM Live Partition Mobility or the creation of a new partition. Therefore, you can install two HMCs in a redundant configuration so that one HMC is always operational, even when maintenance is performed on the other HMC, for example.

If redundant HMC function is wanted, the servers can be attached to two separate HMCs to address availability requirements. Both HMCs must have the same level of Hardware Management Console Licensed Machine Code Version 7 (#0962) to manage POWER7 processor-based servers or an environment with a mixture of POWER5, POWER5+, POWER6, POWER6+, and POWER7 processor-based servers. The HMCs provide a locking mechanism so that only one HMC at a time has write access to the service processor. Depending on your environment, you have multiple options to configure the network.

Figure 1-35 on page 64 shows one possible HA HMC configuration managing two servers. These servers have only one CEC and therefore only one service processor. Each HMC is connected to one service processor port of all managed servers. This is the architectural design for the HMC solution component for the Power servers for the SAN Volume Controller Stretched Cluster solution in this book.

Figure 1-35 HA HMC and network architecture

In Figure 1-35, only hardware management networks (LAN1 and LAN2) are highly available to keep simplicity. However, management network (LAN3) can be made highly available by using a similar concept and adding more Ethernet adapters to LPARs and HMCs. Both HMCs must be on a separate VLAN to protect from a network failure. Each HMC can be on a DHCP server for its VLAN.

In a configuration with multiple systems or HMCs, the customer must provide switches or hubs to connect each HMC to the server service processor Ethernet ports in each system. One HMC should connect to the port labeled as HMC Port 1 on the first two CEC drawers of each system; a second HMC should be attached to HMC Port 2 on the first two CEC drawers of each system. This configuration provides redundancy for the HMCs and the service processors.

The solution in this book requires two IBM Power 750 Express servers (8233-E8B); one server for each data center site (Site A and Site B to simulate Production and DR). There are two HMCs implemented for redundancy.

Figure 1-36 on page 65 depicts the architecture for the Power 750 servers, and the VLAN connectivity between the servers and the HMCs, between the HMCs, and between the three required VLANs and the servers and the HMCs.

Figure 1-36 IBM HMC, VLAN and Power 750 Architecture

There are three required VLANs for our solution architecture for this book, as seen in Figure 1-36. The following design also is a recommended HA design and configuration for PowerVM connectivity to the HMCs:

- HMC Management VLAN (Green): This is a required connection between the HMC and the Service processor of each Power 750 server. There are two HMCs, which are redundant of each other. Each HMC can manage each server, and to support this there is connectivity in place via HMC Management VLAN A,\ and HMC Management VLAN B, as seen in Figure 1-36 for System Administrator tasks via these Management VLAN connections. Each HMC has two Ethernet connections. There also is a direct connection that is required from each HMC to each Power 750 server's Service processor.
- Mover Service processor Network VLAN (Red): This MSP network is required for Live Partition Mobility (LPM) movement of LPARs between the two Power 750 servers. This VLAN is on the private network so it does not interfere with public network traffic. It is recommended that the traffic for LPM must be isolated on its own private MSP VLAN network (or subnet).
- VIOS Management/DLPAR: This connection is required between the HMCs, individual LPARs, and VIOS servers for each Power 750 server. The System Administrator requires these connections via the HMC for LPAR creation and management, and VIOS LPAR Server support and maintenance.

1.8 PowerVM technology

This section provides an overview of the PowerVM architecture and technology as it applies to the SAN Volume Controller Stretched Cluster configuration and solution context in this book. It assumes a base understanding of PowerVM architecture.

1.9 PowerVM

By using Power Systems with optional PowerVM virtualization offerings, businesses can consolidate servers and applications, virtualize system resources, and provide a more flexible, dynamic IT infrastructure. PowerVM virtualization technology is a combination of hardware and software that supports and manages virtual environments on IBM POWER5, POWER5+, POWER6, and POWER7 processor-based systems. These systems are available on IBM Power Systems and IBM BladeCenter® servers as optional editions, and are supported by the IBM AIX, IBM i, and Linux operating systems. With this set of comprehensive systems technologies and services, you can aggregate and manage resources with a consolidated, logical view.

By deploying PowerVM virtualization and IBM Power Systems, you can make use of the following benefits:

- Lower energy costs through server consolidation.
- Reduced cost of your existing infrastructure.
- Better management of the growth, complexity, and risk of your infrastructure.
- Deliver services with superior economics by consolidating virtualized workloads.
- Deliver services faster by automating deployment of virtual machines and storage.
- Optimize usage of server and storage resources to control cost and boost ROI.
- Scale your virtualized deployments without paying underlying performance penalties.
- Eliminate scheduled downtime by deploying live mobility between servers.
- Deliver higher quality services by improving virtual resource management.

Complexity can work its way into any IT infrastructure, which is driven by the rollout of new applications and unanticipated change. However, adding servers in response to each demand for new workloads drives the need for more data center space, power, cooling, network cabling, data storage and administrative resources. Such complexity leads to inefficiency. The answer is virtualization, which allows organizations to consolidate multiple operating systems and software stacks on a single platform.

PowerVM provides the industrial-strength virtualization solution for IBM Power Systems servers and blades. Based on more than a decade of evolution and innovation, PowerVM represents the state of the art in enterprise virtualization and is broadly deployed in production environments worldwide by most Power Systems owners.

The Power Systems family of servers includes proven[14] workload consolidation platforms that help clients control costs while improving overall performance, availability, and energy efficiency. With these servers and IBM PowerVM virtualization solutions, an organization can consolidate large numbers of applications and servers, fully virtualize its system resources, and provide a more flexible, dynamic IT infrastructure. Therefore, IBM Power Systems with PowerVM deliver the benefits of virtualization without limits.

[14] PowerVM Case Studies: http://www.ibm.com/systems/power/success/index.html

PowerVM provides industrial-strength virtualization for AIX, IBM i, and Linux environments on IBM POWER® processor-based systems. By using IBM Power Systems servers that are integrated with PowerVM technology, clients can build a dynamic infrastructure that helps them to reduce costs, manage risk, and improve service levels.

PowerVM also offers a secure and resilient virtualization environment, which is built on the advanced reliability, availability and serviceability (RAS) features, extreme scalability, and leadership performance[15]of the IBM Power platform, which is based on the outstanding POWER7 processors.

The POWER Hypervisor™ enables the hardware to be divided into LPARs and ensures isolation between them. The hypervisor orchestrates and manages system virtualization, including creating LPARs and dynamically moving resources across multiple operating environments. The POWER Hypervisor also enforces partition security and can provide inter-partition communication that enables the VIOS virtual SCSI, virtual Fibre Channel, and virtual Ethernet functions.

PowerVM Editions deliver advanced virtualization functions for AIX, IBM i, and Linux clients such as IBM Micro-Partitioning® technology, Virtual I/O Server, Integrated Virtualization Manager, and Live Partition Mobility. PowerVM features provide the ability to virtualize processor, memory, and I/O resources to increase asset usage and reduce infrastructure costs. PowerVM also allows server resources to be dynamically adjusted to meet changing workload demands without a server shutdown.

PowerVM Editions provide Virtual I/O Server technology to facilitate consolidation of LAN and disk I/O resources and minimizes the number of required physical adapters in a Power system. The VIOS actually owns the resources that are shared with clients. A physical adapter that is assigned to the VIOS partition can be shared by one or more other partitions. The VIOS can use virtualized storage and network adapters, which makes use of the virtual SCSI, virtual Fibre Channel with NPIV, and virtual Ethernet facilities.

You can achieve continuous availability of virtual I/O by deploying multiple VIOS partitions (dual VIOS) on an HMC-managed system to provide highly available virtual services to client partitions.

Table 1-2 provides an overview of the key PowerVM features and benefits. The features with an asterisk (*) indicate features that were implemented for the solution in this book.

Table 1-2 PowerVM Features and Benefit

Feature	Benefit
PowerVM Hypervisor*	Supports multiple operating environments on a single system
Micro-Partitioning	Enables up to 20 VMs per processor core[a]
Dynamic Logical Partitioning	Processor, memory, and I/O resources can be dynamically moved between VMs
Shared Processor Pools	► Processor resources for a group of VMs can be capped, which reduces software license costs ► VMs can use shared (capped or uncapped) processor resources ► Processor resources can automatically move between VMs based on workload demands

[15] Power Systems benchmark results: http://www.ibm.com/systems/power/hardware/benchmarks/index.html

Shared Storage Pools	Storage resources for Power Systems servers and VIOS can be centralized in pools to optimize resource usage
Integrated Virtualization Manager	Simplifies VM creation and management for entry Power Systems servers and blades
Live Partition Mobility*	Live AIX, Linux and IBM i VMs can be moved between servers, which eliminates planned downtime
Active Memory Sharing	Intelligently flows memory from one VM to another for increased memory usage
Active Memory Deduplication	Reduces memory usage for Active Memory Sharing configurations by detecting and eliminating duplicate memory pages
NPIV, MPIO, and SDDPCM*	Simplifies the management and improves performance of Fibre Channel SAN environments Increases resiliency and load balancing in support SAN environments for Power Systems
System Planning Tool	Simplifies the planning for and installation of Power Systems servers with PowerVM
VIOS Performance Advisor	Checks performance and health of the VIO Server then makes recommendations for performance improvement.

a. Available on systems with 7.6 firmware and above

For more information about the implementation of this solution, see 2.6, “Software” on page 130, Chapter 3, “SAN Volume Controller Implementation” on page 149, and Chapter 4, “PowerVM and PowerHA Implementation” on page 213 for the implementation of this solution in this book.

For more information, see the following resources:

- *IBM PowerVM Getting Started Guide*, REDP-4815, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/redp4815.html
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/sg247940.html

1.9.1 Architecture and solution context

The challenges of IT revolve around variety, volume, and velocity. The variety of IT options provides the opportunity to bring various new types of information to the system that needs it. The volume of servers is growing exponentially. The velocity of change in an IT Organization is driving the need to implement the proper IT services. AIX OS infrastructures must be open, flexible, on demand, and responsive to change.

To align server usage costs with project requirements, a comprehensive set of virtual servers are defined that are relative to cost, performance, manageability, availability, scalability,

functionality and recoverability. The implementation of PowerVM AIX Virtualization with these features that is described in 1.7.1, “Power 750 system architecture” on page 61 is essential. To further optimize management of AIX Virtualization, automation is required to manage according to polices and minimize human error.

Figure 1-37 shows the PowerVM AIX Virtualization Architecture Model, which is useful to describe the PowerVM technology and to be used as a reference point for key terms throughout this book.

Figure 1-37 PowerVM AIX virtualization reference architectural model

Data access layer

This is where the applications access their data. There is no difference in this application access whether served by a physical or a virtualized server. The data access layer also can be referenced as a presentation layer. The presentation layer can come from the customer’s application premise or the public network (Internet via DMZ).

Application

Examples of the application include Internet Banking, Image workflow, Oracle, and SAP that are running and hosted within the LPAR.

Middleware

Examples of middleware include WebSphere Application Server, WebSphere MQ, and DB2, that is running within the operating system.

LPAR: Logical Partition

The operating system is running inside a virtual machine or LPAR (logical partition) environment.

Virtualization layer: Hypervisor

Virtualization is a technique that is used to decouple a physical configuration from the virtualized one. A virtualized entity provides flexibility and on-demand service attributes to changing business needs. Virtualized IT resources are grouped in resource pools to improve management and usage. Virtualization is a trigger for increased standardization, consolidation, automation, and infrastructure simplification. In PowerVM terminology, the Hypervisor manages the virtualization layer.

The Hypervisor and virtualization layer is shown in Figure 1-38.

Figure 1-38 Architecture of PowerVM Hypervisor

Physical host

The server box is virtualized to contain multiple logical partitions that are running with AIX operating systems. Multiple physical server boxes always are collected into a virtualized resource pool to enable the use of virtualization functions, such as partition mobility or zero planned outage, which cannot be achieved with a single physical box design.

Storage and backup and recovery

This is the SAN and backup and recovery (Tivoli Storage Manager, Network Interface Manager, and so on) infrastructure that a virtualized server environment often is integrated with as a part of Virtualized Management Environment.

Enterprise data center network

This network includes remote access, shared switches, and load balancer components.

HMC

The HMC manages the physical and logical servers in the PowerVM AIX Virtualization environment.

System management

This helps to manage the system efficiently and effectively via automation and monitoring.

Figure 1-39 shows the design of the Power 750 servers for the solution in this book, which includes the LPARs, VIOS Servers, and System Software. This is the server for Site A; the Site B server is a mirror configuration for our solution for HA and to support PowerHA and LPM functionality. IBM WebSphere is hosted in LPARs 1, and 3 and DB2 in LPARs 2, and 4 for resiliency. WebSphere and DB2 are installed and implemented to simulate real-life workloads. We installed a test financial trading application for workload testing the technologies for the solution in this book.

Figure 1-39 Power 750 logical architecture

1.9.2 Virtual I/O Server

The Virtual I/O Server is software that is in a Power Systems Server LPAR. This software facilitates the sharing of physical I/O resources between client logical partitions within the server. The Virtual I/O Server provides virtual SCSI target, virtual Fibre Channel, Shared Ethernet Adapter, and PowerVM Active Memory Sharing capability to client logical partitions within the system. As a result, client logical partitions can share SCSI devices, Fibre Channel adapters, Ethernet adapters, and expand the amount of memory that is available to logical partitions by using paging space devices. The Virtual I/O Server software requires that the logical partition be dedicated solely for its use.

The Virtual I/O Server is part of the PowerVM Editions hardware feature. It is required in our solution in this book because the solution is using LPM.

By using the Virtual I/O Server, the following functions are available:

- Sharing of physical resources between logical partitions on the system.
- Creating logical partitions without requiring more physical I/O resources.
- Creating more logical partitions than there are I/O slots or physical devices available with the ability for logical partitions to have dedicated I/O, virtual I/O, or both.
- Maximizing the use of physical resources on the system.
- Reducing the SAN infrastructure.

The Virtual I/O Server allows the sharing of physical resources between supported AIX, IBM i, and Linux partitions to allow more efficient usage and flexibility for the use of physical storage and network devices.

The Virtual I/O Server also provides functionality for features such as Active Memory Sharing or Suspend/Resume and is a prerequisite if you want to use IBM Systems Director VMControl™. When the PowerVM Standard Edition and PowerVM Enterprise Edition are used, dual Virtual I/O Servers must be deployed to provide maximum availability for client partitions when Virtual I/O Server maintenance is performed. For more information about dual VIOS, see Chapter 4, “PowerVM and PowerHA Implementation” on page 213.

Figure 1-40 shows a basic Virtual I/O Server configuration. This diagram shows only a small subset of the capabilities to show the basic concept of how the Virtual I/O Server works. The physical resources, such as the physical Ethernet adapter and the physical disk adapter, are accessed by the client partition by using virtual I/O devices.

Figure 1-40 Virtual I/O Server configuration

For more information about VIOS, see the following resources:

- *IBM PowerVM Getting Started Guide*, REDP-4815-00, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/redp4815.html

- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-05, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/sg247940.html

1.9.3 Power Systems network virtualization

Power Systems and PowerVM architecture recommends and uses network virtualization from a VIOS and connectivity and adapter perspective.

PowerVM technology includes the following virtual device requirements:

- Virtual devices that do not require a Virtual I/O Server:
  - Virtual Ethernet
  - Partitions on the same system communicate without the use of physical Ethernet adapters

  - Virtual Ethernet adapter is assigned to LPARs
  - POWER Hypervisor enables internal virtual switch and virtual Ethernet protocol
- Virtual devices that do require a Virtual I/O Server:
  - SEA (also requires a physical Ethernet adapter)
  - Layer 2 bridge function connects internal virtual Ethernet with external physical network
  - VSCSI (also requires a physical disk attachment adapter)
  - Virtual I/O Server partition uses files, logical volumes, whole physical disks, or physical optical devices to provide the backing storage for VSCSI devices presented to client partitions
  - Client sees a hdisk# or cd#
- Hardware-assisted virtual device:
  - IVE adapter
  - Partitions can communicate internally and with external physical network without going through virtual
- Ethernet and SEA:
  - This is equivalent to a physical switch, direct communication from LPAR with no hypervisor interaction.
  - This is not supported for Live Partition Mobility

VLAN: Tagged and untagged ports

PowerVM uses a Layer-2 implementation of the Open Systems Interconnection (OSI) model to create logical networks within switches. A single switch can support multiple VLANs, and VLANs can span across switches. The PowerVM Hypervisor implements VLAN based on IEEE 802.1Q VLAN implementation, and can support 4094 VLANs per server. The maximum transmission unit (MTU) size is 65394.

A port on a VLAN-capable switch has a default port virtual LAN ID (PVID) that indicates the default VLAN to which the port belongs. The switch adds the PVID tag to untagged packets that are received by that port from the host. In addition to a PVID, a port might belong to other VLANs and have those VLAN IDs assigned to it that indicate the other VLANs to which the port belongs. A switch port with a PVID only is called an *untagged port*. Untagged ports are used to connect VLAN-unaware hosts.

A port with PVID and other VIDs is called a *tagged port*. Tagged ports are used to connect VLAN-aware hosts. VLAN-aware means that the host is IEEE 802.1Q compatible and can manipulate VLAN tags, and thus can interpret them, add them, and remove them from Ethernet frames.

Figure 1-41 provides an overview of a VLAN from a tagged and untagged port view.

Figure 1-41 VLAN tagged and untagged ports

Virtual Ethernet

Virtual Ethernet provides inter-partition communication facility without the need for a physical adapter. Configuring a virtual Ethernet adapter results in configuring a port of virtual switch of the hypervisor. The hypervisor copies packets between the LPARs memory. Up to 256 virtual Ethernet adapters per partition are supported. One virtual Ethernet adapter can provide access to 21 VLANs (including PVID).

Power Hypervisor Ethernet switch

The POWER Hypervisor is included with all POWER5, POWER6, and POWER7 processor-based systems and includes the following features:

- Implements a software Ethernet switch and virtual Ethernet protocol
- Is consistent with IEEE 802.1Q
- Supports up to 4094 networks (VLAN IDs)
- Uses PVID tag for untagged packets
- Copies packets between partitions (memory-to-memory)

Multiple VSwitches in IBM Power Systems can include the following definitions:

- One VSwitch (VSwitch0) is defined by default
- VSwitches are defined at system level (hypervisor)
- VLANs cannot communicate through different VSwitches.
- VSwitch can be created from HMC GUI.
- By creating a virtual Ethernet adapter and choose VSwitch in the menu

Shared Ethernet Adapter

The Shared Ethernet Adapter (SEA) connects virtual network to physical network through the Virtual I/O Server. The SEA provides a Layer-2 bridge functionality to transport network traffic between a virtual and physical network. The configuration changes in the SEA do not require changes in the client.

SEA redundancy works in active/passive configuration, and different priorities should be set for Virtual Ethernet trunk adapters. A control channel (separate VLAN) must be created to exchange heartbeat packets The control channel adapter should be placed on a dedicated VLAN that is not used for anything else. Finally, the IP address should be set to ping on the SEA.

Figure 1-42 shows an overview of the SEA in a failover configuration for HA.

Figure 1-42 Shared Ethernet Adapter failover configuration

Integrated Virtual Ethernet

Integrated virtual Ethernet (IVE) offers an alternative to the use of virtual Ethernet or the SEA service of the Virtual I/O Server. It is supported only on platforms that are entry, midrange platforms that are POWER6 or above

IVE refers to the collective name of a number of technologies, including hardware, Host Ethernet Adapter (HEA), software components, and Power Hypervisor functions.

IVE offers greater throughput and lower latency than PCIe or PCI-X bus that are connected Ethernet adapters because it is connected to the GX+ bus of the Power System server. IVE also provides special hardware features that enable logical Ethernet adapters, which can communicate directly with logical partitions that are running on the system.

Figure 1-43 shows a comparison between an SEA and IVE design.

Figure 1-43 VIOS SEA versus IVE design

Table 1-3 shows a functional comparison between IVE and VIOS (SEA), which is helpful for determining the best fit solution based on your requirements. For the solution in this book, we implemented VIOS (SEA).

Table 1-3 IVE versus VIOS (SEA) functional comparison

Virtualized function	IVE	VIOS
Basic Functions		
Requires activation of the Advanced POWER Virtualization feature code		X
Virtualize Ethernet resources over a single server	X	X
Provides hardware acceleration	X	
Managed by the HMC as a hardware device	X	
Scalable and flexible for all virtualized requirements		X
Operating system-based; runs in an LPAR		X
Enables advanced server functions, such as LPM		X
Strategic solution for IBM Systems client requirements	X	X
Network Functions		
Can be used as a SEA	X	X
Provides scalable threading for multi-core systems	X	
Provides 10 Gbps wire speed performance with very low latency	X	
Provides secure network connections between LPARs	X	X
Supports industry standard networking attributes, such as VLAN, MTU, and so on	X	X
Does not require a PCI-x or PCIe slot for external communication	X	

For more information on network virtualization, see the following resources:

- PowerVM Information Roadmap:

  http://publib.boulder.ibm.com/eserver/roadmap_powervm.html

- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590-00, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/sg247590.html

- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-00, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/sg247940.html?Open

- *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/redp4340.html?Open

1.10 Live Partition Mobility

This chapter section provides an overview of the IBM Live Partition Mobility (LPM) architecture as it applies to the SAN Volume Controller Split Cluster configuration and solution context in this book.

1.10.1 Overview and concepts

LPM is the next step in the IBM Power Systems virtualization continuum. It can be combined with other virtualization technologies, such as logical partitions, Live Workload Partitions, and the SAN Volume Controller to provide a fully virtualized computing platform that offers the degree of system and infrastructure flexibility that is required by today's production data centers.

A chargeable feature of IBM POWER6 AND POWER7 processor-based servers and available since 2007, the LPM technology provides the capability to dynamically migrate LPARs t from one system to another across or within data centers without interruption to the applications. In concept, it is similar to VMware VMotion. LPM is used to avoid outages for planned server maintenance, for load balancing across multiple servers, and for energy conservation.

LPM uses a simple and automated procedure that transfers the configuration from source to destination without disrupting the hosted applications or the setup of the operating system and applications. LPM provides the administrator greater control over the usage of resources in the data center. It allows a level of reconfiguration that in the past was not possible because of complexity or SLAs that do not allow an application to be stopped for an architectural change.

A partition migration operation can occur when a partition is powered off (inactive), or when a partition is providing service (active). During an active partition migration, there is no disruption of system operation or user service. For example, a partition that is hosting a live production database with normal user activities can be migrated to a second system with no loss of data, no loss of connectivity, and no effect on running transactions. A logical partition can be migrated between two POWER6 or POWER7 based systems, provided the destination system has enough resources to host the partition.

There is no restriction on processing units or partition memory size for inactive or active migration.

1.10.2 LPM prerequisites and requirements

In this section, we describe the necessary hardware infrastructure and other prerequisites that are required to use LPM.

Hardware infrastructure

The migration of a logical partition includes the following primary requirements:

- Two POWER6 (at least POWER6 level; for our solution we are on POWER7) based systems that are controlled by the same HMC. An optional redundant HMC configuration is supported.
- The destination system must have enough CPU and memory resources to host the mobile partition (the partition profile that is running, as alternative production profiles might exist).
- The operating system, applications, and data of the mobile partition must be on virtual storage on an external storage subsystem.

- No physical adapters can be used by the mobile partition.
- The mobile partition's network and disk access must be virtualized by using one or more Virtual I/O Servers.

  The Virtual I/O Servers on both systems must have a shared Ethernet adapter that is configured to bridge to the same Ethernet network that is used by the mobile partition.

  The Virtual I/O Servers on both systems must be capable of providing virtual access to all disk resources the mobile partition is using.
- LPM requires a specific hardware and microcode configuration that is currently available on POWER6 or 7 based systems only. For the best interoperability and compatibility with the solution software and hardware stack, firmware level AL730_099 was implemented.
- The procedure that performs the migration identifies the resource configuration of the mobile partition on the source system and then reconfigures the source and destination systems accordingly. Because the focal point of hardware configuration is the HMC, it was enhanced to coordinate the process of migrating partitions. The same HMC must be used to manage both systems.
- The mobile partition's configuration is not changed during the migration. The destination system must be able to host the mobile partition and must have enough free processor and memory resources to satisfy the partition's requirements before migration is started. There is no limitation on the size of the mobile partition; it can even use all resources of the source system offered by the Virtual I/O Server.
- The operating system and application data must be on external disks of the source system because the mobile partition's disk data must be available after the migration to the destination system is completed. An external, shared access storage subsystem is required.
- The mobile partition must not own any physical adapters and must use the Virtual I/O Server for network and external disk access. The migration of partitions by using multiple Virtual I/O Servers is supported.
- Because the mobile partition's external disk space must be available to the Virtual I/O Servers on the source and destination systems, you cannot use storage pools. Each Virtual I/O Server must create virtual target devices by using physical disks and not logical volumes.
- Virtual network connectivity must be established before the partition migration task is started while virtual disk setup is performed by the migration process.
- The source and target system must have an appropriate shared Ethernet adapter environment to host a moving partition. All virtual networks in use by the mobile partition on the source system must be available as virtual networks on the destination system.
- VLANs that are defined by PVIDs on the VIOS have no meaning outside of an individual server because all packets are bridged untagged. It is possible for VLAN 1 on CEC 1 to be part of the 192.168.1 network while VLAN 1 on CEC 2 is part of the 10.1.1 network. Because two networks are possible, you cannot see if VLAN 1 exists on both servers. You must check if VLAN 1 maps to the same network on both servers.

1.10.3 Architecture and solution context

Figure 1-44 on page 80 shows a basic hardware infrastructure that is enabled for LPM. Each Power based system is configured with a Virtual I/O Server partition. The mobile partition has virtual access only to network and disk resources. The Virtual I/O Server on the destination system is connected to the same network and is configured to access the same disk space that is used by the mobile partition. For illustration purposes, the device numbers are all shown as zero; but in practice, they vary considerably.

Figure 1-44 Hardware Infrastructure enabled for LPM

LPM prerequisites

LPM requires coordinated movement of a partition's state and resources. Migratable partitions move between capable, compatible, and ready systems.

A single HMC can control several concurrent migrations. There are no architectural restrictions on the number of migrations that can be underway at any one time. However, a single mover service partition can handle a maximum of four simultaneous active migrations. It is possible to have several mover service partitions on a system. In practice, the maximum number of concurrent migrations is limited by the processing capacity of the HMC and contention for HMC locks.

The first step of any mobility operation is to validate the capability and compatibility of the source and destination systems. LPM includes the following high-level prerequisites:

- A migration-capable ready source system.
- A migration-capable ready destination system.
- Compatibility between the source and destination systems.
- The source and destination systems must be under the control of the same HMC.
- A migratable, ready partition to be moved from the source system to the destination system. For an inactive migration, the partition must be powered down, but must be capable of booting on the destination system.
- For active migrations, a mover service partition on the source and destination systems.
- One or more SANs that provide connectivity to all of the mobile partition's disks to the Virtual I/O Server partitions on the source and destination servers. The mobile partition accesses all migratable disks through virtual SCSI devices. The LUNs must be zoned and masked to the Virtual I/O Servers on both systems. SCSI reservation must be disabled.
- The mobile partition's virtual disks must be mapped to LUNs; they cannot be part of a storage pool or logical volume on the Virtual I/O Server.

- One or more physical IP networks (LAN) that provide the necessary network connectivity for the mobile partition through the Virtual I/O Server partitions on the source and destination servers. The mobile partition accesses all migratable network interfaces through virtual Ethernet devices.
- An Rational® Method Composer connection to manage inter-system communication.

If any of these elements are missing, a migration cannot occur.

Figure 1-45 shows the LPM and PowerVM architecture for the SAN Volume Controller Stretched Cluster solution in this book.

Figure 1-45 LPM architecture in SAN Volume Controller Stretched Cluster solution

The arrows indicate the direction of the LPAR migration. Notice the VIOS server re-assignment to the P750 Site B server. The Site B server is designed and configured identically to the Site A P750 server. In addition, the network connectivity to each P750 server is identical on Site A and Site B.

For more information about the implementation of LPM in our solution, see 4.11, “LPM” on page 262.

For more information, see the following resources:

- *IBM PowerVM Live Partition Mobility*, SG24-7460-01, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/sg247460.html
- Basic understanding and troubleshooting of LPM:

  http://www.ibm.com/developerworks/aix/library/au-LPM_troubleshooting/index.html

1.11 N_Port ID Virtualization

This section provides an overview of the IBM NPIV architecture as it applies to the SAN Volume Controller Stretched Cluster configuration and solution context in this book. It assumes a base understanding of storage architecture in general.

1.11.1 Overview and concepts

NPIV is a technology that defines how multiple virtual servers can share a single physical Fibre Channel port ID. NPIV is a Fibre Channel industry standard method for virtualizing a physical Fibre Channel port.

NPIV allows a single host bus adapter (HBA) or target port on a storage array to register multiple worldwide port names (WWPNs) and N_Port ID numbers. NPIV allows one F_Port to be associated with multiple N_Port IDs, so a physical Fibre Channel HBA can be shared across multiple guest operating systems in a virtual environment. This configuration allows each virtual server to present a different worldwide name to the SAN, which, in turn, means that each virtual server sees its own storage but no other virtual server's storage. Each partition is identified by a pair of unique WWPNs, which enables you to connect each partition to independent physical storage on a SAN.

On IBM Power Systems, NPIV allows logical partitions to have dedicated N_Port IDs, which give the OS a unique identity to the SAN as if it had dedicated physical HBAs. Each virtual FC HBA has a unique and persistent identity. Each physical NPIV capable FC HBA can support 64 virtual ports.

Figure 1-46 on page 83 provides an overview of the client LPAR, with connectivity from the client virtual Fibre Channel adapter through the Power Hypervisor to the virtual Fibre Channel adapter in the VIOS server through the SAN to the physical storage connections.

Figure 1-46 NPIV overview

Unlike virtual SCSI, only the client partitions see the disk. The Virtual I/O Server acts only as a pass-through managing the data transfer through the POWER Hypervisor. However, based upon your requirements, VIOS can support both vSCSI and NPIV if needed.

NPIV shifts the administrative workload from the AIX or Power Systems Administrator from a VIOS management perspective to the SAN administrator by easing their role. NPIV provides the following key benefits:

- NPIV allows storage administrators use existing tools and techniques for storage management.
- Solutions such as SAN managers, Copy Services, backup and restore, should work immediately.
- Improved storage provisioning and ease-of-use.
- Easier Zoning and LUN masking.
- Better physical to virtual device compatibility.
- Improved management for tape libraries.
- SCSI-2 Reserve and Release and SCSI3 Persistent Reserve capabilities.
- Better support and alignment for clustered and distributed solutions.
- Load balancing (active-active).
- Solutions enablement (HA, Oracle, and so on).
- Large ecosystem support for storage, multipathing, applications, and monitoring, as so on.

Figure 1-47 shows a comparison between VSCSI and NPIV from an administrator perspective, and how VIOS is used in each design. This represents a minimum install of VIOS 2.1 as a starting point. On the storage virtualization side of Figure 1-47, there are more adapters and overhead from a technology perspective than with NPIV.

Figure 1-47 Storage virtualization versus NPIV

NPIV is supported at certain minimum hardware configurations and software levels.

For more information, see the latest Virtual I/O Server documentation and *PowerVM Migration from Physical to Virtual Storage*, SG24-7825-00, which is available at this website:

http://www.redbooks.ibm.com/redbooks/pdfs/sg247825.pdf

NPIV is supported in PowerVM Express, Standard, and Enterprise Editions. POWER6 and POWER7 processor-based servers require a minimum of one of the following components:

- 8 GB Fibre Channel adapter Feature Code 5735
- An NPIV capable SAN switch:
  - A Brocade SAN switch at firmware level 5.3.0 or later
  - A CISCO MDS 9000 SAN with the optional NPIV licensed feature installed
- NPIV is compatible and supported with LPM

Figure 1-48 shows the NPIV and LPM logical architecture, as implemented for the solution in this book.

Figure 1-48 NPIV and LPM

In addition, there are some restrictions on SAN switches and firmware levels. Check with your IBM support representative for the currently supported levels.

In the next section, we describe the NPIV architecture, Multipath I/O architecture, and the supporting Subsystem Device Driver Path Control Module (SDDPCM).

1.11.2 Architecture and solution context

Figure 1-49 on page 86 shows the differences between a VSCSI model and NPIV, from a high-level design and configuration perspective.

Figure 1-49 VSCI Model Compared to NPIV

To enable NPIV, assign the physical NPIV-capable Fibre Channel adapter to a Virtual I/O Server and connect virtual Fibre Channel adapters on the client partition to virtual Fibre Channel adapters on the Virtual I/O Server.

For the solution in our environment, we are using VSCI and NPIV to allow for multipathing, which is described in the next sections.

Figure 1-50 on page 87 shows how this is designed and configured from a VIOS Block diagram perspective.

Figure 1-50 VIOS Block Diagram VSCSI and NPIV

1.11.3 SDDPCM

Complex SAN configurations are prevalent in computer system configurations. SANs enable large numbers of servers to access common storage via an interconnection of switches and cabling. The availability and integrity of this interconnection is critical to the reliable operation of the system, and as such networks are often implemented with redundant routes (paths) between the servers and storage. Such redundant networks, with intelligent multipath drivers installed on the servers, allows for failing commands to be recovered down alternative paths in the network thereby avoiding system outages because of individual path failures.

SDDPCM is a loadable path control module that supports the multipath configuration environment in the following devices:

- IBM TotalStorage Enterprise Storage Server®
- IBM System Storage SAN Volume Controller
- IBM Storwize V7000
- IBM TotalStorage Data Studio family
- IBM BladeCenter S SAS Raid Controller Module (RSSM)

When the supported devices are configured as MPIO-capable devices, SDDPCM is loaded and becomes part of the AIX MPIO FCP FCoE device driver or SAS device driver. The AIX MPIO device driver or SAS device driver with the SDDPCM module enhances the data availability and I/O load balancing.

SDDPCM manages the paths to provide the following benefits:

- HA and load balancing of storage I/O.
- Automatic path-failover protection.
- Concurrent download of Licensed Internal Code.
- Prevention of a single-point-failure that is caused by HBA, FC cable, Ethernet cable, or host-interface adapter on supported storage.

1.11.4 Legacy failure categories

Traditionally, network path failures are viewed as falling into one of two categories: *permanent* and *temporary*. Perhaps the most well understood and easiest to manage are the permanent failures, which result from a catastrophic failure of a network component. These failures are typically persistent failures where all commands routed to the failing path (or paths) fail. Commands are recovered through retry down alternative paths and the failing paths are taken offline. Paths that are taken offline as a result of permanent path failures remain offline because any subsequent polling that is performed to test path availability also fails, which prevents the paths from being automatically restored to service.

The second category of failure is temporary and transient in nature. These failures can arise from numerous sources, including bit flips in electronics because of alpha particle or cosmic rays, electrical noise from intermittent contacts, and microcode defects. These can produce temporary command failures that are recovered through path retry. These conditions might not cause paths to enter into an offline state; however, because these are temporary conditions, paths that were removed from service are brought back into service by the path reclamation function. Both of these traditional failure categories are generally handled well by existing multipath drivers that are available in the industry today and, in fact, rarely result in any adverse effects on the operation of the system.

1.11.5 Emerging failure categories

Unfortunately, as the line-speed, IOPS, and complexity of networks increased over time, a third category of failure emerged. These failures also are temporary like the second category. However, in addition to being temporary, they also are recurring at varying rates. This third category of path failure is called *temporary/recurring*, which is also referred to as the *sick but not dead* condition.

These failures can arise from marginal components or components and network routes that are insufficiently sized or over-subscribed for the volume of network traffic present. Often times these failures are provoked by secondary conditions, such as an instantaneous increase in network traffic or convergence of network traffic. These conditions can lead to congested ports and result in frames being dropped or discarded by switches and storage devices which, in turn, cause command timeouts to be encountered by the attached hosts.

These conditions tend to be more severe when encountered on ISL because of their potential to expand the number of hosts that are required to perform command timeout recovery. Multipath drivers historically cannot uniquely identify the temporary/recurring failure condition and therefore treat them the same as failures that fall into the permanent or temporary categories. Without specific detection and path management for this failure category, the condition is allowed to persist for an extended period and in many cases, the condition persists indefinitely or until manual intervention is performed.

This type of temporary/recurring failure condition drives recursive error recovery by the attached servers, which leads to symptoms ranging anywhere from moderate performance degradation to complete system outage. A common symptom that is seen is paths cycling between the online and offline states as the multipath driver takes paths offline for command failures that are later returned to service following the successful completion of periodic path health check command.

IBM has introduced new path management policies that are targeted at addressing the temporary/recurring failure condition. These policies enable new failure detection and path management algorithms that detect and respond to temporary/recurring failure conditions and thereby improve availability.

For the solution in this book, we implemented SDDPCM v2.6.3.2 to ensure interoperability and compatibility with the software stack and hardware, and for optimal response and handling of temporary/recurring path failures. SDDPCM 2.6.3.2 introduces a new function that further improves handling of temporary/recurring failure conditions and reduce the symptoms and associated performance degradation. This function includes a new device attribute timeout_policy which can be defined to enable new path management options.

The timeout_policy device attribute supports the following options:

- retry_path
- fail_path
- disable_path

The timeout_policy influences the path recovery behavior for paths that are set to FAILED state (with exception of last remaining path) because of I/O failures for timeout condition. The recovery of failed paths varies depending on the recovery algorithm that is specified by the following timeout_policy attribute:

- retry_path: A path is set to FAILED state if one of the following conditions are met:
  - Error threshold is reached for recoverable I/O errors.
  - Permanent I/O error.

  Paths are recovered immediately upon successful completion of a single healthcheck command. The retry_path algorithm works in the same way as the SDDPCM versions released before 2.6.3.0.
- fail_path: A path is set to FAILED state if one of the following conditions are met:
  - Error threshold is reached for recoverable I/O errors (all errors except for timeout)
  - Single command failed because of timeout
  - Permanent I/O error
- disable_path: A path failed because of timeout is set to DISABLED (OFFLINE) state if one of the following conditions (thresholds) are met:
  - Three consecutive commands fail because of timeout.
  - Three commands fail because of timeout within a 15 minute interval.

  The source of failing commands for these conditions can be host I/O or healthcheck I/O. It is the cumulation of command failures from both of these sources that count towards the defined thresholds.

After a path is set to DISABLED, it is not returned to service until manually recovered by using the **pcmpath set device m path n online** command.

A new **pcmpath** command is added to allow the user to dynamically change the timeout_policy attribute setting, as shown in the following example:

```
pcmpath set device <n>/<n> <m> timeout_policy <option>
```

Important The default setting for timeout_policy is fail_path.

For more information about the **pcmpath** command, see the SDDPCM 2.6.3.2 readme file.

A new AIX error log entry also was added in SDDPCM 2.6.3.2 that is generated when a path is disabled because of the disable_path policy.

Example 1-1 shows the error log entry that appears in the **errpt -a** command output.

Example 1-1 AIX error log entry for disable_path

```
LABEL: SDDPCM_PATH_DISABLE
IDENTIFIER: 6509813E

Date/Time: Tue Apr 10 16:56:53 MST 2012
Sequence Number: 15954
Machine Id: 00F757EA4C00
Node Id: arcp8205jmt1p2
Class: H
Type: PERM
WPAR: Global
Resource Name: hdisk8
Resource Class: disk
Resource Type: 2145
Location: U8205.E6C.1057EAR-V6-C14-T1-W500507680140AD91-L7000000000000

VPD:
      Manufacturer................IBM
      Machine Type and Model......2145
      ROS Level and ID............0000
      Device Specific.(Z0)........0000063268181002
      Device Specific.(Z1)........0200606
      Serial Number...............600507680181853D9000000000000163

Description
PATH HAS DISABLED

Probable Causes
ADAPTER HARDWARE OR CABLE
DASD DEVICE

Failure Causes
EF00

      Recommended Actions
      PERFORM PROBLEM DETERMINATION PROCEDURES
      CHECK PATH

Detail Data
PATH ID
      1

SENSE DATA
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0001 8000 0011 0000
0006 0000 000F 0000 0972 0000 0000 0000 0000 0000 0000 B002 B6B0 0000 0000 0058
FA8B 0000 0000 005B B9AA 0000 0001 0000 0003
```

Special consideration must be given to the use of the disable_path policy because disabled paths are not automatically returned to service. Periodic monitoring of path state must be performed following events such as concurrent code updates and maintenance to ensure full path redundancy.

To avoid conditions where paths can be permanently taken offline during scheduled events (such as code updates and maintenance), it is advisable that the timeout_policy be temporarily changed to fail_path before this activity and then changed back to disable_path at the completion. If this is not done, it is imperative that the path state be checked throughout such activity.

Test results show that the exposure to paths disabling during concurrent code updates and maintenance to be minimal. Even so, it is an exposure and therefore must be considered and protected against. It is generally recommended that clients have automation in place to monitor path status. However, this becomes critically important when the disable_path policy is used because DISABLED paths remain offline until a manual action is performed to return them to service. The underlying condition that caused the paths to become disabled must be dealt with promptly and the paths reactivated to restore full redundancy to the network. Failure to promptly detect and restore disabled paths can compromise the fault tolerance of the system and expose the system to further impact if a secondary fault were to occur.

For more information, see the following resources:

- *Multipath Subsystem Device Driver User's Guide*, GC52-1309-03, which is available at this website:

  http://www-01.ibm.com/support/docview.wss?uid=ssg1S7000303&aid=1

- *Configuring SDDPCM for High Availability*, REDP-4928-00, which is available at this website:

  http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp4928.html

1.11.6 Multi-path I/O

In computer storage, multipath I/O (MPIO) is a fault-tolerance and performance enhancement technique in which there is more than one physical path between the CPU in a computer system and its mass storage devices through the buses, controllers, switches, and bridge devices that are connecting them.

A simple example is a SCSI disk that is connected to two SCSI controllers on the same computer or a disk that is connected to two Fibre Channel ports. Should one controller, port, or switch fail, the operating system can route I/O through the remaining controller transparently to the application with no changes visible to the applications, other than perhaps incremental latency.

Multipath software layers can use the redundant paths to provide the following performance enhancing features:

- Dynamic load balancing
- Traffic shaping
- Automatic path management
- Dynamic reconfiguration

With MPIO, a device can be uniquely detected through one or more physical connections or paths.

Figure 1-51 shows the logical NPIV architecture for our solution from the heterogeneous multipathing perspective from the P750 server to the SAN Directors, SAN Volume Controller, and storage.

Figure 1-51 Heterogeneous multipathing

An MPIO-capable device driver can control more than one type of target device. A path-control module (PCM) provides the path management functions and can support one or more specific devices. Therefore, one device driver can be interfaced to multiple PCMs that control the I/O across the paths to each of the target devices.

Figure 1-52 on page 93 shows the interaction between the different components that make up an MPIO solution. In Figure 1-52 on page 93, the MPIO device driver controls multiple types of target devices, each requiring a different PCM. (In Figure 1-52 on page 93, “KE” refers to Kernel Extension and “RTL” refers to Runtime Loadable).

Figure 1-52 MPIO component interaction

Before a device can make use of MPIO, the device's driver, methods, and predefined attributes in the Object Data Manager (ODM) must be modified to support detection, configuration, and management of multiple paths. The parallel SCSI and Fibre Channel disk device drivers and their device methods support MPIO disk devices. Beginning with AIX 5L™ Version 5.2 with the 5200-04 Recommended Maintenance package, iSCSI disk devices are supported as MPIO devices. Beginning with AIX 5.3, the Fibre Channel tape device driver and its device methods were modified to support MPIO tape devices. Also, the predefined attributes for some devices in the ODM were modified for MPIO.

The AIX PCM consists of the PCM RTL configuration module and the PCM KE kernel extension. The PCM RTL is a runtime loadable module that enables the device methods to detect more PCM KE device-specific or path ODM attributes that the PCM KE requires. The PCM RTL is loaded by a device method. One or more routines within the PCM RTL are then accessed to perform specific operations that initialize or modify PM KE variables.

The PCM KE supplies path-control management capabilities to any device driver that supports the MPIO interface. The PCM KE depends on device configuration to detect paths and communicate that information to the device driver. Each MPIO-capable device driver adds the paths to a device from its immediate parent or parents. The maintenance and scheduling of I/O across different paths is provided by the PCM KE and is not apparent to the MPIO-capable device driver.

The PCM KE can provide more than one routing algorithm, which can be selected by the user. The PCM KE also helps collect information that can be used to determine and select the best path for any I/O request. The PCM KE can select the best path based on various criteria, including load balancing, connection speed, and connection failure.

The AIX PCM has a health-check capability that can be used to perform the following tasks:

- Check the paths and determine which paths are currently usable for sending I/O.
- Enable a path that was previously marked failed because of a temporary path fault (for example, when a cable to a device was removed and then reconnected).
- Check currently unused paths that would be used if a failover occurred (for example, when the algorithm attribute value is failover, the health check can test the alternative paths).

Not all disk devices and tape devices can be detected and configured by using the AIX default PCMs. The AIX default PCMs consist of two path control modules, one to manage disk devices and another to manage tape devices. If your device is not detected, check with the device vendor to determine if a PCM is available for your device.

For the solution in this book, the IBM MPIO FCP Disk Device v1.0.0.23 was implemented for compatibility and interoperability with the software and hardware stack. This Host Attachment allows AIX MPIO device driver configure methods to properly identify and configure ESS, SAN Volume Controller, IBM Storwize V7000, DS6000™, DS8000, DS4000, DS5000, DS5020, and DS3950.

The host attachment for SDDPCM adds ESS(2105), SAN Volume Controller(2145), IBM Storwize V7000(2145), DS6000(1750), DS8000(2107), DS4000(1724/1814/1722/1742/1815), DS5000(1818), DS5020(1814), DS3950(1814) and RSSM(1820) device information to allow AIX to properly configure hdisks. These device information allows AIX to perform the following tasks:

- Identify the hdisk(s) as a 2105/2145/1750/2107/1724/1814/1722/1742/1815/1818/1820 hdisk.
- Set default hdisk attributes, such as queue_depth and timeout values.
- Indicate to AIX device driver configure method to configure 2105/2145/1750/2107/1724/1814/1722/1742/1815/1818/1820 hdisk as MPIO-capable devices.

For more information, see this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S4000203

1.12 PowerHA

This chapter section provides an overview of the IBM PowerHA architecture as it applies to the SAN Volume Controller Stretched Cluster configuration and solution context in this book.

1.12.1 Overview and concepts

PowerHA for AIX is the IBM premier HA solution for IBM Power Systems that are running AIX (it was formerly known as HACMP). Versions before PowerHA 5.5 are referred to by its original name of HACMP. PowerHA is a key component in this solution. This technology in the solution provides for automated failover from the primary site to the secondary site for the p750 server or LPAR, and for the applications that are hosted in the LPARs. In this case, the applications that are hosted are IBM WebSphere Application Server and IBM DB2. PowerHA also restarts the applications in the secondary site server and LPAR.

PowerHA SystemMirror monitors numerous soft and hard errors within the cluster from various event sources, including problems that are severe enough to immobilize the system (such as a process failure or exhaustion of system resources). With PowerHA V7, monitoring and event management were moved into the kernel of the OS, which provides a robust foundation that is not prone to job scheduling issues or other events related to OS operations. Cluster awareness enables OS-based functions to operate harmoniously with PowerHA. PowerHA SystemMirror is integrated with cluster-aware AIX and makes use of the OS features by extending them across the cluster, which enables efficient centralized management.

The PowerHA V7 graphical interface makes cluster management simpler and more productive by providing a secure, centralized point of operations. An administrator can readily deploy clusters and resource groups by using the associated wizard. It provides live updates by graphically displaying the status of resource groups, nodes, sites, and clusters. With a single view, you can see the health of the entire enterprise at a glance. The interface was further enhanced to support the deployment and management of Federated Security and the SAP LiveCache Host and by environments and with the V7 Enterprise Edition, geographically dispersed PowerHA SystemMirror clusters. You can now use your Director-based GUI to deploy, monitor, and manage your enterprise across multiple data centers.

Figure 1-53 on page 96 shows a view of the PowerHA Management Interface with an example PowerHA cluster topology view.

Figure 1-53 PowerHA interface: Cluster topology

PowerHA SystemMirror V7 provides Smart Assists for easier, ready-to-use HA setup and application management for many well-known middleware products. The smart assists are HA agents for application deployment and management. Smart Assists are used to define HA policies through resource groups and to discover software that is deployed within the cluster. Discovery-based information aids in defining the HA policy and provides periodic health monitoring, which enables middleware and resource dependencies to be restarted via the specified policy. With PowerHA SystemMirror V7 the Smart Assist portfolio supports the following applications:

- DB2, WebSphere
- WebSphere MQ
- Oracle
- SAP
- SAP MaxDB
- Enterprise Content Manager
- IBM Tivoli Storage Manager
- Lotus® Domino® Server
- IBM LDAP
- IBM HTTP printers
- FileNet®

PowerHA SystemMirror also includes Smart Assists to deploy and manage the SAP LiveCache Hot Standby Solution.

Figure 1-54 on page 97 shows the PowerHA Management Interface (Resource Groups tab) where HA policies for many popular middleware offerings can be set up for cluster management. HA polices and resource groups also can be established for custom workloads.

Figure 1-54 PowerHA: Add a Resource Group

PowerHA also provides a multiprocessing component that manages multiple hardware and software resources to provide complex application functionality and better resource usage. Taking advantage of the multiprocessing component depends on careful planning and design in the early stages of implementation to efficiently use all resources, including shared storage that is robust and easy to manage.

PowerHA is only one of the HA technologies and builds on the increasingly more reliable operating systems, hot swappable hardware, and increasingly more resilient applications by offering monitoring and automated response.

The IBM HA solution for Power Systems offers the following distinct benefits:

- Proven solution (more than 18 years of product development)
- Flexibility (virtually any application that is running on a stand-alone AIX system can be protected with PowerHA)
- Use of off-the-shelf hardware components
- Proven commitment for supporting our customers

When you are planning to implement a PowerHA solution, the following aspects must be considered:

- Thorough HA design and detailed planning from end-to-end
- Elimination of single points of failure
- Selection of appropriate hardware
- Correct implementation (do not take shortcuts)
- Disciplined system administration practices and change control
- Documented operational procedures
- Comprehensive test plan and thorough testing

An HA solution that is based on PowerHA provides automated failure detection, diagnosis, application recovery, and node reintegration. With an appropriate application, PowerHA also can provide concurrent access to the data for parallel processing applications, thus offering excellent horizontal, and vertical scalability (with the addition of dynamic LPAR management capabilities).

For more information about IBM Power Systems HA and DR Best Practices from the IBM STG High Availability Center of Competency (which also provides specific best practices for PowerHA environments), see the following website:

https://w3-connections.ibm.com/wikis/form/anonymous/api/wiki/b7b38f1c-91b5-43ac-88d8-18f55aeb8abc/page/9b72d7a9-a33f-4607-b554-e2fac21b5cc3/attachment/5d46d62d-af79-4709-9021-a2cec7e72df0/media/HACoC%20POWER_HA%20and%20DR%20Best%20Practices.pdf

For illustrative purposes only, a typical PowerHA environment is shown in Figure 1-55. Though a serial network is shown and is still available today, most implementations provide non-IP heartbeat networks by using disk heartbeating.

Figure 1-55 PowerHA cluster

PowerHA SystemMirror monitors numerous soft and hard errors within the cluster from various event sources, including problems that are severe enough to immobilize the system (such as a process failure or exhaustion of system resources). With PowerHA V7, monitoring and event management is moved into the kernel of the OS, which provides a robust foundation not prone to job scheduling issues or other events that are related to OS operations. Cluster awareness enables OS-based functions to operate harmoniously with PowerHA. PowerHA SystemMirror is integrated with cluster-aware AIX and makes use of the OS features by extending them across the cluster, which enables efficient, centralized management.

1.13 Architecture and solution context

This section provides a high-level overview of the PowerHA SystemMirror Architecture and the specific solution context for this book. It is assumed that the reader has a base understanding of the PowerHA architecture. Links to other IBM Redbooks for PowerHA are provided in this book for reference.

1.13.1 Reliable Scalable Cluster Technology

Reliable Scalable Cluster Technology (RSCT) is a set of software components that provide a comprehensive clustering environment for AIX, Linux, Solaris, and Microsoft Windows. RSCT is the infrastructure that is used by various IBM products to provide clusters with improved system availability, scalability, and ease of use.

This section provides an overview of RSCT, its components, and the communication paths between these components. Several helpful IBM manuals, white papers, and IBM Redbooks publications are available about RSCT. This section focuses on the components that affect PowerHA SystemMirror.

To find the most current documentation for RSCT, see the RSCT library in the IBM Cluster Information Center at this website:

http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=%2Fcom.ibm.cluster.rsct.doc%2Frsctbooks.html

RSCT has the following main components:

- Topology Services

  This component provides node and network failure detection.

- Group Services

  This component provides cross-node or process coordination on some cluster configurations. For more information about how Group Services work, see *IBM Reliable Scalable Cluster Technology: Group Services Programming Guide*, SA22-7888, which is available at this website:

  http://publibfp.boulder.ibm.com/epubs/pdf/a2278889.pdf

- RSCT cluster security services

  This component provides the security infrastructure that enables RSCT components to authenticate the identity of other parties.

- Resource Monitoring and Control (Rational Method Composer) subsystem

  This subsystem is the scalable, reliable backbone of RSCT. It runs on a single machine or on each node (OS image) of a cluster. It also provides a common abstraction for the resources of the individual system or the cluster of nodes. You can use Rational Method Composer for single system monitoring or for monitoring nodes in a cluster. However, in a cluster, Rational Method Composer provides global access to subsystems and resources throughout the cluster. Therefore, it provides a single monitoring and management infrastructure for clusters.

- Resource managers

  A resource manager is a software layer between a resource (a hardware or software entity that provides services to some other component) and Rational Method Composer. A resource manager maps programmatic abstractions in Rational Method Composer into the actual calls and commands of a resource. For more information about the RSCT components, see *IBM Reliable Scalable Cluster Technology: Administration Guide*, SA22-7889, which is available at this website:

  http://publibfp.boulder.ibm.com/epubs/pdf/22788919.pdf

  RSCT version 3.1 is the first version that supports Cluster Aware AIX (CAA). For more information about CAA, see 1.2, “Cluster Aware AIX” on page 7 of the IBM PowerHA SystemMirror 7.1 for AIX book, which is available at this website:

  http://www.redbooks.ibm.com/redbooks/pdfs/sg247845.pdf

For more information, see IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update, SG24-8030-00, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg248030.html?Open

Figure 1-56 shows a high-level architectural view of how IBM HA applications PowerHA, IBM Tivoli System Automation for Multiplatforms, and VIOS Clustered Storage use the RSCT and CAA architecture. This is noted because thus far we described these technologies as the Power Systems High Availability Building Blocks to achieve higher levels of resiliency for service availability for Power Systems.

Figure 1-56 HA applications that use the RSCT and CAA architecture

Figure 1-57 shows the new CAA-based communication paths of PowerHA, RSCT, and CAA. You use this architecture when you have PowerHA v7.1 or later (as in this SAN Volume Controller Stretched Cluster book solution). It is the same architecture for AIX 6.1 TL 6 and AIX 7.1 or later.

Figure 1-57 RSCT with Cluster Aware AIX (CAA)

1.13.2 SAN Volume Controller Stretched Cluster solution context: PowerHA clusters and workload

We implemented WebSphere Application Server and the DB2 database in the P750 LPARs. We also implemented a test application called DayTrader, which runs in WebSphere Application Server, and its database schema is installed in DB2. We use this application to test our SAN Volume Controller Stretched Cluster, PowerHA, and LPM aspects of the solution in this book. It offers a mix of read and write activity, which we tested in various outage scenarios for the technology component stack in our solution architecture. For more information about testing, results and conclusions, see Chapter 5, “Stretched Cluster Diagnostic and Recovery Guidelines” on page 293.

Figure 1-58 on page 102 provides a high-level overview of the DayTrader logical application architecture, which is run and hosted by WebSphere.

Figure 1-58 DayTrader Logical Application Architecture

For the solution testing in the book, we run and test PowerHA to restart WebSphere and DB2 services for a real-life outage situation should the P750 LPAR or server be affected.

Figure 1-59 shows the solution component architecture for PowerHA and DB2 in the solution. The server on the right side is the Site A P750 server (for example, Production).

Figure 1-59 PowerHA DB2 Cluster

Figure 1-60 shows what occurs during the outage on the Site A P750 server, and how PowerHA reconnects over to the Site B P750 server and restarts the DB2 services.

Figure 1-60 PowerHA DB2 cluster and outage failover

For more information about the PowerHA implementation details of the solution, see Chapter 4, “PowerVM and PowerHA Implementation” on page 213.

1.14 SAN Volume Controller, Layer-2 IP Network, Storage Networking infrastructure and LPM integration

This section provides an overview of the SAN Volume Controller Stretched Cluster architecture as it applies to the SAN Volume Controller, Layer-2 IP Network, IBM b-type SAN Infrastructure, and LPM solution context in this book.

1.14.1 Overview and concepts

Virtualization is now recognized as a key technology for improving the efficiency and cost effectiveness of a company's IT infrastructure. As a result, critical business applications are being moved to virtualized environments, which bring with them requirements for higher availability, protection of critical business data, and the ability to fail over and continue supporting business operations in the case of a local outage or a widespread disaster.

LPM is a feature of IBM Power Systems servers that allows the live migration of LPARs from one Power server to another with no application downtime. Typically, LPM was used within a data center to keep IT environments running, which gives unprecedented flexibility and availability to meet the increasing demands for data. However, it is possible to migrate LPARs between data centers with no downtime or user disruption.

Because IT can now perform a secure migration of a live virtualized application and its associated storage between data centers with no downtime or user disruption, IT managers can realize the benefits of disaster avoidance and recovery, load balancing between data centers, better usage of a cloud infrastructure, optimization of power consumption, and maintaining the right level of performance for applications.

LPM over distance, spanning data centers, or geographical boundaries, requires a specialized infrastructure and environment. The following areas are key to this configuration:

- Data synchronization between data centers, which allows servers (regardless of their location) to access that data anytime.
- The network infrastructure that provides high performance, high reliability, and proper Layer-2 extension capabilities to interconnect the data centers.
- IP traffic management of client network access to the site where the application server is.

The solution that is described in this book addresses these key areas through the combination of Power Systems LPM with SAN Volume Controller Stretched Cluster capabilities, which runs over Layer 2 IP Network and storage infrastructure.

Continuous access to data is provided by SAN Volume Controller Stretched Cluster configuration and SAN Volume Controller Volume Mirroring capability.

The Layer 2 IP Network and storage networking infrastructure provides a reliable and high-performance, end-to-end solution with network adapters, edge, aggregation, and core switching, and high-performance application delivery controllers. This combination provides a flexible infrastructure that results in simplification, reduced costs, higher resource usage, and most importantly, data protection and resiliency.

The combination of LPM, with SAN Volume Controller Stretched Cluster solution the Layer 2 IP networking infrastructure, enables the design and implementation of a robust business continuity, disaster avoidance, and recovery solution for virtualized application environments.

1.14.2 Application mobility over distance

Various testing scenarios for each of the solution architecture components (SAN Volume Controller, Stretched Cluster, SAN Directors, PowerHA, and LPM) were included to simulate the following distances between Site A (Production) and Site B (DR). We used a Network Latency simulator at the Brocade Labs in the testing to simulate different distances:

- Less than 5 KM (3.10 miles), as an example of Local HA or partitioned single data center location
- 5 KM (3.10 miles), as an example of a Campus data center architecture
- 15 KM (9.32 miles), as an example of a Metro Mirror data center solution and architecture
- 75 KM (46.60 miles), as an example of a Metro Mirror data center solution and architecture
- 150 KM (93.20 miles), as an example of a Global Mirror/Out of Region DR data center solution and architecture
- 300 KM (186.41 miles), as an example of a Global Mirror/Out of Region DR data center solution and architecture

The SAN Volume Controller component testing is included a varying mix of Read, Write, and Read and Write I/O activity against the different distance categories. We also tested outage scenarios for the Site A single SAN Director loss, Site A two SAN Director loss, Site A Server and LPAR loss, Site A, and Storwize V7000 loss with a mix of Read and Write I/O Activity against the six distance categories.

Finally, we tested LPM against the six distance categories with varying application DB2 sizes of 45 GB, 200 GB, and 500 GB. Additional testing is planned to simulate Jumbo Frames across the distances with DB2 sizes of greater than 500 GB (1 TB, 1 TB +).

For more information about the testing scenarios, results, conclusions, and solution recovery guidelines, see Chapter 5, “Stretched Cluster Diagnostic and Recovery Guidelines” on page 293.

1.14.3 IBM, PowerVM, and Layer 2 IP Network solution

Building the right storage and network infrastructure to enable data and application mobility requires a data center infrastructure that can provide optimal storage extension capabilities and end-to-end advanced network functionality. It is imperative that a complete, comprehensive solution that comprises storage, network, and server infrastructure be designed and implemented to facilitate the movement of applications across data centers. The following products from IBM and VMware were validated in the solution:

- PowerVM with Enterprise Edition licensing to support LPM.
- IBM SAN Volume Controller Stretched Cluster to ensure the availability of access to the storage in both data centers.
- IBM PowerHA SystemMirror.
- Layer 2 IP Network switch, as described in Chapter 2, “Hardware and Software Description” on page 115.
- IBM SAN FC switch and director family products for FC connectivity and FC/IP connectivity, as described in Chapter 2, “Hardware and Software Description” on page 115.

1.15 Architecture and solution context

This chapter section provides an overview of the overall SAN Volume Controller Stretched Cluster architecture and solution context and topology as implemented in our lab.

1.15.1 IP and Layer 2 network architecture

End-to-end data center network design is outside the scope of this book. However, we review the network architecture as seen in the context of our solution.

Figure 1-61 shows the high-level network architecture.

Figure 1-61 High-level Brocade network architecture

As shown in Figure 1-61, the following IP networking components are described in this book:

- Layer 2 switches

  The AIX hosts connect directly to Layer 2 switches. The Layer 2 network can be any standards-compliant switch such as any IBM RackSwitch™ family switch or the Brocade VDX series of switches. In a three-tier network design, there often is an Edge that is made up of Top of Rack switches that connect to a higher-density Aggregation switch over Layer 2 running a version of Spanning Tree Protocol. With higher-density Top of Rack switches, scalable logical switch capabilities, such as stacking or Virtual Switch Clustering, or passive patch paneling in each rack to a higher density End of Row switch, the overall data center network might be collapsed into a flat Layer 2 Edge that connects directly into the IP Core over Layer 2 or Layer 3.

- IP Core

  The Layer 2 network connects into the IP core which then connects out to the WAN Edge or Internet. In our configuration, we connect from the Layer 2 switches to the IP core over Layer 2. The IP core that is made up of Brocade MLXe devices provides aggregation and routing capabilities between the various subnetworks.

- Brocade MLXe/CER acting as the data center interconnect or FileNet Process Engine router

For Layer 2 extension, we use standards-compliant Layer 2 VPN technology that uses MPLs/VPLS/VLL. Although these services are provided by a service provider and are more advanced, we review the configuration in our setup.

Our solution architecture and topology is shown in Figure 1-62.

Figure 1-62 IBM SAN Volume Controller Stretched Cluster Lab solution architecture

A closer view of data center A (Site 1 or Site A) connection is seen in Figure 1-63.

Figure 1-63 BM SAN Volume Controller Stretched Cluster - data center A topology

The architecture solution features the following components:

- An IBM Power 750 server that is AIX host Power 750 A. This host has two 10 GbE ports and four16 Gbps FC ports.
- The 16 Gbps FC ports are connected to two separate IBM SAN768B-2 chassis, which forms two separate, air-gapped FC SAN fabrics.
- The IBM SAN768B-2 Directors are attached to a single IBM Storwize V7000.
- The two 10 GbE ports are connected to two separate 10 GbE VDX switches. There are a total of four VDX switches in a Layer 2 network. These four VDX switches are clustered together by using VCS fabric technology, which to other network entities makes them appear to be a single logical switch. The IBM System Networking RackSwitch G8264CS Top of Rack switch offers the benefits of a converged infrastructure. As part of its forward-thinking design, this switch has flexibility for future growth and expansion. This switch is ideal for clients who are looking to connect to existing SANs and clients who want native Fibre Channel connectivity in addition to support for such protocols as Ethernet, FCoE, and iSCSI.

For further information, see this website:

https://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/tips0970.html?Open

- VDX switches are connected over Layer 2 to two MLXe routers that are acting as the IP Core. The MLXe routers provide routing between the various subnetworks and aggregate the various connections.
- The Brocade ADX GSLB/SLB is connected to the MLXe.
- The data center interconnect links (which are Brocade CES routers) are connected to the MLXs. The Brocade CES routers provide the Layer 2 VPN extension capabilities by using VLL, which act as FileNet Process Engine routers.
- The Brocade MLXe IP Core also is connected to simulated clients that might come in outside the data center from the Internet.
- Also connected to the Brocade MLXe are two 10 GbE FCIP connections from each of the IBM SAN768B chassis. IBM SAN768B-A1 has two links that are connected to DC1-Core1-MLX4: one link forms the Private SAN FCIP tunnel and the second link forms the Public SAN FCIP tunnel. Similarly, IBM SAN768B-A2 has two links that are connected to DC1-Core2-MLX4.

For more information about the SAN Volume Controller Stretched Cluster solution implementation, see Chapter 3, “SAN Volume Controller Implementation” on page 149.

1.16 Open data center Interoperable Network

Data centers are undergoing a major transition toward a smarter, more dynamic infrastructure. More flexible IT architectures are emerging to address the demands of current business applications and new areas, such as cloud computing, multitenancy, bring your own device (BYOD), big data, and analytics.

As part of the dynamic infrastructure trend, the role of data center networks also is changing. It is causing businesses to re-evaluate their current networks, which were never designed to handle modern workloads and applications. In light of these changes, the networking industry is experiencing a major discontinuity.

The IBM viewpoint is that the most cost-effective approach to building a best-of-breed data center network is based on open industry standards that are supported by multiple vendors. An open data center with an interoperable network delivers the following noticeable benefits:

- Scales to tens of thousands of virtual machines, which reduces capital and operating expense 15% - 25%.
- Optimizes usage by virtualizing network resources and moving virtual machines freely within and between data centers.
- Integrates pre-optimized solutions seamlessly with the rest of your data center.
- Automates network provisioning, management, and troubleshooting so that it becomes more agile and less error prone.
- Creates flatter, application-aware networks that need to be wired once.

This section provides an overview of the key concepts for the ODIN architecture as it applies to this book.

1.16.1 Overview and concepts

Over the past several years, progressive data centers have undergone fundamental and profound architectural changes. Nowhere is this more apparent than in the data center network infrastructure. The conventional networks are designed for traffic flows traverse multiple network tiers, adding latency and degrading performance; Open data center Interoperable Network (ODIN) promotes a flat, two-tier network that is optimized for traffic between servers.

ODIN promotes scaling the network to thousands of physical ports and tens of thousands of virtual machines, without high oversubscription and without requiring large numbers of switches that take up space, use energy, and add cost. ODIN also promotes virtual machine mobility through large Layer 2 domains and best practices for low latency, even when extended distance connectivity is used between multiple data centers.

ODIN promotes software-defined networking and network overlays that enable virtualized, wire-once networks and overcome traditional topology and bandwidth limits. ODIN describes equal cost multipath spine-leaf architectures, which help scale networks in a cost and performance efficient manner. ODIN promotes high availability, energy efficiency, automated management, network virtualization, and other features intended to provide lower capital and operating expense.

IBM took the lead to articulate the issues that are faced by agile data center networks and promote best practices by creating a roadmap based on industry standards, known as ODIN.

ODIN is a flat, converged, virtualized data center network that is based on open industry standards, as shown in Figure 1-64. Instead of under-used devices, multitier networks, and complex management environments, the modern data center is characterized by highly used servers that are running multiple VMs, flattened, lower latency networks and automated, integrated management tools.

Figure 1-64 An overview of an ODIN driven network

New software-defined network approaches (including overlay networks and OpenFlow standards) greatly simplify the implementation of features such as dynamic workload provisioning, load balancing, and redundant paths for high availability and network reconfiguration. Further, as shown in Figure 1-64, on high-bandwidth links between virtualized data center resources can extend across multiple data center locations to provide business continuity and backup and recovery of mission-critical data.

1.16.2 Architecture and solution context

The ODIN reference architecture is a work in progress that addresses best practices and helps you interpret the networking standards that are vital to efficient data center operations. The information about these methods and standards help you make the transition from discrete, special-purpose networks that each have their own management tools, to a converged, flattened network that uses a common set of management tools.

IBM believes that the practical, cost-effective evolution of data center networks should be based on open industry standards. This can be a challenging and often confusing proposition because there are so many different emerging network architectures, standard and proprietary. The ODIN materials that are created by IBM currently address issues such as virtualization and Virtual Machine (VM) migration across flat Layer 2 networks, lossless Ethernet, Software-Defined Networking (SDN) and OpenFlow, and extended distance WAN connectivity (including low latency).

SDN is used to simplify network control and management, automate network virtualization services, and provide a platform from which to build agile network services. By abstracting the control and management aspects of a network into a logical software program, SDN allows real-time programmability and manageability of networks comparable to what is achieved on computers. It can use a centralized logical network view that is easily manipulated via software to implement complex networking rules. This allows networks to achieve unprecedented levels of scalability and flexibility, and dynamic behaviors matching cloud service-oriented dynamics. Figure 1-65 presents an overview of SDN.

Figure 1-65 Overview of SDN

ODIN features the following benefits:

- Customer choice and future-proof designs for data center networks, including vendor-neutral RFQs.
- Lower total cost of ownership by enabling a multi-vendor network.
- Avoiding confusion in the marketplace between proprietary and vendor-neutral solutions.
- Providing best practices, relative maturity, and interpretation of networking standards.

IBM developed solutions that are based on the following open industry standards:

- Fibre Channel and FC-IP storage technologies
- Lossless Ethernet
- Flat, Layer 2 Networks
- Distance extension options that use dark fiber WDM and MPLS/VPLS

For more information about ODIN technical briefs, see these resources:

- ODIN networking solutions:

  http://www.ibm.com/systems/networking/solutions/odin.html

- *Building a World-class data center Network Based on Open Standards*, REDP-4933-00, which is available at this website:

  http://www.redbooks.ibm.com/abstracts/redp4933.html?Open

Figure 1-66 shows typical ODIN solutions.

Figure 1-66 Typical ODIN solutions

The following major targets are associated with ODIN:

- Support customer choice and future-proof network design choices, including vendor-neutral RFQs
- Referring to *"Gartner Group, "Debunking the myth of the single-vendor network", 17 November 2010"* enable lower TCO by 15% - 25%
- Avoid confusion in the marketplace between proprietary and vendor-neutral solutions
- Provide best practices, relative maturity, and interpretation of the many networking standards which have recently emerged
- Give clients a single, trusted source for understanding standards-based networking requirements, and a voice in what those requirements will look like in the future

All of these major targets are satisfied with the SAN Volume Controller Stretched Cluster solution.

2

Hardware and Software Description

In this chapter we describe the hardware and software that is needed to implement an IBM storage area network (SAN) Volume Controller Stretched Cluster solution on Live Partition Mobility (LPM). We describe LPM and some features that are useful when implementing a SAN Volume Controller Stretched Cluster.

All of the products that are described provide the functionality that is needed to implement a SAN Volume Controller Stretched Cluster. It is the responsibility of the implementor to choose the most suitable product for the environment.

When implementing SAN Volume Controller Stretched Cluster, the following considerations must be made in terms of possible hardware needs:

- Distance and latency between data centers
- Connectivity between data centers
- Bandwidth of sent data
- Customer budget
- Current customer infrastructure

All of these considerations can result in greatly varied hardware requirements. In this section, we describe some hardware possibilities and what features are recommended to purchase with the hardware.

© Copyright IBM Corp. 2013. All rights reserved.

This chapter includes the following topics:

- IBM System Storage SAN Volume Controller
- SAN directors and switches
- FCIP routers
- Ethernet switches and routers
- P750 and Hardware Management Console
- Software
- IBM Power Systems operating system
- IBM PowerVM: VIOS, NPIV, and SDDPCM/MPIO
- PowerHA
- IBM DB2
- IBM WebSphere Application Server

2.1 IBM System Storage SAN Volume Controller

As described in Chapter 1, “Introduction” on page 1, this solution requires the use of the IBM System Storage SAN Volume Controller. The SAN Volume Controller provides an active-active storage interface that can allow for simple fail over and fail back capabilities in case of a site disruption or failure.

To implement SAN Volume Controller Stretched Cluster solution over 4 km (2.48 miles), it is recommended that the 2145-CF8 or 2145-CG8 hardware models of the SAN Volume Controller controllers (as shown in Figure 2-1) are used because of their increased node capabilities. Also, depending on the architecture that you want to deploy, you must run a minimum level of firmware. Check with your IBM contact or see Chapter 3, “SAN Volume Controller Implementation” on page 149 to ensure that the SAN Volume Controller node model and firmware version are supported with the solution you want to implement.

Figure 2-1 SAN Volume Controller CF8 nodes

2.2 SAN directors and switches

To implement an IBM System Storage SAN Volume Controller Stretched Cluster solution, any SAN fabrics must be extended across two data centers or failure domains. How you want to extend this fabric depends on the distance between failure domains. Several choices of architecture are described in Chapter 3, “SAN Volume Controller Implementation” on page 149.

We do not describe any particular Wavelength-Division Multiplexing (WDM) devices. We also do not describe any Ethernet infrastructure options (other than FCIP devices). All of the hardware that is described in this section is compatible with CWDM devices by using colored small form-factor pluggables (SFPs), dense wavelength division multiplexing (DWDM) devices, and Fibre Channel over IP (FCIP) routers.

2.2.1 SAN384B-2 and SAN768B-2 directors

The IBM System Storage SAN384B-2 and SAN768B-2 directors provide scalable, reliable, and high-performance foundations for virtualized infrastructures.

As shown in Figure 2-2 on page 118, the SAN768B-2 and SAN384B-2 fabric backbones increase business agility while providing nonstop access to information and reducing infrastructure and administrative costs. With 16 Gbps Fibre Channel capabilities, they deliver a new level of scalability and advanced capabilities to this robust, reliable, high-performance technology.

Figure 2-2 SAN768B-2 and SAN384B-2 fabric backbones

Both directors are capable of 16, 10, 8, 4, and 2 Gbps connections with the capability to have up to 512 or 256 ports. Included with the purchase of the directors is an enterprise software bundle that has the Extended Fabrics and Trunking features. The Extended Fabrics feature is essential for implementing Stretched Cluster solutions over 10 km (6.21 miles). The Trunking feature is necessary if multiple links are required to accommodate the bandwidth that is used for SAN traffic.

For more information about the Stretched Cluster solution over long distance, see this website:

http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp

For more information about the IBM System Storage SAN384B-2 and SAN768B-2 directors, see this website:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san768b-2/index.html

2.2.2 SAN24B-5, SAN48B-5, and SAN80B-4 switches

IBM System Storage offers various Fibre Channel switches to suit various client data center needs and budgets. The IBM System Storage SAN24B-5, SAN48B-5, and SAN80B-4 switches support highly virtualized environments while maintaining excellent cost-performance ratios.

SAN24B-5 switch

The IBM System Networking SAN24B-5 switch, as shown in Figure 2-3, can be configured with 12 or 24 active ports and is capable of 2, 4, 8, and 16 Gbps speeds in a 1-U form factor. This switch is suited for smaller environments and for environments where a small performance switch is needed for SAN Volume Controller node traffic.

Figure 2-3 SAN24B-5 Switch

When you are implementing SAN Volume Controller Stretched Cluster solutions over 10 km (6.21 miles) with the SAN24B-5, it is necessary to purchase the Extended Fabrics feature to allow the switch to extend the distance of links.

For more information about the IBM System Networking SAN24B-5 switch, see this website:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san24b-5/index.html

SAN48B-5 switch

The IBM System Storage SAN48B-5 switch, as shown in Figure 2-4, is configurable with 24, 32, or 48 active ports and is capable of 2, 4, 8, 10, and 16 Gbps speeds in a 1-U form factor. The performance, reliability, and price of this switch make it a suitable candidate for an edge switch in large to mid-sized environments.

Figure 2-4 SAN48B-5 Switch

When implementing SAN Volume Controller Stretched Cluster solutions over 10 km (6.21 miles) with the SAN48B-5, it is necessary to purchase the Extended Fabrics feature to allow the switch to extend the distance of links.

For more information about the IBM System Storage SAN48B-5 switch, see this website:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san48b-5/index.html

SAN80B-4 switch

The IBM System Storage SAN80B-4 switch, as shown in Figure 2-5, is configurable with 48, 64, or 80 active ports and is capable of 1, 2, 4, and 8 Gbps speeds. High availability (HA) features make this an ideal candidate for a core switch in medium-sized environments and an edge switch in larger enterprise environments.

Figure 2-5 SAN80B-4 switch

The Extended Fabric feature is enabled on the SAN80B-4 by default, which makes it even more useful for SAN Volume Controller Stretched Cluster configurations over 10 km (6.21 miles). Because this switch is suited to larger environments, it is possible that the Trunking Activation license must be purchased to ensure bandwidth between failure domains.

For more information about the IBM System Storage SAN80B-4 switch, see this website:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san80b-4/index.html

2.3 FCIP routers

When implementing an IBM System Storage SAN Volume Controller Stretched Cluster over long distances, it is not always possible or feasible to extend SAN fabrics by using direct Fibre Channel connectivity or WDM. This is because the distance between the two failure domains is over 10 km (6.21 miles) or that it is too expense to lay cable or hire dark fibre.

Numerous dual data center environments already have existing IP connections between data centers. These connections allow FCIP technologies to be used to enable the SAN fabric to extend across data centers while using existing infrastructure. When SAN Volume Controller Stretched Cluster solutions with FCIP are implemented, it is important to note that there are minimum bandwidth requirements that must be met so that the solutions are supported. For more information, see Chapter 1, “Introduction” on page 1.

2.3.1 8 Gbps Extension Blade

The 8 Gbps Extension Blade is an FCIP blade, as shown in Figure 2-6 on page 120, that can be placed into the SAN384B-2 and SAN768B-2 SAN directors. This blade uses 8 Gbps Fibre Channel, FCIP, and 10 GbE technology to enable fast, reliable, and cost-effective remote data replication, backup, and migration with existing Ethernet infrastructures.

Figure 2-6 8 Gbps Extension Blade

The 8 Gbps Extension Blade has 12 8-Gbps Fibre Channel ports and 10 1-GbE Ethernet ports by default. With the 8-Gbps Extension Blade 10GbE Activation feature on the SAN384B-2 and SAN768B-2 directors, we have the option to have two 10-GbE ports or 10 1-GbE Ethernet ports and one 10-GbE port on the 8-Gbps Extension Blade. Also, when this blade is ordered, the 8-Gbps Advanced Extension Activation on the SAN384B-2 and SAN 768B-2 directors feature must be ordered.

2.3.2 SAN06B-R extension switch

The IBM System Storage SAN06B-R extension switch (as shown in Figure 2-7) optimizes backup, replication, and data migration over a range of distances by using Fibre Channel and FCIP networking technologies.

Figure 2-7 SAN06B-R extension switch

The SAN06B-R extension switch provides up to 16 8-Gbps Fibre Channel ports and six 1-GbE ports to enable FCIP routing. To enable FCIP routing on the switch, the R06 Trunking Activation feature, R06 8 Gbps Advanced Extension feature, or the R06 Enterprise Package must be ordered.

For more information about the IBM System Storage SAN06B-R Extension switch, see this website:

http://www-03.ibm.com/systems/networking/switches/san/b-type/san06b-r/index.html

2.4 Ethernet switches and routers

To support LPM over long distances, a scalable and robust IP network must be available to the VIOS hosts for data connectivity and SAN FCIP Extension devices for storage traffic over FCIP. Layer 2 extension between the data centers also is required to enable LPM support. This can be accomplished readily with standards-compliant MPLS/VPLS/VLL technology.

2.4.1 IBM system networking switches

IBM System Networking RackSwitch G8124E

The IBM RackSwitch G8124E, as seen in Figure 2-8, is a 10-Gigabit Ethernet switch that is specifically designed for the data center and provides a virtualized, cooler, and easier network solution. Designed with top performance in mind, the G8124E provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data and large data center-grade buffers to keep traffic moving.

Figure 2-8 IBM RackSwitch G8124E

The G8124E offers 24 10-Gigabit Ethernet ports in a high-density, 1-U footprint, which makes it ideal for data center Top of Rack switching.

IBM System Networking RackSwitch G8264CS

Designed with top performance in mind, the IBM RackSwitch G8264CS (as seen in Figure 2-9), is ideal for today's big data, cloud, and optimized workloads. The IBM System Networking RackSwitch G8264CS Top of Rack switch offers the benefits of a converged infrastructure. As part of its forward-thinking design, this switch has flexibility for future growth and expansion. This switch is ideal for clients who want to connect to existing SANs and those who want native Fibre Channel connectivity in addition to support for such protocols as Ethernet, Fibre Channel over Ethernet (FCoE), and iSCSI.

The RackSwitch G8264CS includes the following features:

- Lossless Ethernet, Fibre Channel, and FCoE in one switch
- 36 SFP+ ports supporting 1-Gb or 10-Gb Ethernet
- Flexibility with 12 IBM Omni Ports that support 10-Gb Ethernet or 4/8-Gb Fibre Channel connections
- Can aggregate FC or FCoE converged traffic with the ability to connect to existing SANs
- Designed for future expansion with four 40-Gb QSFP+ ports
- Low-cost, low-complexity, and simpler deployment and management

Figure 2-9 IBM System Networking RackSwitch G8264CS

The IBM RackSwitch G8264CS simplifies deployment with its innovative and flexible IBM Omni Port technology. The 12 Omni Ports on the G8264CS give clients the flexibility to choose a 10-Gb Ethernet, or 4/8 Gb Fibre Channel (or both) for upstream connections. In FC mode, Omni Ports provide convenient access to FC storage. The Omni Port technology that is provided on the G8264CS helps consolidate enterprise storage, networking, data, and management onto a manageable, efficient, and cost-effective single fabric.

For more information see *IBM System Networking RackSwitch G8264CS*, TIPS0970, which is available at this website:

https://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/tips0970.html?Open

2.4.2 Brocade IP routers and Layer 4 - 7 application delivery controllers

Brocade MLX series routers

As shown in Figure 2-10, the Brocade MLX Series of routers provides a rich set of high-performance IPv4, IPv6, and Multiprotocol Label Switching (MPLS) capabilities and advanced Layer 2 switching capabilities. The Brocade MLX series routers provide a high-density, highly available, and scalable IP network aggregation in the data center core that enables Layer 2 data center extension by using MPLS/VPLS/VLL.

Figure 2-10 Brocade MLX series routers

The Brocade MLX Series routers are available in 4-, 8-, 16-, and 32-slot options and are designed with a fully distributed, non-blocking architecture with up to 15.36 Tbps fabric capacity. This capacity provides packet forwarding rates of approximately 5 billion packets per second. The 32-slot chassis supports up to 1,536 1 GbE, 256 10 GbE, and 32 100 GbE wire-speed ports.

Brocade NetIron CES2000

As seen in Figure 2-11 on page 124, the Brocade NetIron CES 2000 Series is a family of compact 1-U, multi-service edge and aggregation switches that combine powerful capabilities with high performance and availability. The switches provide a broad set of advanced Layer 2, IPv4, IPv6, and MPLS capabilities in the same device.

Figure 2-11 Brocade NetIron CES2000

The Brocade NetIron CES 2000 Series are available in 24-port and 48-port 1 GbE configurations with two 10 GbE uplinks in Hybrid Fiber (HF) and RJ45 versions to suit versatile deployment needs.

Brocade ADX Series L4 - L7 application deliver controllers

The Brocade ServerIron ADX Series, as seen in Figure 2-12, enables high-speed application delivery by making use of a purpose-built architecture that is designed with high-core density and embedded application accelerators. With the support of advanced traffic management, an open application scripting engine, and extensible application programming interfaces (APIs), the Brocade ADX Series of application delivery switches optimizes service delivery.

Figure 2-12 Brocade ADX Series L4-L7 Application Deliver Controllers

The Brocade ADX Series provides Global Server Load Balancing (GSLB), which provides host and application health checks and directs new client connections to the correct data center location after a VM moves and then works in tandem with Brocade Application Resource Broker (ARB).

The ARB is a plug-in for vSphere and enables automated provisioning of VMs and simplifies the management of application resources and ensures optimal application performance by dynamically adding and removing application resources within globally distributed data centers. Brocade ARB provides these capabilities through real-time monitoring of application resource responsiveness, traffic load information, and infrastructure capacity information from server infrastructures.

The Brocade ServerIron ADX Series is available in 1 RU devices with up to 24 1-GbE ports and two 10-GbE uplinks, or a 4- and 8-slot chassis with the largest supporting up to 48 1-GbE ports and 16 10-GbE ports.

IBM 3759 switch family

As IT organizations look for better ways to build clouds and virtualized data centers, they are turning to high-performance networking solutions that increase flexibility through leading-edge technologies. IBM is partnering with Brocade, an industry-leading SAN Switch provider, to deliver converged data center solutions. Converging FC storage and regular Ethernet traffic on one platform allows administrators to reduce the number of adapters, simplify management, and protect storage investments by using the existing SAN infrastructure.

IBM delivers interoperability by providing end-to-end FCoE testing with the latest System x servers, Brocade VDX 6730 Converged switches, IBM System Storage SAN b-type and Brocade SAN switches, and IBM Storage. IBM System Networking OEMs the Brocade VDX 6730 Converged switches and offers the following products:

- Brocade VDX 6730-32 Converged Switch for IBM (3759-C32) 1-U model with 8 Gbps FC ports and 24 10-GbE Ethernet ports.
- Brocade VDX 6730-76 Converged Switch for IBM (3759-C76) 2-U model with 16 8-Gbps FC ports and 60 10-GbE Ethernet ports.
- Brocade VDX 6730 Converged switches connect to Fibre Channel SANs and FCoE, iSCSI, and NAS storage, which provides unified Ethernet storage connectivity options.
- Organizations can deploy the IBM 3759 for 10 GbE server access and 1 GbE Top-of-Rack aggregation with a broad set of Layer 2 features.
- Configuration with command-line interface (CLI) and integrated management with the Brocade Network Advisor V11 IP Pro version.
- Brocade Fabric Watch innovative switch health monitoring and critical switch component failure notification.

IBM 3759 switches are specifically designed to improve network usage, maximize application availability, increase scalability, and dramatically simplify network architecture in virtualized data centers. IBM 3759 switches with Brocade VCS Fabric technology enable organizations to revolutionize the design of data center networks and provide an intelligent foundation for cloud computing.

IBM 3759 switches provide a flexible choice for building an Ethernet fabric by using two switches initially and scaling to add switches as demand increases. IT managers can mix and match 1 G and 10 G fixed switches in an access-layer fabric by using the IBM 3759 6710 for cost-effective 1 GbE connectivity and the IBM 3759 and the IBM 3759 for high-bandwidth 10 GbE connectivity.

IBM 3759 converged switches with Brocade VCS fabric technology enable organizations to build data center Ethernet fabrics revolutionizing the design of Layer 2 networks and providing an intelligent foundation for cloud computing. Without disrupting classic Ethernet architectures, Brocade VCS technology enables organizations to preserve existing network designs and cabling and to gain active-active server connections without employing Spanning Tree Protocol (STP). For scale-out fabric architectures, Brocade VCS technology allows organizations to flatten network designs, provide VM mobility without network reconfiguration, and manage the entire fabric more efficiently.

For more information about the VCS fabric technology, see this website:

http://www.brocade.com/downloads/documents/white_papers/Introducing_Brocade_VCS_WP.pdf

Figure 2-13 shows the VDX Series switches.

Figure 2-13 IBM 3759 switches: Brocade reseller agreement versions

2.5 P750 and Hardware Management Console

In this section, we describe the P750 and Hardware Management Console (HMC).

2.5.1 P750

To implement our IBM SAN Volume Controller Stretched Cluster solution on LPM, the solution requires the use of two Power 750 Express server (8233-E8B). Figure 2-14 shows the server model that we implemented.

Figure 2-14 Power 750 Machine model

The Power 750 Express server (8233-E8B) supports up to four 6-core 3.3 GHz or four 8-core 3.0 GHz, 3.3 GHz, and 3.55 GHz POWER7 processor cards in a rack-mount drawer configuration. The POWER7 processors in this server are 64-bit, 6-core, and 8-core modules that are packaged on dedicated processor cards with 4 MB of L3 cache per core and 256 KB of L2 cache per core.

The Power 750 Express server supports a maximum of 32 DDR3 DIMM slots, eight DIMM slots per processor card. Memory features (two memory DIMMs per feature) supported are 8 GB, 16 GB, and 32 GB, and run at speeds of 1066 MHz. A system with four installed processor cards has a maximum memory capacity of 512 GB.

The Power 750 Express server provides great I/O expandability. For example, with 12X-attached I/O drawers, the system can have up to 50 PCI-X slots or up to 41 PCIe slots. This combination can provide over 100 LAN ports or up to 576 disk drives (over 240 TB of disk storage). Extensive quantities of externally attached storage and tape drives and libraries also can be attached.

The Power 750 Express system unit without I/O drawers can contain a maximum of eight small form factor (SFF) SAS disks or eight SFF SAS solid-state drives (SSDs), which provides up to 2.4 TB of disk storage.

All disks and SSDs are direct dock and hot-pluggable. The eight SAS bays can be split into two sets of four bays for greater AIX or Linux configuration flexibility. The system unit also contains a slimline DVD-RAM and a half-high media bay for an optional tape drive or removable disk drive.

Also available in the Power 750 Express system unit is a choice of quad-gigabit or dual-10 Gb integrated host Ethernet adapters. These native ports can be selected at the time of the initial order. Virtualization of these integrated Ethernet adapters is supported.

The front and rear views of the Power 750 Express are shown in Figure 2-15.

Figure 2-15 The front and rear views of the Power 750 Express system

For more information, see *IBM Power 750 and 755 (8233-E8B, 8236-E8C) Technical Overview and Introduction*, REDP-4638, which is available at this website:

http://www.redbooks.ibm.com/redpapers/pdfs/redp4638.pdf

2.5.2 HMC

The IBM HMC for pSeries provides a standard user interface for configuring and operating partitioned and symmetric multiprocessor (SMP) systems. The HMC supports the system with features that enable a system administrator to manage configuration and operation of partitions in a system and to monitor the system for hardware problems. It consists of a 32-bit Intel-based desktop PC with a DVD-RAM drive.

The HMC, which is a member of the IBM Systems Director platform management family, provides systems administrators a tool for planning, deploying, and managing IBM System p and IBM System i® servers. The IBM Systems Director platform management family provides IT professionals with the tools that they must better coordinate and manage all of their virtual and physical resources in the data center.

The HMC is an ordinary PC or 1-U Intel Server with a special serial card controller (for POWER4) or network card connected to the service processors (on POWER5) and the HMC software. The software is written in Java. One HMC can control up to 32 POWER systems. HMC is based on Linux but should be considered as an appliance system. Root access is disabled by default. HMC is used for defining and changing partitioning it is not required to run systems and can be detached or exchanged easily.

Partitioning data is kept on HMC and the machine. It can be restored from HMC to the frame. If the server is replaced, it also can be reread by exchanged HMC.

Beside being an interface to the logical partitioning, HMC also features some other tasks: it works as a central hardware control point, as a service focal point gathering and forwarding hardware errors (calling home), and as a central administration point, which provides virtual terminal access to all partitions on all servers connected. Keyboard Video Mouse (KVM) switches or serial switches are not needed if you have an HMC.

Management can be performed through the HMC GUI or through the CLI (after using Secure Shell to get into the HMC).

There is no serial RS-232 connection between the CWS and the HMC controlled servers or between the HMC and the CWS. Only the HMC has a serial connection to the HMC ports of the servers. For M/T 7039, more RS-422 connections are needed between the Bulk Power Controllers (BPC) and the HMC. Because the HMC has only two integrated RS-232 ports, more multiport adapters must be installed for these connections. An 8-port adapter can be used. This adapter supports RS-232 and RS-422 connections.

In terms of compatibility and interoperability with the other software for this solution in this book, the V7R7.4.0 HMC Build level 20110913.1 HMC version was implemented.

For more information about the HMC overview, architecture, and solution context that is specific to this Redbooks solution, see 1.6.2, “IBM Hardware Management Console” on page 55.

For more HMC reference information, see the following resources:

- HMC Related Technical Information

  http://www-304.ibm.com/webapp/set2/sas/f/hmcl/resources.html

- HMC Support and Downloads

  https://www-304.ibm.com/webapp/set2/sas/f/hmcl/home.html

2.6 Software

When a solution such as a SAN Volume Controller Stretched Cluster is implemented, it is important to provide availability for the application layer of the environment. This maximizes the benefit that can be derived from the storage infrastructure and the host operating systems.

There are many different software stacks that can achieve host availability for applications. However, in this book, we focus on the features that are provided by the IBM PowerVM, PowerHA, and LPM. This section describes this software and some other features that are useful when implementing a Stretched Cluster solution. In addition, this section provides an overview of the network, application, and database software that is implemented in this solution to provide context to real-world usage in customer environments.

Table 2-1 provides the software listing and release levels that were implemented in our solution. This is important information because it provides the software stack information that functions together optimally without interoperability issues because of conflicts between software versions.

Table 2-1 SAN Volume Controller Stretched Cluster solution on Power Systems Software™ stack

Software	Version Level
Brocade Fabric OS	7.0.2a
IBM Network Advisor	11.1.4
IBM 3759 Fabric Ethernet Network NOS Firmware	VDX 3.0.0a
IBM SAN Volume Controller/Storwize V7000	6.4.1.2
IBM Power Systems OS	AIX 7.1 TL7100-01-04
IBM p750 Server Firmware	AL730_099AL730_099
IBM HMC	V7R7.4.0 HMC Build level 20110913.1
IBM PowerVM	PowerVM Enterprise Edition
IBM VIOS	2.2.2.1 w/expansion pack. Fixpack:FP-26
IBM MPIO FCP Disk Device	1.0.0.23
IBM SDDPCM for AIX 7.1	2.6.3.2
IBM PowerHA	7.1.1.2 + FixPack APAR:IV25311
IBM DB2	AESE 9.7
IBM WebSphere Application Server	6.1ND + Supplements Install + Fixpack: 6.1.0-WS-WASEJB3-AixPPC64-FP0000045.pak

2.6.1 IBM Fabric OS

The solution in this book was implemented on Brocade Networking Infrastructure and Software. IBM Fabric OS (FOS) v7.0.2a is a maintenance release that is based on FOS v7.0.2. All hardware platforms and features that are supported in FOS v7.0.2 also are supported in FOS v7.0.2a.

Features

The FOS v7.0.2 release includes the following features and enhancements that are implemented in FOS v7.0.2:

- D_Port Enhancements

  D_Port link saturation capability: The ability to drive test traffic to saturate link usage close to line rate as part of D_Port test.

- Ability to assign individual ports of an optical ICL to logical switches.

  Before FOS v7.0.2 in Virtual Fabrics enabled environment, an entire ICL port (which includes four individual ports as part of a single QSFP) must be associated with a single logical switch. Starting with FOS v7.0.2, the individual ports within an optical ICL can be assigned to different logical switches. This feature is applicable to the DCX8510 platforms.

  Any individual port of a QSFP is part of the Base Switch. The remaining ports of that QSFP cannot be assigned to any other logical switch.

- Bottleneck detection enhancements: Decoupling of latency and congestion alerts.

In FOS v7.x releases before FOS v7.0.2, when users enabled bottleneck alerts, it enabled alerting for congestion and latency bottleneck conditions. Starting with FOS v7.0.2, users can choose to enable alerts only for latency bottleneck while not enabling alerts for congestion bottleneck or vice versa. Users still have the option to enable alerts for congestion and latency bottleneck conditions.

For more information, see *Brocade Fabric OS v7.0.2a Release Notes v1.0*, which is available at this website:

`ftp://service.boulder.ibm.com/storage/san/fos7/v7.0.2a_ReleaseNotes_v1.0.pdf`

2.6.2 IBM Network Advisor

The IBM Network Advisor V11 is a software management platform that unifies network management for SAN and converged networks. It provides a consistent user interface across Fibre Channel and FCoE over data center bridging (DCB), along with custom views and controls based on users' areas of specialization. The IBM Network Advisor offers flexibility in deployment and operational models, including traditional SAN deployments (Fibre Channel, IBM FICON® iSCSI, and NAS) and end-to-end convergence (FCoE, iSCSI, and NAS). IBM Network Advisor is licensed and deployed to manage SAN-only networks.

Although every network type has unique management requirements, most organizations face similar challenges managing their network environments, such as minimizing network downtime, reducing operational expenses, managing application service level agreements (SLAs), and providing robust security. Several tools are available to handle these requirements for different types of networks, yet no single tool provides cohesive management across all of them. To address this issue, the IBM Network Advisor provides comprehensive management for data, storage, and converged networks. It supports Fibre Channel SANs, including 16 Gbps Fibre Channel platforms, FICON, and FCoE networks. As a result, it provides end-to-end visibility across different network types within a single application. In addition, it supports comprehensive lifecycle management capabilities across different networks through a seamless and unified user experience.

SAN management feature enhancements

IBM Network Advisor 11.1.4 includes the following new SAN capabilities:

- Support for the following hardware products:
  - SAN24B-5 24 port 16Gb FC switch, and FC8-32E and FC8-48E 8G FC blades supported in 2499-416 and 2499-816 chassis.
  - Host Connectivity Manager: IBM Network Advisor includes HCM 3.0.3 version for adapter management.
  - ICL Support: Support for up to nine 2499-416 or 816 chassis full mesh connectivity that uses optical ICLs.
- Support for Dynamic Fabric Provisioning

  IBM Network Advisor adds support for Fabric Assigned worldwide name (FA-WWN). To simplify and accelerate server deployment and improve operational efficiency, FOS v7.0.1 provides FA-PWWN capability.

  This feature allows users to create a virtual WWN for a server instead of using the server‘s physical port WWN (PWWN) to create zoning and LUN mapping or masking. When a FA-PWWN capable server is attached to the SAN, this feature allows the fabric to assign this virtual WWN to that server. This feature requires servers to be using Brocade HBAs or adapters. Consult Brocade HBA and adapter driver documentation and Release Notes to confirm minimum requirements for this feature. For IBM Network Advisor support, consult IBM Network Advisor documentation and Release Notes.
- DCB Switches Support: IBM Network Advisor allows users to discover and manage DCB Switches from SAN tab.
- Host Views: Host views are supported on SAN tab. They allow users to discover host by using an HCM agent or a Virtual Machine Manager (VMM), such as vCenter Server.
- Virtualization: Network Advisor now allows users to discover VMs directly from the vCenter.
- Unified VLAN Management: Allows users to configure VLANs across all of the devices that are managed by Network Advisor through a single unified VLAN wizard.
- Unified ACL Management: Allows users to consistently define and deploy ACLs across different types of devices.

IBM Network Advisor also brings forward capabilities that are found in Data Center Fabric Manager (DCFM) for SAN management with the following other features and enhancements:

- Policy Monitor: Policy Monitor allows users to set up best practices for their network and run a test manually or on a scheduled basis to ensure the network is conforming to the best practices.
- Frame Monitor: Frame monitor provides a monitoring mechanism that allows configuring different thresholds for different type of frames and to report events whenever a count for particular frame type crosses its respective threshold.
- D-port: IBM Network Advisor allows users to run optical and electrical loopback and link traffic tests. IBM Network Advisor also provides digital diagnostic information about 10 G/16 G port optics, such as temperature, Rx/Tx power, bias current, and supply voltage.
- Bottleneck Detection: Bottleneck Detection allows users to configure parameters and view statistics. In addition, when a bottleneck is detected, IBM Network Advisor allows users to view the topology and receive information about the affected hosts.
- SAN Size Check: A message is displayed when the discovered port/switch/AG count exceeds the SAN size selected (small, medium, or large).

- MD5 Checksum Validation: MD5 Checksum validation is done before firmware files are imported into Repository.

Organizations can use the IBM Network Advisor to provide a management experience that offers better coordination between storage and data networking administrators for provisioning, troubleshooting, and reporting. IBM Network Advisor provides the following key capabilities:

- Intelligent dashboard: Presents an at-a-glance summary of all discovered b-type devices, including inventory and event summary information that is used to identify problem areas and help prevent network downtime.
- User management: Allows flexible definitions of administrator roles and responsibilities with robust role-based access control (RBAC) for SAN and FCoE management.
- Security management: Offers robust security administration by integrating with leading authentication, authorization, and accounting (AAA) solutions.
- Device configuration wizard: Used to configure and manage dynamically updated groups of devices.
- Flexible reporting: Features a powerful and flexible reporting framework with predefined asset reports.
- Automation: Simplifies firmware downloads and backup and restore of switch configurations across dynamic groups of switches to automate time-consuming activities and enable more rapid provisioning.
- Topology manager: Integrated with discovery and network mapping capabilities that provide Fibre Channel, FCIP tunnels, and FCR links views of managed devices with link usage and track changes support.
- Search tool: Enables fast location of devices based on multiple search criteria.
- FICON Configuration and Merge wizards: Simplifies fabric and switch configuration for FICON environments.

For more information, see *IBM Network Advisor v11.1.4 Release Notes*, which is available at this website

`ftp://index.storsys.ibm.com/san/networkadvisor/IBM_Network_Advisor_v11.1.4_Release Notes.pdf`

2.6.3 IBM 3759 Fabric Ethernet Network Networking Operating System

The IBM 3759 Fabric Ethernet Network Operating System (NOS) Firmware that was implemented for this solution is VDX - v3.0.0a.

Brocade NOS v3.0.0a is a major new release to support the introduction of the VDX 8770 switch, several new features, and improved scalability on all VDX platforms. Brocade VCS Fabric Technology provides a new level of bandwidth scale with the introduction of the VDX 8770 modular switch.

Additionally, NOS v3.0.0a provides greater design flexibility with its enhanced Layer 3 capabilities. The Layer 3 capabilities improve performance by clustering bandwidth and routing traffic more efficiently.

Brocade VCS Fabric technology brings intelligent, efficient routing to VCS fabrics. It enables highly elastic Layer 2 and Layer 3 domains with extremely efficient load balancing and multiple active Layer 3 gateways that are partnered with L2 ECMP and Brocade ISL Trunking. The result is more effective link usage that reduces overall cost, increases network resiliency resulting in greater application uptime, and a more flexible and agile network that helps organizations rapidly adapt to changing business conditions.

The addition of Layer 3 routing to VCS technology is the next logical step in the evolution of VCS fabrics. The addition of routing capabilities has effectively flattened the traditional access and aggregation layers into a single layer. In other words, the traditional three tier architecture is now flattened into two layers. Layer 3 in VCS is supported by both the fixed configuration and modular VDX platforms. Layer 3 support in the Brocade VCS Fabric includes the following features:

- Inter-VLAN routing in the VCS Fabric
- Router port to VE (Virtual Ethernet) routing capabilities in the VCS Fabric
- Router port to Router port routing within the VCS Fabric.

A summary of new features and enhancements NOS v3.0.0a includes the following new features for IBM 3759 products:

- IPv4 and IPv6 Management Services in VCS
- IPv4 unicast routing /Static Routes in VCS
- Open Shortest Path First (OSPF) in VCS
- VRRP and Brocade VRRP-E in VCS
- Layer 3 ACLs in VCS
- Layer 3 QoS Enhancements
- FIPS Snooping Bridge (FSB) Connectivity Support
- Zone Alias Names
- FCoE Over LAG

2.6.4 IBM Storwize V7000

In our Redbooks solution, we implemented one IBM Storwize V7000 each for Site A, and one for Site B. We are using this configuration as back-end storage disk and for the quorum disk in the Stretched Cluster solution. The software level is v6.4.1.2.

Important: The IBM Storwize V7000 cannot be configured as a Stretched Cluster. It is used as disk in this Stretched Cluster solution.

For more information, see 1.2, “Storwize V7000” on page 33, and 3.6, “IBM SAN Volume Controller using Stretched Cluster” on page 182.

2.6.5 IBM SAN Volume Controller

The solution in this book was implemented on IBM SAN Volume Controller that uses a software code level of 6.4.1.2.

IBM System Storage SAN Volume Controller version 6.4

IBM System Storage SAN Volume Controller version 6.4 can help with the following conditions:

- When storage requirements are growing at a fast rate.
- When the costs necessary to manage storage requirements are taking over the IT budget.

For more information, see the following resources:

- 1.3, “SAN Volume Controller” on page 38
- Chapter 3, “SAN Volume Controller Implementation” on page 149.
- V6.4.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller, which is available at this website:

  http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004111

2.7 IBM Power Systems operating system

Businesses today must maximize the return on investment in IT. Their IT infrastructure should have the flexibility to quickly adjust to changing business computing requirements and scale to handle ever expanding workloads, without adding complexity. But providing only flexibility and performance is not enough. The IT infrastructure also must provide rock-solid security and near-continuous availability while managing energy and cooling costs.

These are just some of the reasons why more businesses are choosing the AIX OS that is running on IBM systems that are designed with Power Architecture® technology. With its proven scalability, advanced virtualization, security, manageability, and reliability features, the AIX OS is an excellent choice for building an IT infrastructure. AIX also is the only OS that makes use of decades of IBM technology innovation that is designed to provide the highest level of performance and reliability of any UNIX operating system.

The newest version of IBM AIX Version 7/1 (AIX 7.1) is binary, which is compatible with previous versions of the AIX OS, including AIX 6, AIX 5L and even earlier versions of AIX. This means that applications that ran on earlier versions continue to run on AIX 7.1. IBM ensures that applications (whether written in-house or supplied by an application provider) run on AIX 7.1 if they currently run on AIX 6.1 or AIX 5L, without recompilations or modification. Even 32-bit applications from AIX V4.1, V4.2, and V4.3 run without recompilations.

For more information about binary compatibility, see this website:

http://www-03.ibm.com/systems/power/software/aix/compatibility/

AIX 7.1 is an open standards-based UNIX OS that complies with the Open Group’s Single UNIX Specification Version 4. AIX 7.1 runs on and makes use of systems that are based on POWER4, PPC970, POWER5, POWER6, and POWER7, including the latest generation of POWER7 based technology, the POWER7+™.

Most of the new features of AIX 7.1 are available on the earlier POWER processor-based platforms, but the most capability is delivered on systems that are built with the POWER6 and POWER7 processors. The AIX OS is designed for the IBM Power, System p, System i, System p5®, System i5®, eServer™ p5, eServer pSeries, and eServer i5 server product lines and the IBM BladeCenter blades that are based on Power Architecture technology.

AIX 7.1 extends the capabilities of the AIX OS to expand the vertical scalability of AIX to partitions with 256 processor cores and 1024 threads to handle the largest workloads. To support higher performance for large workloads, AIX 7.1 also includes new Terabyte segment support, which uses memory management capabilities of POWER7 processors that improve memory performance. This Terabyte segment capability is also included in AIX 6 at Technology Level 6, but is not automatically enabled on AIX 6.

The following new enhancements in AIX 7.1 provide continuous availability for your most critical applications and services that are running on Power Systems:

- Firmware Assisted Dump

  The introduction of the POWER6 processor-based systems allowed system dumps to be firmware assisted. When performing a firmware-assisted dump, system memory is frozen and the partition rebooted, which allows a new instance of the OS to complete the dump.

  Firmware-assisted dump is now the default dump type in AIX V7.1 when the hardware platform supports firmware-assisted dump. The traditional dump remains the default dump type for AIX V6.1, even when the hardware platform supports firmware-assisted dump.

  Firmware-assisted dump offers improved reliability over the traditional dump type by rebooting the partition and using a new kernel to dump data from the previous kernel crash.

  Firmware-assisted dump includes the following requirements:

  – A POWER6 processor-based or later hardware platform.
  – The LPAR must have a minimum of 1.5 GB memory.
  – The dump logical volume must be in the root volume group.
  – Paging space cannot be defined as the dump logical volume.

  In the unlikely event that a firmware-assisted system encounters a problem with running, the firmware-assisted dump is substituted by a traditional dump for this instance.

- User Key Enhancements

  AIX 7.1 allows for configuring the number of user storage keys. It also allows a mode where all hardware keys are dedicated to user keys. This helps in developing large applications to use more user keys for application-specific needs.

  Important: By dedicating all of the hardware keys to user keys, kernel storage keys are disabled. However, we do not recommend this configuration because the kernel storage keys cannot help debug the kernel memory problems if they are disabled. Review your application requirements and needs carefully and ensure that you have a thorough and accurate view of your application resource usage requirements.

  Table 2-2 lists the maximum number of supported hardware keys on different hardware platforms.

Table 2-2 Supported storage keys

Power Hardware Platform	Maximized supported hardware keys on AIX
P5++	4
P6	8
P6+	15
P7	31

- Cluster Data Aggregation Tool

  First Failure Data Capture (FFDC) is a technique that ensures that when a fault is detected in a system (through error checkers or other types of detection methods), the root cause of the fault is captured without the need to re-create the problem or run any sort of extended tracing or diagnostics program.

For more information about FFDC, see *IBM AIX Continuous Availability Features*, REDP-4367, which is available at this website:

http://www.redbooks.ibm.com/abstracts/redp4367.html

FFDC was enhanced to provide capabilities for quick analysis and root cause identification for problems that arise in workloads that span multiple systems. FFDC data is collected on each of the configured nodes by the Cluster Data Aggregation Tool.

The Cluster Data Aggregation Tool environment consists of a central node and remote nodes. The central node is where the Cluster Data Aggregation Tool is installed and run from. It hosts the data collection repository, which is a new file system that contains collection of data from multiple remote nodes. The remote nodes are where FFDC data is collected, which is AIX LPARs (AIX 6.1 TL3), VIOS (2.1.1.0 based on AIX 6.1 TL3), or HMC (V7 R 3.4.2). The central node must connect as an administrator user on the remote nodes. There is no need to install the Cluster Data Aggregation Tool on these remote nodes. For creating a secure connection, the SSH package should be installed on these nodes.

- Cluster Aware AIX

  AIX 7.1 also includes a new built-in clustering capability called Cluster Aware AIX. This new technology builds clustering technologies in the AIX base OS. This built-in clustering support provides commands and programming APIs to create a cluster from a group of AIX instances and provides kernel-based heartbeat, monitoring, and event infrastructure. This new infrastructure supports common device naming for storage devices across the cluster. While this new Cluster Aware AIX functionality is primarily intended to provide a reliable, scalable clustering infrastructure for products such as PowerHA SystemMirror and PowerVM, you can directly use the Cluster Aware AIX functionality facilitate management of scale-out computing environments.

  The Cluster Aware AIX (CAA) services help to create and manage a cluster of AIX nodes to build a highly available and ideal architectural solution for a data center. IBM cluster products such as Reliable Scalable Cluster Technology (RSCT) and PowerHA use these services. CAA services can assist in the management and monitoring of an arbitrary set of nodes or in running a third-party Cluster software.

  CAA services are a set of commands and services that the Cluster software can use to provide HA and disaster recovery support to external applications. The CAA services include the following broad classifications:

  – Cluster-wide event management

    The AIX Event Infrastructure allows event propagation across the cluster so that applications can monitor events from any node in the cluster.

  – Cluster-wide storage naming service

    When a cluster is defined or modified, the AIX interfaces automatically create a consistent shared device view across the cluster. A global device name, such as cldisk1, refers to the same physical disk from any node in the cluster.

  – Cluster-wide command distribution

    The **clcmd** command provides a facility to distribute a command to a set of nodes that are members of a cluster. For example, the **clcmd date** command returns the output of the **date** command from each node in the cluster.

– Cluster-wide communication

  Communication between nodes within the cluster is achieved by using multicasting over the IP-based network and by using storage interface communication through Fibre Channel and SAS adapters. A new socket family (AF_CLUST) is provided for reliable, in-order communication between nodes. When all network interfaces are lost, applications that use these interfaces still can run.

  The nodes that are part of the cluster should have common storage devices through the SAN or through the Serial-Attached SCSI (SAS) subsystems.

- SCTP component trace and RTEC adoption

  The AIX enterprise Reliability Availability Serviceability (eRAS) infrastructure defines a component definition framework. This framework supports the following distinct domains:

  – Runtime Error Checking (RTEC)
  – Component Trace (CT)
  – Component Dump (CD)

  The Stream Control Transmission Protocol (SCTP) implementation in AIX V7.1 and AIX V6.1 TL 6100-06 significantly enhances the adoption of the RAS component framework for the RTEC and CT domains. To that extent, the following new trace hooks are defined:

  – Event ID 6590 (0x659) with event label SCTP
  – Event ID 65a0 (0x65a) with event label SCTP_ERR

  The previously existing base component `sctp` of the CT and RTEC component tree is complemented by another subcomponent, `sctp_err`. The integration into the component trace framework enables the memory trace mode (private memory trace) and the user trace mode (system trace) for the base component and its new subcomponent.

- Cluster aware `perfstat` library interfaces

  IBM PowerHA is an HA solution for AIX that provides automated failure detection, diagnosis, application recovery, and node reintegration.

  It consists of the following components:

  – HA: The process of ensuring an application is available for use through the use of duplicated or shared resources.

  – Cluster multiprocessing: Multiple applications that are running on the same nodes with shared or concurrent access to the data.

  This HA solution demands the following important capabilities from the performance monitoring perspective:

  – The ability to collect and analyze the performance data of the entire cluster at the aggregate level (from any node in the cluster).

  – The ability to collect and analyze the performance data of an individual node in the cluster (from any node in the cluster).

  The `perfstat` application programming interface (API) is a collection of C programming language subroutines that run in the user space and use the `perfstat` kernel extension to extract various AIX performance metrics. Beginning with AIX V7.1 and AIX 6.1 TL06, the existing `perfstat` library is enhanced to support performance data collection and analysis for a single node or multiple nodes in a cluster. The enhanced perfstat library provides APIs to obtain performance metrics related to processor, memory, I/O, and others to provide performance statistics about a node in a cluster.

  The `perfstat` library also is updated with a new interface called `perfstat_cluster_total` (similar to the `perfstat_partion_total` interface) that provides cluster level aggregate data. A separate interface called `perfstat_node_list` also is added to retrieve the list of nodes available in the cluster.

New APIs (NODE interfaces) are available that return usage metrics related to a set of components or individual components specific to a remote node in a cluster.

AIX 7.1 also includes new virtualization capabilities that simplify the consolidation of older, AIX V5.3 environments. This new capability, which requires the purchase of the AIX 5.3 Workload Partitions for AIX 7.1 product, allows administrators to back up an existing LPAR that is running AIX 5.3 and restore it into an AIX 7.1 Workload Partition.

AIX 7.1 also includes new security features to improve and simplify security administration. For example, the new Domain Support in Role-Based Access Control is an enhancement to Role-Based Access Control (RBAC) that allows a security policy to restrict administrative access to a specific set of similar resources, such as a subset of the available network adapters. This allows IT organizations that host services for multiple tenants to restrict administrator access to only the resources that are associated with a particular tenant. Domains can be used to control access to Volume Groups, file systems, files, and devices.

FAIX 7.1 also includes new manageability enhancements, such as the AIX Profile Manager. The AIX Profile Manager can manage the configuration of AIX via XML profiles. This capability builds on the Runtime Expert capability that was introduced in AIX 6 Technology Level 4. This new management capability features an IBM Systems Director interface.

This AIX release underscores IBM's firm commitment to long-term UNIX innovations that deliver business value. This release of AIX continues the evolution of the UNIX OS that started in Austin, Texas, with AIX on the RT PC and the RISC Systems/6000 (RS/6000®) over 20 years ago.

In summary, AIX Version 7.1 introduces the following new features:

- Improved Scalability through Terabyte segment support
- Continuous availability enhancements
- Domain RBAC
- Workload partition enhancements
- Topas performance tool enhancements

In terms of compatibility with the other software for this solution in this book, the AIX 7.1TL7100-01-04 AIX version was implemented.

For more reference information that describes the enhancements that are provided by AIX 7.1, see *IBM AIX Version 7.1 Differences Guide*, SG24-7910, which is available at this website:

http://www.redbooks.ibm.com/redbooks/pdfs/sg247910.pdf

2.7.1 IBM server firmware

To avoid incidents, issues, and ongoing problems, it is necessary to proactively work to prevent incidents and reduce problems. High levels of availability cannot be achieved with a continual focus on preventive maintenance. The following key preventive maintenance recommendations are suggested for IBM Servers:

- Install an AIX recommended technology level at least once per year.
- Stay informed about new fixes by subscribing to email alerts for firmware, AIX, and HMC.
- Use Concurrent Firmware Maintenance (CFM) to deploy firmware updates on a running system without rebooting partitions or disturbing applications.

- Use the following standard IBM Power Systems support and preventive maintenance tooling:
  - Fix Level Recommendation Tool (FLRT):

    http://www-304.ibm.com/webapp/set2/sas/f/flrt3/use.html

    > **Important:** Based on the recommendations of the Fix Level Recommendation tool, and to ensure compatibility and interoperability with the other software and hardware in our solution, Firmware Level AL730_099 was implemented.

  - Fix Central:

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

  - System Storage Interoperation Center (SSIC):

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

  - IBM Call Home:

    http://pic.dhe.ibm.com/infocenter/tivihelp/v37r1/index.jsp?topic=%2Fcom.ibm.ia.doc_1.0%2Fic%2Fc_call_home.html

  - Electronic Service Agent™ (ESA):

    http://www-01.ibm.com/support/esa/

  - CEC Hardware Add Repair Maintenance (CHARM) to allow the system to stay in a running state during CEC component repair in hardware:

    http://www-03.ibm.com/systems/power/hardware/whitepapers/770_780_cec.html

- Monitor High Impact Pervasive (HIPER) and Parallel Environment (PE) Fixes regularly

  Receive all HIPERs to have ready at least monthly in case of need.

- Identify reviewers by area of expertise (for example, OS, Network, Storage).
- Pay particular attention to Cluster software maintenance.

  Mismatched Cluster software levels can lie dormant; these are often discovered only when a failover does not run properly.

2.8 IBM PowerVM: VIOS, NPIV, and SDDPCM/MPIO

PowerVM is a combination of hardware enablement and added value software. In this section, we describe the difference between the PowerVM editions and their capabilities. We also identify and describe the specific PowerVM edition and other PowerVM software versions (Virtual I/O Server, Live Partition Mobility, N_Port ID Virtualization, Subsystem Device Driver Path Control Module, and Multi-path I/O) that we implemented for the solution in this Redbooks publication.

For more information and an overview of the PowerVM software, key concepts, architecture, and solution context specific to this solution, see 1.8, “PowerVM technology” on page 66.

2.8.1 PowerVM Editions

This section provides information about the virtualization capabilities of PowerVM. The following versions of PowerVM are available, each suited for various purposes:

- PowerVM Express Edition

  PowerVM Express Edition is designed for customers looking for an introduction to more advanced virtualization features at a highly affordable price.

- PowerVM Standard Edition

  PowerVM Standard Edition provides advanced virtualization functionality for AIX, IBM i, and Linux operating systems. PowerVM Standard Edition is supported on all POWER processor-based servers and includes features designed to allow businesses to increase system usage.

- PowerVM Enterprise Edition

  PowerVM Enterprise Edition includes all the features of PowerVM Standard Edition plus two new industry-leading capabilities called Active Memory Sharing and LPM. It provides the most complete virtualization for AIX, IBM i, and Linux operating systems in the industry. It is possible to upgrade from the Standard Edition to the Enterprise Edition.

Table 2-3 outlines the functional elements of the available PowerVM editions.

Table 2-3 PowerVM Capabilities

PowerVM Editions	Express	Standard	Enterprise
PowerVM Hypervisor	Yes	Yes	Yes
Dynamic Logical Partitioning	Yes	Yes	Yes
Maximum Partitions	3 per server	254 per server	254 per server
Management	VM Control, IVM	VM Control, IVM, HMC	VM Control, IVM, HMC
Virtual I/O Server (VIOS)	Yes	Yes (Dual)	Yes (Dual)
Integrated Virtualization Manager	Yes	Yes	Yes
PowerVM Lx86	Yes	Yes	Yes
Suspend/Resume	No	Yes	Yes
N_Port ID Virtualization (NPIV)	Yes	Yes	Yes
Multiple Shared Processor Pools	No	Yes	Yes
Shared Storage Pools	No	Yes	Yes
Thin Provisioning	No	Yes	Yes
Active Memory Sharing	No	No	Yes
Live Partition Mobility (LPM)	No	No	Yes

For the solution in this book, we implemented PowerVM Enterprise Edition because LPM is required in the solution.

For more information, see the following Redbooks publications:

- *IBM PowerVM Getting Started Guide*, REDP-4815:
  http://www.redbooks.ibm.com/abstracts/redp4815.html
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940:
  http://www.redbooks.ibm.com/abstracts/sg247940.html
- *IBM PowerVM Best Practices*, SG24-8062:
  http://www.redbooks.ibm.com/abstracts/sg248062.html
- *IBM System p Advanced POWER Virtualization (PowerVM) Best Practices*, REDP-4194:
  http://www.redbooks.ibm.com/abstracts/redp4194.html

2.8.2 IBM Virtual I/O Server

This section describes the IBM Virtual I/O Server.

VIOS supported platforms

The Virtual I/O Server can run on any POWER5 or later server which has the PowerVM Standard feature enabled. Also supported are IBM BladeCenter POWER processor-based blade servers. With the PowerVM Standard Edition or the PowerVM Enterprise Edition, Virtual I/O Servers can be deployed in pairs to provide HA.

The VIOS software Level for the solution in this book is: 2.2.2.1 with expansion pack, fix pack FP-26.

For more information about Virtual I/O Server support for physical network and storage devices, see this website:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html

For more reference information about VIOS, see the IBM Information Center at this website:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/iphb1/iphb1_vios_virtualioserveroverview.htm

2.8.3 IBM LPM

The LPM, which is a licensed, chargeable feature offered on POWER6 or 7 based systems since 2007, enables the migration of an entire logical partition (active or from one system to another) for planned server maintenance, load balancing across multiple servers, and for energy conservation. In concept, it is similar to VMware VMotion.

The source and target systems must have access to the same network and SANs but need not be of the same type. The only requirement is they use POWER6 or POWER7 processors. Partitions that are to be relocated must be fully virtualized (that is, have no dedicated I/O adapters) although it is possible to use multi-pathing software to fail over to virtual adapters for the duration of the move.

Any sized partition can be moved; essentially, memory is copied asynchronously from one system to another to create a clone of a running partition, with “dirty” pages re-copied as necessary. When a threshold is reached (that is, when a high percentage of the pages is successfully copied across), the partition is moved to the target machine and any remaining pages are copied across synchronously. The agents that carry out the memory copying are nominated Virtual I/O Servers on each machine; a standard Ethernet network is used for data transmission.

For more information, see 1.8, “PowerVM technology” on page 66, and 4.11, “LPM” on page 262.

2.8.4 N_Port ID Virtualization

N_Port ID Virtualization (NPIV) is a Fibre Channel facility that allows multiple N_Port IDs to share a single physical N_Port. This allows multiple Fibre Channel initiators to occupy a single physical port, which eases hardware requirements in SAN design, especially where virtual SANs are called for. NPIV is defined by the Technical Committee T11 in the Fibre Channel - Link Services (FC-LS) specification.

Normally N_Port initialization includes the following process:

1. N_Port sends FLOGI to address 0xFFFFFE to obtain a valid address.
2. N_Port sends PLOGI to address 0xFFFFFC to register this address with the name server.
3. N_Port sends SCR to address 0xFFFFFD to register for state change notifications.

However, with NPIV, it might continue by using the following process:

1. N_Port sends FDISC to address 0xFFFFFE to obtain another address.
2. N_Port sends PLOGI to address 0xFFFFFC to register this other address with the name server/
3. N_Port sends SCR to address 0xFFFFFD to register for state change notifications.
4. Repeat FDISC/PLOGI/SCR for next address.

For more information, see 1.9, “PowerVM” on page 66, and 4.9.1, “High availability clusters” on page 258.

2.8.5 Subsystem Device Driver Path Control Module

Subsystem Device Driver Path Control Module (SDDPCM) is a loadable path control module that supports the multipath configuration environment on the IBM Storwize V7000, SAN Volume Controller, IBM TotalStorage Data Studio Family, and IBM BladeCenter S SAS RAID Controller Module (RSSM).

To support Multi-path I/O (MPIO) and NPIV in our solution, we implemented SDDPCM version 2.6.3.2.

For more information about and the context of SDDPCM as implemented in our solution, see 1.9.2, “Virtual I/O Server” on page 71.

2.8.6 MPIO

Multipath software layers can use the redundant paths to provide performance enhancing features. With MPIO, a device can be uniquely detected through one or more physical connections or paths that can be configured through priority settings.

A path-control module (PCM) provides the specific path management functions.

An MPIO-capable device driver can manage more than one type of target device. A PCM can support and manage one or more specific devices. Therefore, one device driver can be interfaced to multiple PCMs that control the I/O across the paths to each of the target devices.

For more information, see 1.9.2, “Virtual I/O Server” on page 71.

To support our solution, we implemented IBM MPIO FCP Disk Device V1.0.0.23.

2.9 PowerHA

PowerHA SystemMirror V7 represents the next generation of solutions for HA and disaster recovery. This offering integrates core clustering technology into the kernel and the OS, which enables our customers to be more productive and deliver higher quality IT services while enjoying a more robust and easier-to-manage clustering environment. The IBM PowerHA SystemMirror for AIX Standard Edition features Cluster Aware AIX, which provides the strategic foundation for Power Systems Software. It enables PowerHA to perform kernel level messaging, event management, and monitoring in coordination with AIX.

PowerHA features a Director-based management interface, a set of Smart Assists, and a number of other enhancements that make the solution more robust and self managing. For the solution in this book, we implemented PowerHA version 7.1.1.2 + FixPack APAR:IV25311 to ensure interoperability and compatibility with the software stack and hardware. In addition, this is the optimal version as part of this complete solution in this book (SAN Volume Controller Stretched Cluster, PowerVM, PowerHA, and LPM), based on where the solution fits based on customer requirements (clustered resource management, cross site clusters) for HA, as described in the Infrastructure Availability Continuum in 1.1.3, “Infrastructure availability continuum” on page 6.

2.9.1 PowerHA SystemMirror 7.1 deprecated features

PowerHA SystemMirror 7.1 has removed support for the following previously available features:

- IP address takeover (IPAT) via IP replacement
- Locally administered address (LAA) for hardware MAC address takeover (HWAT)
- Heartbeat over IP aliases
- `clcomdES` with the `/usr/es/sbin/cluster/etc/rhosts` directory is replaced by the Cluster Aware AIX (CAA) `clcomd` with the `/etc/cluster/rhosts` directory
- The following IP network types:
  - ATM
  - Fiber Distributed Data Interface (FDDI)
  - Token Ring
- The following point-to-point (non-IP) network types:
  - RS232
  - TMSCSI
  - TMSSA
  - Disk heartbeat (diskhb)
  - Multi-node disk heartbeat (mndhb)
- Two-node configuration assistant
- WebSMIT (replaced with the IBM Systems Director plug-in)
- Site support in this version

  Cross-site Logical Volume Manager (LVM) mirroring (available in PowerHA 7.1 SP3)
- IPV6 support in this version

2.9.2 PowerHA SystemMirror 7.1 new features

The new version of PowerHA uses much simpler heartbeat management. This method uses multicasting, which reduces the burden on the customer to define aliases for heartbeat monitoring. By default, it supports dual communication paths for most data center deployments by using the IP network and the SAN connections (available in 7.1 SP3 and later). These communication paths are done through the CAA and the central repository disk.

PowerHA SystemMirror 7.1 introduces the following features:

- SMIT panel enhancements
- The rootvg system event
- Systems Director plug-in
- Resource management enhancements:
  - StartAfter
  - StopAfter
- User-defined resource type
- Dynamic node priority: Adaptive failover
- More disk fencing by CAA
- New Smart Assists for the following products:
  - SAP NetWeaver 7.0 (2004s) SR3
  - IBM FileNet 4.5.1
  - IBM Tivoli Storage Manager 6.1
  - IBM Lotus Domino Server
  - SAP MaxDB v7.6 and 7.7
- The `clmgr` tool

  The `clmgr` tool is the new CLI with which an administrator can use a uniform interface to deploy and maintain clusters

For more reference information for PowerHA, see the following resources:

- *PowerHA for AIX Cookbook*, SG24-7739:

  http://www.redbooks.ibm.com/abstracts/sg247739.html
- IBM PowerHA SystemMirror for AIX:

  http://www-03.ibm.com/systems/power/software/availability/aix/resources.html
- IBM Information Center for PowerHA SystemMirror 7.1:

  http://pic.dhe.ibm.com/infocenter/aix/v7r1/index.jsp?topic=%2Fcom.ibm.aix.powerha.navigation%2Fpowerha.htm
- *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030:

  http://www.redbooks.ibm.com/abstracts/sg248030.html?Open

2.10 IBM DB2

For the purposes of providing a real-life use case and testing the SAN Volume Controller, PowerHA, and LPM aspects of the solution in this book, the software implementation also includes the IBM DB2 database software. For the best interoperability and compatibility with the solution software and hardware stack, DB2 9.7 Advanced Enterprise Server Edition for Linux, UNIX, and Windows was implemented.

DB2 is installed in both of the p750 servers in an LPAR on each server. The DB2 LPAR is moved from one server to another via LPM technology to explain, test, and show a real-life scenario wherein customers can use LPM for DB2 workloads that are hosted on IBM Power Systems LPARs for planned maintenance. IBM PowerHA can restart and manage the DB2 services through the SmartAssists functionality.

DB2 9.7 Advanced Enterprise Server Edition integrates key DB2, Optim™, and InfoSphere® technologies in a single edition for companies that must accelerate and simplify their business' use of data. This new edition includes the cost- and-time savings features of DB2 Enterprise Server Edition, the ease of administration and development of key Optim products, and the federation and replication technologies of InfoSphere.

Today's IT departments are increasingly challenged to reduce capital and labor costs and increase the value of information by simplifying access to critical business systems. IBM DB2 V9.7 Advanced Enterprise Server Edition brings together the key functionality leading companies must effectively manage their data. The new release includes the following features:

- Industry-leading pureXML® and Oracle compatibility with DB2 Enterprise Server Edition
- Integrated federation between Oracle and DB2 data sources allowing quick integration between your key applications and databases
- Reduced storage costs with DB2 Deep Compression
- Simplified administration, development, and tuning with Optim Database Administrator, Optim Development Studio, Optim Performance Manager, and DB2 Workload Management
- Flexible HA or disaster recovery with Q-Replication between two DB2 databases

For more reference information about DB2 9.7 Advanced Enterprise Server Edition, see this website:

http://www-01.ibm.com/common/ssi/rep_ca/7/897/ENUS210-377/ENUS210-377.PDF

For more information about High Availability and Disaster Recovery Options for DB2 for Linux, UNIX, and Windows, see *High Availability and Disaster Recovery Options for DB2 for Linux, UNIX, and Windows*, SG24-7363, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247363.html

2.11 IBM WebSphere Application Server

For the purposes of providing a real-life use case PowerHA and LPM aspects of the solution in this book, the software implementation also includes the IBM WebSphere Application Server software. For the best interoperability and compatibility with the overall solution software and hardware stack, IBM WebSphere Application Server v6.1 Network Deployment (ND)+ Supplements Install + Fixpack: 6.1.0-WS-WASEJB3-AixPPC64-FP0000045.pak was implemented.

WebSphere Application Server is installed in both of the p750 servers in an LPAR on each server. The WebSphere Application Server LPAR can be moved from one server to another via LPM technology wherein customers can use LPM for WebSphere Application Server workloads that are hosted on IBM Power Systems LPARs for planned maintenance. IBM PowerHA can restart and manage the WebSphere Application Server services through the SmartAssists functionality.

WebSphere Application Server V6.1 builds on and extends the proven runtime and systems management model of WebSphere Application Server version 6.0 and is focused on improvements in several key areas. The first is a focus on simplifying the development and deployment experience, providing several ease-of-use features to help accelerate time to value. Support for the Java 5 SDK gives developers access to new interfaces and data types that simplify application development. The increasingly robust Application Server Toolkit builds on its application assembly and deployment features by adding new application development tools, including wizard-driven development of Java Platform, Enterprise Edition and web services components. The Installation Factory tool enables you to build reusable customized installation packages that include product maintenance, enterprise applications, and customized configurations, greatly simplifying repetitive installations.

This release is also focused on improving the security, scalability, and availability of your environment. Several new features make it easier to secure your WebSphere Application Server environment, including the option to automatically secure administrative access during installation and new user repository options. Performance also is improved because of the new IBM Virtual Machine and performance enhancements in several application server components.

Version 6.1 also improves on version 6.0 by extending the WebSphere Application Server programming model, which enables you to develop new and innovative types of applications. These new types include support for Portlet applications and new types of web services applications. It also includes support for Session Initiation Protocol, which is increasingly used in rich media applications, such as instant messaging, voice over IP, and video conferencing.

This release also has an increased focus on administrative and usability enhancements. Many tasks were simplified in the administrative console and the wsadmin command-line tool introduced many new high-level commands that simplify the administrative scripting experience. These improvements, based on extensive research and customer feedback, improve the efficiency of managing and maintaining a WebSphere Application Server environment which allows you to focus your efforts on innovation.

For more information about WebSphere Application Server v6.1, see these websites:

- IBM Education Assistant for WebSphere software products:

  http://publib.boulder.ibm.com/infocenter/ieduasst/v1r1m0/index.jsp

- *WebSphere Application Server V6.1: System Management and Configuration*, SG24-7304:

  http://www.redbooks.ibm.com/redbooks/pdfs/sg247304.pdf

3

SAN Volume Controller Implementation

This chapter provides information about how this solution is implemented and includes the following topics:

- Test environment
- ADX: Application Delivery Controller
- IP networking configuration
- IBM System Networking Rack Switch G8264CS
- IBM FC SAN
- IBM SAN Volume Controller using Stretched Cluster
- SAN Volume Controller volume mirroring
- SAN Volume Controller I/O operations on mirrored volumes
- Read operations
- Write operations
- SAN Volume Controller quorum disk
- Quorum disk requirements and placement
- Automatic SAN Volume Controller quorum disk selection
- Failure scenarios in SAN Volume Controller Split Cluster configuration
- SAN Volume Controller Split Cluster configurations
- Fibre Channel settings for distance
- Back-end storage allocation to the SAN Volume Controller Cluster
- Volume allocation

© Copyright IBM Corp. 2013. All rights reserved.

3.1 Test environment

Figure 3-1 shows the environment on which this book is based.

Figure 3-1 IBM SAN and SAN Volume Controller Stretched Cluster solution

In the topics that follow, we show the products that we configured to create this solution. Where an IBM product does not exist, we chose the best available product in the marketplace at the time of this writing to complete the solution.

3.2 ADX: Application Delivery Controller

The Brocade ServerIron ADX provides virtual machine (VM) aware application delivery by using Global Server Load Balancing (GSLB) technology with Network Address Translation (NAT) and the Application Resource Broker (ARB) plug-in. This configuration offers seamless access for clients that are connecting to VMs that migrate between data centers. In our configuration, we use ServerIron ADX v12.4.0.

For each VM or group of VMs (also known as Real Servers) that are in a data center, a Virtual IP (VIP) address is created for client access. Each ServerIron ADX in each data center has a different VIP for the same set of Real Servers. In our setup, we have a VM (Real Server) ITSO_VM_1 with an IP address of 192.168.201.101. On the ADX in Data Center A, we create a VIP of 177.20.0.88 that we associated with ITSO_VM_1 while on the ADX in Data Center B, we created a VIP of 178.20.0.88 that we also associated with ITSO_VM_1. For Data Center B ADX configuration, we also disabled GSLB from seeing the associated ITSO_VM_1 Real Server configuration so in the GSLB view, the only Real Server that is online is the one seen in Data Center A.

Figure 3-2 shows the VIP to Real Server configuration.

Figure 3-2 VIP and Real Server configuration for ITSO_VM_1

GSLB enables a ServerIron ADX to add intelligence to authoritative Domain Name System (DNS) servers by serving as a proxy to the VMs and providing optimal IP addresses to the querying clients. As a DNS proxy, the GSLB ServerIron ADX evaluates the IP addresses in the DNS replies from the authoritative DNS server for which the ServerIron ADX is a proxy and places the best host address for the client at the top of the DNS response.

In this solution, the best host address is the VIP for the data center that the VM is currently active in. For example, if ITSO_VM_1 is active in Data Center A, GSLB directs clients to VIP 177.20.0.88, which uses NAT on the back-end to connect to ITSO_VM_1.

If ITSO_VM_1 performs a vMotion migration to Data Center B, the ARB vCenter plug-in automatically detects this migration and updates the GSLB configuration on the ADXs in both data centers to direct clients to access the VIP in Data Center B, 178.20.0.88.

> **Important:** Existing client connections to the Real Servers (VMs) are not affected with this configuration and client traffic can traverse the WAN interconnects for a period of time. There are other methods, such as route-injection that might be used to re-direct client requests more immediately. Also, upstream DNS caching might affect the period before new client requests are directed to the correct data center. To help mitigate this, the ServerIron ADX uses a DNS record TTL value of 10 seconds.

The following components are configured:

- VIP address assignment that outside clients use and the corresponding Real Server configuration that is behind the VIP
- GSLB configuration
- ARB server installation
- ADX registration in ARB plug-in
- VM mobility enablement in ARB plug-in within vCenter

3.2.1 VIP and real server configuration

For the ServerIron ADX in Data Center A, we create a VIP of 177.20.0.88 for the VM ITSO_VM_1 that is located at a real IP address of 192.168.201.101. In our example, we choose port 50 to bind for ITSO_VM_1.

Example 3-1 Data Center A: ServerIron ADX, VIP, and real server configuration

```
telnet@DC1-SLB1-ADX(config)#server real itso_vm_1 192.168.201.101
telnet@DC1-SLB1-ADX(config-rs-itso_vm_1)#source-nat
telnet@DC1-SLB1-ADX(config-rs-itso_vm_1)#port 50
telnet@DC1-SLB1-ADX(config-rs-itso_vm_1)#exit
telnet@DC1-SLB1-ADX(config)#server virtual dca-itso_vm_1 177.20.0.88
telnet@DC1-SLB1-ADX(config-vs-dca-itso_vm_1)#port 50
telnet@DC1-SLB1-ADX(config-vs-dca-itso_vm_1)#bind 50 itso_vm_1 50
```

For the ServerIron ADX in Data Center B, we create a VIP of 178.20.0.88 for the same VM, ITSO_VM_1. However, because the ITSO_VM_1 is active in Data Center A, we also configure `gslb-disable` on the real server configuration in Data Center B. This is so GSLB sees only that the real server (VM) is online in Data Center A and direct requests to that VIP.

Example 3-2 Data Center B: ServerIron ADX, VIP, and real server configuration

```
telnet@DC2-SLB1-ADX(config)#server real itso_vm_1 192.168.201.101
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#source-nat
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#port 50
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#gslb-disable
telnet@DC2-SLB1-ADX(config-rs-itso_vm_1)#exit
telnet@DC2-SLB1-ADX(config)#server virtual dcb-itso_vm_1 178.20.0.88
telnet@DC2-SLB1-ADX(config-vs-dcb-itso_vm_1)#port 50
telnet@DC2-SLB1-ADX(config-vs-dcb-itso_vm_1)#bind 50 itso_vm_1 50
```

3.2.2 GSLB configuration

The ServerIron ADX can act as proxy for local or remote DNS servers or be populated with host information to with which it can respond. In our setup, we configure the ServerIron ADX in Data Center A to respond with DNS request without needing to query another local or remote DNS server for the itso.com domain.

To do so, we must configure a VIP with the **dns-proxy** command for DNS clients to access. Additionally, we must define a list of hosts and corresponding IP addresses for the itso.com domain, as shown in Example 3-3.

Example 3-3 Configuration ServerIron ADX in Data Center A as a DNS server

```
telnet@DC1-SLB1-ADX(config)#server virtual dns-proxy 177.20.0.250
telnet@DC1-SLB1-ADX(config-dns-proxy)#port dns
telnet@DC1-SLB1-ADX(config-dns-proxy)#exit
telnet@DC1-SLB1-ADX(config)#gslb dns zone itso.com
telnet@DC1-SLB1-ADX(config-gslb-dns-itso.com)#host-info itso_vm_1 50
telnet@DC1-SLB1-ADX(config-gslb-dns-itso.com)#host-info itso_vm_1 ip-list
177.20.0.88
telnet@DC1-SLB1-ADX(config-gslb-dns-itso.com)#host-info itso_vm_1 ip-list
178.20.0.88
```

Next, we configure the ServerIron ADX in Data Center A for GSLB, as shown in Example 3-4.

Example 3-4 ServerIron ADX in Data Center A: GSLB configuration

```
telnet@DC1-SLB1-ADX(config)#gslb protocol
telnet@DC1-SLB1-ADX(config)#gslb site DataCenterA
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterA)#weight 50
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterA)#si DC1-SLB1-ADX 192.168.1.2
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterA)#gslb site DataCenterB
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterB)#weight 50
telnet@DC1-SLB1-ADX(config-gslb-site-DataCenterB)#si DC2-SLB1-ADX 192.168.1.3
```

Next, we configure the ServerIron ADX in Data Center B for GSLB, as shown in Example 3-5.

Example 3-5 ServerIron ADX in Data Center B: GSLB configuration

```
telnet@DC2-SLB1-ADX(config)#gslb protocol
telnet@DC2-SLB1-ADX(config)#gslb site DataCenterA
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterA)#weight 50
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterA)#si DC1-SLB1-ADX 192.168.1.2
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterA)#gslb site DataCenterB
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterB)#weight 50
telnet@DC2-SLB1-ADX(config-gslb-site-DataCenterB)#si DC2-SLB1-ADX 192.168.1.3
```

We can use the **show gslb dns detail** command to view the status of our configuration. In Example 3-6 on page 154, we see the itso.com zone with two VIPs. Although there is one Active Binding for each VIP, only VIP 177.20.0.88 (corresponding to Data Center A) is Active because the Real Server definition for ITSO_VM_1 behind VIP 178.20.0.88 was initially disabled.

Example 3-6 Checking our GSLB status using show gslb dns detail

```
telnet@DC1-SLB1-ADX(config)#show gslb dns detail

ZONE: itso.com

ZONE: itso.com
HOST: itso_vm_1:
(Global GSLB policy)
GSLB affinity group: global
                                            Flashback    DNS resp.
                                            delay        selection
                                            (x100us)     counters
                                            TCP  APP     Count (%)
* 177.20.0.88    : cfg v-ip    ACTIVE N-AM      0    0    ---
                  Active Bindings: 1
                  site: DataCenter1, weight:  50, SI: DC1-SLB1-ADX (192.168.1.2)
                  session util:   0%, avail. sessions: 7999944
                  preference: 128
* 178.20.0.88    : cfg v-ip    DOWN   N-AM     --   --    ---
                  Active Bindings: 1
                  site: DataCenter2, weight:  50, SI: DC2-SLB1-ADX (192.168.1.3)
                  session util:   0%, avail. sessions: 31999728
                  preference: 128

telnet@DC1-SLB1-ADX(config)#
```

3.3 IP networking configuration

There are several IP networking components that are used within this solution. Although end-to-end, data center network design is outside the scope of this Redbooks publication, we review the configuration of our switches as seen in the context of our setup. Figure 3-3 shows the IP networking areas that we review.

Figure 3-3 High-level IP network architecture

3.4 IBM System Networking Rack Switch G8264CS

IBM System Networking Rack Switch G8264CS is an end-to-end Fibre Channel over Ethernet (FCoE) solution with full Fibre Channel (FC) fabric services. This configuration is an alternative solution to the existing Brocade VDX. Figure 3-4 shows the front view of the IBM Rack Switch G8264CS.

Detailed information about the switch can be found at this website:

http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=ca&infotype=an&appname=iSource&supplier=897&letternum=ENUS113-019

Figure 3-4 IBM Rack Switch G8264CS: Front view

We describe the following IP networking components:

- Layer 2 switches

  The AIX hosts connect directly to Layer 2 switches. The Layer 2 network can be any standards-compliant switch, such as any IBM RackSwitch family switch or the Brocade VDX series of switches. In a three-tier network design, there often is an “Edge” made up of Top of Rack switches that connect to a higher-density Aggregation switch over Layer 2 that is running a version of Spanning Tree Protocol (STP). With higher-density Top of Rack switches, scalable logical switch capabilities, such as stacking or Virtual Switch Clustering, or passive patch paneling in each rack to a higher density End of Row switch, the overall data center network might be collapsed into a flat Layer 2 “Edge” that connects directly into the IP Core over Layer 2 or Layer 3.

- IP Core

  The Layer 2 network connects into the IP core, which then connects out to the wide area network (WAN) Edge/Internet. In our configuration, we connect from the Layer 2 switches to the IP core over Layer 2. The IP core that is made up of Brocade MLXe devices provides aggregation and routing capabilities between the various subnetworks.

- Brocade MLXe/CER that is acting as the data center interconnect or Provider Edge (PE) router.

  For Layer 2 extension, we use standards-compliant L2 VPN technology that uses MPLs/VPLS/VLL and we review the configuration in our setup.

Our actual lab configuration is shown in Figure 3-5.

Figure 3-5 BM SAN Volume Controller Stretched Cluster lab architecture

A closer view of the Data Center A (Site 1) connection is shown in Figure 3-6.

Figure 3-6 BM SAN Volume Controller Stretched Cluster - Data Center A topology

Our lab configuration includes the following configurations and components:

- We have an IBM Power 750 server that is AIX host Power 750 A. This host has two 10 GbE ports and four 16 Gbps FC ports.
- The 16 Gbps FC ports are connected to two separate IBM SAN768B-2 chassis that form two separate air-gapped FC SAN fabrics.
- The two 10 GbE ports are connected to two separate 10 GbE VDX switches. There are a total of four VDX switches in a Layer 2 network. These four VDX switches are clustered together by using VCS fabric technology, which to other network entities makes them appear to be a single logical switch.
- VDX switches are connected over Layer 2 to two MLXe routers that are acting as the IP Core. The MLXe routers provide routing between the various subnetworks and aggregate the various connections.
- The Brocade ADX GSLB/SLB is connected to the MLXe.

- The data center interconnect links (which are Brocade CES routers) are connected to the MLXes. The Brocade CES routers provide the Layer 2 VPN extension capabilities that use VLL, acting as PE routers.
- The Brocade MLXe IP Core also is connected to simulated clients that might come in outside the data center from the Internet.
- Also connected to the Brocade MLXe are two 10 GbE FCIP connections from each of the IBM SAN768B chassis. IBM SAN768B-A1 has two links connected to DC1-Core1-MLX4: one link forms the Private SAN FCIP tunnel and the second link forms the Public SAN FCIP tunnel. Similarly, IBM SAN768B-A2 has two links connected to DC1-Core2-MLX4.

3.4.1 Layer 2 switch configuration

This section describes the Layer 2 switch configuration.

Layer 2 loop prevention

Within the Layer 2 network, loop prevention is required. Typically, a version of STP is used, usually PVRSTP, MSTP, or RSTP, depending on the environment. The following guidelines should be considered when STP is used:

- Configure the Bridge Priority to ensure proper Root Bridge placement, typically at one of the Aggregation switches that is closest to the Core.
- Configure switch-to-switch links as point-to-point and edge ports that are connecting to the AIX hosts as edge ports.
- Consider enabling BPDU guard on edge ports that are facing the AIX hosts to prevent loops (be aware that this configuration eliminates access to all LPARs behind that port if it is triggered).

VLAN configuration

VLANs also must be created for the various networks that vSphere and VMs are using. As outlined in Chapter 1, “Introduction” on page 1 we have four VLANs that must be defined all on switches in the Layer 2 network. This configuration is shown in Example 3-7.

Example 3-7 Defining VLANs on a switch

```
S1_RB4(config)# int vlan 700
S1_RB4(config-Vlan-700)# description Management
S1_RB4(config-Vlan-700)# int vlan 701
S1_RB4(config-Vlan-701)# description VM_Traffic
S1_RB4(config-Vlan-701)# int vlan 702
S1_RB4(config-Vlan-702)# description vMotion
S1_RB4(config-Vlan-702)# int vlan 703
S1_RB4(config-Vlan-703)# description Fault_Tolerant
```

The edge ports on the switch that are connected to the AIX host and all traditional Ethernet links between switches also must be configured as VLAN trunk ports that allow VLANs 700 - 703. A configuration example is shown in Example 3-8.

Example 3-8 Configuring switch to ESXi host and switch-to-switch links to carry VLANs

```
S1_RB4(config)# int ten 4/0/15
S1_RB4(conf-if-te-4/0/15)# switchport
S1_RB4(conf-if-te-4/0/15)# switchport mode trunk
S1_RB4(conf-if-te-4/0/15)# switchport trunk allowed vlan add 700-703
```

Link Aggregation Group configuration

In our configuration, AIX is using “route based on originating virtual port” load balancing so we do not require configuring a Link Aggregation Group (LAG) between the two switches that are connected to S1_RB3 and S1_RB4.

However, we are creating a LAG between S1_RB1 and S1_RB2 to the two MLXs that are acting as the IP Core. This configuration is possible because our four Layer 2 switches are in a single logical cluster, S1-VCS-1, that uses VCS fabric technology. The two MLXs also are clustered together by using Multi-Chassis Trunking (MCT) technology.

To create a port-channel group on the switches in S1-VCS-1, we first put all four ports in the same port channel, as shown in Example 3-9 and Example 3-10.

Example 3-9 Switch S1_RB1: Putting ports connected to the Core MLXs in channel-group 2

```
S1_RB1(config)# int ten 1/0/14
S1_RB1(conf-if-te-1/0/14)# channel-group 2 mode active type standard
S1_RB1(conf-if-te-1/0/14)# int ten 1/0/15
S1_RB1(conf-if-te-1/0/15)# channel-group 2 mode active type standard
```

Example 3-10 Switch S1_RB2 - Putting ports connected to the Core MLXs in channel-group 2

```
S1_RB1(config)# int ten 2/0/14
S1_RB1(conf-if-te-2/0/14)# channel-group 2 mode active type standard
S1_RB1(conf-if-te-2/0/14)# int ten 2/0/15
S1_RB1(conf-if-te-2/0/15)# channel-group 2 mode active type standard
```

Now that the port channels are created on both switches, we configure the interface definitions at the port-channel level, as shown in Example 3-11 and Example 3-12.

Example 3-11 Switch S1_RB1 - Interface Port-Channel 2 configuration

```
S1_RB1(config)# int port-channel 2
S1_RB1(config-Port-channel-2)# switchport
S1_RB1(config-Port-channel-2)# switchport mode trunk
S1_RB1(config-Port-channel-2)# switchport trunk allowed vlan add 701-703
```

Example 3-12 Switch S1_RB2 - Interface Port-Channel 2 configuration

```
S1_RB2(config)# int port-channel 2
S1_RB2(config-Port-channel-2)# switchport
S1_RB2(config-Port-channel-2)# switchport mode trunk
S1_RB2(config-Port-channel-2)# switchport trunk allowed vlan add 701-703
```

Other considerations

The maximum transfer unit (MTU) size can be increased on all Layer 2 ports to support jumbo frames. This can help with traffic efficiency when IP-based storage or Live Partition Mobility (LPM) are used.

> **Important:** Make sure that the MTU size on all interfaces within the Layer 2 domain, including the AIX host configuration, is the same; otherwise, fragmentation can occur.

Other basic considerations when you are setting up your network include the following tasks:

- Hardening the switch with RBAC (e.g. RADIUS/TACACS+), user accounts, disabling Telnet, and restricting login to certain VLANs or IP ranges.

- Configuring SNMP, syslog, or sFlow monitoring servers.
- Configuring a Network Time Protocol (NTP) server.
- Enabling LLDP, Tivoli Continuous Data Protection for Files, or additional network visibility.
- Configuring quality of service for specific traffic classes or trusting the 802.1p settings that are sent by AIX.
- Configuring security and Access Control Lists (ACLs).

3.4.2 IP Core (MLXe) configuration

VLAN configuration

We begin by defining the VLANs that the MLXe are carrying traffic over. These VLANs are the same four VLANS (700 - 703), that we defined in the Layer 2 switch network. However, the MLXe also is carrying the following FCIP traffic that we defined:

- VLAN 705 - Fabric A - Private SAN traffic
- VLAN 706 - Fabric A - Public SAN traffic
- VLAN 707 - Fabric B - Private SAN traffic
- VLAN 708 - Fabric B - Public SAN traffic

Within our lab topology, there are three different types of network traffic access patterns for traffic that is passing through the MLXe. We show the following example of the different network traffic types and interfaces that are applicable on DC1-Core1-MLXe4:

- Traffic that needs access only to the internal network. These are ports that are connected to the Layer 2 network (eth 1/4, eth 2/4), the MCT link between the two MLXes (eth 1/1, eth 2/1), and the data center interconnect link to the CES (eth 2/5). Traffic types that fall in this category are the Management, LPM, and Fault Tolerant traffic. Additionally, while the Management traffic might need to be routed to a different subnetwork (hence the creation of a virtual interface in that VLAN), the LPM and Fault Tolerant traffic never should be need to be routed.
- Traffic that needs access to the internal network and also the external client traffic (eth 1/6) and to the ServerIron ADX GSLB/SLB (eth 1/7). This is the VM Traffic, which also might need to be routed.
- FCIP traffic that requires access to directly connected IBM SAN768B-2 (eth 1/3, eth 2/6), the MCT link between the two MLXes (eth 1/1, eth 2/1), and the data center interconnect link to the CES (eth 2/5).

An example of creating these VLANs and tagging the appropriate interfaces is shown in Example 3-13.

Example 3-13 Creating VLANs and virtual routing interfaces on DC1-Core1-MLXe4

```
telnet@DC1-Core1-MLXe(config)#vlan 700 name I-MGMT
telnet@DC1-Core1-MLXe(config-vlan-700)#tag eth 1/1 eth 1/4 eth 2/1 eth 2/4 to 2/5
telnet@DC1-Core1-MLXe(config-vlan-700)#router-interface ve 70

telnet@DC1-Core1-MLXe(config-vlan-700)#vlan 701 name I-VM_Traffic
telnet@DC1-Core1-MLXe(config-vlan-701)#tag ethe 1/1 eth 1/4 eth 1/6 to 1/7 eth 2/1
eth 2/4 to 2/5
telnet@DC1-Core1-MLXe(config-vlan-701)#router-interface ve 71

telnet@DC1-Core1-MLXe(config-vlan-701)#vlan 702 name I-vMotion
telnet@DC1-Core1-MLXe(config-vlan-702)#tag eth 1/1 eth 1/4 eth 2/1 eth 2/4 to 2/5
```

```
telnet@DC1-Core1-MLXe(config-vlan-702)#vlan 703 name I-Fault_Tolerant
telnet@DC1-Core1-MLXe(config-vlan-703)#tag ethe 1/1 eth 1/4 eth 1/6 to 1/7 eth 2/1
eth 2/4 to 2/5

telnet@DC1-Core1-MLXe(config-vlan-703)#vlan 705 name I-FCIP-Priv-FabA
telnet@DC1-Core1-MLXe(config-vlan-705)#untag eth 1/3
telnet@DC1-Core1-MLXe(config-vlan-705)#tag eth 1/1 eth 2/1 eth 2/5

telnet@DC1-Core1-MLXe(config-vlan-705)#vlan 706 name I-FCIP-Pub-FabA
telnet@DC1-Core1-MLXe(config-vlan-706)#untag eth 2/6
telnet@DC1-Core1-MLXe(config-vlan-706)#tag ethe 1/1 ethe 2/1 ethe 2/5

telnet@DC1-Core1-MLXe(config-vlan-706)#vlan 707 name I-FCIP-Priv-FabB
telnet@DC1-Core1-MLXe(config-vlan-707)#tag ethe 1/1 ethe 2/1 ethe 2/5

telnet@DC1-Core1-MLXe(config-vlan-707)#vlan 708 name I-FCIP-Pub-FabB
telnet@DC1-Core1-MLXe(config-vlan-708)#tag ethe 1/1 ethe 2/1 ethe 2/5
```

The connection to the data center interconnects, DC1-DCI1-CER and DC1-DCI2-CER, are a Layer 2 link. In this configuration, the MLXes are acting as the Customer Edge routers while the CER are configured with MPLS and VLL acting as the Provider Edge routers.

VRRPe: Layer 3 Gateway configuration

Each VLAN that requires routing had a virtual router interface created (specifically, the Management VLAN has Virtual Ethernet 70 created), and the LPM Traffic VLAN has Virtual Ethernet 71 created within those VLANs. Those virtual interfaces must be configured with an IP address to route traffic. We are using Virtual Router Redundancy Protocol-Extended (VRRPe) to provide active-active Layer 3 gateways for our LPMs on each of those virtual routing interfaces.

A VRRPe instance (VRID) is created with a single VIP that VMs use as their gateway address. Within the VRID, one or more router interfaces also are configured as the actual paths that traffic passes through. With VRRPe, there are router interfaces within a VRID that can route traffic instead of switching through a single designated Master interface, in the case of regular VRRP.

Example 3-14 shows how to enable VRRPe on a router.

Example 3-14 Enable VRRPe protocol on a router

```
telnet@DC1-Core1-MLXe(config)#router vrrp-extended
telnet@DC1-Core1-MLXe(config-vrrpe-router)#exit
telnet@DC1-Core1-MLXe(config)#
```

The process to configure VRRPe on interface Virtual Ethernet 70 includes the following steps:

1. An IP address must be configured for interface Virtual Ethernet 70. In this case, we use 192.168.200.250/24.
2. We configure VRRPe and choose an arbitrary VRID value of 70.

   In the context of VRRPe, all VRRPe interfaces are designated as backup interfaces, there is no one Master interface.
3. The VIP of the VRID also is defined, in this case, 192.168.200.1.
4. We enable short-path-forwarding, which allows any VRRPe interface to route traffic instead of having to switch to a designated Master interface.

5. We activate the VRRPe configuration.

Example 3-15 shows how to configure VRRPe on interface Virtual Ethernet 70 on DC1-Core1-MLXe.

Example 3-15 Configuring VRRPe on DC1-Core1-MLXe4, interface Virtual Ethernet 70

```
telnet@DC1-Core1-MLXe(config)#interface ve 70
telnet@DC1-Core1-MLXe(config-vif-70)#port-name I-Internal-Mgmt
telnet@DC1-Core1-MLXe(config-vif-70)#ip address 192.168.200.250/24
telnet@DC1-Core1-MLXe(config-vif-70)#ip vrrp-extended vrid 70
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#backup
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#advertise backup
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#ip-address 192.168.200.1
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#short-path-forwarding
telnet@DC1-Core1-MLXe(config-vif-70-vrid-70)#activate
```

A similar configuration is done on DC1-Core2-MLXe4, and the two routers in Data Center B: DC2-Core1-MLXe16 and DC2-Core2-MLXe4. A different or unique IP address for each virtual interface on those routers is selected, though the VRRPe VRID and VIP remains the same.

We show an example configuration of DC1-Core2-MLXe4 in Example 3-16.

Example 3-16 Configuring VRRPe on DC1-Core2-MLXe4, interface Virtual Ethernet 70

```
telnet@DC1-Core2-MLXe(config)#interface ve 70
telnet@DC1-Core2-MLXe(config-vif-70)#port-name I-Internal-Mgmt
telnet@DC1-Core2-MLXe(config-vif-70)#ip address 192.168.200.251/24
telnet@DC1-Core2-MLXe(config-vif-70)#ip vrrp-extended vrid 70
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#advertise backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#ip-address 192.168.200.1
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#short-path-forwarding
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#activate
```

A third example configuration of DC2-Core1-MLXe16 is shown in Example 3-17.

Example 3-17 Configuring VRRPe on DC2-Core1-MLXe16, interface Virtual Ethernet 70

```
telnet@DC1-Core2-MLXe(config)#interface ve 70
telnet@DC1-Core2-MLXe(config-vif-70)#port-name I-Internal-Mgmt
telnet@DC1-Core2-MLXe(config-vif-70)#ip address 192.168.200.252/24
telnet@DC1-Core2-MLXe(config-vif-70)#ip vrrp-extended vrid 70
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#advertise backup
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#ip-address 192.168.200.1
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#short-path-forwarding
telnet@DC1-Core2-MLXe(config-vif-70-vrid-70)#activate
```

MCT Trunking configuration

The two MLXes are also configured together in a Multi-Chassis Trunk (MCT) cluster. The first task we must do is create a LAG of two ports between DC1-Core1-MLX4 and DC1-Core2-MLX4 over ports eth 1/1 and eth 2/1 to use as the MCT Inter-Chassis Link (ICL). We give an example of a static LAG configuration from the DC1-Core1-MLX4 chassis in Example 3-18; similar configuration can be done on DC1-Core2-MLX4.

Example 3-18 MLXe: Creating a static LAG

```
telnet@DC1-Core1-MLXe(config)#lag "ICL" static id 1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#ports eth 1/1 eth 2/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#primary-port 1/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#deploy
telnet@DC1-Core1-MLXe(config-lag-ICL1)#port-name "ICL" ethernet 1/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#port-name "ICL" ethernet 2/1
telnet@DC1-Core1-MLXe(config-lag-ICL1)#int eth 1/1
telnet@DC1-Core1-MLXe(config-if-e10000-1/1)#enable
```

After the LAG is created and is between the two MLXes, we configure a VLAN and a virtual router interface to carry MCT cluster communication traffic, as shown in Example 3-19.

Example 3-19 Creating a VLAN for MCT cluster communication traffic

```
telnet@DC1-Core1-MLXe(config)#vlan 4090 name MCT_SESSION_VLAN
telnet@DC1-Core1-MLXe(config-vlan-4090)#tag ethe 1/1
telnet@DC1-Core1-MLXe(config-vlan-4090)#router-interface ve 1
```

Important: Only eth 1/1 must be tagged because it is the primary port in the LAG.

Next, we configure the virtual routing interface with an IP address that the two MLXes use for MCT cluster communication, as shown in Example 3-20.

Example 3-20 Assigning an IP address to a virtual routing interface

```
telnet@DC1-Core1-MLXe(config)#interface ve 1
telnet@DC1-Core1-MLXe(config-vif-1)#port-name MCT-Peer
telnet@DC1-Core1-MLXe(config-vif-1)#ip address 1.1.1.1/24
```

Finally, we configure the cluster. For more information about each step, see *Brocade MLX Series and Brodcade NetIron XMR Hardware Installation Guide*, which is available at this website:

http://www.brocade.com/downloads/documents/product_manuals/B_NetIron/NetIronXMRMLXMLXe_05200b_InstallGuide.pdf

Example 3-21 Completing the MCT Cluster configuration

```
telnet@DC1-Core1-MLXe(config)#cluster MCT_CLUSTER 1
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#rbridge-id 10
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#session-vlan 4090
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#member-vlan 100 to 999
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#icl ICL ethernet 1/1
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#peer 1.1.1.2 rbridge-id 20 icl
ICL
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#deploy
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#client DC1-SLB1-ADX
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#rbridge-id
100
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#client-inter
face ethernet 1/7
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#deploy
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-DC1-SLB1-ADX)#exit
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#client VCS1
```

```
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#rbridge-id 200
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#client-interface
ethernet 1/4
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#deploy
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-VCS1)#exit
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER)#client CER_DCI
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-CER_DCI)#rbridge-id 300
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-CER_DCI)#client-interface
ethernet 2/5
telnet@DC1-Core1-MLXe(config-cluster-MCT_CLUSTER-client-CER_DCI)#deploy
```

Other considerations

The Brocade MLXe routers are high-performance routers that are capable of handling the entire Internet routing table with advanced MPLS/VPLS/VLL features and other high-end Service Provider-grade capabilities. However, it is beyond the scope of this book to describe IP network design, routing protocols, and so on.

Other references

For more information about the Brocade MLXe, see *Brocade MLX Series and NetIron Family YANG Guide supporting r05.4.00a*, which is available at this website:

http://www.brocade.com/downloads/documents/product_manuals/B_NetIron/Brocade_05400a_YangGuide.pdf

3.4.3 Data Center Interconnect configuration

The Layer 2 network must be extended from Data Center A to Data Center B to support VM mobility. This can be done through the use of a Layer 2 VPN technology that uses standards-based MPLS/VPLS/VLL technology. This is often a service that is provided by the Service Provider.

However, there might be situations where it can be beneficial to extend the Layer 2 VPN deeper into the data center, or perhaps there the two data centers are connected point-to-point over dark fiber and MPLS can be used for advanced QoS control or other purposes. The Brocade MLXe chassis and Brocade NetIron CER 1 RU switches both support MPLS/VPLS/VLL capabilities.

It is outside of the intended scope of this book to describe this technology in-depth because it is fairly complex, though we provide the configurations between two of the CER interconnects in our lab setup for reference in Example 3-22 and Example 3-23 on page 167.

Example 3-22 DC1-DCI1-CER configuration

```
router ospf
 area 0
!
interface loopback 1
 ip ospf area 0
 ip address 80.80.80.10/32
!
interface ethernet 2/1
 enable
 ip ospf area 0
 ip ospf network point-to-point
```

```
 ip address 200.10.10.1/30
!
interface ethernet 2/2
 enable
!
router mpls

 mpls-interface e1/1
  ldp-enable

 mpls-interface e2/1
  ldp-enable

 vll DCIvl700 700
  vll-peer 90.90.90.10
  vlan 700
   tagged e 2/2

 vll DCIvl701 701
  vll-peer 90.90.90.10
  vlan 701
   tagged e 2/2

 vll DCIvl702 702
  vll-peer 90.90.90.10
  vlan 702
   tagged e 2/2

 vll DCIvl703 703
  vll-peer 90.90.90.10
  vlan 703
   tagged e 2/2

 vll DCIvl705 705
  vll-peer 90.90.90.10
  vlan 705
   tagged e 2/2

 vll DCIvl706 706
  vll-peer 90.90.90.10
  vlan 706
   tagged e 2/2

 vll DCIvl707 707
  vll-peer 90.90.90.10
  vlan 707
   tagged e 2/2

 vll DCIvl708 708
  vll-peer 90.90.90.10
  vlan 708
   tagged e 2/2
```

Example 3-23 DC2-DCI1-CER configuration

```
router ospf
 area 0
!
interface loopback 1
 ip ospf area 0
 ip address 90.90.90.10/32
!
interface ethernet 2/1
 enable
 ip ospf area 0
 ip ospf network point-to-point
 ip address 200.10.10.2/30
!
interface ethernet 2/2
 enable
!
router mpls

 mpls-interface e1/1
  ldp-enable

 mpls-interface e2/1
  ldp-enable

 vll DCIvl700 700
  vll-peer 80.80.80.10
  vlan 700
   tagged e 2/2

 vll DCIvl701 701
  vll-peer 80.80.80.10
  vlan 701
   tagged e 2/2

 vll DCIvl702 702
  vll-peer 80.80.80.10
  vlan 702
   tagged e 2/2

 vll DCIvl703 703
  vll-peer 80.80.80.10
  vlan 703
   tagged e 2/2

 vll DCIvl705 705
  vll-peer 80.80.80.10
  vlan 705
   tagged e 2/2

 vll DCIvl706 706
  vll-peer 80.80.80.10
  vlan 706
   tagged e 2/2
```

```
vll DCIvl707 707
 vll-peer 80.80.80.10
 vlan 707
  tagged e 2/2

vll DCIvl708 708
 vll-peer 80.80.80.10
 vlan 708
  tagged e 2/2
```

Other references

For more information about the Brocade CER, see *Brocade MLX Series and NetIron Family YANG Guide supporting r05.4.00a*, which is available at this website:

http://www.brocade.com/downloads/documents/product_manuals/B_NetIron/Brocade_05400a_YangGuide.pdf

3.5 IBM FC SAN

We assume that the reader is familiar with general FC SAN design and technologies. Typical best practice SAN design includes servers and storage that is connecting into dual, redundant fabrics. In our lab configuration, we implement a redundant fabric design that uses two IBM SAN768B-2 chassis at each data center site. Each IBM SAN768B-2 is equipped with the following components:

- IBM FC 8 Gbps FCIP Extension blade in Slot 1
- IBM FC 16 Gbps 48-port blade in Slot 8

IBM SAN Volume Controller Stretched Cluster requires the following types of SAN:

- Public SAN: Dedicated for host attachment, storage system attachment, IBM SAN Volume Controller, and links for Metro or Global Mirror.
- Private SAN: Dedicated for IBM SAN Volume Controller node-to-node communication.

Each IBM SAN768B-2 splits into two logical chassis to implement segregated Private and Public SANs on the same chassis. Figure 3-7 shows the various Public and Private SAN connections.

Figure 3-7 Public and Private SAN connections at each data center site

Our actual lab environment is shown in Figure 3-8.

Figure 3-8 IBM SAN Volume Controller Stretched Cluster lab topology

3.5.1 Creating the logical switches

We configure a total of four fabrics, each with two different logical switches.

Figure 3-9 shows the SAN port topology in Data Center A.

Figure 3-9 Data Center A: SAN port topology

Figure 3-10 shows the SAN port topology in Data Center B.

Figure 3-10 Data Center B: SAN port topology

We create logical switches on the SAN768B-2s and map them according to Table 3-1.

Table 3-1 Fabric to Physical Switch to Logical Switch mappings

Fabric Name	Physical Switch	Logical Switch	LS #	Ports
Fabric-Public-1	SAN768B-2_A1	Public_A1	111	1/12, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9, 8/12, 8/13
	SAN768B-2_B1	Public_B1		1/12, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9, 8/12, 8/13
Fabric-Public-2	SAN768B-2_A2	Public_A2	113	1/12, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9, 8/12, 8/13
	SAN768B-2_B2	Public_B2		1/12, 8/2, 8/4, 8/5, 8/6, 8/7, 8/8, 8/9, 8/12, 8/13
Fabric-Private-1	SAN768B-2_A1	Private_A1	112	1/22, 8/3
	SAN768B-2_B1	Private_B1		1/22, 8/3
Fabric-Private-2	SAN768B-2_A2	Private_A2	114	1/22, 8/3
	SAN768B-2_B2	Private_B2		1/22, 8/3

Port 1/12 corresponds to one of the FCIP tunnels on 1/xge0 while port 1/22 corresponds to one of the FCIP tunnels on 1/xge1.

We show two examples of creating a logical switch: once through the CLI and again through IBM Network Advisor.

For more information about Virtual Fabrics and creation, see *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg246116.html?Open

Example 3-24 shows the creation of the logical switch Public_A1 on SAN768B-2_A1.

Example 3-24 Creating the logical switch Public_A1

```
SAN768B-2_A1:FID128:admin> lscfg --create 111
About to create switch with fid=111. Please wait...
Logical Switch with FID (111) has been successfully created.

Logical Switch has been created with default configurations.
Please configure the Logical Switch with appropriate switch
and protocol settings before activating the Logical Switch.

SAN768B-2_A1:FID128:admin> lscfg --config 111 -slot 1 -port 12
This operation requires that the affected ports be disabled.
Would you like to continue [y/n]?: y
Making this configuration change.  Please wait...
Configuration change successful.
Please enable your ports/switch when you are ready to continue.

SAN768B-2_A1:FID128:admin> lscfg --config 111 -slot 8 -port 2
This operation requires that the affected ports be disabled.
Would you like to continue [y/n]?: y
Making this configuration change.  Please wait...
Configuration change successful.
Please enable your ports/switch when you are ready to continue.

SAN768B-2_A1:FID128:admin> lscfg --config 111 -slot 8 -port 4-9
This operation requires that the affected ports be disabled.
Would you like to continue [y/n]?: y
Making this configuration change.  Please wait...
Configuration change successful.
Please enable your ports/switch when you are ready to continue.

SAN768B-2_A1:FID128:admin> setcontext 111
Please change passwords for switch default accounts now.
Use Control-C to exit or press 'Enter' key to proceed.

Password was not changed. Will prompt again at next login
until password is changed.
switch_111:FID111:admin> switchname Public_A1
Done.
switch_111:FID111:admin>
switch_111:FID111:admin> setcontext 128
Please change passwords for switch default accounts now.
Use Control-C to exit or press 'Enter' key to proceed.

Password was not changed. Will prompt again at next login
until password is changed.
SAN768B-2_A1:FID128:admin> setcontext 111
Public_A1:FID111:admin> switchshow
switchName:     Public_A1
switchType:     121.3
switchState:    Offline
switchMode:     Native
switchRole:     Disabled
switchDomain:   1 (unconfirmed)
```

```
switchId:       fffc01
switchWwn:      10:00:00:05:33:b5:3e:01
zoning:         ON (ITSO_Public1)
switchBeacon:   OFF
FC Router:      OFF
Allow XISL Use: ON
LS Attributes:  [FID: 111, Base Switch: No, Default Switch: No, Address Mode 0]

Index Slot Port Address Media  Speed  State       Proto
=====================================================
  12    1   12   01fcc0   --    --   Offline     VE  Disabled
 194    8    2   01cf40   id    N16   In_Sync     FC  Disabled
 196    8    4   01cec0   id    N16   In_Sync     FC  Disabled
 197    8    5   01ce80   id    N16   In_Sync     FC  Disabled
 198    8    6   01ce40   id    N16   In_Sync     FC  Disabled
 199    8    7   01ce00   id    N16   In_Sync     FC  Disabled
 200    8    8   01cdc0   id    N16   In_Sync     FC  Disabled
 201    8    9   01cd80   id    N16   In_Sync     FC  Disabled
 204    8   12   01ccc0   id    N16   In_Sync     FC  Disabled
 205    8   13   01cc80   id    N16   In_Sync     FC  Disabled
Public_A1:FID111:admin> switchenable
Public_A1:FID111:admin> switchshow
switchName:     Public_A1
switchType:     121.3
switchState:    Online
switchMode:     Native
switchRole:     Subordinate
switchDomain:   2
switchId:       fffc01
switchWwn:      10:00:00:05:33:b5:3e:01
zoning:         ON (ITSO_Public1)
switchBeacon:   OFF
FC Router:      OFF
Allow XISL Use: ON
LS Attributes:  [FID: 111, Base Switch: No, Default Switch: No, Address Mode 0]

Index Slot Port Address Media  Speed  State       Proto
=====================================================
  12    1   12   01fcc0   --    --   Online      VE  VE-Port
10:00:00:05:33:97:a5:01 "Public_B1" (upstream)
 194    8    2   01cf40   id     N8   Online      FC  F-Port
50:05:07:68:01:40:b1:3f
 196    8    4   01cec0   id     N8   Online      FC  F-Port
50:05:07:68:02:10:00:ef
 197    8    5   01ce80   id     N8   Online      FC  F-Port
50:05:07:68:02:20:00:ef
 198    8    6   01ce40   id     N8   Online      FC  F-Port
50:05:07:68:02:10:00:f0
 199    8    7   01ce00   id     N8   Online      FC  F-Port
50:05:07:68:02:20:00:f0
 200    8    8   01cdc0   id     N8   Online      FC  F-Port
50:05:07:68:02:10:05:a8
 201    8    9   01cd80   id     N8   Online      FC  F-Port
50:05:07:68:02:10:05:a9
```

```
 204    8   12   01ccc0   id     N8   Online      FC  F-Port  1 N Port + 4 NPIV
public
 205    8   13   01cc80   id     N8   Online      FC  F-Port  1 N Port + 4 NPIV
public
```

In our next example, we create the logical switch Public_B1 on SAN768B-2_B1 by using IBM Network Advisor. First we find the Chassis Group that represents the discovered physical switches for SAN768B-2_B1. We can right-click the switch, then go to **Configuration** → **Logical Switches**, as shown in Figure 3-11.

Figure 3-11 Entering the Logical Switches configuration menu

Important: In Figure 3-11, we created most of the logical switches and discovered fabrics in this configuration. However, we are re-creating Fabric-Public-1 as an example. We already created logical switch Public_A1 in Data Center A that is a single-switch fabric that we name Fabric-Public-1A for now.

We now must create Public_B1 in Data Center B and then the corresponding FCIP tunnels for this fabric to merge.

In the Logical Switches window, we check to make sure that the Chassis we selected is the SAN768B-2_B1. Then we click **Undiscovered Logical Switches** in the right-side window and click **New Switch**, as shown in Figure 3-12 on page 175.

Figure 3-12 Creating a new Logical Switch on SAN768B-2_B1

In the New Logical Switch window, we set the Logical Fabric ID to 111. This configuration is chassis-local and does not have to be the same as the Public_A1 switch that we eventually merge with, but we set it the same to remain consistent, as shown in Figure 3-13.

Figure 3-13 Setting the Logical Fabric ID

Next, we select the Switch tab and name the logical switch we create Public_B1, as shown in Figure 3-14 on page 176. Make sure that the domain ID of the logical switch is different than the logical switch Public_A1. If the names are the same, logical switches issue an error message of domain conflict and do not merge while the FCIP tunnels are created.

Figure 3-14 Setting the switch name for Public_B1

Now that the logical switch is created, we move the ports that we want from SAN768B-2_B1 to Public_B1, as shown in Figure 3-15.

Figure 3-15 Selecting the ports from SAN768B-2_B1 to move to logical switch Public_B1

In the next window, we select **Start** to begin the process. After some time, we receive a message that indicates the process was completed successfully, as shown in Figure 3-16 on page 177.

Figure 3-16 Successfully creating Public_B1

The new switch is not yet discovered, so we select the Discovery menu and start a discovery process on SAN768B-2_B1 to add the new logical switch as a monitored switch in the fabric. The IP address for SAN768B-2_B1 is a previously discovered IP address so we select it, as shown in Figure 3-17.

Figure 3-17 Re-discovering SAN768B-2_B1 to add Public_B1

In the Fabric Discovery window, we name the new fabric Fabric-Public-1B for now, as shown in Figure 3-18.

Figure 3-18 Discovering Fabric-Public-1B

We select that we only want to discover and monitor Public_B1 and not the default SAN768B-2_B1 switch, as seen in Figure 3-19.

Figure 3-19 Selecting to monitor only Public_B1

Fabric-Public-1B with Public_B1 should now be visible in the SAN view, as seen in Figure 3-20.

Figure 3-20 Fabric-Public-1B successfully discovered

3.5.2 Creating FCIP tunnels

In this section, we show an example of creating the FCIP tunnel connectivity through the CLI.

For more information, see *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation*, SG24-7544, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247544.html?Open

An FCIP Tunnel or FCIP Trunk is a single, logical ISL, or VE_port. On the IBM FC 8 Gbps FCIP Extension blade, an FCIP tunnel can have one or up to eight circuits. A circuit is an FCIP connection between two unique IP addresses.

The process that is used to manually create an FCIP Tunnel includes the following steps:

1. Create an IP interface on the physical Ethernet port. This is done in the default switch context.
2. Validate connectivity between the SAN768B-2 chassis IP interfaces by using ping.
3. Within the logical switch to which the VE_port belongs, create an FCIP tunnel off the IP interface that was created in the default switch context.

First we create the IP interface on Public_A1. This is created by using interface 1/xge1, which VE_port 1/12 can use. We create an IP interface with an IP address of 192.168.76.10, netmask 255.255.255.0, with an MTU of 1500 bytes, as shown in Example 3-25.

Example 3-25 Creating the IP interface on Public_A1

```
SAN768B-2_A1:FID128:admin> portcfg ipif 1/xge1 create 192.168.76.10 255.255.255.0
1500
Operation Succeeded
```

We create a similar interface but with an IP address of 192.168.76.20/24 on Public_B1, as shown in Example 3-26.

Example 3-26 Creating the IP interface on Public_B1

```
SAN768B-2_B1:FID128:admin> portcfg ipif 1/xge1 create 192.168.76.20 255.255.255.0
1500
Operation Succeeded
```

Next, we perform some validation tests from Public_A1 to make sure that connectivity through these interfaces works, as shown in Example 3-27.

Example 3-27 Validating IP connectivity between Public_A1 and Public_B1

```
SAN768B-2_A1:FID128:admin> portcmd --ping 1/xge1 -s 192.168.76.10 -d 192.168.76.20
Pinging 192.168.76.20 from ip interface 192.168.76.10 on 1/xge1 with 64 bytes of
data
Reply from 192.168.76.20: bytes=64 rtt=2ms ttl=20
Reply from 192.168.76.20: bytes=64 rtt=0ms ttl=20
Reply from 192.168.76.20: bytes=64 rtt=0ms ttl=20
Reply from 192.168.76.20: bytes=64 rtt=0ms ttl=20

Ping Statistics for 192.168.76.20:
        Packets: Sent = 4, Received = 4, Loss = 0 ( 0 percent loss)
        Min  RTT = 0ms, Max RTT = 2ms Average = 0ms

SAN768B-2_A1:FID128:admin> portcmd --traceroute 1/xge1 -s 192.168.76.10 -d
192.168.76.20
Traceroute to 192.168.76.20 from IP interface 192.168.76.10 on 1/xge1, 30 hops max
 1  192.168.76.20  0 ms  0 ms  0 ms
Traceroute complete.
```

Now that basic IP connectivity is established, we change our logical switch context to 111 where VE_port 1/12 is to create the FCIP tunnel. We set a minimum bandwidth of 622 Kbps and turn on compression with standard settings, as shown in Example 3-28.

> **Important:** FastWrite is not needed because IBM SAN Volume Controller uses a different algorithm to improve transfer over distance.
>
> Also, IPsec is supported and can be turned on, if needed.

Example 3-28 Creating the FCIP tunnel on Public_A1

```
SAN768B-2_A1:FID128:admin> setcontext 111
Public_A1:FID111:admin> portcfg fciptunnel 1/12 create 192.168.76.20 192.168.76.10
-b 622000 -B 1000000 -c 1
Operation Succeeded

Public_A1:FID111:admin> portcfgshow fciptunnel 1/12
-------------------------------------------
Tunnel ID: 1/12
   Tunnel Description:
   Compression: On (Standard)
   Fastwrite: Off
   Tape Acceleration: Off
   TPerf Option: Off
```

```
   IPSec: Disabled
   QoS Percentages: High 50%, Med 30%, Low 20%
   Remote WWN: Not Configured
   Local WWN: 10:00:00:05:33:b5:3e:01
   Flags: 0x00000000
   FICON: Off

Public_A1:FID111:admin> portcfgshow fcipcircuit 1/12
   -------------------------------------------
   Circuit ID: 1/12.0
      Circuit Num: 0
      Admin Status: Enabled
      Connection Type: Default
      Remote IP: 192.168.76.20
      Local IP: 192.168.76.10
      Metric: 0
      Min Comm Rt: 622000
      Max Comm Rt: 1000000
      SACK: On
      Min Retrans Time: 100
      Max Retransmits: 8
      Keepalive Timeout: 10000
      Path MTU Disc: 0
      VLAN ID: (Not Configured)
      L2CoS: (VLAN Not Configured)
      DSCP:  F:  0 H:  0 M:  0 L:  0
      Flags: 0x00000000
Public_A1:FID111:admin>
```

Finally, we configure an FCIP tunnel on Public_B1 with the same settings and validate that the tunnel was established, as shown in Example 3-29.

Example 3-29 Creating a FCIP tunnel on Public_B1 and validating establishment

```
SAN768B-2_B1:FID128:admin> setcontext 111
Public_B1:FID111:admin> portcfg fciptunnel 1/12 create 192.168.76.10 192.168.76.20
-b 622000 -B 1000000 -c 1
Operation Succeeded

Public_B1:FID111:admin> portshow fcipcircuit all
--------------------------------------------------------------------------------
 Tunnel Circuit  OpStatus  Flags    Uptime  TxMBps  RxMBps ConnCnt CommRt  Met
--------------------------------------------------------------------------------
 1/12   0 1/xge1  Up      ---4--s    1m27s    0.00    0.00    1   622/1000  0
--------------------------------------------------------------------------------
 Flags: circuit: s=sack v=VLAN Tagged x=crossport 4=IPv4 6=IPv6
                 L=Listener I=Initiator

Public_B1:FID111:admin>
```

3.6 IBM SAN Volume Controller using Stretched Cluster

The SAN Volume Controller code that was used for this book is based on code level 6.4.1.2 (build 75.0.1211301000). The back-end storage that we used was on a Storwize V7000 that was running a code level of 6.4.1.2 (build 75.0.1211301000). The third site, the quorum storage, was on a Storwize V7000 that was running a code level of 6.4.1.2 (build 75.0.1211301000).

For more information and a full list of supported extended Quorum devices, see this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004111

Figure 3-21 shows the components that are used for the SAN Volume Controller Stretched Cluster and the SAN connectivity.

Figure 3-21 SAN Volume Controller Stretched Cluster diagram with SAN connectivity

This book does not include more information about the physical installation nor the initial configuration. Rather, this book is intended to supplement the *SVC V6.3.0 Configuration Guidelines for Extended Distance Split-System Configurations for IBM System Storage SAN Volume Controller*, S7003701, which is available at this website:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

We assume that the reader is familiar with the major concepts of SAN Volume Controller clusters such as nodes, I/O groups, managed disks (MDisks), and quorum disks. If you are not, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

For more information, see this website:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp

3.7 SAN Volume Controller volume mirroring

Stretched Cluster uses SAN Volume Controller volume mirroring functionality. As shown in Figure 3-22, by using volume mirroring, you can create one volume with two copies of MDisk extents; there are not two volumes that contain the same data.

Figure 3-22 describes a mirrored volume. The two data copies should be in different MDisk groups. Thus, volume mirroring can minimize the effect to volume availability if one or more MDisks fails. The re-synchronization between both copies is incremental; SAN Volume Controller SAN Volume Controller starts the re-synchronization process automatically. A mirrored volume has the same functionality and behavior as a standard volume. In the SAN Volume Controller software stack, volume mirroring is below the cache and copy services. Therefore, FlashCopy, Metro Mirror, Global Mirror have no awareness that a volume is mirrored. Everything that can be done with a volume also can be done with a mirrored volume, including migration, expand and shrink, and so on. Like a standard volume, each mirrored volume is owned by one I/O group with a preferred node. Thus, the mirrored volume goes offline if the whole I/O group goes offline. The preferred node performs all IO operations, reads, and writes. The preferred node can be set manually.

Important: While a mirrored copy of a volume is created, if `-syncrate` is not specified, expansion of a volume is not possible until both the copies are in sync state. Use the `lsvdisksyncprogress` command to check the sync progress.

Figure 3-22 Mirrored Volumes on the IBM SAN Volume Controller

3.7.1 Volume mirroring prerequisites

The three quorum disk candidates keep the status of the mirrored volume. The last status and the definition of primary and secondary volume copy (for read operations) are saved there. Thus, an active quorum disk is required for volume mirroring. To ensure data consistency, SAN Volume Controller disables mirrored volumes if there is no longer any access to any quorum disk candidate. Therefore, quorum disk availability is an important point with volume mirroring and split I/O group configuration. Furthermore, it is required to allocate bitmap memory space prior usage of volume mirroring.

Use the **chiogrp** command, as shown in the following example:

```
chiogrp -feature mirror -size memory_size io_group_name|io_group_id
```

The volume mirroring grain size is fixed at 256 KB (for example, one bit of the synchronization bitmap represents 256 KB of virtual capacity). Thus, bitmap memory space of 1 MB is required for each 2 TB of mirrored volume capacity. Example 3-30 shows the memory capacity on the IBM SAN Volume Controller Stretched Cluster for IOgrp 0.

Example 3-30 Memory capacity on the IBM SAN Volume Controller Stretched Cluster for IOgrp 0

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsiogrp 0
id 0
name ITSO_SVC_SPLIT
node_count 2
vdisk_count 54
host_count 12
flash_copy_total_memory 20.0MB
flash_copy_free_memory 20.0MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 20.0MB
mirroring_total_memory 20.0MB
mirroring_free_memory 19.2MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
maintenance no
compression_active no
accessible_vdisk_count 54
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

3.8 SAN Volume Controller I/O operations on mirrored volumes

There are two architectural behaviors with the SAN Volume Controller that must be considered for SAN Volume Controller mirrored volumes in a SAN Volume Controller Stretched Cluster configuration: preferred node and primary copy.

3.8.1 Preferred node

The preferred node performs all write de-staging to both copies of the volume mirror. Additionally, for hosts with multipath drivers that support SAN Volume Controller preferred node, the host directs all read and write activity to the preferred node. For these hosts, the preferred node performs all read caching and write cache de-staging. For hosts that do not support SAN Volume Controller preferred node, the node that is selected for read and write operations is determined by the host multipath driver.

The preferred node is queried by supporting multipath drivers through the SCSI **`Report Target Port Groups`** command.

> **Important:** AIX hosts support SAN Volume Controller preferred node. Therefore, multipathing software IBM Subsystem Device Driver Device Specific Module (SDDDSM) or IBM Subsystem Device Driver Path Control Module (SDDPCM) must be installed and configured.

3.8.2 Primary copy

The primary copy of the volume mirror is used for all read operations. When volumes are created, the first copy that is created becomes the primary copy. The primary copy can be changed after the initial volume is created without incurring any interruption in I/O processing.

3.9 Read operations

Volume mirroring uses a single read algorithm with one copy that is designated as the primary. SAN Volume Controller reads the data from the primary copy and does not automatically distribute the read requests across both copies. The first copy that is created becomes the primary by default, which can be changed by using the **`chvdisk -primary`** command. The secondary copy is read from only if the primary copy is unavailable. This is independent of which node the host is reading from. Because of this behavior, for optimal performance the primary copy of the mirror should be in the same failure domain as the host that is accessing the mirrored volume. The host should be accessing the mirrored volume through the node that also is in the same failure domain as the host. This configuration ensures that read operations are occurring locally and therefore are not subjected to the latency that is incurred by traversing the long-distance links.

Figure 3-23 on page 186 shows the data flow for a read command. The green line represents an optimal configuration where the host in Failure Domain 1 is accessing a volume that has its primary copy set to the node that also is in Failure Domain 1. The red line represents a non-optimal configuration where the host in Failure Domain 2 is accessing a volume that has its primary copy set to the node in Failure Domain 1.

Figure 3-23 Read operation

Best practice: The primary copy of the mirror should be in the same failure domain as the host that is accessing the mirrored volume. The primary copy can be changed by using the **`chvdisk`** command.

Also, the host should access the mirrored volume through the SAN Volume Controller node that is in the same failure domain as the host.

3.10 Write operations

Write operations are run on all copies. The storage system with the lowest performance determines the response time between SAN Volume Controller and the storage system back-end. The SAN Volume Controller cache can hide this to the server up to a certain level. If a back-end write fails or a copy goes offline, a bitmap is used to keep track of out of sync grains, like with other SAN Volume Controller copy services. When the missing copy is back, SAN Volume Controller evaluates the change bitmap to perform an automatic resynchronization of both copies. The resynchronization process has a similar performance effect on the system like a FlashCopy background copy or a volume migration. The resynchronization bandwidth can be controlled by using the **`chvolume -syncrate`** command. Host access to the volume continues during that time. This behavior can cause difficulties in case of a site failure. Since version 6.2, SAN Volume Controller provides the volume attribute **`-mirrorwritepriority`** to prioritize between strict data redundancy (**`-mirrorwritepriority redundancy`**) and best performance (**`-mirrorwritepriority latency`**) for the volume. Figure 3-24 shows the flow of data from the server to the volumes on the SAN Volume Controller.

Best practice: The host should access the mirrored volume through the SAN Volume Controller node that is in the same failure domain as the host.

Figure 3-24 Data flow from server to the volumes on the SAN Volume Controller

The default setting is `-mirrorwritepriority latency` to maintain compatibility with an earlier version. The recommended setting in split cluster environment is `-mirrorwritepriority redundancy`. Table 3-2 describes the differences between latency and redundancy while using `mirrorwritepriority`.

Table 3-2 Differences between -mirrorwritepriority redundancy and latency

-mirrorwritepriority redundancy	**-mirrorwritepriority latency**
Let host writes wait until “red” copy is updated	Acknowledge host write because “green” copy is up to date
Acknowledge host write as successful	Release cache pages. Mark copies in bitmap as not synchronized
Release cache pages	Try again later to reach the “red” copy and resynchronize from “green”
Sync=100%; Performance impact	Sync <= 100%; Performance smoothed

3.10.1 `-mirrorwritepriority` settings

We must consider two situations for each volume. The site failure can occur before or after write cache destaging and SAN Volume Controller behaves differently depending on the setting of `-mirrorwritepriority`.

`-mirrorwritepriority latency`

If the site failure occurs before write cache destaging, write data is still in the cache of the other node and needs to be destaged as soon as possible. The volume remains online when “red” back-end storage system is online.

If the site failure occurs after write cache destaging, data is destaged to the “green” copy. Cache pages are released, but the “red” copy is not yet synchronized. Thus, the “green” copy not available and the “red” copy not in sync. The volume goes offline.

`-mirrorwritepriority redundancy`

If the site failure occurs before write cache destaging, write data is still in the cache of the other node, and needs to be destaged as soon as possible. The volume remains online when “red” back-end storage system is online.

If the failure occurs after write cache destaging, both copies are updated and are in sync. The volume remains online (when “red” back-end storage system is online).

3.10.2 Best Practices

We recommend the following best practices for volume mirroring in a split I/O group cluster:

- Select the primary copy of a mirrored volume in the same location as the preferred node of that volume.
- The setting `-mirrorwritepriority redundancy` is the preferred option for each volume. Default option is `-mirrorwritepriority latency`.
- Be sure to follow best practices on SAN and back-end storage systems.

3.11 SAN Volume Controller quorum disk

The quorum disk fulfills the following functions for cluster reliability:

- Act as tiebreaker in split brain scenarios.
- Saves critical configuration metadata.

The SAN Volume Controller quorum algorithm distinguishes between the active quorum disk and quorum disk candidates. There are three quorum disk candidates. At most, one of these candidates acts as the active quorum disk. The other two candidates are reserved to become active if the current active quorum disk fails.

All three quorum disks are used to store configuration metadata, but only the active quorum disk acts as tiebreaker and is used in T3 recovery.

Important: Each quorum disk should be placed in one site (failure domain). The quorum disk in the third site (quorum site) should be set as active quorum disk.

Important: A quorum disk must be placed in each of the three failure domains. The quorum disk in the third failure domain should be set as the active quorum disk. Losing all quorum disks results in a loss of data.

3.12 Quorum disk requirements and placement

Because of the quorum disk's role in the voting process, quorum functionality is not supported for internal drives on SAN Volume Controller nodes. Inside a SAN Volume Controller node, the quorum disk cannot act as a tie-breaker. Therefore, only MDisks from external storage system are selected as SAN Volume Controller quorum disk candidates. Distributing quorum disk candidates across storage systems in different Failure Domains eliminates the risk of losing all three quorum disk candidates because of an outage of a single storage system or site. Up to SAN Volume Controller 6.1, SAN Volume Controller selects the first three MDisks from external storage systems as quorum disk candidates and reserves some space on each of these disks per default. SAN Volume Controller does not verify whether the MDisks are from the same disk controller or from different disk controllers. To achieve that the quorum disk candidates and the active quorum disk are in their proper sites, you should change the quorum disk candidates by using the **`chquorum`** command.

Important: In case of failure of the IBM SAN Volume Controller nodes in Failure Domain 2, the quorum disk must be accessible within at least 80 milliseconds by the IBM SAN Volume Controller nodes in Failure Domain 1. If the quorum is not accessible with in 80 milliseconds, the SAN Volume Controller cluster goes offline.

Starting with SAN Volume Controller 6.2, the quorum disk selection algorithm has changed. SAN Volume Controller reserves space on each MDisk and dynamically selects the quorum disk candidates and the active quorum disk. Thus, the location of the quorum disk candidates and the active quorum disk might change unexpectedly. Because of this, make sure you disable the dynamic quorum selection in a split I/O group cluster by using the **`-override`** flag for all three quorum disk candidates.

Example CLI command: **`chquorum -override yes -mdisk mdisk_id|mdisk_name`**

The storage system that provides the quorum disk in a split I/O group configuration at the third site must be supported as extended quorum disk. For more information about storage systems that provide extended quorum support, see this website:

http://www.ibm.com/storage/support/2145

Guaranteed bandwidth: Quorum disk storage controllers must be Fibre Channel or FCIP attached and provide less than 80 ms response times with a guaranteed bandwidth of greater than 2 Mbps.

Important: Quorum disk candidate includes the following requirements for SAN Volume Controller Split Cluster configuration:

- Requires three quorum disk candidates with one quorum disk candidate placed in each of the three Failure Domains.
- The active quorum disk must be assigned to Failure Domain 3.
- Dynamic quorum selection must be disabled by using the **`chquorum`** command.
- Quorum disk candidates and the active quorum disk assignment must be done manually by using the **`chquorum`** command.

3.13 Automatic SAN Volume Controller quorum disk selection

In the output from the CLI that is shown in Example 3-31, you see that the SAN Volume Controller cluster initially has automatically assigned the quorum disks.

Example 3-31 Quorum disks assigned

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name   active
object_type override
0            online 1  ITSO_V7K_SITEA_SAS0 1             ITSO_V7K_SITEA_N2 yes
mdisk       no
1            online 2  ITSO_V7K_SITEA_SAS1 1             ITSO_V7K_SITEA_N2 no
mdisk       no
2            online 3  ITSO_V7K_SITEA_SAS2 1             ITSO_V7K_SITEA_N2 no
mdisk       no
```

To change from automatic selection to manual selection, the commands that are shown in Example 3-32 were run.

Example 3-32 Changing from automatic to manual selection

```
IBM_2145:ITSO_SVC_SPLIT:superuser>chquorum -override yes -mdisk 110
IBM_2145:ITSO_SVC_SPLIT:superuser>chquorum -override yes -mdisk 1 1
IBM_2145:ITSO_SVC_SPLIT:superuser>chquorum -override yes -mdisk 6 2
```

When we run the **`lsquorum`** command, you see the results that are shown in Example 3-33.

Example 3-33 Quorum changed

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name
active object_type override
0           online 10 ITSO_V7K_SITEC_Q    5            ITSO_V7K_SITEC_Q_N2 yes
mdisk       yes
1           online 6  ITSO_V7K_SITEB_SAS0 0            ITSO_V7K_SITEB_N2   no
mdisk       yes
2           online 4  ITSO_V7K_SITEA_SAS3 1            ITSO_V7K_SITEA_N2   no
mdisk       yes
```

You see that the controller named `ITSO_V7K_SITEC_Q` that is in Failure Domain 3/site 3 is now the active quorum disk.

You also can choose to assign the quorum disk manually from the GUI. From the GUI, click **Pools** → **Pools by MDisk**, as shown Figure 3-25.

Figure 3-25 Open the MDisk by Pools view

You might need to expand the pools to view all of the MDisks. You can now select the MDisks that you want to use for quorum disks by pressing Ctrl while you choose the three candidates. When the candidates are selected, right-click and select **Quorum** → **Edit Quorum**, as shown in Figure 3-26.

Figure 3-26 Yellow color shows the MDisks that were selected for Quorum

Next you are asked if the manual quorum selection is to be used for Split Site Configuration or a Single Site Configuration. In this case, we choose the Split Site Configuration where you must choose the MDisk that is in site 3 (Failure Domain 3), as shown in Figure 3-27.

Figure 3-27 MDisk selected for site 3 (Failure Domain 3)

Finally, you must click **Edit**, and the quorum disks are manually assigned in the GUI, as shown in Figure 3-28 on page 194.

Figure 3-28 Quorum disk assigned

3.13.1 SAN Volume Controller cluster state and voting

The cluster state information on the active quorum disk is used to decide which SAN Volume Controller nodes survive if exactly half the nodes in the cluster should fail at the same time. Each node has one vote; the quorum disk has half votes for determining cluster quorum.

The SAN Volume Controller cluster manager implements a dynamic quorum. This means that following loss of a node or nodes, if the cluster can continue to operate, it dynamically moves the quorum disk to allow more node failure to be tolerated. This improves the availability of the central cluster metadata, which enables servicing of the cluster.

The cluster manager determines the dynamic quorum from the current voting set and a quorum disk, if available. If nodes are added to a cluster, they are added to the voting set; when nodes are removed, they are also removed from the voting set. Over time, the voting set, and hence the nodes in the cluster, can completely change so that the cluster has migrated onto a completely separate set of nodes from the set on which it started. Within a SAN Volume Controller cluster, the quorum is defined in one of the following ways:

- More than half the nodes in the voting set are available.
- Exactly half the nodes in the voting set and the quorum disk from the voting set or when there is no quorum disk in the voting set, exactly half of the nodes in the voting set if that half includes the node which appears first in the voting set (configuration node).

- When there is no quorum disk in the voting set, exactly half of the nodes in the voting set if that half includes the node which appears first in the voting set (configuration node).

In a SAN Volume Controller Split Cluster configuration, the voting set is distributed across the failure domains. Figure 3-29 shows the behavior of the SAN Volume Controller Stretched Cluster as a result of failures affecting the failure domains.

Failure Domain 1 Node 1	Failure Domain 2 Node 2	Failure Domain 3 Quorum Disk	Cluster status
Operational	Operational	Operational	Operational, Optimal
Failed	Operational	Operational	Operational, Write cache disabled (Write through mode)
Operational	Failed	Operational	Operational, Write cache disabled (Write through mode)
Operational	Operational	Failed	Operational, Active Quorum disk moved
Operational, Link to Failure domain 2 has failed split brain	Operational, Link to Failure domain 1 has failed split brain	Operational	The node that access the active quorum disk first remains active and the partner node goes offline
Operational	Failed same time with site 3	Failed same time with site 2	Stopped
Failed same time with site 3	Operational	Failed same time with site 1	Stopped

Figure 3-29 SAN Volume Controller Split Cluster behavior

Important: No manual fail over or fail-back activities are required. The SAN Volume Controller is an active-active architecture that does not perform fail over or fail-back operations.

3.14 Failure scenarios in SAN Volume Controller Split Cluster configuration

We use Figure 3-30 on page 196 to describe the following failure scenarios in a split cluster. The blue lines represent local I/O traffic and the green lines represent I/O traffic between failure domains:

- Power off FC Switch SAN768B-A1 in Failure Domain 1: FC Switch SAN768B-A2 takes over the load and routes I/O to SAN Volume Controller Node 1 and SAN Volume Controller Node 2.
- Power off SAN Volume Controller Node 1 in Failure Domain 1: SAN Volume Controller Node 2 takes over the load and continues processing host I/O. SAN Volume Controller Node 2 changes the cache mode to write-through to avoid data loss in case SAN Volume Controller Node 2 also fails.
- Power off Storage System V7000-A: SAN Volume Controller waits a short time (15 - 30 seconds), pauses Volume copies on Storage System V7000-A and continues I/O operations by using the remaining Volume copies on Storage System V7000-B.
- Power off Failure Domain 1: I/O operations can continue from Failure Domain 2.

Figure 3-30 SAN Volume Controller Split Cluster configuration

As Table 3-3 shows, SAN Volume Controller can handle every kind of single failure automatically without affecting applications.

Table 3-3 Failure scenarios

Failure scenario	SAN Volume Controller Split Cluster behavior	Server and Application impact
Single switch failure	System continues to operate using an alternative path in the same failure domain to the same SAN Volume Controller node	None
Single data storage failure	System continues to operate using the secondary data copy	None
Single quorum storage failure	System continues to operate on the same data copy, SAN Volume Controller selects another quorum disk candidate as the active quorum disk.	None
Single SAN Volume Controller node failure	System continues to operate using the secondary SAN Volume Controller node, SAN Volume Controller write cache is disabled	None Disabled cache can cause performance degradation

Failure scenario	SAN Volume Controller Split Cluster behavior	Server and Application impact
Failure of either Failure Domain 1 or 2. (containing SAN Volume Controller nodes)	System continues to operate on the remaining failure domain containing SAN Volume Controller nodes.	Servers without HA functionality in the failed site stop; servers in the other site continue to operate. Servers with HA software functionality are restarted from the HA software. The same disks are seen with the same UIDs in the surviving failure domain. SAN Volume Controller cache is disabled, which might degrade performance.
Failure of Failure Domain 3 (containing active quorum disk)	System continues to operate on both Failure Domains 1 and 2. SAN Volume Controller selects another quorum disk.	None
Access loss from one site with SAN Volume Controller node (Site 1 or 2) to the site with active quorum disk (Site 3)	System continues to operate on the same data copy. Usage of an alternative path to the active quorum disk.	None
Access loss between Failure Domains 1 and 2 (containing SAN Volume Controller nodes)	System continues to operate, the failure domain with SAN Volume Controller configuration node continues with operation, the SAN Volume Controller in the other failure domain stops.	Servers without HA functionality in the failed site stop; servers in the other site continue to operate. Servers with HA software functionality are restarted from the HA software. The same disks are seen with the same UIDs in the surviving failure domain. SAN Volume Controller cache are disabled, which can degrade performance.

3.15 SAN Volume Controller Split Cluster configurations

The SAN Volume Controller nodes of a Split Cluster configuration must be connected to each other by Fibre Channel or FCIP links. These links provide paths for SAN Volume Controller node-to-node communication and for host access to SAN Volume Controller nodes. SAN Volume Controller Split Cluster supports two different approaches for node-to-node intra-cluster communication between failure domains.

The following Split Cluster configurations are available:

- Attach each SAN Volume Controller node to the Fibre Channel switches in the local and in the remote failure domain directly. Thus, all node-to-node traffic can be done without traversing ISLs. This approach is referred to as Split Cluster *No ISL configuration.*
- Attach each SAN Volume Controller node only to local Fibre Channel switches and configure ISLs between failure domains for SAN Volume Controller node-to-node traffic. This approach is referred to as Split Cluster *ISL configuration*.
- Attach each SAN Volume Controller node only to local Fibre Channel switches and configure FCIP between failure domains for SAN Volume Controller node-to-node traffic support for FCIP was introduced in SAN Volume Controller 6.4. This approach is referred to as Split Cluster *FCIP configuration.*

We describe the following Split Cluster configurations with their associated attributes to assist with the selection of the appropriate configuration to meet individual solution requirements.

- No ISL configuration
- ISL configuration
- FCIP configuration

The maximum distance between failure domains without ISLs is limited to 40 km (24.85 miles). This is to ensure that any burst in I/O traffic that can occur does not use all of the buffer-to-buffer credits. The link speed also is limited dependent on the cable length between nodes. Table 3-4 lists the supported distances for each of the SAN Volume Controller Split Cluster configurations with their associated SAN Volume Controller code level and ports speed requirements.

Table 3-4 Supported distances

Configuration	SAN Volume Controller Code Level	Maximum Length	Maximum Link Speed
No ISL	5.1 or Later	< 10 km (6.21 miles)	8 Gbit/s
No ISL	6.3 or Later	< 20 km (12.42 miles)	4 Gbit/s
No ISL	6.3 or Later	< 40 km (24.85 miles)	2 Gbit/s
ISL	6.3 or Later	< 300 km (186.41 miles)	2,4,8 Gbit/s
FCIP	6.4 or Later	< 300 km (186.41 miles)	2,4,8 Gbit/s

3.15.1 No ISL configuration

This configuration is similar to a standard SAN Volume Controller environment with the main difference being that nodes are distributed across two failure domains. Figure 3-31 on page 200 shows the No ISL configuration. Failure Domain 1 and Failure Domain 2 contain the SAN Volume Controller nodes with customer data, and Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

Advantages

This configuration includes the following advantages:

- HA solution is distributed across two independent data centers.
- Configuration is similar to a standard SAN Volume Controller cluster.
- Limited hardware effort; WDM devices can be used, but are not required.

Requirements

This configuration includes the following requirements:

- Four dedicated Fibre Channel links per I/O group between failure domains.
- ISLs are not used between SAN Volume Controller nodes.
- Passive wave division multiplexing (WDM) devices can be used between failure domains (up to SAN Volume Controller version 6.2).
- Active or passive WDM can be used between failure domains (SAN Volume Controller version 6.3 and later)
- Long Wave small form-factor pluggables (SFPs) are required to reach 10 km (6.21 miles) without WDM.
- The supported distance is up to 40 km (24.85 miles) with WDM.
- Two independent fibre links between site 1 and site 2 must be configured with WDM connections.
- Third failure domain is required for quorum disk placement.
- Quorum disk storage system must be Fibre Channel attached.

Figure 3-31 on page 200 shows the SAN Volume Controller Split Cluster No ISL configuration. Failure Domain 1 and Failure Domain 2 contain the SAN Volume Controller nodes along with customer data, and Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

Figure 3-31 SAN Volume Controller Split Cluster: No ISL configuration

Zoning Requirements

The following zoning requirements for the SAN Volume Controller Split Cluster No ISL configuration are the same as with a standard SAN Volume Controller configuration:

- Servers access only SAN Volume Controller nodes; there is no direct access from servers to back-end storage.
- Separate zone is configured for node-to-node traffic.
- SAN Volume Controller nodes of the same I/O group do not communicate by using ISLs.
- Zones should not contain multiple back-end disk systems.

Figure 3-32 on page 201 shows the SAN Volume Controller Split Cluster No ISL configuration with passive WDM connections between failure domains 1 and 2.

Figure 3-32 SAN Volume Controller Split Cluster: No ISL configuration with WDM

3.15.2 ISL configuration

This configuration is similar to a standard SAN Volume Controller configuration except that nodes are distributed across two failure domains and SAN Volume Controller node-to-node communication between failure domains is performed over ISLs. SAN Volume Controller support for ISLs was introduced in SAN Volume Controller 6.3.

The use of ISLs increases the supported distance for SAN Volume Controller Split Cluster configurations to 300 km (186.41 miles). Although the maximum supported distance is 300 km (186.41 miles), there are instances where host-dependent I/O are required to traverse the long distance links multiple times. Because of this, it is possible that the associated performance degradation might exceed acceptable levels. To mitigate this exposure, it is recommended that the distance between failure domains be limited to 150 km (93.20 miles).

Best practice: Limiting the distance between failure domains to 150 km (93.20 miles) minimizes the risk of encountering elevated response times.

Advantages

This configuration includes the following advantages:

- ISLs enable longer distances greater than 40 km (24.85 miles) between failure domains.
- Active and passive WDM devices can be used between failure domains.
- The supported distance is up to 300 km (186.41 miles) with WDM.

Requirements

This configuration includes the following requirements:

- Four dedicated Fibre Channel links per I/O group between failure domains.
- The use of ISLs for node-to-node communication requires configuring the following separate SANs:
  - One SAN is dedicated for SAN Volume Controller node-to-node communication. This SAN is referred as the *private SAN*.
  - One SAN is dedicated for host and storage controller attachment. This SAN is referred to as the *public SAN*.
- Third failure domain is required for quorum disk placement.
- Storage controllers that contain quorum disks must be Fibre Channel attached.
- A guaranteed minimum bandwidth of 2 MBs is required for node-to-quorum traffic.
- No more than one ISL hop is supported for connectivity between failure domains.

Note: Private and public SANs can be implemented by using any of the following approaches:

- Dedicated Fibre Channel switches for each SAN
- Switch partitioning features
- Virtual or logical fabrics

Figure 3-33 on page 203 shows the ISL configuration. Failure Domain 1 and Failure Domain 2 contain the SAN Volume Controller nodes with customer data. Failure Domain 3 contains the storage subsystem that provides the active quorum disk.

Zoning Requirements

The SAN Volume Controller Split Cluster ISL configuration requires private and public SANs. The two SANs must be configured according to the following rules:

- Two ports of each SAN Volume Controller node are attached to the public SANs.
- Two ports of each SAN Volume Controller node are attached to the private SANs.
- A single trunk between switches is required for the private SAN.
- Hosts and storage systems are attached to fabrics of the public SANs. Links used for SAN Volume Controller Metro Mirror or Global Mirror must be attached to the public SANs.
- Failure Domain 3 (quorum disk) must be attached to the public SAN.
- ISLs belonging to the private SANs must not be shared with other traffic and must not be over-subscribed.

For more information, see the SAN Volume Controller Information Center or *V6.3.0 Configuration Guidelines for Extended Distance Split-System Configurations for IBM System Storage SAN Volume Controller*, S7003701, which is available at this website:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

Figure 3-33 on page 203 shows the SAN Volume Controller Split Cluster ISL configuration. The private and public SANs are represented as logical switches on each of the four physical switches.

Figure 3-33 SAN Volume Controller Split Cluster: ISL configuration

Figure 3-34 on page 204 shows the SAN Volume Controller Split Cluster ISL configuration with active or passive WDM between failure domains 1 and 2. The private and public SANs are represented as logical switches on each of the four physical switches.

Figure 3-34 SAN Volume Controller Split Cluster: ISL configuration with WDM

3.15.3 FCIP configuration

In this configuration, FCIP links are used between failure domains. SAN Volume Controller support for FCIP was introduced in SAN Volume Controller 6.4. This configuration is variation of the ISL configuration that was described in this chapter and therefore many of the same requirements apply.

Advantage

This configuration uses existing IP networks for extended distance connectivity.

Requirements

This configuration includes the following requirements:

- At least two FCIP tunnels between failure domains.
- The use of ISLs for node-to-node communication requires configuring the following separate SANs:
  - One SAN is dedicated for SAN Volume Controller node-to-node communication. This SAN is referred as the `private SAN`.
  - One SAN is dedicated for host and storage controller attachment. This SAN is referred to as the *public SAN*.
- Third failure domain is required for quorum disk placement.

- Failure domain 3 (quorum disk) must be either Fibre Channel or FCIP attached.

  If FCIP attached, the response time to the quorum disk cannot exceed 80 ms.
- Storage controllers containing quorum disks must be Fibre Channel or FCIP attached.
- A guaranteed minimum bandwidth of 2 MBps is required for node-to-quorum traffic.
- No more than one ISL hop is supported for connectivity between failure domains.

Zoning Requirements

The SAN Volume Controller Split Cluster *FCIP configuration* requires private and public SANs. These SANs must be configured according to the following rules:

- Two ports of each SAN Volume Controller node are attached to the public SANs.
- Two ports of each SAN Volume Controller node are attached to the private SANs.
- A single trunk between switches is required for the private SAN.
- Hosts and storage systems are attached to fabrics of the public SANs. Links used for SAN Volume Controller Metro Mirror or Global Mirror must be attached to the public SANs.
- Failure Domain 3 (quorum disk) must be attached to the public SAN.
- ISLs belonging to the private SANs must not be shared with other traffic and must not be over-subscribed.

For more information, see the SAN Volume Controller Information Center or *V6.3.0 Configuration Guidelines for Extended Distance Split-System Configurations for IBM System Storage SAN Volume Controller*, S7003701, which is available at this website:

http://www-01.ibm.com/support/docview.wss?&uid=ssg1S7003701

Figure 3-35 on page 206 shows the SAN Volume Controller Split Cluster FCIP configuration.

Figure 3-35 FCIP configuration

3.16 Fibre Channel settings for distance

Usage of Long Wave (LW) SFPs is an appropriate method to overcome long distances. Starting with SAN Volume Controller 6.3, active or passive DWDM/CWDM technology is supported.

Passive WDM devices cannot change wavelengths; colored SFPs are required and must be supported by the switch vendor.

Active WDM devices can change wavelengths. All active WDM components that already are supported with SAN Volume Controller Metro Mirror also are supported with SAN Volume Controller Split Cluster configurations.

Buffer credits, also called buffer-to-buffer (BB) credits, are used for Fibre Channel flow control and represent the number of frames a port can store. Each time a port transmits a frame, that port's BB credit is decremented by one; for each R_RDY received, that port's BB credit is incremented by one. If the BB credit is zero, the corresponding port cannot transmit until an R_RDY is received back.

Thus BB credits are necessary to have multiple Fibre Channel frames in flight, as shown in Figure 3-36 on page 207. An appropriate number of BB credits are required for optimal performance. The number of buffer credits to achieve maximum performance over a specified distance depends on the speed of the link.

Figure 3-36 Buffer credits in flight

The calculation assumes that the other end of the link starts transmitting the acknowledgement frame R_RDY in the same moment when the last bit of the incoming frame arrives at the receiver, which is not the case. So the guidelines above give the minimum numbers. The performance drops dramatically if there are not enough buffer credits according to link distance and link speed, as shown in Table 3-5.

Table 3-5 Buffer-to-buffer credits

FC link speed	B2B credits for 10 km (6.21 miles)	Distance with 8 credits
1 Gbit/s	5	16 km (9.94 miles)
2 Gbit/s	10	8 km (4.97 miles)
4 Gbit/s	20	4 km (2.48 miles)
8 Gbit/s	40	2 km (1.24 miles)

The number of BB credits that are provided by a SAN Volume Controller Fibre Channel host-bus adapter (HBA) is limited. An HBA of a model 2145-CF8 node provides 41 buffer credits, which is sufficient for 10 km (6.21 miles) distance at 8 Gbit/s. The HBAs in all earlier SAN Volume Controller models provide only eight buffer credits, which is enough only for 4 km (2.48 miles) distance with 4 Gbit/s link speed. These numbers are determined by the HBA's hardware and cannot be changed.

Best practice: The use of 2145-CF8 or CG8 nodes for distances longer than 4 km (2.48 miles) is recommended to provide enough BB credits.

FC switches have default settings for the BB credits (in Brocade switches: eight BB credits for each port). Although the SAN Volume Controller HBAs provide 41 BB credits, the switch stays on the default value. Thus, it is necessary to adjust the switch BB credits manually. For Brocade switches, the port buffer credits can be changed by using the **`portcfgfportbuffers`** command.

3.17 Back-end storage allocation to the SAN Volume Controller Cluster

In this section, we describe the back-end storage allocation. For Failure Domain 1 Site 1 and Failure Domain 2 Site 2, we chose to use Storwize V7000 as our back-end storage. Both Storwize V7000 are configured the same way. Failure Domain 3 Site 3 has a Storwize V7000 acting as the active Quorum disk.

For more information about how to implement Storwize V7000, see the following resources:

- *Implementing the IBM Storwize V7000 V6.3*, SG24-7938
- The Storwize V7000 information center website:

  http://pic.dhe.ibm.com/infocenter/storwize/ic/index.jsp

Figure 3-37 shows the MDisks that were created in Failure Domain 1.

Figure 3-37 Mdisks that were created in the V7000

Figure 3-38 shows the volumes that were assigned to the SAN Volume Controller Split Cluster host.

Figure 3-38 Volume assignment to the SAN Volume Controller Split cluster Host

3.18 Volume allocation

The volume allocation that uses the SAN Volume Controller Split Cluster solution is described here and all volume assignments are based on the local-to-local policy. This means that if a host is in Failure Domain 1 Site 1, the preferred node also must be in Failure Domain 1 Site 1.

Copy 0 of the volume mirror (also referred to as the *primary copy*) also must be in Failure Domain 1 Site 1. This configuration ensures that during normal operations, there are no unnecessary roundtrips for the I/O operations.

Figure 3-39 shows the volumes that are assigned to host `itsosjcvios1`.

Figure 3-39 Volumes assigned to Host itsosjcvios1

Figure 3-40 shows that the preferred node is from the local site, and that Step 1 - Step 6 do not affect the application latency. Step 7 - Step 10 might affect performance indirectly. It also shows that the front-end latency is one round trip and the back-end latency is two round trips.

Figure 3-40 Split Cluster Local I/O

Figure 3-41 shows that the preferred node is at the remote site and that Step 1 - Step 6 affect the application latency. Step 7 - Step 10 might affect performance indirectly. It also shows that the front-end latency is three round trips and the back-end latency is two round trips.

Figure 3-41 Remote I/O diagram

We describe the implementation of the solution further in the following chapters.

4

PowerVM and PowerHA Implementation

In this chapter, we show the steps we took to create our PowerVM and PowerHA environment. We also show the various other PowerVM technologies and the considerations for their use in the SAN Volume Controller Stretched Cluster solution.

PowerVM can be used to improve server usage and I/O resources through sharing. It provides a simple infrastructure where workload is independent of hardware. It also provides a flexible environment where applications are made continuously available and can be changed dynamically.

There are various PowerVM technologies that are provided by Power Hypervisor or VIOS, or both. There also is some virtualization provided by AIX and Power hardware that is described in this chapter.

PowerVM is available in different editions. For the SAN Volume Controller Stretched Cluster solution, PowerVM Enterprise Edition is recommended with the Live Partition Mobility (LPM) feature enabled.

For more information about the PowerVM editions, see 2.8.1, “PowerVM Editions” on page 141.

You also can upgrade between editions, but you must ensure that the LPM feature is enabled.

References: For the purpose of brevity, it is beyond the intended scope of this book to include each and every step of a PowerVM install. We point the reader to the relevant information that is contained in other books and web resources to enable a successful implementation of the required PowerVM components.

© Copyright IBM Corp. 2013. All rights reserved.

This chapter includes the following topics:

- PowerVM overview
- PowerVM considerations
- PowerVM initial configuration
- Installing VIOS
- Installing AIX and creating VIOS clients
- NPIV initial configuration
- Live Partition Mobility implementation
- LPM operations
- Combining LPM with other features
- LPM configuration
- LPM
- Error identification and resolution
- LAM considerations
- PowerHA SystemMirror initial configuration and other considerations
- Advanced configuration

4.1 PowerVM overview

Businesses are turning to PowerVM virtualization to consolidate multiple workloads onto fewer systems, increase server usage, and reduce cost. PowerVM technology provides a secure and scalable virtualization environment for AIX, IBM i, and Linux applications, which are built upon the advanced reliability, availability, and serviceability features and the leading performance of the Power platform.

For more information about PowerVM, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

http://www.redbooks.ibm.com/redpieces/abstracts/sg247940.html?Open

For SAN Volume Controller Stretched Cluster solution, we deployed PowerVM with NPIV and dual virtual I/O (VIOS) with LPM configuration, which is described in this chapter.

4.2 PowerVM considerations

In this section, we describe the planning considerations for SAN Volume Controller Stretched Cluster solution with respect to PowerVM. Presented here is a list of recommended steps for preparing for this solution. However, there might be other features of PowerVM you might want to deploy in this solution that requires other considerations, which can require extra planning.

At a high level, the following steps must be considered:

- Physical environment requires hardware installation in different sites that are connected through switches. Switches also must be configured for necessary ports, which must be planned carefully, as shown in the following examples:
  - Site planning with respect to installing hardware in the respective sites.
  - Power distribution units (PDU) must be carefully planned to take into account power outages and uninterruptible power supply (UPS) installation.
  - Networks and switches appropriately planned for on the various subnets, and the virtual local area network (VLAN) and private and public network configuration thoroughly researched.
  - SANs and SAN switches must be N-Port ID Virtualization (NPIV) enabled and properly connected through FCIP routers.
- Central electrical complex (CEC) Hot Add Repair Maintenance (CHARM) function allows the expansion of the system processors, memory, and I/O hub capacity, and their repair with limited disruption to the system operation.
- Hardware Management Console (HMC) planning involves the plugging in of power cords, updating the management system password, starting ASMI to change the system configuration and Flexible Service Processor (FSP) setup, checking the firmware levels, updating as required, and powering on the system and creating partitions. The following distinctions should be made between private and public networks concerning HMC:
  - HMC and the private network: Private network means the use of a selected range of non-routable IP addresses. It is intended that the only devices on the HMC private network are the HMC and the managed servers. The HMC is connected to the flexible service processor (FSP) of the managed server or servers.
  - HMC and the public network: Public or open network means a network connection from the HMC to the logical partitions and, potentially, to other systems on your regular network infrastructure. If wanted, the open network also can be connected to a suitable firewall or router for connection to the Internet. Such a connection to the Internet enables the HMC to call home when there are any hardware errors to report. The HMC provides its own firewall (by using iptables) on each of its network Interfaces. The basic firewall configuration is automatically setup when the HMC code is installed, but further configuration might be necessary.
- The following general planning guidelines must be considered for highly available systems to make all partitions fault tolerant:
  - Advanced PowerVM planning requires an enterprise edition license
  - Micro-partition and virtual CPU enables efficient managing of the CPU
  - Virtual I/O Server LPAR sizing
  - VIOS implementation in terms of dual VIO server to provide network and storage redundancy

  - Networks:
    - Ethernet adapter connection settings
    - VLAN must be properly planned for private and public communication
    - Hypervisor IEEE VLAN-compatible virtual switches for Virtual Ethernet
- Logical partitioning
- Virtualized processors
- Virtual Fibre Channel (FC) adapters
- Number of partitions:
  - Number of physical and virtual devices
  - Number of VIO servers created
  - Number of NPIV adapters created
  - Partition profiles, for instance, the maximum memory that is specified on the profile
- LPM has specific requirements in terms of the operating system level, firmware level, storage layout, and network interfaces. A successful migration, regardless of whether it is active, requires careful deployment planning. Sometimes you can qualify a partition for mobility by taking more steps, such as removing physical adapters (non-virtual adapters) or by using a dynamic logical partitioning (DLPAR) operation.
- Software and operating system planning involves checking supported versions
- High availability planning involves network and storage planning for shared communication
- Planning for PowerHA needed to purchase standard edition license with smart assist functionality
- Operating system considerations, checking supported versions compatible with PowerHA standard edition 7.1.1
- Planning cluster networks involves enabling multicast communication in switches. Proper planning for management and private communication to separate the traffic
- Planning storage requirements, making disks shared across partitions across the sites
- Application planning, checking supported version of DB2, WebSphere Application Server for PowerHA enterprise edition 7.1.1
- Planning for resource groups, separate each application with separate resource group
- Developing a cluster test plan, test basic failover scenario
- Change and problem management; use the problem determination tool and cluster test tools of PowerHA to test the setup

For more information about licensing and other operating system requirements and the minimum hardware requirements, see Chapter 2, “Hardware and Software Description” on page 115 and 2.6, “Software” on page 130.

PowerVM features are supported on most of the Power Systems offerings; however, there are some exceptions. For more information about which features are available on which server models, see this website:

http://www.ibm.com/systems/power/software/virtualization/editions/features.html

Because individual Power Systems offerings are capable of hosting many system images, the importance of isolating and handling service interruptions becomes greater. These service interruptions can be planned or unplanned. Carefully consider interruptions for systems maintenance when planning system maintenance windows and the following factors:

- Environmental, including cooling and power
- System firmware
- Operating systems, for example, AIX, IBM i and Linux
- Adapter microcode

For availability planning in PowerVM, we used LPM for planned workload movement.

PowerHA SystemMirror was used to provide continuous availability for applications. Dual VIOS server is used to provide continuous availability of client LPAR for scheduled maintenance of the machine.

For more information about PowerHA, see *Exploiting IBM PowerHA SystemMirror Enterprise Edition*, SG24-7841, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247841.html

For more information about how to use PowerHA in a PowerVM environment, see this website:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/index.jsp?topic=%2Fcom.ibm.aix.powerha.navigation%2Fpowerha.htm

To activate the Virtual I/O Server, the PowerVM Editions (or PowerVM) hardware feature is required. A logical partition with enough resources to share with other logical partitions also is required.

For more information about the requirements for configuring Virtual I/O Server, see section 10.1 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

http://www.redbooks.ibm.com/redpieces/abstracts/sg247940.html?Open

4.2.1 NPIV planning

NPIV is an industry-standard technology that allows an NPIV-capable Fibre Channel adapter to be configured with multiple virtual worldwide port names (WWPNs). This technology is sometimes referred to as *virtual Fibre Channel*. Similar to the virtual SCSI functionality, virtual Fibre Channel is another way of securely sharing a physical Fibre Channel adapter among multiple Virtual I/O Server client partitions.

For more information about hardware and software requirements for using NPIV, see section 10.2.2 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

http://www.redbooks.ibm.com/redpieces/abstracts/sg247940.html?Open

For more information about planning security in PowerVM, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

http://www.redbooks.ibm.com/redpieces/abstracts/sg247940.html?Open

4.3 PowerVM initial configuration

In this section, we describe the basic steps that are taken to install and configure PowerVM.

4.3.1 Power Systems management console

These management consoles are available for PowerVM:

- Hardware Management Console
- Systems Director VMControl

The HMC is a dedicated Linux-based appliance that you use to configure and manage IBM Power System servers. The HMC provides access to logical partitioning functions, service functions, and various system management functions through a browser-based interface and a command-line interface (CLI). Because it is a separate stand-alone system, the HMC does not use any managed system resources and you can maintain it without affecting system activity.

If you are deploying dual HMCs for redundancy, we recommend not to cable them both into the same network switch, because this moves the single point of failure from the HMC to the network. Implementing a single switch also can cause unpredictable results if both HMCs are providing Dynamic Host Configuration Protocol (DHCP) services.

For more information about redundant HMCs, see *IBM System p5 Approaches to 24x7 Availability Including AIX 5L*, SG24-7196, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247196.html?Open

The network configuration for HMC must accommodate all the attached managed systems. If the HMC is acting as a DHCP server, its configured DHCP IP range must be large enough. A similar consideration exists if your HMCs, CECs and LPARs are all hosted on the same network (VLAN) or subnet (IP segment). If you have enough CECs to employ multiple HMCs, will you be using separate networks (one per HMC) or a shared network? Consider the preceding implications and decide which is the most appropriate option for your environment.

If you are using CSM with the Cluster-Ready Hardware Server (CRHS) feature, there are a number of documented prerequisites and environment requirements that must be adhered to. Implementing CRHS mandates a single service network and DHCP server. Because the single network is shared by the CSM Management Server, HMCs, and CEC connections, you must decide carefully which machine is the DHCP server.

Also, when you are considering any of these points, take into account any potential future growth. Careful planning avoids disrupting the configuration in the future.

HMC code: You can download or order the latest HMC code from the IBM Fix Central, which is available at this website:

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

Check the HMC software level for compatibility with the entire configuration by using the IBM Fix Level Recommendation Tool (FLRT), which is available at this website:

http://www14.software.ibm.com/webapp/set2/flrt/home

Although concurrent update can be done, we advise that you update in a negotiated maintenance window with a scheduled change.

For more information about other management consoles, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

http://www.redbooks.ibm.com/redpieces/abstracts/sg247940.html?Open

We chose to use HMC because we are familiar with it and so are most of the users that are likely to implement this solution.

4.4 Installing VIOS

Important: For brevity, it is beyond our intended scope to show how to install VIOS. In the sections that follow, we used *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is available at this website:

http://www.redbooks.ibm.com/redpieces/abstracts/sg247940.html?Open

to successfully install our environment. In the sections that follow we point the reader to the relevant sections in that book that must be successfully implemented.

Also refer to section 5.2 for planning steps.

4.4.1 Creating a Virtual I/O Server

Section 12.1 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, describes the steps we used to create a Virtual I/O Server logical partition on the HMC, install the Virtual I/O Server software, and configure the Virtual I/O Server for providing virtualized devices to its client partitions.

Be certain to follow those steps and check for successful completion at each stage.

Figure 4-1 shows the VIO server `itsosjc1vios1` and `itsosjc1vios2` profile successfully created after following the steps that are described in *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. Here, two VIO servers were created for dual VIO configuration. Select physical and virtual adapters appropriately, an as shown in the example in the next sections of this chapter.

Figure 4-1 HMC view after VIOS is created

Our VIOS is now created.

Figure 4-2 shows the list of virtual adapters that we created for VIO server by following section 12.1of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. This shows that we now have four server virtual Fibre Channels created and six Ethernet adapters created for two SEA adapters that we create later. Each adapter details are described in the following sections.

Figure 4-2 Virtual adapter list for VIO server

Figure 4-3 shows the VIO server profile properties. Note that Mover service partition must be selected for LPM. This is required because LPM identifies each VIO server as a mover service partition, which is necessary for communication between the source and target managed system because at least one mover service partition in both (source and target CEC or managed system) assists LPM communication.

Figure 4-3 VIO server profile properties

Figure 4-4 shows the physical adapters that are connected to VIO server. Each VIO server required at least one physical Fibre Channel adapter and one physical Ethernet adapter to create virtual adapters for clients. Depending on requirements, more adapters also might be needed.

Figure 4-4 Physical adapters selected for VIO server

Figure 4-5 shows the Host Ethernet Adapters (HEA) configuration for VIO server. The HEA adapter port must be used in promiscuous mode to use this adapter for creating the shared Ethernet adapter. As shown in Figure 4-5, the first adapter physical port is selected for `vios1` and the other port for `vios2`. This is a two-port HEA adapter.

Figure 4-5 HEA adapter configuration for managed system

To create a shared Ethernet adapter (SEA) that uses the HEA adapter, HEA ports must be used in promiscuous mode, which means that the physical port is dedicated for SEA traffic. We cannot create or use HEA virtual ports in promiscuous mode. We need a high-speed 10 Gbps network traffic for private cluster communications, so we used this high speed HEA adapter.

We have now successfully created the VIOS in our environment VIOS `itsosjc1vios1` and `itsosjc1vios2` is connected to p7 server in site A.

4.4.2 Installation of Virtual I/O Server

Installation of the Virtual I/O Server partition is performed from DVD-ROM installation media that is provided to clients that order the PowerVM feature. The Virtual I/O Server software is only supported in Virtual I/O Server partitions. For more information, see section 12.2 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

The Virtual I/O Server DVD-ROM installation media can be installed in the following ways:

- Media (assigning the DVD-ROM drive to the partition and booting from the media).
- The HMC (inserting the media in the DVD-ROM drive on the HMC and running the **`installios`** command, or installing from a media image copied to the HMC). If you run the **`installios`** command without any flags, a wizard is started and you are prompted to manually enter the information that is contained in the flags. The default is to use the optical drive on the HMC for the Virtual I/O Server installation media, but you can specify a remote file system instead.

  For more information about how to install the Virtual I/O Server from the HMC, see the IBM Systems Hardware Information Center at this website:

  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hb1/iphb1_vios_configuring_installhmc.htm

> **Requirement:** A network adapter with connection to the HMC network is required for Virtual I/O Server installation through the HMC.

- By using the DVD-ROM media with the Network Installation Management (NIM) server and running the **`smitty installios`** command (the Secure Shell must be working between NIM and HMC).
- NIM by copying the DVD-ROM media to the NIM server and generating NIM resources for installation.

4.4.3 Configuring VIOS for client network

In a PowerVM environment, it is the Virtual I/O Server that provides the link between the internal virtual and external physical LANs. This configuration can introduce an increased level of complexity because of multiple VLANs within a server that must be connected to multiple VLANs outside the server in a secure manner. Therefore, the Virtual I/O Server must be connected to all of the VLANs and not allow packets to move between the VLANs. For more information, see 4.2, “PowerVM considerations” on page 216.

In this scenario, the following requirements must be met:

- All client partitions must communicate with other client partitions on the same VLAN.
- All client partitions must communicate to a single virtual Ethernet adapter in the Virtual I/O Server. This is achieved by using IEEE 802.1Q on the Virtual I/O Servers virtual Ethernet adapter to allow more than one VLAN ID to be accepted at the virtual Ethernet adapter.
- VLAN tags must not be stripped from arriving packets from the client partitions. Stripped VLAN tags results in Virtual I/O Server not being able to forward the packets to the correct external VLAN.
- SEA must be enabled to allow packets from multiple VLANs.

This implementation requires that the enterprise security policy encompasses the following considerations:

- The enterprise security policy recognizes IEEE 802.1Q VLAN tagging.

  The IEEE 802.1Q VLAN tagging is implemented in the PowerVM hypervisor firmware. The Virtual I/O Server can have up to 21 VLANs per SEA. However, to use these SEAs, the physical network port must support the same number of VLANs. The physical VLAN policy within the enterprise determines the virtual VLAN policy.
- The enterprise security policy allows a network switch to have multiple VLANs.

  The enterprise security policy allows multiple VLANs to share a network switch (non-physical security). If it is a security requirement that a network switch only has one VLAN, every VLAN requires a separate SEA or Virtual I/O Server. If you make a separate Virtual I/O Server in a managed system, the hypervisor firmware acts like one switch with multiple VLANs, which in this case is not allowed by the security policy outside the Virtual I/O Server.

Various virtual Ethernet configurations can be used with PowerHA and it depends on the usage and functionality as to which to choose. We recommend the following virtual Ethernet configuration:

- Two VIOSs per physical server.
- Use the servers that are already configured virtual Ethernet settings because no special modification is required. For a VLAN-tagged network, the preferred solution is to use SEA failover; otherwise, use the network interface backup.
- One client-side virtual Ethernet interface simplifies the configuration; however, PowerHA misses network events.

- Two virtual Ethernet interfaces must be on the cluster LPAR to enable PowerHA to receive the network events. This configuration results in a more stable cluster.

> **Important:** The VIOS provides Ethernet network redundancy by using the SEA failover or Network Interface Backup (NIB) feature. However, the SEA failover and the NIB do not check the reachability of the specified ping address through the backup interface if the primary interface is up and running. So, you do not know that your backup path is working until the primary interface fails.

You might want two Ethernet interfaces in the cluster LPARs so that PowerHA can track the network changes, similar to physical network cards. Ensure that the two virtual Ethernet cards use different SEAs, VIOSs, and virtual Ethernet switches as their primary path.

In the following section, we describe some of the possible virtual Ethernet solutions and their properties. The only common component in the configurations is the dual VIOSs.

Figure 4-6 on page 227 shows the virtual Ethernet configuration of the environment. You can see that there are two SEAs created: one is for the public or management network with VLAN 701, and the other is for the private network with VLAN 150/160/170. Private network requires high-speed traffic, so we used the HEA adapter for configuring SEA. VLAN 199 and 299 are used for the heartbeat between two VIOS to check the health of each other so that an automatic failover of network load occurs in case of failures.

Figure 4-6 Virtual Ethernet configuration

SEA failover with two virtual Ethernet interfaces in the cluster LPAR

This configuration features the following characteristics:

- There are two virtual switches.
- Each virtual adapter connects to its own virtual switch and SEAs. There are no common devices.
- It works with VLAN tagging.
- SEA shared mode provides load sharing between VLANs.
- When you use SEA shared mode, the default path can change after every reboot.
- Two virtual Ethernet adapters exist per network.
- Dual redundancy is available with SEA failover and PowerHA.
- PowerHA tracks the network events.

For more information, see 16.3 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

In the sections that follow, we show the shared Ethernet adapter configuration in our environment. We used two SEA, one for the private network and the other for the public or management network. SEAs for multiple VLANs were set up for load sharing. In a dual VIOS setup, if both VIOS are up, SEA shares the load. If one of the VIOS is down, all network traffic is transferred to the other VIOS. Also, the trunk priority for VIOS1 should be 1, and the priority for VIOS2 should be 2. There is no trunk priority that must be set for the control channel adapters that are used for heartbeating.

Figure 4-7 shows the server Ethernet adapter that was created with VLAN 150. The trunk priority here is set to 1 because this is first VIO server. IEEE 802.1q was selected for VLAN tagged traffic with managed system LPARs, which is also called vswitch. Also, we have a different vswitch for different SEA traffic. Here we selected ETHERNET0, which is for VLAN 150/160/170. Access external network is also set to true because this adapter is used to accept the traffic from outside the network or to bridge the network from outside physical switch or port.

Figure 4-7 Virtual Ethernet adapter for VLAN 150

Figure 4-8 on page 229 shows the server Ethernet adapter that was created with VLAN 160. The trunk priority here is set to 1 because this is first VIO server. IEEE 802.1q was selected for VLAN tagged traffic with managed system LPARs, which is called vswitch. We have a different vswitch for different SEA traffic. Here we selected ETHERNET0, which is for VLAN 150/160/170. Access external network is also set to true because this adapter is used to accept the traffic from outside network or to bridge the network from outside physical switch or port.

Figure 4-8 Virtual Ethernet adapter for VLAN 160

Figure 4-9 shows the server Ethernet adapter that was created with VLAN 170. The trunk priority here is set to 1 because this is first VIO server. IEEE 802.1q was selected for VLAN tagged traffic with managed system LPARs, which is called vswitch. We have a different vswitch for different SEA traffic. Here we selected ETHERNET0, which is for VLAN 150/160/170. Access external network is also set to true because this adapter is used to accept the traffic from outside network or to bridge the network from outside physical switch or port.

Figure 4-9 Virtual Ethernet adapter for VLAN 170

Figure 4-10 on page 230 shows the server Ethernet adapter that was created with VLAN 299. This adapter is created for heartbeat network among VIO servers. We should not select the Access external network option for these adapters. This is internal to vswitch. Because we do not select the Access external network option for these adapters, trunk priority does not make any sense for it. However, in the properties it shows 1, but that should be ignored.

Figure 4-10 Virtual Ethernet adapter for heartbeat in ETHERNET0

Figure 4-11 shows the server Ethernet adapter that was created with VLAN 701. The trunk priority here is set to 1 because this is the first VIO server. IEEE 802.1q was selected for VLAN tagged traffic with managed system LPARs, which is also called vswitch. Also, we have a different vswitch for different SEA traffic. Here we selected ETHERNET0, which is for VLAN 701. Access external network is also set to true, because this adapter is used to accept the traffic from outside network or to bridge the network from outside physical switch or port.

Figure 4-11 Virtual Ethernet adapter for VLAN 701

Figure 4-12 on page 231 shows the server Ethernet adapter that was created with VLAN 199. This adapter is created for the heartbeat network among VIO servers. We should not select the Access external network option for these adapters. This is internal to vswitch. Because we do not select access external network (False) for these adapters, trunk priority does not make any sense for it. However, in the properties that are shown you can see it shows as 1, but that should be ignored.

Figure 4-12 Virtual Ethernet adapter for heartbeat in ETHERNET1

In Example 4-1, we show examples from our environment where we configured two SEA. It shows all the adapters that are available in the VIO server. By using the **mkvdev** command, SEA was created and the respective IP address was configured on it.

Example 4-1 SEA example

```
$ lsdev|grep ent
ent0             Available   Logical Host Ethernet Port (lp-hea)
ent1             Available   2-Port 10/100/1000 Base-TX PCI-X Adapter (14108902)
ent2             Available   2-Port 10/100/1000 Base-TX PCI-X Adapter (14108902)
ent3             Available   Virtual I/O Ethernet Adapter (l-lan)
ent4             Available   Virtual I/O Ethernet Adapter (l-lan)
ent5             Available   Virtual I/O Ethernet Adapter (l-lan)
ent6             Available   Virtual I/O Ethernet Adapter (l-lan)
ent7             Available   Virtual I/O Ethernet Adapter (l-lan)
ent8             Available   Virtual I/O Ethernet Adapter (l-lan)

$ mkvdev -sea ent0 -vadapter ent3,ent5,ent8 -default ent3 -defaultid 150 -attr
ha_mode=sharing ctl_chan=ent4
ent9 Available
$ mkvdev -sea ent1 -vadapter ent6 -default ent6 -defaultid 701 -attr ha_mode=auto
ctl_chan=ent7
ent10 Available
# netstat -in
Name  Mtu   Network     Address            Ipkts Ierrs    Opkts Oerrs  Coll
en10  1500  link#2      0.21.5e.8a.6.8a   9770068     0     3204     0     0
en10  1500  10.17.80    10.17.90.17       9770068     0     3204     0     0
en9   1500  link#3      0.14.5e.92.a5.e0  5785884     0        5     0     0
en9   1500  192.168.150 192.168.150.101   5785884     0        5     0     0
lo0   16896 link#1                         266584     0   266584     0     0
lo0   16896 127         127.0.0.1          266584     0   266584     0     0
lo0   16896 ::1%1                          266584     0   266584     0     0
```

4.4.4 Configuring VIOS for client storage

PowerHA 7.1.1 supports a virtualized environment. For example, you can use virtual Ethernet, virtual SCSI disks, and FC NPIV. The PowerHA is indifferent to the device that you use, be it a real hardware device or virtual adapter. However, the following considerations are important about the use of VIOS:

- Have two VIOS on each frame for redundancy.
- Configure SEA failover or network interface backup in the VIOS.
- You must configure the `netmon.cf` file to check the status of the network behind the virtual switch.
- Have two independent Ethernet switches for fully redundant network access.
- Configure multipath I/O (MPIO) for disk access for the VIOS.
- Configure MPIO for disk access for the PowerHA LPAR.

This configuration is common for all virtualized environments, so you are likely already using these features in your server. Before you start installing and configuring the PowerHA cluster, check and test that the VIOSs, SEA failover, and MPIO work correctly.

For more information, see section 16.2 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Figure 4-13 shows the server virtual Fibre Channel adapter properties. Here you can see that to properly map this to a specific client adapter, we must select the client LPAR name and the client adapter ID. The type of adapter is Server, which means this is for the VIOS server.

Figure 4-13 Server virtual Fibre Channel adapter

Figure 4-14 on page 233 shows the server virtual Fibre Channel adapter properties. Here you can see that to properly map this to a specific client adapter, we must select the client LPAR name and the client adapter ID. The type of adapter is Server, which means this is for the VIOS server.

Figure 4-14 Server virtual Fibre Channel adapter

Figure 4-15 shows the client virtual Fibre Channel adapter properties. Here you can see that we select the virtual Fibre Channel adapter ID with the server name to specify which server adapter we want to connect to. The type of adapter is Client, which means that this is for a VIOS client. Also, there are two WWPNs showing: one is active at any point, and the other is used only while LPM is performed. So, we can use both WWPNs to perform advanced zoning so that when this LPAR moves from one managed system to the other, the zoning path is already available.

Figure 4-15 Client virtual Fibre Channel adapter

Figure 4-16 shows the client virtual Fibre Channel adapter properties. Here you can see we select the virtual Fibre Channel adapter ID with the server name to specify which server adapter we want to connect to. The type of adapter is Client, which means that this is for a VIOS client.

Figure 4-16 Client virtual Fibre Channel adapter

Example 4-2 shows that after the virtual Fibre Channel adapter is created, we used the **vfcmap** command to create the mapping from server to client. Two client adapters in each client receive one adapter, or mapping, from each VIOS in our dual VIOS setup.

Example 4-2 Virtual Fibre Channel adapter mapping example

```
VIOS1:
$ vfcmap -vadapter vfchost4 -fcp fcs0

$ lsmap -vadapter vfchost4 -npiv
Name          Physloc                           ClntID ClntName       ClntOS
------------- ---------------------------------- ------ -------------- -------
vfchost4      U8233.E8B.0617C6P-V2-C10               6 itsosjc1lpar4  AIX

Status:LOGGED_IN
FC name:fcs0                    FC loc code:U78A0.001.DNWHS36-P1-C2-T1
Ports logged in:1
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs1            VFC client DRC:U8233.E8B.0617C6P-V6-C11

VIOS2:
$ vfcmap -vadapter vfchost4 -fcp fcs1

$ lsmap -vadapter vfchost4 -npiv
Name          Physloc                           ClntID ClntName       ClntOS
------------- ---------------------------------- ------ -------------- -------
vfchost4      U8233.E8B.0617C6P-V1-C10               6 itsosjc1lpar4  AIX

Status:LOGGED_IN
FC name:fcs1                    FC loc code:U78A0.001.DNWHS36-P1-C3-T2
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0            VFC client DRC:U8233.E8B.0617C6P-V6-C10
```

Figure 4-17 shows the volume to host mapping from the SAN Volume Controller in which two active WWPNs are shown. The other two WWPNs are used only for LPM when a workload moves to the other managed system.

Figure 4-17 SAN Volume Controller configuration shows host mapping for active WWPNs

Figure 4-18 shows inactive WWPNs mapped to the same volumes so that while LPM is performed, it is activated through the target CEC. This is required to be mapped for LPM in advance so that in case the LPAR moves from one managed system to another, this mapping is already in place to access the logical unit numbers (LUNs).

Figure 4-18 SAN Volume Controller configuration shows host mapping for inactive WWPNs

Figure 4-19 shows the volumes that are shared across different WWPNs of different LPARs to allow us to share the disks for setting up the cluster. In this example, we chose `itsosjc1lpar4` and `itsosjc2lpar4` configuration to go through. Because these two LPARs are part of one cluster, we shared the LUNs among these LPARs. However, `rootvg` LUNs are different for each LPAR.

Figure 4-19 SAN Volume Controller configuration for sharing disk in cluster nodes

4.5 Installing AIX and creating VIOS clients

The client partition definitions are similar to the creation of our Virtual I/O Server partition, but instead of selecting VIO Server, choose AIX or Linux, or IBM i. For more information, see section 13.1 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Figure 4-20 shows the client LPARs profile `itsosjc1lpar#` that was created by using the steps that are described in section 13.1 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. Memory and processing units to be considered for these LPARs depends on the planning and the types of application to be used.

Figure 4-20 VIO client LPAR profiles

Figure 4-21 shows the list of virtual Ethernet adapters that were created. We created one for the public or management network, and another for the private network. We can have many adapters here for VLANs 150/160/170 for each interface.

Figure 4-21 LPAR adapter list

Figure 4-22 on page 239 shows the client Ethernet adapter properties that are created. Here you can see that we selected this for VLAN 701 and the vswitch selected is ETHERNET1 because this vswitch was used earlier to create a SEA in the VIO server. So, the same vswitch must be used for VLAN 701. We selected IEEE 802.1q compatible as True for VLAN tagged traffic in managed system. This is required by vswitch to uniquely identify the traffic.

Figure 4-22 Client virtual Ethernet adapter

Figure 4-23 shows the client Ethernet adapter properties that are created. Here you can see that we selected this for VLAN 150 and the vswitch selected is ETHERNET0 because this vswitch was used earlier to create a SEA in VIO server. So, the same vswitch must be used for VLAN 150. We selected IEEE 802.1q compatible as True for VLAN tagged traffic in the managed system. This is required by vswitch to uniquely identify the traffic.

Figure 4-23 Client virtual Ethernet adapter

Now that we created client LPARs, we must install AIX on clients.

4.5.1 AIX client partition installation

This section describes the method to install AIX onto a previously defined client partition. You can choose your preferred method, but for our basic scenario, we opted to install the NIM_server from CD and then install the DB_server partition by using NIM from the NIM_server. We also are going to use the virtual Ethernet adapters for network booting and the virtual FC disks that were previously allocated to client partitions for rootvg. For more information, see section 13.2 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

In Figure 4-24, a SAN Volume Controller Stretched Cluster solution that uses PowerVM is shown. This configuration is already explained in Chapter 1, “Introduction” on page 1 and Chapter 2, “Hardware and Software Description” on page 115. For more information, see 4.2, “PowerVM considerations” on page 216.

Figure 4-24 SAN Volume Controller Stretched Cluster PowerVM example configuration

This is a dual VIOS configuration in which each client has storage available through different switches of the private SAN to maintain redundancy. For Ethernet, there also is a dedicated network for HMC communication and for cluster communication. Each client LPAR has the SDDPCM driver compatible to SAN Volume Controller installed.

LPM zoning

For LPM, the zoning is important. Example 4-3 on page 241 shows a zoning configuration. In this example, you can see that all virtual WWPN number were zoned to the respective controller from each public switch in a different site. This is required for LPM to work.

If we look closely at Example 4-3 on page 241, all four WWPNs of the LPAR are connected to the controllers that are visible to both fabrics. 50:05:07:68:01:10:b1:3f and 50:05:07:68:01:10:b0:c6 are visible to switch 113 and 0:05:07:68:01:40:b1:3f and 50:05:07:68:01:40:b0:c6 are visible to switch 111.

Zones can be created by using the **`zonecreate`** command.

Example 4-3 Zone example

```
SAN768B-2_A2:FID128:admin> setcontext 113
Public_A2:FID113:admin>
zone:  itsosjc2lpar4_P1_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0c
                50:05:07:68:01:10:b1:3f
                50:05:07:68:01:10:b0:c6
zone:  itsosjc2lpar4_P2_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0d
                50:05:07:68:01:10:b1:3f
                50:05:07:68:01:10:b0:c6
zone:  itsosjc2lpar4_P3_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0e
                50:05:07:68:01:10:b1:3f
                50:05:07:68:01:10:b0:c6
zone:  itsosjc2lpar4_P4_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0f
                50:05:07:68:01:10:b1:3f
                50:05:07:68:01:10:b0:c6

SAN768B-2_A1:FID128:root> setcontext 111
Public_A1:FID111:root>
zone:  itsosjc2lpar4_P1_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0c
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:b0:c6
zone:  itsosjc2lpar4_P2_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0d
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:b0:c6
zone:  itsosjc2lpar4_P3_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0e
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:b0:c6
zone:  itsosjc2lpar4_P4_ITSO_SAN Volume Controller_SPLIT
                c0:50:76:02:ba:dc:00:0f
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:b0:c6
```

SAN zoning

For SAN Volume Controller Stretched Cluster solution, we must create a virtual fabric by using the physical fabric. There is a public fabric for back-end storage controller connection and AIX host connection and a private fabric for the heartbeat across the sites between the SAN Volume Controllers.

Example 4-4 on page 242 shows the following Public zoning parameters:

- SAN Volume Controller_INTRA lists the SAN Volume Controller controllers
- SAN Volume Controller_V700A lists the Storwize V7000 controllers that are used as back-end storage in site A that is zoned with the SAN Volume Controller controllers
- SAN Volume Controller_V7000B lists the Storwize V7000 controller in site B zoned with the SAN Volume Controller controllers

- SAN Volume Controller_V7000C_Q lists the Storwize V7000 controllers in site C for the quorum zoned with the SAN Volume Controller controllers

Private configuration is only for the heartbeat between the SAN Volume Controllers in different sites to check the health and react accordingly in case of failure.

Example 4-4 Public zoning

```
Public zoning
 zone:  SAN Volume Controller_INTRA
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:37:e5
                50:05:07:68:01:40:27:e2
                50:05:07:68:01:40:b0:c6
 zone:  SAN Volume Controller_V7000A
                50:05:07:68:01:40:b1:3f
                50:05:07:68:02:20:00:ef
                50:05:07:68:02:10:00:ef
                50:05:07:68:02:10:00:f0
                50:05:07:68:02:20:00:f0
                50:05:07:68:01:40:37:e5
                50:05:07:68:01:40:27:e2
                50:05:07:68:01:40:b0:c6
 zone:  SAN Volume Controller_V7000B
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:37:e5
                50:05:07:68:02:20:54:cb
                50:05:07:68:02:10:54:cb
                50:05:07:68:02:10:54:ca
                50:05:07:68:02:20:54:ca
                50:05:07:68:01:40:27:e2
                50:05:07:68:01:40:b0:c6
 zone:  SAN Volume Controller_V7000C_Q
                50:05:07:68:01:40:b1:3f
                50:05:07:68:01:40:37:e5
                50:05:07:68:02:30:05:a8
                50:05:07:68:02:30:05:a9
                50:05:07:68:01:40:27:e2
                50:05:07:68:02:10:05:a9
                50:05:07:68:02:10:05:a8
                50:05:07:68:01:40:b0:c6

Private
Effective configuration:
 cfg:   ITSO_Private1
 zone:  ITSO_SAN Volume Controller_Intracluster
                50:05:07:68:01:30:b1:3f
                50:05:07:68:01:30:b0:c6
                50:05:07:68:01:30:37:e5
                50:05:07:68:01:30:27:e2

Private_A1:FID112:admin>

zone:  SAN Volume Controller_V7000A
50:05:07:68:01:40:b1:3f SAN Volume Controller 1
0:05:07:68:02:20:00:ef
```

```
50:05:07:68:02:10:00:ef V7000
50:05:07:68:02:10:00:f0
50:05:07:68:02:20:00:f0
50:05:07:68:01:40:37:e5 SAN Volume Controller 2
50:05:07:68:01:40:27:e2 SAN Volume Controller 3
50:05:07:68:01:40:b0:c6 SAN Volume Controller 4
```

4.5.2 AIX HOST path configuration

An important consideration to take into account here is that AIX clients that use a Subsystem Device Driver Path Control Module (SDDPCM) in a SAN Volume Controller Stretched Cluster solution should not have more than four paths showing.

In Example 4-5, an example configuration is shown where only four paths exist. More than four paths might lead to performance issues; therefore, a four-paths configuration is recommended.

Example 4-5 Four-path configuration

```
# pcmpath query device

Total Dual Active and Active/Asymmetric Devices: 8


DEV#:   0 DEVICE NAME: hdisk0  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF800000000000038
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0           fscsi0/path0          CLOSE   NORMAL        165          0
    1*          fscsi0/path1          CLOSE   NORMAL        147          0
    2           fscsi1/path2          CLOSE   NORMAL        151          0
    3*          fscsi1/path3          CLOSE   NORMAL        147          0

DEV#:   1  DEVICE NAME: hdisk1  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF800000000000022
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0          fscsi1/path12           OPEN   NORMAL    2573827          0
    1*         fscsi1/path13           OPEN   NORMAL         40          0
    2          fscsi0/path14           OPEN   NORMAL    2571208          0
    3*         fscsi0/path15           OPEN   NORMAL         40          0

DEV#:   2  DEVICE NAME: hdisk2  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF80000000000005B
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0*          fscsi1/path8           OPEN   NORMAL         14          0
    1           fscsi1/path9           OPEN   NORMAL   14756399          0
    2*         fscsi0/path10           OPEN   NORMAL         14          0
    3          fscsi0/path11           OPEN   NORMAL   14752856          0

DEV#:   3  DEVICE NAME: hdisk3  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF800000000000062
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
```

```
    0           fscsi0/path0           OPEN   NORMAL        848          0
    1*          fscsi0/path1           OPEN   NORMAL         98          0
    2           fscsi1/path2           OPEN   NORMAL        849          0
    3*          fscsi1/path3           OPEN   NORMAL         98          0

DEV#:   4  DEVICE NAME: hdisk4  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF80000000000005E
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0           fscsi0/path0           OPEN   NORMAL    8238335          0
    1*          fscsi0/path1           OPEN   NORMAL         98          0
    2           fscsi1/path2           OPEN   NORMAL    8242033          0
    3*          fscsi1/path3           OPEN   NORMAL         98          0

DEV#:   5  DEVICE NAME: hdisk5  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF800000000000063
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0*          fscsi0/path0          CLOSE   NORMAL        154          0
    1           fscsi0/path1          CLOSE   NORMAL        175          0
    2*          fscsi1/path2          CLOSE   NORMAL        154          0
    3           fscsi1/path3          CLOSE   NORMAL        156          0

DEV#:   6  DEVICE NAME: hdisk6  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF800000000000064
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0           fscsi0/path0          CLOSE   NORMAL        162          0
    1*          fscsi0/path1          CLOSE   NORMAL        147          0
    2           fscsi1/path2          CLOSE   NORMAL        154          0
    3*          fscsi1/path3          CLOSE   NORMAL        147          0

DEV#:   7  DEVICE NAME: hdisk7  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 600507680183053EF800000000000065
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0*          fscsi0/path0          CLOSE   NORMAL        147          0
    1           fscsi0/path1          CLOSE   NORMAL        166          0
    2*          fscsi1/path2          CLOSE   NORMAL        147          0
    3           fscsi1/path3          CLOSE   NORMAL        150          0
```

4.6 NPIV initial configuration

When FC is used in a typical SAN environment, the LUNs are created by using the physical storage and then mapped to physical host bus adapters (HBAs). Each physical port on each physical FC adapter is identified by its own unique worldwide port name (WWPN).

NPIV is an FC adapter virtualization technology. By implementing NPIV, you can configure your system so that each LPAR has independent access to the storage system that shares a physical adapter.

To enable NPIV in your managed system, you must install one or more virtual I/O servers. By creating servers' and clients' virtual FC adapters, the Virtual I/O Server provides a pass through to enable the client virtual server to communicate with the storage subsystem by using a shared HBA.

By using NPIV, you can have FC redundancy by using MPIO or by mirroring at the client partition. Because the Virtual I/O Server is just a pass through, the redundancy completely occurs in the client.

Figure 4-25 shows how the NPIV configuration looks in PowerVM and how it works. A separate server and client adapter are created for each client and server adapter works as pass through. Physical adapters in VIO server are used to map server virtual Fibre Channel adapter with client virtual Fibre Channel adapter. Then client WWPNs are zoned with storage controller to see the LUNs. Power Hypervisor takes care to pass I/O from virtual adapter to physical adapter.

Figure 4-25 Typical NPIV configuration

Example 4-6 shows a SAN Volume Controller Stretched Cluster environment NPIV example configuration. In Example 4-6, you can see FC Name, which relates to the physical adapter. Also, VFC client name shows the adapter name in the client machine. For more information about how to create this mapping, see this website:

http://www.ibm.com/developerworks/aix/library/au-NPIV/index.html

Example 4-6 NPIV configuration

```
$ lsmap -all -npiv
Name          Physloc                          ClntID ClntName       ClntOS
------------- ---------------------------------- ------ -------------- -------
vfchost0      U8233.E8B.0617C6P-V2-C4                 3 itsosjc1lpar1  AIX

Status:LOGGED_IN
FC name:fcs0                    FC loc code:U78A0.001.DNWHS36-P1-C2-T1
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0            VFC client DRC:U8233.E8B.0617C6P-V3-C10

Name          Physloc                          ClntID ClntName       ClntOS
------------- ---------------------------------- ------ -------------- -------
vfchost1      U8233.E8B.0617C6P-V2-C5                 4 itsosjc1lpar2  AIX

Status:LOGGED_IN
FC name:fcs0                    FC loc code:U78A0.001.DNWHS36-P1-C2-T1
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0            VFC client DRC:U8233.E8B.0617C6P-V4-C10

Name          Physloc                          ClntID ClntName       ClntOS
------------- ---------------------------------- ------ -------------- -------
vfchost2      U8233.E8B.0617C6P-V2-C6                 5 itsosjc1lpar3  AIX

Status:LOGGED_IN
FC name:fcs0                    FC loc code:U78A0.001.DNWHS36-P1-C2-T1
Ports logged in:3
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0            VFC client DRC:U8233.E8B.0617C6P-V5-C10
```

Figure 4-26 on page 247 shows that a typical configuration for NPIV in a Dual VIOS environment has two adapters from each fabric, which this leads to more than eight paths for a client. This is not recommended for an SAN Volume Controller SDDPCM MPIO solution. So, in this solution, we maintain redundancy by connecting each virtual adapter with one pass through physical adapter that is connected to a different switch. By using this configuration, the client has two paths from the first switch and two paths from the other switch.

Figure 4-26 SAN Volume Controller Stretched Cluster NPIV configuration for dual VIOS

4.7 Live Partition Mobility implementation

The use of LPM offers many advantages, including the ability to migrate a partition without having to stop the application. For more information about LPM, see *IBM PowerVM Live Partition Mobility*, SG24-7460. This publication also provides the reasons to use LPM and the advantages of LPM. The advantages include server consolidation, energy savings, the ability to perform scheduled maintenance of the server, such as firmware upgrades, and the ability to handle growth where the current system cannot handle the load. For example, you can move a workload from a POWER6 550 to a POWER 780 by using LPM.

In this section, we describe the LPM implementation and configuration that is specific to our Stretched Cluster solution.

4.7.1 LPM components

Inactive and active partition migration from one physical system to another is achieved through interaction between several components, as shown in Figure 4-27.

Figure 4-27 Live Partition Mobility components

These components and their roles are described in the following sections.

Hardware Management Console

The HMC is the central point of control. It coordinates administrator initiation and setup of the subsequent migration command sequences that flow between the various partition migration components.

The HMC provides a GUI wizard interface and a CLI to control migration. The HMC interacts with the service processors and POWER Hypervisor on the source and destination servers, the mover service partitions, the Virtual I/O Server partitions, and the mobile partition.

Resource Monitoring and Control

Resource Monitoring and Control (RMC) is a distributed framework and architecture that allows the HMC to communicate with a managed logical partition.

Dynamic LPAR resource manager

The Dynamic LPAR resource manager is an RMC daemon that runs inside the AIX, Linux, and Virtual I/O Server partitions. The HMC uses this capability to remotely run partition-specific commands.

Virtual asynchronous services interface

The source and destination mover service partitions use virtual asynchronous services interface (VASI), which is the new virtual device type, to communicate with the POWER Hypervisor to gain access to partition state.

The VASI device must be configured on the source and the destination Virtual I/O Servers for the mobile partition to participate in an active partition migration. Because the VASI device is used only by the mover service partition, it is required only for active partition migration.

Mover service partition

Mover service partition (MSP) is a new *attribute* of the Virtual I/O Server partition. It is a Virtual I/O Server partition that includes a function that asynchronously extracts, transports, and installs partition state. Two MSPs are involved in an active partition migration: one on the source system, the other on the destination system. MSPs are not used for inactive migrations.

An MSP must contain a single VASI device. This can be added or removed from a partition's profile or moved in or out of a running Virtual I/O Server partition by using dynamic reconfiguration.

POWER Hypervisor

Active partition migration requires server hypervisor support to process informational and action requests from the HMC and to transfer a partition state through the VASI in the mover service partitions.

VIOS

Only virtual adapters can be migrated with a partition. The physical resources that back the mobile partition's virtual adapters must be accessible by the VIOS on the source and destination systems.

Partition profiles

The HMC copies all of the mobile partition's profiles without modification to the target system as part of the migration process.

The HMC creates a migration profile that contains the partition's current state. Unless you specify a profile name when the migration is started, this profile replaces the existing profile that was last used to activate the partition. If you specify an existing profile name, the HMC replaces that profile with the new migration profile. Therefore, if you do not want the migration profile to replace any of the partition's existing profiles, you must specify a new, unique profile name when the migration is started. All profiles that belong to the mobile partition are deleted from the source server after the migration is complete.

If the mobile partition's profile is part of a system profile on the source server, it is automatically removed after the source partition is deleted. It is not automatically added to a system profile on the target server.

Time reference

Time reference is a new attribute of partitions, including Virtual I/O Server partitions. This partition attribute is supported only on managed systems that are capable of active partition migration.

Synchronizing the time of day clocks for the source and destination Virtual I/O Server partitions is optional for active and inactive partition migration. However, it is a recommended step for active partition migration. If you choose not to complete this step, the source and destination systems synchronize the clocks while the mobile partition is moving from the source system to the destination system.

The time reference partition (TRP) setting was introduced to enable the POWER Hypervisor to synchronize the mobile partition's time of day as it moves from one system to another. It uses Coordinated Universal Time (UTC) that is derived from a common network time protocol (NTP) server with NTP clients on the source and destination systems. More than one TRP can be specified per system. The POWER Hypervisor uses the longest running time reference partition as the provider of authoritative system time. It can be set or reset through POWER Hypervisor while the partition is running.

4.7.2 Other components affecting LPM

Though not considered to be part of LPM, there are other System p server components that can influence (or are influenced by) the mobility of a partition.

Integrated Virtual Ethernet adapter

A host Ethernet adapter uses a two or four-port integrated Ethernet adapter, which is directly attached to the POWER Hypervisor. The hypervisor can create up to 32 logical Ethernet ports that can be given to one or more logical partitions. This configuration provides a partition with a virtual Ethernet communications link, which is seen as a logical host Ethernet adapter (LHEA), without recourse to a shared Ethernet adapter in a Virtual I/O Server.

Performance monitor API

The performance monitor API (PMAPI) is an AIX subsystem that is consists of commands, libraries, and a kernel extension that controls the use of the POWER performance registers.

Barrier synchronization registers

Barrier synchronization registers (BSRs) provide a fast, lightweight barrier synchronization between CPUs.

This facility is intended for use by application programs that are structured in a single instruction, multiple data (SIMD) manner. Such programs often proceed in phases where all tasks synchronize processing at the end of each phase. The BSR is designed to accomplish this efficiently. Barrier synchronization registers cannot be migrated or reconfigured dynamically.

4.8 LPM operations

Partition migration can be performed as an inactive (non-running LPAR) or an active ("live" running LPAR) operation.

The steps of inactive and active partition migration use the following same four-step sequence:

1. Preparation: Ready the infrastructure to support Live Partition Mobility.
2. Validation: Check the configuration and readiness of the source and destination systems.
3. Migration: Transfer of partition state from the source to destination takes place. The same command is used to start inactive and active migrations. The HMC determines the appropriate type of migration to use based on the following state of the mobile partition:
   - If the partition is in the Running state, the migration is active.
   - If partition is in the Not Activated state, the migration is inactive.
4. Completion: Free unused resources on the source system and the HMC.

We describe these steps in the next sections.

4.8.1 Inactive migration

The following basic steps are required for inactive migration:

1. Prepare the mobile partition for migration (if required), such as remove adapters that are not supported and ensure applications support mobility.
2. Shut down the mobile partition.
3. Perform the migration validation procedure that is provided by the HMC to verify that the migration can be performed successfully.
4. Start the inactive partition migration by using the HMC. The HMC connects to source and destination systems and performs the following migration steps:
   - Transfers the mobile partition's configuration from source to destination, including all partition profiles.
   - Updates the destination Virtual I/O Server to provide virtual SCSI access to the mobile partition.
   - Updates the source Virtual I/O Server to remove resources that were used to provide virtual SCSI access to the mobile partition.
   - Removes the mobile partition configuration on the source system.
5. When migration is complete, the mobile partition can be activated on the destination system.

The steps that are run are similar to those an administrator would follow when they are performing a manual migration. These tasks normally require accurate planning and a system-wide knowledge of the configuration of the two systems because virtual adapters and virtual target devices must be created on the destination system and follow virtualization configuration rules.

The inactive migration task takes care of all planning and validation and performs the required tasks without user action. This mitigates the risk of human error and runs the movement in a timely manner.

Figure 4-28 on page 252 shows the inactive migration workflow.

Figure 4-28 Inactive migration workflow

This workflow includes the following features:

- The HMC inhibits any changes to the source system and the mobile partition that might invalidate the migration.
- The HMC extracts the virtual device mappings from the source Virtual I/O Servers and uses this to generate a source-to-destination virtual adapter migration map. This map ensures that there is no loss of multipath I/O capability for virtual SCSI and virtual Ethernet. The HMC fails the migration request if the device migration map is incomplete.
- The HMC creates a compatible partition shell on the destination system.
- The HMC creates a migration profile for the mobile partition's current (last-activated) profile. If the mobile partition was last activated with profile `my_profile` and resources were moved in to or out of the partition before the partition was shut down, the migration profile will differ from that of `my_profile`.
- The HMC copies over the partition profiles. This includes all the existing profiles that are associated with the mobile partition on the source system and the migration profile. The existing partition profiles are not modified at all during the migration; the virtual devices are not re-mapped to the new system.
- The HMC creates the required virtual SCSI adapters in the Virtual I/O Servers on the destination system and completes the LUN to virtual adapter mapping.
- On completion of the transfer of state, the HMC sets the migration state to completed and informs the POWER hypervisor on the source and destination.

When the migration is complete, the following unused resources are deleted:

- The source Virtual I/O Servers remove the virtual adapter slots and virtual target devices that are used by the mobile partition. The HMC removes the virtual slots from the source Virtual I/O Server's profile.
- The HMC deletes the partition on the source server.

Important: Virtual slot numbers can change during migration. When a partition is moved to a server and then back to the original, it does not have the same slot numbers. If this information is required, you should record slot numbers.

4.8.2 Active migration

The active partition migration function provides the capability to move a running operating system, hosted middleware, and applications between two systems without disrupting the service that is provided. Databases, application servers, network and SAN connections, and user applications are all transferred in a manner not apparent to users. The mobile partition retains its name, active state, nonvolatile random access memory (NVRAM), profiles, and current configuration. Its virtual I/O resources are assigned and remapped to the appropriate Virtual I/O Server partitions on the destination system.

The following basic steps are required for an active migration:

1. Prepare the mobile partition for migration, keeping it active, such as remove adapters that are not supported and ensure applications support mobility.
2. Perform the migration validation procedure that is provided by the HMC to verify that the migration can be performed successfully.
3. Initiate the active partition migration by using the HMC.

   The HMC connects to source and destination systems and performs the following migration steps:

   – Transfers the mobile partition's configuration from source to destination, including all the partition's profiles.
   – Updates the destination Virtual I/O Server to provide virtual SCSI access to the mobile partition.
   – Activates the mover service partition function on the source and destination Virtual I/O Servers. The mover service partitions copy the mobile partition's state from the source to the destination system.
   – Activates the mobile partition on the destination system.
   – Updates the source Virtual I/O Server to remove resources that are used to provide virtual SCSI access to the mobile partition.
   – Removes the mobile partition configuration on the source system.

Active migration performs similar steps to inactive migration, but it also copies physical memory to the destination system. It keeps applications running, regardless of the size of the memory that is used by the partition. Also, the service is not interrupted, the I/O continues to disk, and network connections keep transferring data.

In addition to the partition definition and resource configuration, active migration involves the transfer of active runtime state, which includes the following features:

- Partition's memory
- Hardware page table (HPT)
- Processor state
- Virtual adapter state
- NVRAM
- Time of day (ToD)
- Partition configuration
- State of each resource
- Partition configuration

It is the mover service partitions on the source and destination (under the control of the HMC) that move this state between the two systems.

After you create the Virtual I/O Servers, mover service partitions, and VASI devices, you must prepare the source and destination systems for migration.You must ensure that the LPM prerequisites are met before the partition is migrated and that the source and target system is adequately prepared. For more information, see 1.10, “Live Partition Mobility” on page 78.

In summary, the following preparation tasks are needed:

1. Synchronize the time of day clocks on the mover service partitions by using an external time reference, such as the NTP. This is an optional step that increases the accuracy of time measurement during migration. It is not required by the migration mechanisms and even if this step is omitted, the migration process correctly adjusts the partition time. Time never moves backwards on the mobile partition during a migration.
2. Prepare the partition for migration.

   Use dynamic reconfiguration on the HMC to remove all dedicated I/O, such as PCI slots, GX slots, and HEA from the mobile partition.

   Remove the partition from a partition workload group.
3. Prepare the destination Virtual I/O Server.

   Configure the shared Ethernet adapter as necessary to bridge VLANs.

   Configure the SAN such that requisite storage devices are available.
4. Initiate the partition migration by using the GUI or CLI on the HMC.

   Select the partition to migrate and destination system and, optionally, the mover service partitions on the source and destination systems.

   If there is only one active mover service partition on the source or the destination server, the mover service partition selection is automatic. If there are multiple active mover service partitions on one or both, you can specify which ones to use or let the HMC choose for you.

   Optionally, select (in GUI) or specify (in CLI) specific virtual device mappings in the destination Virtual I/O Server.

After the source and destination mover service partitions are identified, the HMC performs a series of validation pre-checks (configuration and resource availability) to ensure that the user is performing a valid migration, that there are no high-level blocking problems, and that the environment satisfies the prerequisites for a migration operation.

If all the validation checks pass, the HMC initiates the migration procedure. From this point forward, all state changes are rolled back in the event of an error.

Figure 4-29 on page 255 shows the tasks and workflow of the migration phase of an active migration.

Figure 4-29 Migration phase of an active migration

For active partition migration, the transfer of partition state follows a path as shown in Figure 4-30. This shows the Active migration partition state transfer path between a Source System (Power Server) and Destination System (Power Server).

Figure 4-30 Active migration transfer path

In Figure 4-30, the numbers represent the following paths:

1. From the mobile partition to the source system's hypervisor.
2. From the source system's hypervisor to the source mover service partition.
3. From the source mover service partition to the destination mover service partition.
4. From the destination mover service partition to the destination system's hypervisor.
5. From the destination system's hypervisor to the partition shell on the destination.

The migration process includes the following steps:

1. The HMC creates a compatible partition shell on the destination system. This shell partition is used to reserve the resources that are required to receive the inbound mobile partition.

   The pending values of the mobile partition are not preserved across the migration; the current values of the partition on the source system become the pending and the current values of the partition on the destination system. The configuration of the partition on the source system includes the following configurations:

   – Processor configuration, which is dedicated or shared processors, processor counts, and entitlements (minimum, maximum, and wanted).
   – Memory configuration (minimum, maximum, and wanted)
   – Virtual adapter configuration

   The creation of the partition shell on the destination system ensures that all of the required resources are available for the mobile partition and cannot be taken at some point during the migration. The current partition profile associated with the mobile partition is created on the destination system.

2. The HMC configures the mover service partitions on the source and destination systems. The two movers establish the following components:

   – A connection to their respective POWER hypervisor through the Virtual Asynchronous Services Interface (VASI) adapter.
   – A private, full-duplex communications channel between themselves over a standard TCP/IP connection for transporting the moving partition's state.

3. The HMC issues a prepare command for migration event to the migrating operating system (which is still on the source system) that gives the mobile partition the opportunity to get ready to be moved. The operating system passes this event to registered kernel extensions and applications so that they can take any necessary actions, such as reducing memory footprint, throttling workloads, loosening heartbeats, and other timeout thresholds. The operating system inhibits access to the PMAPI registers and zeros internal counters upon receipt of this event.

   If the partition is not ready to perform a migration at this time, it returns a failure indicator to the HMC, which cancels the migration and rolls back all changes.

4. The HMC creates the virtual target devices and virtual SCSI adapters in each of the Virtual I/O Servers on the destination system that hosts the virtual SCSI client adapters of the mobile partition. This step uses the virtual adapter migration map that was created during the validation phase. The migration is stopped in the event of an error.

5. The mover on the source system starts sending the partition state to the mover on the destination system, copying the mobile partition's physical pages to the physical memory reserved by the partition shell on the destination.

6. Because the mobile partition is still active, with running applications, its state continues to change while the memory is moved from one system to the other. Memory pages that are modified during the transfer of state are marked modified, or *dirty*. After the first pass, the source mover re-sends all the dirty pages. This process is repeated until the number of pages that are marked as dirty at the end of each loop no longer decreases, or is considered sufficiently small, or a timeout is reached. Based on the total number of pages that are associated with partition state and the number of pages left to transmit, the mover service partition instructs the hypervisor on the source system to suspend the mobile partition.

7. The mobile partition confirms the suspension by quiescing all its running threads.

   The partition is now suspended.

8. During the partition suspension, the source mover service partition continues to send partition states to the destination server.

   The partition is now resumed.

9. The mobile partition resumes running on the destination server and re-establishes its operating environment. This is the point of no return; the migration can no longer be rolled back to the source. If the migration fails after this, recovery completes the migration on to the destination system. It is possible that the mobile partition resumes running before all its memory pages are copied to the destination. If the mobile partition requires a page that has not yet been migrated, it is demand-paged from the source system. This technique significantly reduces the length of the pause, during which the partition is unavailable.

10. The mobile partition recovers I/O and retries all pending I/O requests that were not completed while on the source system. It also sends a gratuitous ARP requests on all VLAN virtual adapters to update the ARP caches in the various switches and systems in the external network.

    The partition is now active and visible again.

11. When the destination mover service partition receives the last dirty page from the source system, the migration is complete.

    The period between the end of step 7 and the end of step 10 is called the *suspend window*. It lasts just a few seconds.

The final steps of the migration return all of the resources to the source and destination systems and restore the partition to its fully functional state.

An AIX partition continues running during an active migration. Most AIX features work seamlessly before, during, and after the migration, including the following features:

- System and advanced accounting
- Workload manager
- System trace
- Resource sets (including exclusive use processor resource sets)
- Pinned memory
- Memory affinity
- Kernel and kernel extensions
- Large memory pages (huge memory pages cannot be used)
- Processor binding

  Processes remain bound to the same logical processor throughout the migration.

Performance monitoring tools (such as `topas`, `tprof`, `filemon`, and so on) can run on a mobile partition during an active migration. However, the data that these tools report during the migration process might not be significant because of underlying hardware changes, performance monitor counters that might be reset, and so on.

Although AIX is migration safe, verify that any applications you are running are migration safe or aware.

Application migrations can fail because of dependencies and differences on binary files and compiler sets, and dependencies on Barrier Synchronization Registers. For more information, see 4.7.2, “Other components affecting LPM” on page 250. Also review your specific applications readme file.

4.9 Combining LPM with other features

There are many ways to take advantage of LPM. It can be used to perform actions that were not previously possible because of complexity or time constraints. Migration is a new function that can be combined with existing IBM System p features and software to provide better and more flexible service.

4.9.1 High availability clusters

An environment that has only small windows for scheduled downtime can use LPM to manage many scheduled activities to reduce downtime through inactive migration or to avoid service interruption through active migration.

For example, if a system must be shut down because of a scheduled power outage, its hosted partitions can be migrated to powered systems before power is cut.

LPM is a reliable procedure for system reconfiguration and it can be used to improve the overall system availability.

High availability environments also require the definition of automated procedures that detect software and hardware events and activate recovery plans to restart a failed service as soon as possible.

LPM increases global availability but it is not a high availability solution on its own. It requires source and destination systems be operational and that the partition is not in a failed state. In addition, it does not monitor operating system and application state and it is a user-initiated action.

Unplanned outages still require specific actions that are normally run by cluster solutions such as IBM PowerHA.

Cluster software and LPM provide different functions that can be implemented as Cluster software. While LPM provides different functions that can be used together to improve the availability and uptime of applications, they also can simplify administration, which reduces the related cost.

In Figure 4-31 on page 259, the integrated PowerVM, PowerHA, SAN Volume Controller Stretched Cluster LPM configuration for the solution that is used in this book is shown.

Figure 4-31 LPM configuration in SAN Volume Controller Stretched Cluster

4.10 LPM configuration

In this section, we describe the LPM configuration as it relates to our SAN Volume Controller Stretched Cluster solution.

4.10.1 LPM prerequisites

In this section, we provide links to sources of more information to which we referred to ensure that our environment satisfied all of the necessary prerequisites.

For more information, see the following resources:

- "LPM prerequisites" on page 80.
- *IBM PowerVM Live Partition Mobility*, SG24-7460, which is available at this website:

  http://www.redbooks.ibm.com/redbooks/pdfs/sg247460.pdf
- *LMP IBM Information Center*:

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/iphc3/iphc3.pdf

- *LPM Preparation Checklist*:

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3hmcprepmobilelpar.htm

The LPM Configuration consists of the following categories:

- System Level Preparation
- VIOS Server Preparation
- Client Configuration
- Live Migration Process
- Error Identification and Resolution

We describe each of these in the following topics.

4.10.2 System level preparation

The Power Systems environment should be prepared in advance for LPM migration; at the systems level in terms of logical memory, and validating the CPU levels of the system. This is important, especially in respect to an active partition that can be running at high CPU and memory usage, which can slow down LPM.

Set default logical memory block

Complete the following steps to set the default logical memory block:

1. Log in to the server Advanced System Management (ASM) interface and set the default logical memory block size, based on your requirements. From **HMC** → **Servers**, choose the server that you need to change then click **Operations** → **Launch Advanced Systems Management (ASM)**.
2. Choose OK for the selection of the HMC to connect to the ASM IP interface. Log in to ASM with your user name and password.
3. From the left side menu, click **Performance Setup** → **Logical Memory Block Setup** → **Settings** and set to `256MB`.

Important: This is reflective of our solution needs for testing, and most servers have 32 GB RAM or more in real-life scenarios. Ultimately, this depends on the size of your LPARs and the memory allocation that is based on the database, middleware, or application engine that is hosted on the LPAR.

Validate the CPU levels of the system

Complete the following steps to validate the CPU levels of the system:

1. Log in to HMC Shell via SSH and list the current settings.
2. Browse to the Manage Profiles Menu option in the HMC (select **Server** under **Systems Management** → **Configuration** → **Manage Profiles**).
3. Choose the profile of the client partition (or partitions) you want to be in the LPM cluster and edit their profile.
4. Under the Processors settings, set the processor type to the lowest common CPU compatibility type of any server that you are going to enable in the LPM cluster. This is an important step because the source and target servers in the LPM cluster must be identical in terms of the configuration.

4.10.3 VIOS preparation

Complete the following steps to prepare the VIOS:

1. Enable Mover Service on any and all VIOS partitions that are source or targets for LPM. Choose the system and select **Properties**.
2. Enable Time Sync for VIOS systems to HMC as Master. This is a best practice because the HMC provides central management for the Power Systems environment.
3. For every VIOS partition that is a member in the LPM cluster, you must allocate the LUN (or LUNs) for the client partitions, scan them in, and set the physical volume to allow shared reservations. For more information about how to allocate the LUNs, see the SAN Volume Controller documentation.
4. Log in to the first VIOS system on which the client partition definition is built and list the current LUNs that are known by the VIOS system.
5. Scan for the new LUN for the client partition and list the physical volumes.
6. Show the details of the new `hdisk#` LUN by using the **`lscfg -vl hdisk#`** command to validate that the assigned LUN ID matches the `hdisk#`.

   Important: Convert the hexadecimal number that is shown by the **`lscfg`** command to a decimal value to validate it matches the assigned LUN ID set in the SAN Volume Controller / SAN. For more information, see this website:

   http://www.ascii.cl/conversion.htm

7. After you are sure that the LUN is correct, set the LUN on the VIOS server to `no_reserve`.

4.10.4 Client configuration

Complete the following steps to set the partition profile settings and the shard disk settings in the client configuration.

Partition profile settings

Complete the following steps to set the partition profile settings:

1. Telnet into the Client AIX hosted partition and check if the service connection to HMC is active.
2. Complete the following steps to disable the Redundant Path Management option so as not to have any potential bottlenecks that would interfere with LPM:
   a. From the HMC, choose each partition that participates in LPM and choose **Properties**.
   b. Click the link of the profile that is reflective of the LPAR you want to modify.
   c. Choose the Settings tab and ensure that Enable redundant error path reporting option is not selected. Also ensure it is not a member of any partition workload group.

   Important: If the Enable Redundant Error Path Reporting option was enabled, you must restart the partition to have it take effect.

3. Ensure no virtual serial beyond 0 and 1 (HMC reserved) are enabled on the host.
4. Ensure that the mobile partition is not using barrier synchronization register (BSR) arrays. BSRs cannot be migrated or reconfigured dynamically.

5. Click the link directly from the host's client partition list.
6. Choose **Hardware**, then the Memory tab and ensure BSR=0.
7. At the server level, choose the server and **Properties**.
8. Choose the Advanced tab and ensure the Huge Page Memory is set to 0 (zero).
9. Ensure the client partitions have no physical Adapters or LHEA ports assigned.
10. Ensure the partition name is unique among all Power Systems that participate in the LPM cluster. This configuration provides a specific flagging or identification for the partition so there are no issues and confusion with the other clusters.

Shared disk settings

Complete the following steps to set the shared disk settings:

1. Telnet into the HMC host and check the disk (or disks) that are assigned to each partition are set to non-reserve.
2. Validate the disks have unique signature IDs.

> **Important:** Ensure that the LUN has a disk signature and ID. Try to specify in the ID that the disk is associated with the LPM Partition.

3. Complete the following steps to record the SCSI Slot number that the host assigned:
   a. Open the client partition profile properties.
   b. Click the link for the host profile details.
   c. Choose **Virtual Adapters** and record the host and VIOS slot number of the SCSI adapter.
   d. Review the VIOS server settings to ensure the adapter is present and correlates to the client partition. Ensure that the adapter is not listed as required to boot for the VIOS system. Also ensure it has enough room for more slot expansions to be added.
4. Validate the host's mapping and note that the VIOS (Remote) has mappings to the LUN.

4.11 LPM

For more information about Inactive and Active LPAR Migration Processing, see 4.8, "LPM operations" on page 251.

The information in this section is provided at a lower level of specific procedural detail to how we ran LPM in our Redbooks solution implementation.

4.11.1 Source system validation

Complete the following steps to perform the source system validation:

1. From the HMC, choose the source system and validate migration.
2. Select the target system from the list to which you want the client partition to migrate to.
3. Enter the name of the partition that is generated as part of the target system's profile. Input the host name to maintain consistency (remember the guidelines for identifying or flagging the LPM only assigned partitions). Choose the target VIOS partition with the available slot matching the host's SCSI slot number.

4. Click **VLAN Configuration** and ensure that the VLAN is selected on one of the two VIOS partitions. This ensures that packets can egress the system. Click **Ok** when you are done reviewing.
5. Choose **Validate** to start the validation process.
6. A return back to the base window without error constitutes a valid migration plan.

Figure 4-32 shows how to select the validation menu through HMC.

Figure 4-32 Selecting the validation menu

Figure 4-33 shows the window that opens after Validate is selected. Here we entered the remote HMC IP and user name, then clicked **Refresh Destination System**.

Figure 4-33 Enter remote HMC IP

Figure 4-34 shows the list of managed systems that are available in the remote HMC. We click **Validate**.

Figure 4-34 Managed systems

Figure 4-35 shows that our validation is successful. It also shows the list of VFC and Ethernet configuration it detected. We click **Migrate**.

Figure 4-35 Successful validation

4.11.2 Migration

Now that the validation is complete, click **Migrate** to start the actual migration process and complete the following steps:

1. Watch the progress and look carefully for errors (for more information about common errors and their resolution, see 4.12, “Error identification and resolution” on page 267).
2. After the migration is complete, you see the progress indicator specify 100% and the status is “Success”.
3. Validate that the migration is complete. Select the target host and review the defined partitions to ensure the partition definition exists.
4. Select the original source partition and validate that the partition does not exist.
5. Ensure that you can open a remote shell into the host to validate the VLAN transfer is functioning.

4.12 Error identification and resolution

This section identifies some of the following common errors you might encounter while you are performing LPM and the recommended resolution:

- Error: LPM fails on validation of Virtual Host migration.

  Resolution: In the HMC, for the server you are changing, under the Logical Partition Profile Properties tab (Memory tab), adjust the VIOS partition definition of the wanted memory to an even multiple of RAM.

- Error: When validation is made, it fails on the processor check.

  Resolution: Change the processor type to make it compatible between source and destination CEC. For more information about the steps to perform this task, see this website:

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3hmcpreppcm.htm

- Error: When validation is run on client LPM, the test fails when you are trying to create the Virtual Adapter on the target system.

  Resolution: The physical Volume that is allocated to each VIOS server in the LPM cluster must have the **cfgmgr** command run (see the LUN) and the LUN set to allow sharing via the `chdev -dev hdisk## -attr reserve_policy=no_reserve`.

In the next section, we describe some other considerations for Live Application Mobility (LAM), which are beyond the intended scope of our solution. However, you might want to further investigate based upon any specific requirements in your environment.

4.13 LAM considerations

The IBM PowerVM offering extends the virtualization further to include software virtualization. This function is called *workload partition* (WPAR). An LPAR is a hardware-based partitioning feature that allows you to create multiple independent operating system environments, which are called LPARs. Each LPAR can run a version of AIX, Linux, or IBM i. A WPAR is built within an AIX partition. Therefore, WPARs are software-created, virtualized operating system environments within a single instance of the AIX operating system.

Figure 4-36 shows LAM considerations.

Figure 4-36 LAM considerations

4.14 PowerHA SystemMirror initial configuration and other considerations

PowerHA for AIX (formerly known as HACMP) is the IBM premier high availability solution for Power Systems that are running AIX. In addition to high availability, PowerHA provides the multi-processing component. The multi-processing capability comes from the fact that in a cluster there are multiple hardware and software resources that are managed by PowerHA to provide complex application functionality and better resource usage.

IBM also designed an extended version of PowerHA that provides disaster recovery functionality that is integrated is a solution known as PowerHA Extended Distance (PowerHA/XD), which supports PowerHA functionality between two geographic sites. PowerHA/XD supports a number of distinct methods for replicating the data and is described in later sections of this book.

PowerHA 7.1.1 introduces many improvements from PowerHA 6.1 and older versions. Most of the improvements come from changes and improvements to the base AIX operating system. AIX Version 7 is enhanced to maintain a few nodes by using CAA capabilities. Most of the PowerHA monitoring infrastructure is part of the base operating system. For earlier versions of AIX to take advantage of PowerHA 7.1.1, CAA is included in AIX 6.1 TL6. CAA requires Reliable Scalable Cluster Technology (RSCT) 3.1.

PowerHA features subnet requirements and needs a number of interfaces because PowerHA monitors failures and moves IP addresses to a surviving interface before it moves to the next node. AIX provides methods of monitoring the network adapter by using Etherchannel, which is implemented as link aggregation or a network interface backup. The use of Etherchannel eliminates the need for configuring multiple interfaces because this requirement is taken care of by implementing Etherchannel.

The Virtual I/O Server provides a method of creating a Shared Ethernet Adapter Failover (SEA Failover), which allows the Virtual I/O Server to provide required redundancy. An SEA configuration also removes the need to create multiple interfaces in the PowerHA configuration.

For more information, see *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg248030.html?Open

To get started with PowerHA or to migrate to PowerHA 7.1, see *PowerHA for AIX Cookbook*, SG24-7739, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247739.html?Open

4.14.1 Considerations before PowerHA 7.1.1 is used

You must be aware of the following considerations before PowerHA 7.1.1 is used:

- You cannot change the host name in a configured cluster. After the cluster is synchronized, you cannot change the host name of any of the cluster nodes. Therefore, changing the host name is not supported.
- You cannot change the cluster name in a configured cluster. After the cluster is synchronized, you cannot change the name of the cluster. If you want to change the cluster name, you must completely remove and re-create the cluster.

- You cannot change the repository location or cluster IP address in a configured cluster. After the cluster is synchronized, you cannot change the repository disk or cluster multicast IP address. To change the repository disk or the cluster multicast IP address, you must completely remove and re-create the cluster.
- No IPV6 support is available, which is a restriction from the CAA implementation.

Basic recommendations for preparing the environment

The following steps must be taken before you can configure an PowerHA cluster and make it available.

1. Make sure that the hardware you are using for the two or more servers is as similar as possible. However, this is a recommendation, not a restriction.
2. Size the environment in such a way that each node can manage all the resource groups simultaneously (this step should coincide with the first step). If you decide that you have multiple resource groups running in the cluster, assume a worst-case scenario where one node must run everything at once. Ensure that the servers have adequate processing power to cover everything.
3. If you use SAN disks for storage, the disks for the shared volume groups must be zoned to all nodes. The network VLANs, subnets, and addresses should be connected in the same fashion. Work with your SAN and network administrators to get addresses and disks for the boot, persistent, and service IP addresses.
4. The entire operating system configuration must match between the nodes. The user IDs, third-party software, technology levels, and service packs must be consistent. One of the best ways to make this happen is to build out the intended configuration on one node, make a `mksysb` backup, and use that to build out all subsequent nodes. After the servers are built, consider them joined at the same time.

Important: Make changes on both servers consistently all of the time.

With all of the virtualization technology available today, it is far more worthwhile to use VIO to create a pair of production and development LPARs on the same set of System p servers and hardware resources. Use things such as shared processor weights, maximum transmission unit (MTU) sizes, and RAM allocation to give the production LPARs more clout than the development LPARs. Doing so creates an environment that can handle a failover and assures managers and accountants that finances are used wisely.

4.14.2 PowerHA installation and configuration

In the beginning of the configuration phase, usually limited information about our cluster configuration is available. You might know only the application and its basic requirements: required disk size, mount points, and the designated service IP address. Starting from this information, you can gradually collect all required data for the cluster configuration. Follow the steps that are described in 4.14.3, “High-level PowerHA configuration steps” on page 271 to design and implement your own cluster.

The following concepts are the most important design concepts of the PowerHA cluster:

- Avoid a single point of failure (SPOF)
- Keep it simple

Figure 4-37 shows an overview of the steps to use or deploy PowerHA SystemMirror.

Figure 4-37 PowerHA SystemMirror implementation steps

For more information about hardware and software requirements, see Chapter 3, “SAN Volume Controller Implementation” on page 149.

There are also considerations for VIOS clients and virtual Ethernet. For more information about installing PowerHA cluster, see Chapter 4 of *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg248030.html?Open

For more information about configuring PowerHA cluster, see Chapter 5 of *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030.

4.14.3 High-level PowerHA configuration steps

Following are the high-level steps that we performed to configure the cluster in our test environment.

In this example, we set up a simple two-node cluster for DB2 mutual takeover across two Ethernet networks: one shared volume group on a SAN disk that also uses a second SAN disk for a heartbeat and with an application managed by PowerHA in one resource group.

Important: This process assumes that all IP addresses are predetermined and that the SAN zoning of the disks is complete. Unless otherwise stated, you must run the tasks here on each and every node of the cluster.

Complete the following steps to configure the cluster in the test environment:

1. Install the PowerHA software.

   You can purchase this software from IBM directly. The file sets all start with the word cluster. Use the **installp** command to install the software, much like any other licensed program package (LPP).

2. Edit some flat files.

   Put all of the IP addresses that are associated with the cluster (boot, persistent, and service) into each `/etc/cluster/rhosts` file on each node of the cluster. Verify that the server host names match the appropriate IP addresses. The server's host name also must match the persistent IP address.

3. Configure the boot IP addresses.

   Run the **smitty chinet** command and set the boot IP addresses for each network adapter. Make sure that you can ping and connect freely from node to node on all respective networks. Also, double-check to make sure that the default route is properly configured. If it is not configured, run the **smitty tcpip** command, go to the Minimum Configuration menu, enter the default route for the primary adapter, and press Enter.

4. Define the cluster.

   Go to **smitty hacmp** → Cluster Nodes and Networks → Standard Cluster Deployment → Setup a Cluster, Nodes and Networks then **smitty hacmp** → Cluster Nodes and Networks → Standard Cluster Deployment → Define Repository Disk and Cluster IP address.

5. Define the service IP addresses.

   Run the **smitty cm_config_hacmp_service_ip_labels_addresses_menu_dmn** command, which defines the service IP addresses.

6. Install and configure DB2 with instances.

7. Use smart assist.

   Run the **smitty hacmp** command then click Cluster Applications and Resources → Make Applications Highly Available (Use Smart Assists) → Add an Application to the PowerHA SystemMirror Configuration. Select DB2 UDB non-DPF Smart Assist from the list. DB2 is automatically configured for high availability.

8. Perform a synchronization.

   Synchronize the cluster configuration. Run the **smitty cm_ver_and_sync** command. Set Automatically correct errors found during verification? to Interactive. Correct any problems along the way.

9. Start the cluster.

   At this point, the cluster is ready to start. On one of the nodes, run the **smitty clstart** command and pick that particular node. Our preference is not to have the cluster start on reboot because if there is a PowerHA-related problem on startup, it can be difficult to troubleshoot it. After the node comes up with the resources available, start the cluster on the other node.

10. Perform a test failover.

    One way to test PowerHA's adaptability is to reboot the active node and let things fail over naturally while the **tail -f /tmp/hacmp.out** command is run on the other node to watch as things go over. You also can run the **smitty cl_resgrp_move.node_site** command. Select the Move Resource Groups to Another Node option.

11. Perform failure testing

    If you really want to make sure your cluster is solid, perform testing by disconnecting cables and seeing how the resources move back and forth. The more you test, the more reliable your cluster becomes.

Application and resource group configuration

We have two applications, `app1` and `app2` for DB2 in a mutual takeover configuration. Each application has a service IP address (`instsvc1` and `instsvc2`), a volume group (`app1vg`, `app2vg`), and a file system (`/data1` and `/data2`). The cluster consists of two resource groups, one for each application. The resource groups are named `inst1rg` and `inst2rg`. The resource groups contain the application, related volume groups, file systems, and service IP addresses.

Network configuration

Our main service network is 191.168.150.0/24 and the default gateway is 191.168.150.1. This network is used for application client communication. We configure a persistent and service IP for each node from this subnet. The persistent IP label is the host name. The base IP ranges are 10.1.0.0/23 and 10.10.0.0/23. These subnets are not used anywhere else in our company network; the routers do not forward them. We use the default net_ether_01 name for this service network in the cluster configuration.

Hardware configuration

We use two IBM Power 770 servers. We have two VIOSs on each system. The VIOSs are connected to our Ethernet network by two adapters. For SAN connectivity, we use two Fibre Channel (FC) adapters per VIOS. All LPARs are fully virtualized with virtual Ethernet and FC adapters.

Virtual I/O Server configuration

Our servers have the following VIOS configuration:

- Two VIOSs on each server
- SEA failover for all Ethernet interfaces
- Virtual FC disks for rootvg and datavg
- NPIV port virtualization for SAN connectivity

For the start of the PowerHA SystemMirror installation, we assume that you have the following environment:

- Two servers
- Fully working and tested redundant VIOS configuration
- Working Ethernet connectivity and outside network access
- Storage access, including SAN zones and LUN mapping
- Two installed AIX 6.1/7.1 LPARs
- No IP that is configured on the LPARs
- No default gateway that is set

4.14.4 PowerHA SystemMirror configuration with DB2

PowerHA SystemMirror Smart Assist for DB2 extends an existing PowerHA SystemMirror configuration to include monitoring and recovery support for DB2 Universal Database™ (UDB) Enterprise Server Edition.

PowerHA SystemMirror Smart Assist for DB2 allows you to automatically configure PowerHA SystemMirror in an environment where DB2 already is configured to make non-partitioned DB2 instances highly available.

PowerHA SystemMirror Smart Assist cannot be used to configure a cluster in a partitioned (DB2 UDB DPF) environment.

To make use of PowerHA SystemMirror Smart Assist for DB2, the DB2 UDB software must be installed on all nodes that are going to be part of the cluster, and DB2 instances must be configured on some of the nodes. The shared physical disks that contain the DB2 instances must be configured on all cluster nodes that serve as takeover nodes for the DB2 instances.

PowerHA SystemMirror Smart Assist automatically discovers your DB2 configuration. You can use this to create one of the following cluster configurations:

- A hot standby two-node cluster
- A mutual takeover two-node cluster
- A cluster with more than two nodes and more than two instances where some of the nodes serve as takeover nodes for the configured DB2 instances.

We used mutual takeover configuration. An example DB2 configuration is being used for our test environment is shown in the following section.

For more information about configuring PowerHA SystemMirror with DB2, see Chapter 3 of *High Availability and Disaster Recovery Options for DB2 for Linux, UNIX, and Windows*, SG24-7363, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247363.html?Open

This book also includes information about the extra steps for maintaining data integrity in disaster recovery.

For more information about Smart Assist Configuration with DB2, see *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg248030.html?Open

4.14.5 PowerHA SystemMirror configuration with WebSphere Application Server

PowerHA SystemMirror Smart Assist for WebSphere extends an existing PowerHA SystemMirror configuration to include monitoring and recovery support for the following components in a base configuration of WebSphere:

- WebSphere Application Server
- WebSphere Deployment Manager
- WebSphere Transaction Log
- IBM HTTP Server
- Tivoli Directory Server

PowerHA SystemMirror Smart Assist for WebSphere sets up your PowerHA SystemMirror resource configuration for you. It gathers information from you, the application configuration, and the system configuration, and creates a PowerHA SystemMirror resource group to protect a WebSphere application.

A PowerHA SystemMirror resource group contains a set of resources, such as cluster nodes, volume groups, and IP labels (which are used by server applications to communicate with clients) that PowerHA SystemMirror manages as a unit.

Be aware that the term *application server* can have different meanings. A WebSphere Application Server makes web-based applications available to users. A PowerHA SystemMirror application server identifies the application start and stop scripts that are used with PowerHA SystemMirror.

In addition to creating a resource group for each application to be made highly available, PowerHA SystemMirror Smart Assist for WebSphere performs the following tasks:

- Creates a PowerHA SystemMirror application server for the application.
- Ensures the application has a service IP label that can be transferred to another system, thus keeping the application highly available.
- Creates PowerHA SystemMirror application monitors to detect failures in application processes.
- Provides start and stop scripts for the application (that is, application servers).
- Stores the generated PowerHA SystemMirror configuration in the PowerHA SystemMirror configuration database (ODM).

When PowerHA SystemMirror Smart Assist for WebSphere creates the PowerHA SystemMirror resource configuration, it updates the PowerHA SystemMirror configuration and changes the IP address (for IHS, WebSphere Application Server, and Deployment Manager), or the Transaction Log directory (for the WebSphere cluster).

We used WebSphere Application Server, deployment manager, and IHS configuration. An example WebSphere Application Server configuration that is used for our test environment is shown in the next section.

For more information about configuring PowerHA SystemMirror with WebSphere Application Server, see this website:

http://www.ibm.com/developerworks/aix/library/au-aixwas/index.html

4.14.6 PowerHA SystemMirror configuration examples

The following example cluster configuration, as used in this book, is shown in terms of PowerHA topologies and resource groups.

This is a DB2 cluster with mutual takeover configuration. Each node has one DB2 instance, which is up and running. In case of failover, one of the instances moves to the other node at the other site of the SAN Volume Controller Stretched Cluster.

Example 4-7 shows the example PowerHA cluster configuration that we used for testing. It shows the IP addresses that were used for cluster configuration. It also shows the application-specific configurations that are deployed in this cluster. In the output, each field is clearly explained with name and types.

Example 4-7 PowerHA configuration example

```
# ./cltopinfo
Cluster Name: itsosjc_clstr1
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
```

```
Use Persistent Labels for Communication: No
Repository Disk: hdisk2
Cluster IP address: 228.17.90.22
There are 2 node(s) and 2 network(s) defined
NODE itsosjc1lpar4:
        Network net_ether_01
                itsosjc1lpar4   10.17.90.22
        Network net_ether_02
                clstr1_sip2     192.168.150.3
                clstr1_sip1     192.168.150.2
                node1_if1       2.2.2.1
                node1_if2       2.2.3.1
NODE itsosjc2lpar4:
        Network net_ether_01
                itsosjc2lpar4   10.17.90.32
        Network net_ether_02
                clstr1_sip2     192.168.150.3
                clstr1_sip1     192.168.150.2
                node2_if2       2.2.3.2
                node2_if1       2.2.2.2

Resource Group db2inst2_ResourceGroup
        Startup Policy   Online On Home Node Only
        Fallover Policy  Fallover To Next Priority Node In The List
        Fallback Policy  Fallback To Higher Priority Node In The List
        Participating Nodes      itsosjc2lpar4 itsosjc1lpar4
        Service IP Label                 clstr1_sip2

Resource Group db2inst1_ResourceGroup
        Startup Policy   Online On Home Node Only
        Fallover Policy  Fallover To Next Priority Node In The List
        Fallback Policy  Fallback To Higher Priority Node In The List
        Participating Nodes      itsosjc1lpar4 itsosjc2lpar4
        Service IP Label                 clstr1_sip1

# ./clshowres

Resource Group Name                                 db2inst2_ResourceGroup
Participating Node Name(s)                          itsosjc2lpar4 itsosjc1lpar4
Startup Policy                                      Online On Home Node Only
Fallover Policy                                    Fallover To Next Priority Node
In The List
Fallback Policy                                     Fallback To Higher Priority
Node In The List
Site Relationship                                   ignore
Node Priority
Service IP Label                                    clstr1_sip2
Filesystems                                         ALL
Filesystems Consistency Check                       fsck
Filesystems Recovery Method                         parallel
Filesystems/Directories to be exported (NFSv3)
Filesystems/Directories to be exported (NFSv4)
Filesystems to be NFS mounted
Network For NFS Mount
Filesystem/Directory for NFSv4 Stable Storage
```

```
Volume Groups                                        db2vg1
Concurrent Volume Groups
Use forced varyon for volume groups, if necessary    false
Disks
GMVG Replicated Resources
GMD Replicated Resources
PPRC Replicated Resources
SAN Volume Controller PPRC Replicated Resources
EMC SRDF Replicated Resources
Hitachi TrueCopy Replicated Resources
Generic XD Replicated Resources
AIX Connections Services
AIX Fast Connect Services
Shared Tape Resources
Application Servers                                  db2inst2_ApplicationServer
Highly Available Communication Links
Primary Workload Manager Class
Secondary Workload Manager Class
Delayed Fallback Timer
Miscellaneous Data
Automatically Import Volume Groups                   false
Inactive Takeover
SSA Disk Fencing                                     false
Filesystems mounted before IP configured             false
WPAR Name


Run Time Parameters:

Node Name                                            itsosjc2lpar4
Debug Level                                          high
Format for hacmp.out                                 Standard

Node Name                                            itsosjc1lpar4
Debug Level                                          high
Format for hacmp.out                                 Standard


Resource Group Name                                  db2inst1_ResourceGroup
Participating Node Name(s)                           itsosjc1lpar4 itsosjc2lpar4
Startup Policy                                       Online On Home Node Only
Fallover Policy                                     Fallover To Next Priority Node
In The List
Fallback Policy                                      Fallback To Higher Priority
Node In The List
Site Relationship                                    ignore
Node Priority
Service IP Label                                     clstr1_sip1
Filesystems                                          ALL
Filesystems Consistency Check                        fsck
Filesystems Recovery Method                          parallel
Filesystems/Directories to be exported (NFSv3)
Filesystems/Directories to be exported (NFSv4)
Filesystems to be NFS mounted
Network For NFS Mount
```

```
Filesystem/Directory for NFSv4 Stable Storage
Volume Groups                                       db2vg
Concurrent Volume Groups
Use forced varyon for volume groups, if necessary   false
Disks
GMVG Replicated Resources
GMD Replicated Resources
PPRC Replicated Resources
SAN Volume Controller PPRC Replicated Resources
EMC SRDF Replicated Resources
Hitachi TrueCopy Replicated Resources
Generic XD Replicated Resources
AIX Connections Services
AIX Fast Connect Services
Shared Tape Resources
Application Servers                                 db2inst1_ApplicationServer
Highly Available Communication Links
Primary Workload Manager Class
Secondary Workload Manager Class
Delayed Fallback Timer
Miscellaneous Data
Automatically Import Volume Groups                  false
Inactive Takeover
SSA Disk Fencing                                    false
Filesystems mounted before IP configured            false
WPAR Name


Run Time Parameters:

Node Name                                           itsosjc1lpar4
Debug Level                                         high
Format for hacmp.out                                Standard

Node Name                                           itsosjc2lpar4
Debug Level                                         high
Format for hacmp.out                                Standard


# ./clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
db2inst2_Resou ONLINE                       itsosjc2lpar4
               OFFLINE                      itsosjc1lpar4

db2inst1_Resou ONLINE                       itsosjc1lpar4
               OFFLINE                      itsosjc2lpar4
```

Example 4-8 shows a WebSphere Application Server cluster. This is a four-node cluster in which two nodes are configured for WebSphere Application Server Network Deployment manager and IHS, with each node holding one application. The other two nodes are used for WebSphere Application Server.

Example 4-8 WebSphere Application Server cluster

```
# ./cltopinfo
Cluster Name: itsosjc1lpar2_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: hdisk5
Cluster IP address: 228.17.90.20
There are 4 node(s) and 2 network(s) defined
NODE itsosjc1lpar2:
        Network net_ether_01
                itsosjc1lpar2   10.17.90.20
        Network net_ether_02
                sip3    192.168.150.14
                sip2    192.168.150.13
                sip1    192.168.150.12
                node1_int1      12.2.2.1
                node1_int2      12.2.3.1
NODE itsosjc1lpar3:
        Network net_ether_01
                itsosjc1lpar3   10.17.90.21
        Network net_ether_02
                sip3    192.168.150.14
                sip2    192.168.150.13
                sip1    192.168.150.12
                node2_int1      12.2.2.2
                node2_int2      12.2.3.2
NODE itsosjc2lpar2:
        Network net_ether_01
                itsosjc2lpar2   10.17.90.30
        Network net_ether_02
                sip3    192.168.150.14
                sip2    192.168.150.13
                sip1    192.168.150.12
                node3_int2      12.2.3.3
                node3_int1      12.2.2.3
NODE itsosjc2lpar3:
        Network net_ether_01
                itsosjc2lpar3   10.17.90.31
        Network net_ether_02
                sip3    192.168.150.14
                sip2    192.168.150.13
                sip1    192.168.150.12
                node4_int2      12.2.3.4
                node4_int1      12.2.2.4

Resource Group WAS_DM_Cell01_AS_wasadmin_rg
        Startup Policy   Online On Home Node Only
        Fallover Policy  Fallover To Next Priority Node In The List
```

```
        Fallback Policy  Never Fallback
        Participating Nodes      itsosjc1lpar2 itsosjc2lpar2
        Service IP Label                 sip1

Resource Group IHS_itsosjc1lpar2_rg
        Startup Policy   Online On Home Node Only
        Fallover Policy  Fallover To Next Priority Node In The List
        Fallback Policy  Never Fallback
        Participating Nodes      itsosjc1lpar2 itsosjc2lpar2
        Service IP Label                 sip2

Resource Group WAS_NODE_itsosjc1lpar2Node01_rg
        Startup Policy   Online On Home Node Only
        Fallover Policy  Fallover To Next Priority Node In The List
        Fallback Policy  Never Fallback
        Participating Nodes      itsosjc1lpar3 itsosjc2lpar3
        Service IP Label                 sip3


# ./clshowres

Resource Group Name                                 WAS_DM_Cell01_AS_wasadmin_rg
Participating Node Name(s)                          itsosjc1lpar2 itsosjc2lpar2
Startup Policy                                      Online On Home Node Only
Fallover Policy                                    Fallover To Next Priority Node
In The List
Fallback Policy                                     Never Fallback
Site Relationship                                   ignore
Node Priority
Service IP Label                                    sip1
Filesystems                                         ALL
Filesystems Consistency Check                       fsck
Filesystems Recovery Method                         parallel
Filesystems/Directories to be exported (NFSv3)
Filesystems/Directories to be exported (NFSv4)
Filesystems to be NFS mounted
Network For NFS Mount
Filesystem/Directory for NFSv4 Stable Storage
Volume Groups                                       wasndvg
Concurrent Volume Groups
Use forced varyon for volume groups, if necessary   false
Disks
GMVG Replicated Resources
GMD Replicated Resources
PPRC Replicated Resources
SAN Volume Controller PPRC Replicated Resources
EMC SRDF Replicated Resources
Hitachi TrueCopy Replicated Resources
Generic XD Replicated Resources
AIX Connections Services
AIX Fast Connect Services
Shared Tape Resources
Application Servers                                 WAS_DM_Cell01_AS_wasadmin_as
Highly Available Communication Links
Primary Workload Manager Class
```

```
Secondary Workload Manager Class
Delayed Fallback Timer
Miscellaneous Data
Automatically Import Volume Groups                  false
Inactive Takeover
SSA Disk Fencing                                    false
Filesystems mounted before IP configured            false
WPAR Name


Run Time Parameters:

Node Name                                           itsosjc1lpar2
Debug Level                                         high
Format for hacmp.out                                Standard

Node Name                                           itsosjc2lpar2
Debug Level                                         high
Format for hacmp.out                                Standard


Resource Group Name                                 IHS_itsosjc1lpar2_rg
Participating Node Name(s)                          itsosjc1lpar2 itsosjc2lpar2
Startup Policy                                      Online On Home Node Only
Fallover Policy                                    Fallover To Next Priority Node
In The List
Fallback Policy                                     Never Fallback
Site Relationship                                   ignore
Node Priority
Service IP Label                                    sip2
Filesystems                                         ALL
Filesystems Consistency Check                       fsck
Filesystems Recovery Method                         parallel
Filesystems/Directories to be exported (NFSv3)
Filesystems/Directories to be exported (NFSv4)
Filesystems to be NFS mounted
Network For NFS Mount
Filesystem/Directory for NFSv4 Stable Storage
Volume Groups                                       ihsvg
Concurrent Volume Groups
Use forced varyon for volume groups, if necessary   false
Disks
GMVG Replicated Resources
GMD Replicated Resources
PPRC Replicated Resources
SAN Volume Controller PPRC Replicated Resources
EMC SRDF Replicated Resources
Hitachi TrueCopy Replicated Resources
Generic XD Replicated Resources
AIX Connections Services
AIX Fast Connect Services
Shared Tape Resources
Application Servers                                 IHS_itsosjc1lpar2_as
Highly Available Communication Links
Primary Workload Manager Class
```

```
Secondary Workload Manager Class
Delayed Fallback Timer
Miscellaneous Data
Automatically Import Volume Groups                  false
Inactive Takeover
SSA Disk Fencing                                    false
Filesystems mounted before IP configured            false
WPAR Name


Run Time Parameters:

Node Name                                           itsosjc1lpar2
Debug Level                                         high
Format for hacmp.out                                Standard

Node Name                                           itsosjc2lpar2
Debug Level                                         high
Format for hacmp.out                                Standard


Resource Group Name
WAS_NODE_itsosjc1lpar2Node01_rg
Participating Node Name(s)                          itsosjc1lpar3 itsosjc2lpar3
Startup Policy                                      Online On Home Node Only
Fallover Policy                                    Fallover To Next Priority Node
In The List
Fallback Policy                                     Never Fallback
Site Relationship                                   ignore
Node Priority
Service IP Label                                    sip3
Filesystems                                         ALL
Filesystems Consistency Check                       fsck
Filesystems Recovery Method                         parallel
Filesystems/Directories to be exported (NFSv3)
Filesystems/Directories to be exported (NFSv4)
Filesystems to be NFS mounted
Network For NFS Mount
Filesystem/Directory for NFSv4 Stable Storage
Volume Groups                                       wasasvg
Concurrent Volume Groups
Use forced varyon for volume groups, if necessary   false
Disks
GMVG Replicated Resources
GMD Replicated Resources
PPRC Replicated Resources
SAN Volume Controller PPRC Replicated Resources
EMC SRDF Replicated Resources
Hitachi TrueCopy Replicated Resources
Generic XD Replicated Resources
AIX Connections Services
AIX Fast Connect Services
Shared Tape Resources
Application Servers
WAS_NODE_itsosjc1lpar2Node01_as
```

```
Highly Available Communication Links
Primary Workload Manager Class
Secondary Workload Manager Class
Delayed Fallback Timer
Miscellaneous Data
Automatically Import Volume Groups                  false
Inactive Takeover
SSA Disk Fencing                                    false
Filesystems mounted before IP configured            false
WPAR Name


Run Time Parameters:

Node Name                                           itsosjc1lpar3
Debug Level                                         high
Format for hacmp.out                                Standard

Node Name                                           itsosjc2lpar3
Debug Level                                         high
Format for hacmp.out                                Standard


Below example shows that all the instance is online in different node.

# ./clRGinfo


-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
WAS_DM_Cell01_ OFFLINE                      itsosjc1lpar2
               ONLINE                       itsosjc2lpar2

IHS_itsosjc1lp ONLINE                       itsosjc1lpar2
               OFFLINE                      itsosjc2lpar2

WAS_NODE_itsos ONLINE                       itsosjc1lpar3
               OFFLINE                      itsosjc2lpar3
```

4.15 Advanced configuration

For more information about an advanced configuration, we recommend the following publications:

- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *IBM System p Advanced POWER Virtualization (PowerVM) Best Practices*, REDP-4194

4.15.1 VIOS partition and system redundancy

For this dual VIO Server configuration (as shown in Figure 4-38 on page 285), we use SAN-attached disks that are presented by both VIO Servers to provide redundant paths to the storage for the client partitions. In this simplistic example, we use virtual adapters for the network and storage so that multiple partitions can be created quickly and can share the bandwidth of the Ethernet and Fibre Channel adapters on the dual VIO Servers. By using virtual adapters for network and disks, the only other resources that are required for each client partition are CPU and memory. The other design choices for LVM mirroring of boot disks and SEA Failover are mentioned briefly in other sections of this book.

This sample configuration includes the following features:

- Sharing the SAN connections by multiple client partitions.
- The SAN storage is provided by the SAN Volume Controller to the dual VIO Servers.
- All SAN storage for the client partitions is zoned to both VIO Servers.
- The client partitions will boot from the SAN via the VIO Servers.

In Figure 4-38 on page 285, redundancy for the client storage and network is provided so that if one VIOS goes down, the other can still serve the client IO requests. So virtual storage IO is transferred to the path from the other VIO server if the first VIO server is down. For virtual Ethernet, it is in sharing mode and some VLAN traffic is going from VIOS1 and some are going from VIOS2. In the case of a failover, all VLAN traffic moves to VIOS2.

Figure 4-38 Dual VIOS server example

4.15.2 Virtual storage redundancy

Redundancy configurations help protect your network from physical adapter failures and Virtual I/O Server failures.

With NPIV, you can configure the managed system so that multiple logical partitions can access independent physical storage through the same physical Fibre Channel adapter. Each virtual Fibre Channel adapter is identified by a unique WWPN, which means that you can connect each virtual Fibre Channel adapter to independent physical storage on a SAN.

Similar to virtual SCSI redundancy, virtual Fibre Channel redundancy can be achieved by using MPIO and mirroring at the client partition. The difference between traditional redundancy with SCSI adapters and the NPIV technology that uses virtual Fibre Channel adapters, is that the redundancy occurs on the client because only the client recognizes the disk. The Virtual I/O Server is essentially only a pipe. The second example we show in this section uses multiple Virtual I/O Server logical partitions to add redundancy at the Virtual I/O Server level as well. This example uses Host bus adapter (HBA) failover to provide a basic level of redundancy for the client logical partition.

The SAN connects physical storage to two physical Fibre Channel adapters that are on the managed system. The physical Fibre Channel adapters are assigned to the Virtual I/O Server and support NPIV. The physical Fibre Channel ports are each connected to a virtual Fibre Channel adapter on the Virtual I/O Server. The two virtual Fibre Channel adapters on the Virtual I/O Server are connected to ports on two different physical Fibre Channel adapters to provide redundancy for the physical adapters.

Each virtual Fibre Channel adapter on the Virtual I/O Server is connected to one virtual Fibre Channel adapter on a client logical partition. Each virtual Fibre Channel adapter on each client logical partition receives a pair of unique WWPNs. The client logical partition uses one WWPN to log in to the SAN at any time. The other WWPN is used when you move the client logical partition to another managed system.

The SAN connects physical storage to two physical Fibre Channel adapters that are on the managed system. There are two Virtual I/O Server logical partitions to provide redundancy at the Virtual I/O Server level. The physical Fibre Channel adapters are assigned to their respective Virtual I/O Server and support NPIV.

The physical Fibre Channel ports are each connected to a virtual Fibre Channel adapter on the Virtual I/O Server. The two virtual Fibre Channel adapters on the Virtual I/O Server are connected to ports on two different physical Fibre Channel adapters to provide redundancy for the physical adapters. A single adapter can feature multiple ports.

Each virtual Fibre Channel adapter on the Virtual I/O Server is connected to one virtual Fibre Channel adapter on a client logical partition. Each virtual Fibre Channel adapter on each client logical partition receives a pair of unique WWPNs. The client logical partition uses one WWPN to log in to the SAN at any given time. The other WWPN is used when you move the client logical partition to another managed system.

Further considerations

These examples can become more complex as you add physical storage redundancy and multiple clients, but the concepts remain the same. Consider the following points:

- To avoid configuring the physical Fibre Channel adapter to be a single point of failure for the connection between the client logical partition and its physical storage on the SAN, do not connect two virtual Fibre Channel adapters from the same client logical partition to the same physical Fibre Channel adapter. Instead, connect each virtual Fibre Channel adapter to a different physical Fibre Channel adapter.

  Consider load balancing when mapping a virtual Fibre Channel adapter on the Virtual I/O Server to a physical port on the physical Fibre Channel adapter.

  Consider what level of redundancy already exists in the SAN to determine whether to configure multiple physical storage units.

  Consider the use of two Virtual I/O Server logical partitions. Because the Virtual I/O Server is central to communication between logical partitions and the external network, it is important to provide a level of redundancy for the Virtual I/O Server. Multiple Virtual I/O Server logical partitions require more resources as well, so you should plan accordingly.

- NPIV technology is useful when you want to move logical partitions between servers. For example, in active Partition Mobility, if you use the redundancy configurations we described in this section in combination with physical adapters, you can stop all the I/O activity through the dedicated, physical adapter and direct all traffic through a virtual Fibre Channel adapter until the logical partition is successfully moved. The dedicated physical adapter must be connected to the same storage as the virtual path. Because you cannot migrate a physical adapter, all I/O activity is routed through the virtual path while you move the partition. After the logical partition is moved successfully, you must set up the dedicated path (on the destination logical partition) if you want to use the same redundancy configuration you configured on the original logical partition. Then, the I/O activity can resume through the dedicated adapter by using the virtual Fibre Channel adapter as a secondary path.

4.15.3 Multipathing in Virtual I/O Server

For dual VIO Servers attached to IBM SAN storage (SAN Volume Controller, DS8000, DS6000 and ESS), SDDPCM multipathing software for AIX 6L or 7L is installed on the VIO Servers to provide the load-balancing and redundancy across multiple Fibre Channel connections between the VIO Server (or servers) and the IBM SAN Storage. The SDDPCM software can be downloaded from the IBM Storage website. It has a prerequisite file set (`devices.fcp.disk.ibm.mpio.rte`), which is referred to as the Host Attachment Script for SDDPCM on AIX 6L or 7L and is available from the web. At the time of this writing, SDDPCM V2.6.2.2 was the latest version and the required level for dual VIO Servers with the SAN Volume Controller.

For more information, see *Configuring SDDPCM for High Availability*, REDP-4928, which is available at this website:

http://www.redbooks.ibm.com/abstracts/redp4928.html?Open

4.15.4 Multipathing in client partition

For multipathing support of virtual SCSI devices in the AIX 5L client partition, the SAN LUN must be presented as a physical drive (hdiskx) from the VIO Server to the client partition. It is not possible to provide a large SAN LUN and then further sub-divide it into logical volumes at the VIO server level when dual VIO servers are used. The storage management for this configuration is performed in the SAN so there is a one-to-one mapping of SAN LUNs on the VIO Servers to virtual SCSI drives on the client partition.

The MPIO support over virtual SCSI between the client partition and the dual VIO Servers only supports failover mode. The client partition uses a primary path to one VIO Server and failover to the secondary path to use the other VIO Server. Only one path is used at a time even though both paths can be enabled.

To balance the load of multiple client partitions across dual VIO Servers, the priority can be set to select the primary path and a specific VIO Server for each client partition. The priority is set on a per virtual SCSI disk drive basis so all the disks or alternative disks on a client partition can use one of the VIO Servers. The recommended method is to divide the client partitions between the two VIO Servers.

4.15.5 LVM mirroring considerations

With the use of LVM mirroring in the client partition, each Virtual I/O Server can present a virtual SCSI device that is physically connected to a different disk and then used in a normal AIX mirrored volume group on the client partition. This configuration achieves a higher level of reliability by providing redundancy. Client volume group mirroring also is required when a Virtual I/O Server logical volume is used as a virtual SCSI device on the client. In this case, the virtual SCSI devices are associated with different SCSI disks, each controlled by one of the two Virtual I/O Servers.

Figure 4-39 on page 288 shows that LVM mirroring that can be used in dual VIOS setup (vSCSI and VFC can be used same way).

Figure 4-39 Using dual VIOS for LVM client Mirroring

4.15.6 Virtual Ethernet redundancy

The SEA can use multiple Ethernet ports as an aggregated link to the physical network. The extra network adapters provide more bandwidth over a single network connection and provide some redundancy if there is a port or cable issue with one of the connections. All ports must be connected to the same switch and the switch ports used for aggregation must be configured for Etherchannel. If link aggregation is used, each VIO server is connected to a different switch for redundancy.

A SEA configuration example is shown in Figure 4-40 on page 289. In this example, it can be seen that a different VLAN is used to create a dedicated bandwidth for LPAR communication and cluster configuration communication. There is a private and public VLAN that is configured to isolate the traffic. VLAN 199 and 299 is used for heartbeat traffic, which detects a network failure in one of the VIOS and moves the traffic to the other.

Public VLAN 701is there for HMC communications and Private VLAN 150/160/170 is for cluster communications. For more information, see 4.4.1, “Creating a Virtual I/O Server” on page 220.

Figure 4-40 SEA configuration

4.15.7 IBM PowerHA SystemMirror for AIX virtual I/O client

We expect that proper LPAR and DLPAR planning is part of your overall process before any similar configuration is implemented. It is important to understand not only the requirements and how to implement them, but also the overall effects that each decision has on the overall implementation.

Other considerations

When planning a cluster to include DLPAR operations, the following considerations apply:

- Encountering possible config_too_long during DLPAR events
- Mix of LPARs and non-LPAR systems
- Mix of POWER6 and POWER67
- CoD provisioning

While HMC versions might not support all POWER platforms, in general, PowerHA does, meaning that a POWER7 production system might failover to a POWER6 system.

As with any cluster, the configuration must be tested thoroughly. This includes anything that can be done to simulate or produce a real work load for the most realistic test scenarios as possible.

For more information about configuring DLPAR with PowerHA for load balancing, see 9.3.3 of *PowerHA for AIX Cookbook*, SG24-7739, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247739.html?Open

You also can see this book for other Disaster Recovery for PowerHA XD considerations.

PowerHA SystemMirror best practices

PowerHA must be the central point of control for NFS mounts and exports that are related to PowerHA managed volume groups. Consider the following best practices:

- Analyze NFS mounts and exports that preclude takeover (no hard mount outside of PowerHA).
- Analyze volume group setup including quorum.
- Use Cluster Single Point of Control to let PowerHA manage and be aware of cluster-wide changes specifically LVM changes.
- Ensure that PowerHA takeover model (for example, cascading versus rotational) supports availability requirements. Cascading requires failback, which doubles downtime requirements. Active/Passive failover in a rotational failover model minimizes downtime for planned and unplanned outages.
- Heartbeats should not be dependent on IP or any other contingencies that do not allow for adequate notification of inability to failover (for example, IP heartbeat).
- Cluster verification methods should be comprehensive and complete. The use of the **clverify** command is a start for validating clusters, but environment-specific extensions must be used for situation-specific issues that are tailored to your environment.
- Change control procedures and change management must be cluster-aware and nondisruptive to cluster failover ability. Special attention to keeping all nodes of a cluster in sync is warranted.
- PowerHA failover and failback testing must be conducted whenever change occurs on the cluster. If this is not done at the time of change, test clusters are a second choice or alternative. Test environments should have PowerHA.
- PowerHA (with firmware and software) must have reasonable currency.

Application awareness

Be aware of the following application considerations:

- Do not hardcode IP addresses on the application side.
- Avoid application dependencies (such as hardcoded host name) that prevent successful changeover.
- Establish a procedure to verify that the application and all related interfaces failed over successfully.
- Applications must be stopped and started via script (and lpar startup and shutdown).
- Use enhanced concurrent volume groups where possible for quicker failover and improved cluster aware LVM changes.

- The monitoring environment that supports the PowerHA supported infrastructure must be robust and allow for an early-warning system of impending errors and must capture cluster state, disruptive events, and application state
- Use SmartAssists where applicable.

For more information about best practices, see the following publications:

- *Virtualization and Clustering Best Practices Using IBM System p Servers*, SG24-7349:

  http://www.redbooks.ibm.com/abstracts/sg247349.html?Open
- *High Availability and Disaster Recovery Planning: Next-Generation Solutions for Multiserver IBM Power Systems Environments*, REDP-4669:

  http://www.redbooks.ibm.com/abstracts/redp4669.html?Open
- *Exploiting IBM PowerHA SystemMirror Enterprise Edition*, SG24-7841:

  http://www.redbooks.ibm.com/abstracts/sg247841.html?Open
- *IBM PowerVM Best Practices*, SG24-8062:

  http://www.redbooks.ibm.com/abstracts/sg248062.html?Open

5

Stretched Cluster Diagnostic and Recovery Guidelines

In this chapter we focus on the cluster diagnostics and recovery guidelines to understand what might be happening in your Stretched Cluster environment after a critical event, and to put you in the best possible position to take the correct decisions to alleviate the situation. Such corrective action might include waiting until the failure in one of the two sites is fixed or declaring a disaster and starting the recovery action that is described in this chapter.

Also, we describe the tested methodologies and simulations of real live workload scenarios here to show the solution implementation simulation in real production environments.

This chapter includes the following topics:

- SAN Volume Controller Stretched Cluster solution testing
- Solution recovery planning
- SAN Volume Controller recovery planning
- PowerVM recovery planning
- PowerHA SystemMirror recovery planning
- SAN Volume Controller Diagnosis and recovery guidelines
- PowerVM diagnosis and recovery guidelines
- PowerHA SystemMirror diagnosis and recovery guidelines

© Copyright IBM Corp. 2013. All rights reserved.

5.1 SAN Volume Controller Stretched Cluster solution testing

To achieve the best results for the SAN Volume Controller Stretched Cluster solution and formalize the implementation or fit this in a possible customer-like scenario, proper testing examples and results are described in this section. As described in 1.11.1, “Overview and concepts” on page 82, various testing scenarios for each of the solution architecture components, including SAN Volume Controller, Stretched Cluster, SAN Directors, PowerHA, and Live Partition Mobility (LPM) are included to simulate distances as follows between Site A (Production) and Site B (DR). We used a Network Latency simulator at the Brocade Labs in the testing to simulate different distances. The following testing categories or distance frames were included:

- Less than 5 KM (3.11 miles), as an example of Local HA or partitioned single data center location
- 5 KM (3.11 miles), as an example of a “Campus” data center architecture
- 15 KM (9.32 miles), as an example of a “Metro-Mirror” data center solution and architecture
- 75 KM (46.6 miles), as an example of a “Metro-Mirror” data center solution and architecture
- 150 KM (93.2 miles), as an example of a “Global Mirror/Out of Region DR” data center solution and architecture
- 300 KM (186.41 miles), as an example of a “Global Mirror/Out of Region DR” data center solution and architecture

The SAN Volume Controller component testing included a mix of Read Only, Write Only, and Read and Write I/O activity against the different distance categories. We also tested outage scenarios for the Site A single SAN Director loss, Site A two SAN Director loss, Site A Server and LPAR loss, Site A and Storwize V7000 loss with a mix of Read and Write I/O Activity against the six distance categories.

Finally, we tested LPM against with application DB2 sizes of 45 GB, to simulate the migration across a single data center (for example, Local HA). More testing is planned to simulate Jumbo Frames across the five other distances with DB2 sizes of 200 GB, 500 GB, and 1 TB.

5.1.1 Test environment

The test environment consists of two PowerHA SystemMirror cluster, out of which is a two node active-active cluster in which each DB2 instance is active on each node. In case of failures, either of the DB2 instance move to other node and vice versa.

The other cluster is a four node cluster in which two nodes are deployed to make WebSphere Application Server network development (ND) and IBM HTTP Server highly available for single point of failures. The other two nodes are holding WebSphere Application Server stand-alone application server.

To simulate the workload on these applications, we used IBM benchmark tools DayTrader and Trade application workload. Also, during testing of failure scenarios, we performed concurrent read and writes on raw disks by using the AIX **dd** command, as shown in the following examples:

```
dd if=/dev/urandom of=/dev/hdisk0 bs=1234567 count=100
dd if=/dev/hdisk0 of=/dev/null bs=1234567 count=100
```

We also show the different input/output parameters when failure occurs to give a clear idea of how much I/O activity was going on at the time of failures.

We also show the result for the effect of latency (directly proportional to some constant multiplied by distance in kilometers) in this solution by using a latency simulator device. For the latency and distance simulator, we assumed .01 ms of latency = 1 KM as a standard industry definition. We also used .01 ms as the base input for conversion to KM for distance.

The effect on LPM because of latency in network also is shown here.

The latency simulator was put between the FCIP routers, which are responsible to send or receive all the traffic between SAN Volume Controller cluster node in SiteA and SiteB. This is basically a Transmission Control Protocol (TCP) simulator. It works for Fibre Channel (FC) and for testing this Redbooks solution because we use FCIP routers across sites that tunnel FC packets in TCP frames. We also put our simulator in between two Fibre Channel over IP (FCIP) routers across the sites.

The simulator device is made by Shunra. For more information, see these websites:

- http://media.shunra.com/datasheets/Shunra-NetworkCatcher.pdf
- http://www.shunra.com/products/shunra-networkcatcher

Figure 5-1 shows the SAN Volume Controller architecture of the solution.

Figure 5-1 SAN Volume Controller and zoning configuration

Figure 5-2 on page 296 shows the POWER7 processor-based servers that are configured across the SAN Volume Controller Stretched Cluster sites. Server #1 is configured in siteA and server #2 is configured for siteB. Each client LPAR is accessing logical unit numbers (LUNs) from SAN Volume Controller through two different switches configured in private SAN to maintain redundancy. Each private SAN of each site is configured for public SAN through FCIP router configuration to access the same LUNs through another site in case of failures. The same Ethernet network with same VLAN is configured across the sites to maintain the same IP access.

Figure 5-2 Power Server configuration

Figure 5-3 on page 297 shows the PowerHA SystemMirror configuration on POWER7 processor-based servers across the sites. The primary node (or nodes) is picked from SiteA and the failover node (or nodes) is picked from SiteB. Disks are shared across the cluster nodes. The DB2 instances and WebSphere Application Server and IBM HTTP servers are made highly available by using PowerHA SystemMirror across the sites. The cluster is standard because the SAN Volume Controller Stretched Cluster solution provides the ability to share the same disks across different site nodes.

Figure 5-3 PowerHA SystemMirror cluster configuration

Figure 5-4 on page 298 shows the workload we are running on the WebSphere Application Server and DB2 instance. DayTrader is an IBM benchmark tool that simulates the trading workload. We also run AIX **dd** command to perform concurrent read and write I/O to simulate some applications that perform these kinds of I/O activity.

Figure 5-4 The workload configuration

Figure 5-5 on page 299 shows the LPM configuration in which LPAR1 from server #1 moves completely to server #2 and accesses the same LUNs and same network in the destination server. In this same way, the client LPAR can move back to server #1.

Figure 5-5 Live Partition Mobility

Figure 5-6 shows the latency simulator in which it shows that it was placed in the communication link between site A and site B.

Figure 5-6 Latency simulator configuration

Figure 5-7 shows the input panel for the latency simulator.

Figure 5-7 Shunra Latency Simulator input panel

Important: The latency field in msec is one-way latency. In terms of distance, the end-to-end distance is multiplied by 2 to compute round-trip distance. For TCP/FCIP/FC protocol, the latency is round-trip calculation. Therefore, if 1 ms is entered, that is 50 km one way, 100 km round trip.

5.1.2 Test scenarios and results

Figure 5-8 on page 301 shows a graphical comparison between the first seven test scenarios. Each scenario is explained in Table 5-1 on page 301. Any further LPM testing across the simulated distance categories with 200 GB, 500 GB, and 1 TB to simulate Jumbo Frames might occur as an update to this book.

Figure 5-8 SAN Volume Controller Stretched Cluster and PowerVM Test Results

Table 5-1 provides a summary of the test scenarios and the results.

Table 5-1 Test scenarios and results

Test Case and Parameter with Conditions	Results
1. Component Loss - SAN Volume Controller - Site A with “failover” to Site B SAN Volume Controller Read I/O Activity (23,000 IOPS) from SAN Volume Controller and 20 Mbps Read on Host	Failover to Site B: 0 KM = 50 sec. 5 KM = 54 sec. 15 KM = 57 sec. 75 KM = 58 sec. 150 KM = 60 sec. 300 KM = 61 sec.
2. Component Loss - SAN Volume Controller - Site A with “failover” to Site B SAN Volume Controller Write I/O Activity (14,000 IOPS) from SAN Volume Controller and 8 Mbps Write on Host	Failover to Site B: 0 KM = 69 sec. 5 KM = 70 sec. 15 KM = 71 sec. 75 KM = 75 sec. 150 KM = 79 sec. 300 KM = 80 sec.
3. Component Loss - SAN Volume Controller - Site A with “failover” to Site B SAN Volume Controller Read I/O Activity (14,000 IOPS), Write I/O Activity (4,000 IOPS) from SAN Volume Controller and 12 Mbps Read and Write on Host	Failover to Site B: 0 KM = 54 sec. 5 KM = 56 sec. 15 KM = 57 sec. 75 KM = 59 sec. 150 KM = 61 sec. 300 KM = 61 sec.
4. Component Loss - Single SAN Director/Switch - Site A Read I/O Activity (30,000 IOPS), Write I/O Activity (1300 IOPS) from SAN Volume Controller and 12 Mbps Read and Write on Host	Failover to Second Site A Director/Switch: 0 KM = 36 sec. 5 KM = 39sec. 15 KM = 40 sec. 75 KM = 41 sec. 150 KM = 43 sec. 300 KM = 45 sec.

5. Component Loss - Both SAN Directors - Site A failover to Site B Read I/O Activity (30,000 IOPS), Write I/O Activity (1300 IOPS) from SAN Volume Controller and 12 Mbps Read and Write on Host	Failover to Site B: 0 KM = 115 sec. 5 KM = 119 sec. 15 KM = 121 sec. 75 KM = 122 sec. 150 KM = 125 sec. 300 KM = 127 sec.
6. Component Loss - Storage (Storwize V7000) Site A failover to Site B Storage Read I/O Activity (11,500 IOPS), Write I/O Activity (1200 IOPS) from SAN Volume Controller and 12 Mbps Read and Write on Host	Failover to Site B: 0 KM = 67 sec. 5 KM = 78 sec. 15 KM = 79 sec. 75 KM = 80 sec. 150 KM = 85 sec. 300 KM = 86 sec.
7. Component Loss - Server (p750) -Site A - failover to Site B Read I/O Activity (11,500 IOPS), Write I/O Activity (1200 IOPS) from SAN Volume Controller and 12 Mbps Read and Write on Host	Failover to Site B: 0 KM = 130 sec. 5 KM = 132 sec. 15 KM = 133 sec. 75 KM = 139 sec. 150 KM = 141 sec. 300 KM = 144 sec.
8. LPM Read I/O Activity (11,500 IOPS), Write I/O Activity (1200 IOPS) from SAN Volume Controller and 12 Mbps Read and Write on Host	Live Partition Mobility (Site A to Site B): 0 KM = 210 sec. 5 KM = TBD 15 KM = TBD 75 KM = TBD 150 KM = TBD 300 KM = TBD

5.1.3 Test conclusions

This section describes the test conclusions.

Key supporting notes

The test scenarios resulted in the following key supporting notes:

- Workload for all tests: DayTrader and Trade workload with DB2 and WebSphere Application Server (DB2 size is 45 GB). Raw read/write on RAW disk that uses AIX **dd** command concurrently from all the nodes of SystemMirror cluster
- Tested to simulated 0 - 300 km distance by setting latency of 3 ms round-trip time. From the latency/distance simulator, assumed.01 ms of latency = 1 KM.
- Round Trip Time (RTT): Fibre Channel data transfer is limited by the speed of light (increasing the latency) and by Buffer-to-Buffer credits (reducing the bandwidth). The speed of light in glass fiber is approximately 300,000 kmps. This means there is an additional latency of 0.00625 ms per kilometer per round trip. If we add other delay factors on the path, the formula is 0.01 ms/kilometers per round trip.
- Write latency (WL) in the test is because of an increase in response time to the host. That means it increases as the distance is increasing throughout the test. This write is actually the write that SAN Volume Controller performs in background to keep the primary and secondary volumes in sync and maintain the SAN Volume Controller cache. SAN Volume Controller Write Latency was consistent through all scenarios with the following results:
  - 5 KM = 2 ms
  - 15 KM = 3 ms

- 75 KM = 5 ms
- 150 KM = 8 ms
- 300 KM = 14 ms

- As another consideration, more volumes should be added to further validate SAN Volume Controller write latency increases and potential further increases in failover times that are based on your specific workload requirements.
- For the SAN Volume Controller Component loss testing (scenarios 1 and 4 in Table 5-1 on page 301) the testing involved disabling and enabling SAN Volume Controller port from switch.
- Functional tests were performed to validate the quorum's availability on the Storwize V7000 disk and noticed that failure of active quorum successfully moved to the quorum candidate on the alternative Storwize V7000 disk. Quorum's availability in the existing configuration is vital. Quorum's unavailability leads to a SAN Volume Controller cluster reboot. For more information about the Quorum disk concepts, see Chapter 3, "SAN Volume Controller Implementation" on page 149.
- When IOPS is increased, the write increases on the SAN Volume Controller with less impact on the host (overhead is on the SAN Volume Controller side). For the load on SAN Volume Controller, storage uses this parameter to enforce the capacity or capability; for example, Storage Site A works well (no bottleneck) up through 180 K IOPS. This increases when we increase the load (RW request) from host side and on the number of volumes.
- Write latency is increasing by nearly a 2X factor, but not failover. From our test observations, added volume continues to increase Write Latency, but might not dramatically increase failover. There is more transactional activity in this type of test case with write activity, therefore, failover is increased because of higher impact in relation to increase in synchronization between the volumes (primary and secondary).
- Scenario 5 in Table 5-1 on page 301 (Site Failure, both SAN Directors lost) results in the following breakdown: 15 seconds (SAN Volume Controller to Site B Directors/Switches) and 115 Seconds for PowerHA (including DB2 Restart: 1 - 2 GB) to Site B. SAN Volume Controller Change Node from A to B.
- Scenario 6 in Table 5-1 on page 301 (Storwize V7000 loss): IOPS on Storage side: 30000 Read, 1300 Write. Test simulated by disabling and enabling V7000 controller ports.
- Scenario 7 in Table 5-1 on page 301 (Component Loss: Server p750): Site A failover to Site B, Power off P7 server through HMC, Site B server restarted via PowerHA.

Conclusions

The testing results provided the following conclusions for our solution:

- The end-to-end solution with the technologies deployed is a valid end-to-end high availability solution that can be used for critical workloads. It falls between clustered and cross-site solutions on the Infrastructure Availability Continuum and Power Systems High Availability Building Blocks as described in Chapter 1, "Introduction" on page 1.
- Based on our testing parameters, the end-to-end SAN Volume Controller Stretched Cluster solution, sizings, and conditions can safely function up to 300 KM (186.41 miles).
- In terms of the SAN Volume Controller Stretched Cluster component and node-specific testing, the SAN Volume Controller Stretched Cluster is more sensitive to write I/O activity (for example, OLTP type activity). Failover times were higher for Write I/O activity as compared to Read Activity. This should be factored in when considering what types of workload should be set up in the I/O group for the Stretched Cluster node.

- SAN Volume Controller node testing (in our test cases 1 node = 1 entire SAN Volume Controller loss), shows failover times to increment fairly consistently between 5 and 75 KM (3.1 - 46.6 miles) anywhere from 3 - 5 seconds longer from 5 - 75 KM (3.1 - 46.6 miles) across all test cases. Then, if write I/O activity is involved, the times increase another 4 - 5 seconds going to 150 KM or 300 KM (186.41 miles). The failover times to 150 - 300 KM (93.2 - 186.41 miles) were only 1 - 2 seconds more if Read I/O activity is involved.
- For the single SAN Director/Site A Only loss (switch to the other Site A SAN Director), it was assumed the failover times would be consistent from 5 - 300 KM (3.1 - 186.41 miles) with little to no increase in the times. The results appear to be driven by multipathing software on the host, and configuration of the paths between the host, directors, and SAN Volume Controller. Also, load balancing can affect failover, which is more sensitive to failover.

In the test case where both SAN Directors fail on Site A, this would be equivalent to a complete site loss. Thus, the failover times are much larger than a single component (SAN Volume Controller node, Storwize V7000, P750 server, or LPAR loss). Because this simulates a site failure, there is a switch of more technology (application) to failover and restart. Because this simulates a site failure, there are more components to failover and restart.

- The same patterns (as with scenario 4 in Table 5-1 on page 301) emerge with the Storwize V7000 loss testing. The failover times from 5 - 75 KM (3.1 - 46.6 miles) increase only by 1 - 2 seconds, but then jump to 5 seconds more when 150 KM (93.2 miles) and 300 KM (186.41 miles) is tested.
- PowerHA testing is consistent and less prone to varying failover times as distance increases; increments from 5 - 300 KM (3.1 - 186.41 miles) increase only by 2 seconds on average.
- The initial LPM test completed with a 45 GB sized DB2 database, in a single data center “Local HA” distance scenario proved to be reasonable (210 seconds) and meets most customer requirements for critical workloads. Over the additional distance scenarios, checkpointing and restoring might take more time.
- LPM time is affected by CPU and memory workload. So to test this further, a considerable amount of usage (for example, 80% CPU usage and 80% memory usage through application) is required on the LPAR. Additionally, an AIX 'yes' command can be used to put stress.

Other future testing considerations

Other SAN Volume Controller Stretched Cluster solution testing was considered for the following areas:

- Test a 2-node and 4-node SAN Volume Controller Stretched Cluster configuration against the same test cases.
- Further volume and stress testing of all scenarios 100 - 500 vdisks (from SAN Volume Controller) in use for AIX hosts.
- Test with DB2 DB size variations of 200 GB, 500 GB and 1 TB of database or more and place more CPU and memory workload on the P750 servers and LPARs by using benchmark tools.
- Test Oracle Automatic Storage Management (ASM) and Real Application Clusters (RAC) with PowerHA and this overall solution.

The results of these tests are targeted to be made available in future IBM publications.

5.2 Solution recovery planning

In the context of the Stretched Cluster environment, solution recovery planning is more application-oriented, and therefore any plan should be made together with the client application owner. Hence, in each customer environment, when a business continuity or disaster recovery solution is designed, a solution recovery plan also should be incorporated.

It is imperative to identify those high priority applications that are critical to the nature of the business and create a plan to recover those in tandem with the other elements that are described in this chapter.

5.3 SAN Volume Controller recovery planning

To achieve the most benefit from the SAN Volume Controller Stretched Cluster configuration, post installation planning must include several important steps. These steps ensure that your infrastructure can be recovered with the same or a different configuration in one of the surviving sites with minimal effect on the client applications.

Proper planning and configuration backup also helps minimize possible downtime.

We can categorize the recovery in the following ways:

- Recover a fully redundant SAN Volume Controller configuration in the surviving site without Stretched Cluster.
- Recover a fully redundant SAN Volume Controller configuration in the surviving site with Stretched Cluster implemented in the same site or with a new remote site.
- Recovery of one of these scenarios with a fallback chance on the original recovered site after the critical event.

Independently from which scenario you face, the following best practices are valid and must be applied.

5.3.1 Planning for SAN Volume Controller recovery first steps

Planning the SAN Volume Controller recovery requires that you perform the following steps:

1. Collect a detailed SAN Volume Controller configuration as shown in the following steps:
   a. Run a daily-based SAN Volume Controller configuration backup with the SAN Volume Controller CLI commands, as shown in Example 5-1.

Example 5-1 Saving SAN Volume Controller configuration

```
IBM_2145:ITSO_SVC_SPLIT:superuser>svcconfig backup
.............................................................................
CMMVC6155I SVCCONFIG processing completed successfully
IBM_2145:ITSO_SVC_SPLIT:superuser>lsdumps
id filename
0  151580.trc.old
.
.
24 SVC.config.backup.xml_151580
```

b. Save the `.xml` file that is produced in a safe place, as shown in Example 5-2.

Example 5-2 Copying configuration

```
C:\Program Files\PuTTY>pscp -load SVC split iogrp
admin@10.17.89.251:/tmp/SVC.config.backup.xml_151580 c:\temp\configbackup.xml
configbackup.xml | 97 kB | 97.2 kB/s | ETA: 00:00:00 | 100%
```

c. Save the output of the SAN Volume Controller CLI commands in `.txt` format, as shown in Example 5-3.

Example 5-3 List of SAN Volume Controller commands to issue

```
lsystem
lsnode
lsnode node <nodes name>
lsnodevpd <nodes name>
lsiogrp
lsiogrp <iogrps name>
lscontroller
lscontroller <controllers name>
lsmdiskgrp
lsmdiskgrp <mdiskgrps name>
lsmdisk
lsquorum
lsquorum <quorum id>
lsvdisk
lshost
lshost <host name>
lshostvdiskmap
```

From the output of these commands and the `.xml` file, we have a complete picture of the Stretched Cluster infrastructure. We also know what the SAN Volume Controller FC ports worldwide node name (WWNN) were so that they can be reused during the recovery operation, as described in 5.6.3, “SAN Volume Controller recovery guidelines” on page 328.

Example 5-4 shows that we must re-create a Stretched Cluster environment after a critical event, which is contained in the `.xml` file.

Example 5-4 .xml configuration file

```
<object type="node" >
    <property name="id" value="1" />
    <property name="name" value="node_151580" />
    <property name="UPS_serial_number" value="100014P293" />
    <property name="WWNN" value="500507680110B13F" />
    <property name="status" value="online" />
    <property name="IO_Group_id" value="0" />
    <property name="IO_Group_name" value="io_grp0" />
    <property name="partner_node_id" value="2" />
    <property name="partner_node_name" value="node_151523" />
    <property name="config_node" value="yes" />
    <property name="UPS_unique_id" value="2040000044802243" />
    <property name="port_id" value="500507680140B13F" />
    <property name="port_status" value="active" />
    <property name="port_speed" value="8Gb" />
    <property name="port_id" value="500507680130B13F" />
    <property name="port_status" value="active" />
    <property name="port_speed" value="8Gb" />
    <property name="port_id" value="500507680110B13F" />
```

```
    <property name="port_status" value="active" />
    <property name="port_speed" value="8Gb" />
    <property name="port_id" value="500507680120B13F" />
    <property name="port_status" value="active" />
.
lines ommitted for brevity
.
<property name="service_IP_address" value="10.17.89.251" />
    <property name="service_gateway" value="10.17.80.1" />
    <property name="service_subnet_mask" value="255.255.240.0" />
    <property name="service_IP_address_6" value="" />
    <property name="service_gateway_6" value="" />
    <property name="service_prefix_6" value="" />
```

We also can get the Stretched Cluster information from the `.txt` command output, as shown in Example 5-5.

Example 5-5 lsnode example output command

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode 1
id 1
name ITSO_SVC_NODE1_SITE_A
UPS_serial_number 100006B119
WWNN 500507680100B13F
status online
IO_group_id 0
IO_group_name io_grp0
partner_node_id 2
partner_node_name ITSO_SVC_NODE1_SITE_B
config_node yes
UPS_unique_id 2040000006481049
port_id 500507680140B13F
port_status active
port_speed 8Gb
port_id 500507680130B13F
port_status active
port_speed 8Gb
port_id 500507680110B13F
port_status active
port_speed 8Gb
port_id 500507680120B13F
port_status active
port_speed 8Gb
hardware CF8
iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea
iscsi_alias
failover_active no
failover_name ITSO_SVC_NODE1_SITE_B
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1siteb
failover_iscsi_alias
panel_name 151580
enclosure_id
canister_id
enclosure_serial_number
service_IP_address 10.17.89.253
service_gateway 10.17.80.1
service_subnet_mask 255.255.240.0
service_IP_address_6
service_gateway_6
service_prefix_6
service_IP_mode static
```

```
service_IP_mode_6
```

For more information about how the backup of your configuration, see this website:

http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp?topic=%2Fcom.ibm.storage.svc.console.doc%2Fsvc_svconfigbackup_1esjw7.html

It is recommended to have an up-to-date, high-level copy of your configuration in which all elements and connections are described.

It is recommended to have a standard labeling schema and name convention for your FC or Ethernet cabling and ensure it is fully documented.

2. Back up your SAN zoning. The zoning backup can be done by using your FC switch or director command-line interface (CLI) or GUI.

   The essential zoning configuration data, domain ID, zoning, alias, configuration, or zone set can be saved in a `.txt` file by using the output from the CLI commands or by using the appropriate utility to back up the entire configuration.

   Example 5-6 shows what is saved in a `.txt` file by using CLI commands.

Example 5-6 Zoning example

```
Public_A1:FID111:admin> switchshow
switchName:     Public_A1
switchType:     121.3
switchState:    Online
switchMode:     Native
switchRole:     Principal
switchDomain:   11
switchId:       fffc0b
switchWwn:      10:00:00:05:33:b5:3e:01
zoning:         ON (ITSO_Public1)
switchBeacon:   OFF
FC Router:      OFF
Allow XISL Use: OFF
LS Attributes:  [FID: 111, Base Switch: No, Default Switch: No, Address Mode 0]

Index Slot Port Address Media  Speed  State       Proto
=====================================================
 12  1 12 0bfcc0 -- -- Online  VE VE-Port 10:00:00:05:33:97:a5:01
"Public_B1" (downstream)
 192    8    0   0bcfc0   id    N16   No_Light    FC
 193    8    1   0bcf80   id    N16   No_Light    FC
 194    8    2   0bcf40   id     N8   Online      FC  F-Port  50:05:07:68:01:40:b1:3f
 196    8    4   0bcec0   id     N8   Online      FC  F-Port  50:05:07:68:02:10:00:ef
 197    8    5   0bce80   id     N8   Online      FC  F-Port  50:05:07:68:02:20:00:ef
 198    8    6   0bce40   id     N8   Online      FC  F-Port  50:05:07:68:02:10:00:f0
 199    8    7   0bce00   id     N8   Online      FC  F-Port  50:05:07:68:02:20:00:f0


Public_A1:FID111:admin> fabricshow
Switch ID   Worldwide Name           Enet IP Addr    FC IP Addr      Name
-------------------------------------------------------------------------
 11: fffc0b 10:00:00:05:33:b5:3e:01 10.17.85.251    0.0.0.0        >"Public_A1"
 21: fffc15 10:00:00:05:33:97:a5:01 10.17.85.195    0.0.0.0         "Public_B1"

The Fabric has 2 switches

Public_A1:FID111:admin> cfgshow
Defined configuration:
 cfg:   ITSO_Public1
```

```
                V7K_SITEB; SVCN2P1_DS5100Cont1; SVCN2P1_DS5100Cont2;
                V7K_SITEA; SVCN1P1_DS5100Cont2; SVCN2P1_V7KSITEB;
                SVCN1P1_DS5100Cont1; SVCN2P1_V7KSITEA; SVCN1P1_V7KSITEB;
                SVCN1P1_V7KSITEA
 zone:  SVCN1P1_DS5100Cont1
                ITSO_SVC_N1_P1; ITSO_DS5100_Cont1_P3; ITSO_DS5100_Cont1_P1
 zone:  SVCN1P1_DS5100Cont2
                ITSO_SVC_N1_P1; ITSO_DS5100_Cont2_P1; ITSO_DS5100_Cont2_P3
.
lines omitted for brevity
.
zone:  V7K_SITEA
                ITSO_V7K_SITEA_N1_P2; ITSO_V7K_SITEA_N1_P1;
                ITSO_V7K_SITEA_N2_P1; ITSO_V7K_SITEA_N2_P2
 zone:  V7K_SITEB
                ITSO_V7K_SITEB_N2_P1; ITSO_V7K_SITEB_N1_P2;
                ITSO_V7K_SITEB_N1_P1; ITSO_V7K_SITEB_N1_P4
 alias: ITSO_DS5100_Cont1_P1
                20:16:00:A0:B8:47:39:B0
 alias: ITSO_DS5100_Cont1_P3
                20:36:00:A0:B8:47:39:B0
 alias: ITSO_DS5100_Cont2_P1
                20:17:00:A0:B8:47:39:B0
.
lines omitted for brevity
.
alias: ITSO_V7K_SITEA_N2_P2
                50:05:07:68:02:20:00:F0
 alias: ITSO_V7K_SITEB_N1_P1
                50:05:07:68:02:10:54:CA
 alias: ITSO_V7K_SITEB_N1_P2
                50:05:07:68:02:20:54:CA
 alias: ITSO_V7K_SITEB_N1_P4
                50:05:07:68:02:40:54:CA
 alias: ITSO_V7K_SITEB_N2_P1
                50:05:07:68:02:10:54:CB

Effective configuration:
 cfg:   ITSO_Public1
 zone:  SVCN1P1_DS5100Cont1
                50:05:07:68:01:40:b1:3f
                20:36:00:a0:b8:47:39:b0
                20:16:00:a0:b8:47:39:b0
 zone:  SVCN1P1_DS5100Cont2
                50:05:07:68:01:40:b1:3f
                20:17:00:a0:b8:47:39:b0
                20:37:00:a0:b8:47:39:b0
 zone:  SVCN1P1_V7KSITEA
                50:05:07:68:02:20:00:ef
                50:05:07:68:01:40:b1:3f
                50:05:07:68:02:10:00:ef
                50:05:07:68:02:10:00:f0
                50:05:07:68:02:20:00:f0
.
lines omitted for brevity
.
zone:  V7K_SITEA
                50:05:07:68:02:20:00:ef
                50:05:07:68:02:10:00:ef
                50:05:07:68:02:10:00:f0
```

```
                50:05:07:68:02:20:00:f0
zone:  V7K_SITEB
                50:05:07:68:02:10:54:cb
                50:05:07:68:02:20:54:ca
                50:05:07:68:02:10:54:ca
                50:05:07:68:02:40:54:ca
```

As a best practice, we suggest that during the implementation you consider the use of WWNN zoning, and during the recovery phase after a critical event, to reuse if possible the same domain ID and port number that was used in the failing site. Zoning is propagated on each switch or director because of SAN extension with ISL. For more information, see 5.6.3, “SAN Volume Controller recovery guidelines” on page 328.

For more information about how to back up your FC switch or director zoning configuration, see your switch vendors documentation.

3. Back up your backend storage subsystems configuration.

   In your Stretched Cluster implementation, you can use different vendors storage subsystems. Those storage subsystems should be configured by following the SAN Volume Controller best practices to be used for Volume Mirroring.

   It is recommended to back up your storage subsystem configuration to be in the best position to re-create the same environment. This is necessary in case of a critical event when you re-establish your Stretched Cluster infrastructure in a different site with new storage subsystems.

   For more information, see 5.6.3, “SAN Volume Controller recovery guidelines” on page 328.

   Complete the following steps to backup your configuration:

   a. As an example, for DS3XXX, DS4XXX, or DS5XXX storage subsystems, you save a copy of an up-to-date subsystem profile is a safe place, as shown in Figure 5-9.

Figure 5-9 DSXYYY backup configuration example

b. For the DS8000 storage subsystem, we suggest that you save in `.txt` format the output of the SAN Volume Controller CLI commands, as shown in Example 5-7.

Example 5-7 DS8000 commands

```
lsarraysite -l
lsarray -l
lsrank -l
lsextpool -l
lsfbvol -l
lshostconnect -l
lsvolgrp -l
showvolgrp -lunmap <SVC vg_name>
```

c. For the XIV storage subsystem, we suggest saving in `.txt` format the output of the XCLI commands, as shown in Example 5-8.

Example 5-8 XIV commands

```
host_list
host_list_ports
mapping_list
vol_mapping_list
pool_list
vol_list
```

d. For Storwize V7000, collect configuration files and output report as was done for SAN Volume Controller, as described in 5.3.1, “Planning for SAN Volume Controller recovery first steps” on page 305.

e. For any other supported storage vendor, see their documentation to save details where it is relatively easy to find the SAN Volume Controller MDisk configuration and mapping.

5.4 PowerVM recovery planning

PowerVM recovery planning mainly concentrates around backing up system profiles and Virtual I/O Server configuration. The LPAR can be taken by using `mksysb` image and store it in the Network Installation Management (NIM) server, which can be later restored by using NIM. In this section, we describe this process.

5.5 PowerHA SystemMirror recovery planning

PowerHA SystemMirror is a disaster recovery solution in which a workload fails over to another node in configuration that can be a surviving site or recovered site. In a SAN Volume Controller Stretched Cluster solution, the surviving site contains the other node of the cluster that takes over the workload in case of failures. A failed PowerHA SystemMirror node can be recovered in case of catastrophic failures by using `mksysb`.

5.6 SAN Volume Controller Diagnosis and recovery guidelines

In this section, we provide some guidelines regarding how to diagnose a critical event in one of the two sites where the Split I/O Group was implemented.

By using these guidelines, you are in a position to understand the extent of any damage, what is still running, and what can be recovered with which impact on the performance to the application and service level agreement.

5.6.1 Critical event scenarios and complete domain failure

In a Stretched Cluster environment, we can face different critical event scenarios. Some of the scenarios are handled as Business As Usual recovery procedure. In this chapter, we describe the operations that are required to recover from a complete site failure.

The following scenarios and their required recovery procedures can be encountered:

- Backend storage box failure in one failure domain: Business As Usual because of SAN Volume Controller VolumeMirroring.
- Partial SAN failure in one failure domain: Business As Usual because of SAN resilience.
- Total SAN failure in one failure domain: Business As Usual because of SAN resilience but pay attention to performance impact and eventually take appropriate action to minimize application effects.
- SAN Volume Controller node failure in one failure domain: Business As Usual because of SAN Volume Controller High Availability.
- Complete site failure in one failure domain: This is the scope of chapter and all the required operations are described in 5.6.2, “SAN Volume Controller diagnosis guidelines” on page 312.

5.6.2 SAN Volume Controller diagnosis guidelines

In this section, we provide some guidelines regarding how to diagnose a critical event in one of the two sites where the Stretched Cluster was implemented.

We are considering a Stretched Cluster configuration, as shown in Figure 5-10 on page 313.

Figure 5-10 Environment diagram

We are assuming that the configuration that is implemented is consistent with one of the supported configurations that were described in this book.

Up and running scenario analysis

We are starting with a scenario where everything is up and running, all the best practices were applied, and the solution is implemented as shown in the SAN Volume Controller CLI command output in Example 5-9.

Example 5-9 Running example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>svcinfo lscluster
id                 name           location partnership bandwidth id_alias
0000020060C14FBE ITSO_SVC_SPLIT local                             0000020060C14FBE

IBM_2145:ITSO_SVC_SPLIT:superuser>lsiogrp
id name            node_count vdisk_count host_count
0  ITSO_SVC_SPLIT  2          0           0
1  io_grp1         0          0           0
2  io_grp2         0          0           0
3  io_grp3         0          0           0
4  recovery_io_grp 0          0           0

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode
id name                  UPS_serial_number WWNN             status IO_group_id
IO_group_name  config_node UPS_unique_id   hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number
```

```
1  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F online 0
ITSO_SVC_SPLIT yes         2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE1_SITE_B 100006B074        500507680100B0C6 online 0
ITSO_SVC_SPLIT no          20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1siteb             151523

IBM_2145:ITSO_SVC_SPLIT:superuser>lscontroller
id controller_name     ctrl_s/n             vendor_id            product_id_low
product_id_high
0  ITSO_V7K_SITEB_N2   2076                 IBM                  2145
1  ITSO_V7K_SITEA_N2   2076                 IBM                  2145
2  ITSO_V7K_SITEA_N1   2076                 IBM                  2145
3  ITSO_V7K_SITEB_N1   2076                 IBM                  2145
5  ITSO_V7K_SITEC_Q_N2 2076                 IBM                  2145
6  ITSO_V7K_SITEC_Q_N1 2076                 IBM                  2145

IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdiskgrp
id name       status mdisk_count vdisk_count capacity extent_size free_capacity
virtual_capacity used_capacity real_capacity overallocation warning easy_tier
easy_tier_status compression_active compression_virtual_capacity
compression_compressed_capacity compression_uncompressed_capacity
0  V7000SITEA online 5          0          2.22TB   256        2.22TB       0.00MB
0.00MB       0.00MB       0             80      auto      active          no
0.00MB                       0.00MB                          0.00MB
1  V7000SITEB online 5          0          2.22TB   256        2.22TB       0.00MB
0.00MB       0.00MB       0             80      auto      active          no
0.00MB                       0.00MB                          0.00MB
2  V7000SITEC online         1           0           512.00MB 256         512.00MB
0.00MB          0.00MB       0.00MB        0             80      auto      inactive
no                0.00MB                       0.00MB                          0.00MB
0.00MB


IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdisk
id name                status mode    mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name   UID                                                            tier
0  ITSO_V7K_SITEA_SSD0 online managed 0            V7000SITEA     277.3GB  0000000000000000
ITSO_V7K_SITEA_N2 6005076802890002680000000000000800000000000000000000000000000000
generic_ssd
1  ITSO_V7K_SITEA_SAS0 online managed 0            V7000SITEA     500.0GB  000000000000000A
ITSO_V7K_SITEA_N2 6005076802890002680000000000000400000000000000000000000000000000
generic_hdd
2  ITSO_V7K_SITEA_SAS1 online managed 0            V7000SITEA     500.0GB  000000000000000B
ITSO_V7K_SITEA_N2 6005076802890002680000000000000500000000000000000000000000000000
generic_hdd
3  ITSO_V7K_SITEA_SAS2 online managed 0            V7000SITEA     500.0GB  000000000000000C
ITSO_V7K_SITEA_N2 6005076802890002680000000000000600000000000000000000000000000000
generic_hdd
4  ITSO_V7K_SITEA_SAS3 online managed 0            V7000SITEA     500.0GB  000000000000000D
ITSO_V7K_SITEA_N2 6005076802890002680000000000000700000000000000000000000000000000
generic_hdd
5  ITSO_V7K_SITEB_SSD1 online managed 1            V7000SITEB     277.3GB  0000000000000001
ITSO_V7K_SITEB_N2 600507680282018b3000000000000004000000000000000000000000000000000
generic_ssd
6  ITSO_V7K_SITEB_SAS0 online managed 1            V7000SITEB     500.0GB  0000000000000014
ITSO_V7K_SITEB_N2 600507680282018b3000000000000000000000000000000000000000000000000
generic_hdd
```

```
7  ITSO_V7K_SITEB_SAS1 online managed 1            V7000SITEB     500.0GB  0000000000000015
ITSO_V7K_SITEB_N2 600507680282018b3000000000000001000000000000000000000000000000000
generic_hdd
8  ITSO_V7K_SITEB_SAS2 online managed 1            V7000SITEB     500.0GB  0000000000000016
ITSO_V7K_SITEB_N2 600507680282018b3000000000000002000000000000000000000000000000000
generic_hdd
9  ITSO_V7K_SITEB_SAS3 online managed 1            V7000SITEB     500.0GB  0000000000000017
ITSO_V7K_SITEB_N2 600507680282018b3000000000000003000000000000000000000000000000000
generic_hdd
10 ITSO_V7K_SITEC_Q    online         managed 2            V7000SITEC     1.0GB
0000000000000001 ITSO_V7K_SITEC_Q_N2
600507680283801ac8000000000000050000000000000000000000000000000000 generic_hdd

IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk
id name             IO_group_id IO_group_name  status mdisk_grp_id mdisk_grp_name capacity
type    FC_id FC_name RC_id RC_name vdisk_UID                        fc_map_count
copy_count fast_write_state se_copy_count RC_change compressed_copy_count
0  ESXi-01-DCA-HBA1 0           ITSO_SVC_SPLIT online 0            V7000SITEA     1.00GB
striped                            600507680183053EF800000000000000 0           1
empty            0             no        0
1  ESXi-01-DCA-HBA2 0           ITSO_SVC_SPLIT online 0            V7000SITEA     1.00GB
striped                            600507680183053EF800000000000001 0           1
empty            0             no        0
2  ESXi-02-DCB-HBB1 0           ITSO_SVC_SPLIT online 1            V7000SITEB     1.00GB
striped                            600507680183053EF800000000000002 0           1
empty            0             no        0
3  ESXi-02-DCB-HBB2 0           ITSO_SVC_SPLIT online 1            V7000SITEB     1.00GB
striped                            600507680183053EF800000000000003 0           1
empty            0             no        0
4  ESXI_CLUSTER_01  0           ITSO_SVC_SPLIT online many         many           256.00GB
many                               600507680183053EF800000000000004 0           2
empty            0             no        0
5  ESXI_CLUSTER_02  0           ITSO_SVC_SPLIT online many         many           256.00GB
many                               600507680183053EF800000000000005 0           2
empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT online many         many
256.00GB many                                600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT online many         many
256.00GB many                                600507680183053EF800000000000007 0
2          empty            0             no        0


IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name     active
object_type override
0           online 10 ITSO_V7K_SITEC_Q    5            ITSO_V7K_SITEC_Q_N2 yes    mdisk
yes
1           online 6  ITSO_V7K_SITEB_SAS0 0            ITSO_V7K_SITEB_N2   no     mdisk
yes
2           online 4  ITSO_V7K_SITEA_SAS3 1            ITSO_V7K_SITEA_N2   no     mdisk
yes
```

From the SAN Volume Controller CLI command output that is shown in Example 5-9 on page 313, you can see the following results:

- The SAN Volume Controller clustered system is accessible through the CLI.
- The SAN Volume Controller nodes are online and one of them is the configuration node.
- The I/O Groups are in the correct state.

- The subsystem storage controllers are connected.
- The Managed Disk (MDisks) Groups are online.
- The MDisks are online.
- The volumes are online.
- The three Quorum disks are in the correct state.

Now we can check the Volume Mirroring status by running a single SAN Volume Controller CLI command against each single volume, as shown in Example 5-10.

Example 5-10 Volume mirroring status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk ESXI_CLUSTER_01
id 4
name ESXI_CLUSTER_01
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
status online
mdisk_grp_id many
mdisk_grp_name many
capacity 256.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 600507680183053EF800000000000004
throttling 0
preferred_node_id 1
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 95
copy_count 2
se_copy_count 0
filesystem
mirror_write_priority latency
RC_change no
compressed_copy_count 0
access_IO_group_count 1

copy_id 0
status online
sync yes
primary yes
mdisk_grp_id 0
mdisk_grp_name V7000SITEA
type striped
mdisk_id
mdisk_name
.
lines omitted for brevity
.
easy_tier_status active
tier generic_ssd
tier_capacity 0.00MB
tier generic_hdd
```

```
tier_capacity 256.00GB
compressed_copy no
uncompressed_used_capacity 256.00GB

copy_id 1
status online
sync yes
primary no
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
type striped
mdisk_id
mdisk_name
uncompressed_used_capacity 256.00GB
.
lines omitted for brevity
.
```

From the SAN Volume Controller CLI command output that is shown in Example 5-10 on page 316, you can see the following results:

- The volume is online.
- The storage pool name and the MDisk name is “many”, which means Volume Mirroring is in place.
- The Copy id 0 is online, in sync and it is the primary.
- The Copy id 1 is online, in sync and it is the secondary.

If you have several volumes to check, you might create a customized script directly from the SAN Volume Controller shell.

Some example useful scripts can be found at this website:

https://www.ibm.com/developerworks/mydeveloperworks/wikis/home/wiki/W1d985101fbfa_4ae7_a090_dc535355ae7e/page/Show%20Volume%20Copies?lang=en

Critical event scenario analysis

We review our SAN Volume Controller environment following a critical event that caused the complete loss of one of the sites, Site 1.

Complete the following steps as a guideline or checklist to gain a complete view of any damage and to gather enough decision elements to determine your next recovery actions:

1. Is SAN Volume Controller system management available through GUI or CLI?
   a. Is SAN Volume Controller system login possible?
      - If yes: SAN Volume Controller system is online, continue with Step 2
      - If no: SAN Volume Controller system might be offline or suffering connection problems.

      i. Check your connections, cabling, node front panel event messages.

      ii. Verify the SAN Volume Controller system status by using Service Assistant menu or node front panel. For more information, see *IBM System Storage SAN Volume Controller Troubleshooting Guide,* GC27-2284.

      iii. Bring a part of the SAN Volume Controller system online for further diagnostics.

iv. By using a browser, connect to one of the SAN Volume Controller node's service IP addresses as shown in the following example:

```
https://<service_ip_add>/service/
```

Log in with your SAN Volume Controller Cluster GUI password. After the login, you are redirected to the Service Assistant menu, as shown in Figure 5-11 and Figure 5-12 on page 318.

Figure 5-11 SAN Volume Controller Service Assistant Tool login

Figure 5-12 Service Assistant menu

From the Service Assistant menu, you can bring at least a part of the SAN Volume Controller Clustered system online for further diagnostics.

For more information about Service Assistant menu, see *Implementing the IBM System Storage SAN Volume Controller V6.3* SG24-7933.

2. If the SAN Volume Controller system management is available:

   a. Check the status by using the SAN Volume Controller CLI by running the commands that are shown in Example 5-11.

Example 5-11 lscluster example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>svcinfo lscluster
id               name           location partnership bandwidth id_alias
0000020060C14FBE ITSO_SVC_SPLIT local                           0000020060C14FBE

IBM_2145:ITSO_SVC_SPLIT:superuser>svcinfo lscluster ITSO_SVC_SPLIT
id 0000020060C14FBE
name ITSO_SVC_SPLIT
location local
partnership
bandwidth
.
lines ommitted for brevity
.
total_mdisk_capacity 4.4TB
space_in_mdisk_grps 4.4TB
space_allocated_to_vdisks 0.00MB
total_free_space 4.4TB
statistics_status on
statistics_frequency 15
required_memory 0
gm_max_host_delay 5
tier generic_ssd
tier_capacity 554.50GB
tier_free_capacity 554.50GB
tier generic_hdd
tier_capacity 3.90TB
tier_free_capacity 3.90TB
email_contact2
email_contact2_primary
email_contact2_alternate
total_allocated_extent_capacity 1.50GB
has_nas_key no
auth_service_type tip
layer replication
rc_buffer_size 48
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

As you can see, the SAN Volume Controller clustered system appears to be accessible from the GUI, as shown in Figure 5-13.

Figure 5-13 GUI example

b. Check the status of the nodes, as shown in Example 5-12.

Example 5-12 Node status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode
id name                  UPS_serial_number WWNN             status  IO_group_id
IO_group_name config_node UPS_unique_id   hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number
1  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F offline 0
ITSO_SVC_SPLIT no          2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online  0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode ITSO_SVC_NODE1_SITE_A
id 1
name ITSO_SVC_NODE1_SITE_A
UPS_serial_number 100006B119
WWNN 500507680100B13F
status offline
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
partner_node_id 2
partner_node_name ITSO_SVC_NODE2_SITE_B
config_node no
UPS_unique_id 2040000006481049
port_id 500507680140B13F
port_status inactive
port_speed 8Gb
port_id 500507680130B13F
port_status inactive
port_speed 8Gb
port_id 500507680110B13F
```

```
port_status inactive
port_speed 8Gb
port_id 500507680120B13F
port_status inactive
port_speed 8Gb
hardware CF8
iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea
iscsi_alias
failover_active no
failover_name ITSO_SVC_NODE2_SITE_B
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb
failover_iscsi_alias
panel_name 151580
enclosure_id
canister_id
enclosure_serial_number
service_IP_address 10.17.89.253
service_gateway 10.17.80.1
service_subnet_mask 255.255.240.0
service_IP_address_6
service_gateway_6
service_prefix_6
service_IP_mode static
service_IP_mode_6

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode ITSO_SVC_NODE2_SITE_B
id 2
name ITSO_SVC_NODE2_SITE_B
UPS_serial_number 100006B074
WWNN 500507680100B0C6
status online
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
partner_node_id 1
partner_node_name ITSO_SVC_NODE1_SITE_A
config_node yes
UPS_unique_id 20400000064801C4
port_id 500507680140B0C6
port_status active
port_speed 8Gb
port_id 500507680130B0C6
port_status active
port_speed 8Gb
port_id 500507680110B0C6
port_status active
port_speed 8Gb
port_id 500507680120B0C6
port_status active
port_speed 8Gb
hardware CF8
iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb
iscsi_alias
failover_active no
failover_name ITSO_SVC_NODE1_SITE_A
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea
failover_iscsi_alias
panel_name 151523
enclosure_id
canister_id
enclosure_serial_number
```

```
service_IP_address 10.17.89.254
service_gateway 10.17.80.1
service_subnet_mask 255.255.240.0
service_IP_address_6
service_gateway_6
service_prefix_6
service_IP_mode static
service_IP_mode_6
```

As you can see from the Example 5-12 on page 320:

- The configuration node role moved from node 1 to node 2.
- Node 1 is offline.
- FC ports in node 1 are inactive.
- Node 2 is online.
- FC ports in node 2 are still online.

This means that we lost 50% of the SAN Volume Controller clustered system resources, but the system is still up and running.

By using the GUI, we can see the same information as shown in Figure 5-14.

Figure 5-14 Node status with GUI

c. Check the IO Group status, as shown in Example 5-13.

Example 5-13 I/O Group status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsiogrp
id name            node_count vdisk_count host_count
0  ITSO_SVC_SPLIT  2          8           2
1  io_grp1         0          0           0
2  io_grp2         0          0           0
3  io_grp3         0          0           0
4  recovery_io_grp 0          0           0
```

As you can see, the I/O Group still reports two nodes per I/O Group.

d. Check the quorum status, as shown in Example 5-14 on page 323.

Example 5-14 Quorum status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name     active
object_type override
0           online 10 ITSO_V7K_SITEC_Q    5            ITSO_V7K_SITEC_Q_N2 yes    mdisk
yes
1           online 6  ITSO_V7K_SITEB_SAS0 0            ITSO_V7K_SITEB_N2   no     mdisk
yes
2           online 5  ITSO_V7K_SITEB_SSD1 0            ITSO_V7K_SITEB_N2   no     mdisk
ignored
```

As you can see, the active quorum disk is still active because it was not affected by the critical event. However, the quorum index 1 that is in the site that suffered the power failure flagged it with override ignored. This is because if the original resource that is in DS3400_09 went offline and another resource is used instead, the override field in `lsquorum` shows as ignored.

e. Check the controllers status as shown in Example 5-15 and you see that some controllers might be offline as expected or might have `path_count = 0` because of the site failure.

Example 5-15 Controller status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lscontroller
id controller_name     ctrl_s/n             vendor_id            product_id_low
product_id_high
0  ITSO_V7K_SITEB_N2   2076                  IBM                   2145
1  ITSO_V7K_SITEA_N2   2076                  IBM                   2145
2  ITSO_V7K_SITEA_N1   2076                  IBM                   2145
3  ITSO_V7K_SITEB_N1   2076                  IBM                   2145
5  ITSO_V7K_SITEC_Q_N2 2076                  IBM                   2145
6  ITSO_V7K_SITEC_Q_N1 2076                  IBM                   2145

IBM_2145:ITSO_SVC_SPLIT:superuser>lscontroller ITSO_V7K_SITEA_N2
id 1
controller_name ITSO_V7K_SITEA_N2
WWNN 50050768020000F0
mdisk_link_count 5
max_mdisk_link_count 5
degraded yes
vendor_id IBM
product_id_low 2145
product_id_high
product_revision 0000
ctrl_s/n 2076
allow_quorum yes
WWPN 50050768023000F0
path_count 0
max_path_count 6
WWPN 50050768024000F0
path_count 0
max_path_count 6
WWPN 50050768021000F0
path_count 0
max_path_count 6
WWPN 50050768022000F0
path_count 0
max_path_count 6
```

As you can see in the output, some controllers are still accessible from the SAN Volume Controller system and others are no longer accessible because the power loss in site 1 has affected the SAN Volume Controller node, the storage subsystem, and the FC SAN switches.

The same information can be ascertained from the GUI, as shown in Figure 5-15.

Figure 5-15 Online controllers

The same information also can be ascertained from the offline controller, as shown in Figure 5-16 on page 324.

Figure 5-16 Offline controller

f. Check the MDiskgroup status, as shown in Example 5-16.

Example 5-16 MDiskgroup status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdiskgrp
id name       status  mdisk_count vdisk_count capacity extent_size free_capacity
virtual_capacity used_capacity real_capacity overallocation warning easy_tier
easy_tier_status compression_active compression_virtual_capacity
compression_compressed_capacity compression_uncompressed_capacity
0  V7000SITEA offline 5          6          2.22TB   256        1.22TB       1.00TB
1.00TB       1.00TB       45           80      auto      active          no
0.00MB                       0.00MB                           0.00MB
```

```
1  V7000SITEB online  5          6          2.22TB   256        1.22TB       1.00TB
1.00TB       1.00TB       45           80      auto      active          no
0.00MB                       0.00MB                           0.00MB
2  V7000SITEC online  1          0          512.00MB 256        512.00MB     0.00MB
0.00MB       0.00MB       0            80      auto      inactive        no
0.00MB                       0.00MB                           0.00MB
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

As you can see from the output and because of the critical event, some MDiskgroups are offline and others are still online. The offline MDiskgroups are those that had space allocated on the storage subsystem that suffered the critical event.

g. Check the MDisk status, as shown in Example 5-17.

Example 5-17 MDisk status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdisk
id name               status  mode    mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name     UID                                                                tier
0  ITSO_V7K_SITEA_SSD0 offline managed 0            V7000SITEA     277.3GB
0000000000000000 ITSO_V7K_SITEA_N2
6005076802890002680000000000000800000000000000000000000000000000 generic_ssd
1  ITSO_V7K_SITEA_SAS0 offline managed 0            V7000SITEA     500.0GB
000000000000000A ITSO_V7K_SITEA_N2
6005076802890002680000000000000400000000000000000000000000000000 generic_hdd
2  ITSO_V7K_SITEA_SAS1 offline managed 0            V7000SITEA     500.0GB
000000000000000B ITSO_V7K_SITEA_N2
6005076802890002680000000000000500000000000000000000000000000000 generic_hdd
3  ITSO_V7K_SITEA_SAS2 offline managed 0            V7000SITEA     500.0GB
000000000000000C ITSO_V7K_SITEA_N2
6005076802890002680000000000000600000000000000000000000000000000 generic_hdd
4  ITSO_V7K_SITEA_SAS3 offline managed 0            V7000SITEA     500.0GB
000000000000000D ITSO_V7K_SITEA_N2
6005076802890002680000000000000700000000000000000000000000000000 generic_hdd
5  ITSO_V7K_SITEB_SSD1 online  managed 1            V7000SITEB     277.3GB
0000000000000001 ITSO_V7K_SITEB_N2
600507680282018b3000000000000004000000000000000000000000000000 generic_ssd
6  ITSO_V7K_SITEB_SAS0 online  managed 1            V7000SITEB     500.0GB
0000000000000014 ITSO_V7K_SITEB_N2
600507680282018b3000000000000000000000000000000000000000000000 generic_hdd
7  ITSO_V7K_SITEB_SAS1 online  managed 1            V7000SITEB     500.0GB
0000000000000015 ITSO_V7K_SITEB_N2
600507680282018b3000000000000001000000000000000000000000000000 generic_hdd
8  ITSO_V7K_SITEB_SAS2 online  managed 1            V7000SITEB     500.0GB
0000000000000016 ITSO_V7K_SITEB_N2
600507680282018b3000000000000002000000000000000000000000000000 generic_hdd
9  ITSO_V7K_SITEB_SAS3 online  managed 1            V7000SITEB     500.0GB
0000000000000017 ITSO_V7K_SITEB_N2
600507680282018b3000000000000003000000000000000000000000000000 generic_hdd
10 ITSO_V7K_SITEC_Q    online  managed 2            V7000SITEC     1.0GB
0000000000000001 ITSO_V7K_SITEC_Q_N2
600507680283801ac8000000000000050000000000000000000000000000000 generic_hdd
```

As you can see from the output and because of the critical event, some MDisks are offline and others are still online. The offline MDisks are those that had space allocated on the storage subsystem that suffered the critical event.

We can get the same information from the GUI, as shown in Figure 5-17 on page 326.

Figure 5-17 MDisk status with GUI

h. Check the Volume status, as shown in Example 5-18.

Example 5-18 Volume status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk
id name                     IO_group_id IO_group_name  status   mdisk_grp_id mdisk_grp_name
capacity type    FC_id FC_name RC_id RC_name vdisk_UID                        fc_map_count
copy_count fast_write_state se_copy_count RC_change compressed_copy_count
0  ESXi-01-DCA-HBA1         0           ITSO_SVC_SPLIT offline  0            V7000SITEA
2.00GB   striped                              600507680183053EF800000000000000 0
1          not_empty        0             no        0
1  ESXi-01-DCA-HBA2         0           ITSO_SVC_SPLIT offline  0            V7000SITEA
2.00GB   striped                              600507680183053EF800000000000001 0
1          not_empty        0             no        0
2  ESXi-02-DCB-HBB1         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000002 0
1          empty            0             no        0
3  ESXi-02-DCB-HBB2         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000003 0
1          empty            0             no        0
4  ESXI_CLUSTER_01          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000004 0
2          empty            0             no        0
5  ESXI_CLUSTER_02          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000005 0
2          empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000007 0
2          empty            0             no        0
```

As you can see from the output in Example 5-18 on page 326, although we lost 50% of the resources because of a loss of power in site 1, the volumes are not offline but in a degraded state. This is because Volume Mirroring ensured business continuity and the volumes are still accessible from the hosts that are still running in site 2 where power is still present.

Some other volumes are offline because they are not protected by the Volume Mirror feature, but this was known during the planning of this environment.

In this case, it might be helpful to use the **filtervalue** option in the SAN Volume Controller CLI command to reduce the number of lines produced and volumes to check, as shown in Example 5-19.

Example 5-19 Volume status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk -filtervalue status=degraded
id name                     IO_group_id IO_group_name  status   mdisk_grp_id mdisk_grp_name
capacity type    FC_id FC_name RC_id RC_name vdisk_UID                         fc_map_count
copy_count fast_write_state se_copy_count RC_change compressed_copy_count
2  ESXi-02-DCB-HBB1         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000002 0
1          empty            0             no        0
3  ESXi-02-DCB-HBB2         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000003 0
1          empty            0             no        0
4  ESXI_CLUSTER_01          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000004 0
2          empty            0             no        0
5  ESXI_CLUSTER_02          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000005 0
2          empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000007 0
2          empty            0             no        0
```

As you can see from the output in Example 5-19 on page 327, one copy of each volume is offline. You also can see to which this storage pool is related.

We can get the same information from the GUI, as shown in Figure 5-18.

Figure 5-18 Volume status with GUI

As you can see from Figure 5-18, it is pretty easy to understand which resources are online and offline and why we have a degraded status related to each volume.

3. Check path status.

   Check the status of the storage paths from your hosts point of view by using your multipathing software commands.

   For SAN Volume Controller, it is recommended to use Subsystem Device Driver (SDD) as multipath software.

For more information about SDD commands, see the following resources:

– http://www-01.ibm.com/support/docview.wss?rs=540&context=ST52G7&uid=ssg1S7000303

– *Multipath Subsystem Device Driver User's Guide* GC52-1309-03.

You also can verify the SDD vpath device configuration by running the **lsvpcfg** or **datapath query device** commands.

All of these steps also are valid for a limited failure in which we are facing a failure with limited effect in one of the sites.

In a case of a limited failure, it might be helpful to use the following steps to verify the status of your Split I/O Group infrastructure:

1. Check your SAN by using the FC switch, director CLI, or web interface to verify any failure.
2. Check the FC and FC/IP connection between the two sites by using the FC switch, director CLI, or web interface to verify any eventual partial failure.
3. Check the storage subsystem status by using its own management interface to verify any failure.

After performing these steps and identifying what was the root cause and the effect of the event on your infrastructure, you have all of the information to take one of the following strategic decisions:

- Wait until the failure in one of the two sites is fixed.
- Declare a disaster and start with the recovery actions that are described in 5.6.3, “SAN Volume Controller recovery guidelines” on page 328.

If you decide to wait until the failure in one of the two sites is fixed, and when the resources that are affected become available again, the SAN Volume Controller Split I/O Group is fully operational and the following results occur:

- Automatic Volume Mirroring resynchronization takes place.
- Missing nodes rejoin the SAN Volume Controller clustered system.

If the effect of the failure is more serious and you are forced to declare a disaster, you must take a more strategic decision, which is described in 5.6.3, “SAN Volume Controller recovery guidelines” on page 328.

5.6.3 SAN Volume Controller recovery guidelines

In this section, we describe some recovery scenarios. Regardless of the different scenarios, the common starting point is the complete loss of site 1 or site 2, which is caused by a severe critical event.

After an initial analysis phase of the event, one of the following strategic decisions must be made:

- Wait until the lost site is restored.
- Start a recovery procedure so that the surviving site configuration is rebuilt such that it provides the same performance and availability characteristics as it did before the event.

If the recovery times are too long and the lost site cannot wait for its eventual return to online, you must take the appropriate recovery actions.

What must you supply to recover your Stretched Cluster configuration?

If you arrived at this point, it is because you cannot wait for the site to be brought back online in a reasonable time. Therefore, you must take some recovery actions.

The answers to the following questions determine the appropriate recovery action:

- Where do you want to recover to? In the same site or in a new site?
- Is it a temporary or permanent recovery?
- If it is a temporary recovery, do we need to plan a failback scenario?
- Does the recovery action address performance issues or business continuity issues?

It is almost certain we need more storage space, more SAN Volume Controller nodes, and more SAN components. For planning purposes, the following questions should be considered:

- Do we plan to use brand new nodes supplied by IBM?
- Do we plan to reuse other, existing SAN Volume Controller node, that might be used for non-business-critical applications (test environment) at the moment?
- Do we plan to use new FC SAN switches or directors?
- Do we plan to reconfigure FC SAN switches or directors to host newly acquired SAN Volume Controller nodes and storage?
- Do we plan to use new back-end storage subsystems?
- Do we plan to configure some free space on the surviving storage subsystem (or subsystems) to host the space that is required for Volume Mirroring?

The answers to these questions directs the recovery strategy, investment, and monetary steps to take, which cannot be improvised. Instead, they must be part of a recovery plan to create a minimal impact to applications and therefore service levels.

We describe what the recovery guidelines are, assuming that we already answered the previous questions and we decided to recover a fully redundant configuration in the same surviving site, and supply new SAN Volume Controller nodes, storage subsystems, and FC SAN devices.

We also give some indication about how to reuse SAN Volume Controller nodes that are already available, storage or SAN devices, and guidelines on how to plan a failback scenario.

If you must recover your Split I/O Group infrastructure, we recommend that you contact IBM Support as early as possible.

Recovery guidelines for our configuration

In this section we describe the recovery guidelines assuming that we decided to recover a fully redundant configuration in the same surviving site by supplying new SAN Volume Controller nodes, storage subsystems, and FC SAN devices.

This recovery action is based on a decision to recover the Stretched Cluster infrastructure to supply the same performance characteristics as before the critical event, but with limited business continuity because the Stretched Cluster is recovered only at one site.

We are assuming that we already received and physically installed a new SAN Volume Controller node, FC switches, and backend storage subsystems.

Figure 5-19 on page 330 shows the new recovery configuration.

Figure 5-19 New recovery configuration in surviving site

We decided to recover the configuration exactly as it was by using passive WDM, even if it was recovered in the same site. This makes it easier to implement this configuration over distance in the future when a new site is provided by completing the following steps:

1. Disconnect the FCIP links between Failure Domains.
2. Uninstall or reinstall all the new devices in the new site.
3. Reconnect the FCIP links between Failure Domains.

The following steps must be performed to recover your Split I/O Group configuration as it was before the critical event in the same site and after you installed the new devices:

1. Restore your back-end storage subsystem configuration. LUN masking can be done in advance because the SAN Volume Controller node's WWNN already is known.
2. Restore your SAN configuration exactly as it was before the critical event. This might be done by configuring the new switches with the same domain ID as before and connecting them to the surviving switches through the passive WDM. In this way, the WWPN zoning propagation automatically propagate to the new switches.
3. If possible, connect the new storage subsystems to the same FC switch ports as before the critical event. SAN Volume Controller to storage zoning must be reconfigured to see the new storage subsystems WWNN. Old WWNNs can be removed, but take care to remove the correct WWNNs because at this time we have only one volume copy active.
4. Do not connect the SAN Volume Controller node FC yet. Wait until you are directed to do so by the SAN Volume Controller nodes WWNN change procedure.

5. Remove the offline node from the SAN Volume Controller system configuration with the SAN Volume Controller CLI commands, as shown in Example 5-20.

Example 5-20 Remove node command

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode
id name                  UPS_serial_number WWNN             status  IO_group_id
IO_group_name  config_node UPS_unique_id   hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number
1  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F offline 0
ITSO_SVC_SPLIT no          2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online  0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523

IBM_2145:ITSO_SVC_SPLIT:superuser>rmnode ITSO_SVC_NODE1_SITE_A

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode
id name                  UPS_serial_number WWNN             status IO_group_id
IO_group_name  config_node UPS_unique_id   hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online 0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523
```

You also can use the GUI, as shown in Figure 5-20.

Figure 5-20 Remove node using GUI

6. Complete the following steps to remove the Volume Mirroring definitions:

   a. Identify which copy Id is offline for each volume by using the SAN Volume Controller CLI command, as shown in Example 5-21.

Example 5-21 Identifying Volume copy id

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk
id name                    IO_group_id IO_group_name  status   mdisk_grp_id mdisk_grp_name
capacity type    FC_id FC_name RC_id RC_name vdisk_UID                        fc_map_count
copy_count fast_write_state se_copy_count RC_change compressed_copy_count
0  ESXi-01-DCA-HBA1         0           ITSO_SVC_SPLIT offline  0            V7000SITEA
2.00GB   striped                              60050768018305 3EF800000000000000 0
1          not_empty        0             no        0
```

```
1  ESXi-01-DCA-HBA2         0           ITSO_SVC_SPLIT offline  0            V7000SITEA
2.00GB   striped                              600507680183053EF800000000000001 0
1          not_empty        0             no        0
2  ESXi-02-DCB-HBB1         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000002 0
1          empty            0             no        0
3  ESXi-02-DCB-HBB2         0           ITSO_SVC_SPLIT degraded 1            V7000SITEB
2.00GB   striped                              600507680183053EF800000000000003 0
1          empty            0             no        0
4  ESXI_CLUSTER_01          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000004 0
2          empty            0             no        0
5  ESXI_CLUSTER_02          0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000005 0
2          empty            0             no        0
6  ESXi_Cluster_DATASTORE_A 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000006 0
2          empty            0             no        0
7  ESXI_Cluster_DATASTORE_B 0           ITSO_SVC_SPLIT degraded many         many
256.00GB many                                 600507680183053EF800000000000007 0
2          empty            0             no        0

IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk ESXi_Cluster_DATASTORE_A
id 6
name ESXi_Cluster_DATASTORE_A
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
status degraded
mdisk_grp_id many
mdisk_grp_name many
capacity 256.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 600507680183053EF800000000000006
throttling 0
preferred_node_id 0
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 95
copy_count 2
se_copy_count 0
filesystem
mirror_write_priority redundancy
RC_change no
compressed_copy_count 0
access_IO_group_count 1

copy_id 0
status offline
sync no
primary yes
mdisk_grp_id 0
```

```
mdisk_grp_name V7000SITEA
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 256.00GB
real_capacity 256.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
se_copy no
easy_tier on
easy_tier_status active
tier generic_ssd
tier_capacity 15.75GB
tier generic_hdd
tier_capacity 240.25GB
compressed_copy no
uncompressed_used_capacity 256.00GB

copy_id 1
status online
sync yes
primary no
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 256.00GB
real_capacity 256.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
se_copy no
easy_tier on
easy_tier_status active
tier generic_ssd
tier_capacity 14.00GB
tier generic_hdd
tier_capacity 242.00GB
compressed_copy no
uncompressed_used_capacity 256.00GB
```

You also can use the GUI, as shown in Figure 5-21.

ITSO_SVC_SPLIT > Volumes > Volumes

New Volume | Actions

Name	Status	Capacity	Compression Savings	Storage Pool	UID	Host Mappings	Pr...
ESXi-01-DCA-HBA1	Offline	2.00 GB		V7000SITEA	600507680183053EF800000000000000	Yes	0
ESXi-01-DCA-HBA2	Offline	2.00 GB		V7000SITEA	600507680183053EF800000000000001	Yes	0
ESXi-02-DCB-HBB1	Degraded	2.00 GB		V7000SITEB	600507680183053EF800000000000002	Yes	2
ESXi-02-DCB-HBB2	Degraded	2.00 GB		V7000SITEB	600507680183053EF800000000000003	Yes	2
ESXI_CLUSTER_01	Degraded	256.00 GB		V7000SITEA	600507680183053EF800000000000004	Yes	0
Copy 0*	Offline	256.00 GB		V7000SITEA	600507680183053EF800000000000004	Yes	0
Copy 1	Online	256.00 GB		V7000SITEB	600507680183053EF800000000000004	Yes	0
ESXI_CLUSTER_02	Degraded	256.00 GB		V7000SITEB	600507680183053EF800000000000005	Yes	2
Copy 0*	Online	256.00 GB		V7000SITEB	600507680183053EF800000000000005	Yes	2
Copy 1	Offline	256.00 GB		V7000SITEA	600507680183053EF800000000000005	Yes	2
ESXi_Cluster_DATASTORE_A	Degraded	256.00 GB		V7000SITEA	600507680183053EF800000000000006	Yes	0
Copy 0*	Offline	256.00 GB		V7000SITEA	600507680183053EF800000000000006	Yes	0
Copy 1	Online	256.00 GB		V7000SITEB	600507680183053EF800000000000006	Yes	0
ESXI_Cluster_DATASTORE_B	Degraded	256.00 GB		V7000SITEB	600507680183053EF800000000000007	Yes	2
Copy 0*	Online	256.00 GB		V7000SITEB	600507680183053EF800000000000007	Yes	2
Copy 1	Offline	256.00 GB		V7000SITEA	600507680183053EF800000000000007	Yes	2

Figure 5-21 Identify the offline copy id with GUI

b. Remove each identified offline Volume Mirroring copy with the SAN Volume Controller CLI command, as shown in Example 5-22.

Example 5-22 rmvdiskcopy

```
IBM_2145:ITSO_SVC_SPLIT:superuser>rmvdiskcopy -copy 0 ESXi_Cluster_DATASTORE_A

IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk ESXi_Cluster_DATASTORE_A
id 6
name ESXi_Cluster_DATASTORE_A
IO_group_id 0
IO_group_name ITSO_SVC_SPLIT
status degraded
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
capacity 256.00GB
type striped
formatted no
mdisk_id
mdisk_name
FC_id
FC_name
RC_id
RC_name
vdisk_UID 600507680183053EF800000000000006
throttling 0
preferred_node_id 0
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 95
copy_count 1
se_copy_count 0
filesystem
mirror_write_priority redundancy
RC_change no
compressed_copy_count 0
access_IO_group_count 1

copy_id 1
```

```
status online
sync yes
primary yes
mdisk_grp_id 1
mdisk_grp_name V7000SITEB
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 256.00GB
real_capacity 256.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
se_copy no
easy_tier on
easy_tier_status active
tier generic_ssd
tier_capacity 14.00GB
tier generic_hdd
tier_capacity 242.00GB
compressed_copy no
```

You also can use the GUI as shown in Figure 5-22.

Figure 5-22 Delete copy with GUI

7. Power on the new node and leave the FC cable disconnected.
8. Change the new node WWNN by powering on the replacement node from the front panel with the Fibre Channel cables and the Ethernet cable disconnected.

   After the node is booted, you might receive error 540, “An Ethernet port has failed on the 2145” or error 558, “The 2145 cannot see the Fibre Channel fabric or the Fibre Channel card port speed might be set to a different speed than the Fibre Channel fabric”. This is to be expected as the node was booted with no fiber optic cables connected and no LAN connection.

   If you see Error 550, “Cannot form a cluster due to a lack of cluster resources”, this node still thinks it is part of an SAN Volume Controller clustered system. If this is a new node from IBM, this error should not occur.

Change the WWNN of the replacement node to match the WWNN that you recorded earlier by completing the following steps:

a. From the front panel of the new node, press the button until the “Node:” panel is displayed and then use the right or left navigation button to display the “Node WWNN:” panel.

   Press and hold the Down button, press and release the Select button, and then release the Down button. Online one should be “Edit WWNN:” and online two are the last five numbers of this new node’s WWNN.

b. Press and hold the Down button, press and release the Select button, and then release the Down button to enter WWNN edit mode. The first character of the WWNN is highlighted.

   Important: When changing the WWNN is changed, you might receive error 540, “An Ethernet port has failed on the 2145” or error 558, “The 2145 cannot see the FC fabric or the FC card port speed might be set to a different speed than the Fibre Channel fabric”. This is to be expected because the node was booted with no fiber optic cables connected and no LAN connection. However, if this error occurs while you are editing the WWNN, you are taken out of edit mode with partial changes saved. You must reenter edit mode by starting again at step b.

c. Press the Up or Down button to increment or decrement the character that is displayed.

   The characters wrap F - 0 or 0 - F.

d. Press the left navigation button to move to the next field or the right navigation button to return to the previous field and repeat step d for each field. At the end of this step, the characters that are displayed must be the same as the WWNN you recorded in step a.

e. Press the select button to retain the characters that you have updated and return to the WWNN panel.

f. Press the select button again to apply the characters as the new WWNN for the node.

   Important: You must press the select button twice as instructed in steps f and g. After step f, it might appear that the WWNN was changed, but it is in step g that the change is applied.

g. Ensure the WWNN has changed by following steps a, which gives the WWNN.

9. Connect the node to the same FC switch ports as it was before the critical event.

   This is the key point of the recovery procedure. Connecting the new SAN Volume Controller nodes to the same SAN ports and reusing the same SAN Volume Controller WWNN avoids the need to reboot, rediscover, or reconfigure, and the host continues unaffected.

   Important: Do not connect the new nodes to different ports at the switch or director because this causes port IDs to change, which might affect hosts’ access to volumes or cause problems with adding the new node back into the clustered system.

   If you cannot connect the SAN Volume Controller nodes to the same FC SAN ports as before, consider that you are forced to reboot, rediscover, or reconfigure your host to see the lost disk resources and bring the paths back online.

10. Issue the SAN Volume Controller CLI command as shown in Example 5-23 to verify that the last five characters of the WWNN are correct.

Example 5-23 Verify candidate node with correct WWNN

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsnodecandidate
id                panel_name UPS_serial_number UPS_unique_id    hardware
500507680100B13F 151580     100006B119        2040000006481049 CF8
```

Important: If the WWNN does not match the original node's WWNN exactly as recorded, you must repeat steps 8 b - 8 g.

11. Add the node to the clustered system and ensure it is added back to the same I/O group as the original node with the SAN Volume Controller CLI commands, as shown in Example 5-24.

Example 5-24 Adding node

```
IBM_2145:ITSO_SVC_SPLIT:superuser>addnode -wwnodename 500507680100B13F -iogrp 0
Node, id [3], successfully added

IBM_2145:ITSO_SVC_SPLIT:superuser>lsnode
id name                  UPS_serial_number WWNN             status IO_group_id
IO_group_name  config_node UPS_unique_id   hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number
3  ITSO_SVC_NODE1_SITE_A 100006B119        500507680100B13F online 0
ITSO_SVC_SPLIT no          2040000006481049 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode1sitea             151580
2  ITSO_SVC_NODE2_SITE_B 100006B074        500507680100B0C6 online 0
ITSO_SVC_SPLIT yes         20400000064801C4 CF8
iqn.1986-03.com.ibm:2145.itsosvcsplit.itsosvcnode2siteb             151523
IBM_2145:ITSO_SVC_SPLIT:superuser>
```

You also can use the GUI, as shown in Figure 5-23.

Figure 5-23 Adding node using GUI

12. Verify that all volumes for this I/O group are back online and are not degraded. If the node replacement process is done disruptively, such that no I/O is occurring to the I/O group, you still must wait some time (we recommend 30 minutes) to make sure the new node is back online and available to take over before you do the next node in the I/O group.

    Use the SAN Volume Controller CLI command that is shown in Example 5-25 to verify that all volumes for this I/O group are back online and are not degraded.

    Example 5-25 No longer degraded Volume

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk -filtervalue status=degraded
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisk -filtervalue status=offline
```

13. Because we are now using the CLI, we use that to discover the new MDisk that is supplied by new back-end storage subsystem. They appear as status online with a mode of unmanaged, as shown in Example 5-26.

    Example 5-26 New MDisk discovered

```
IBM_2145:ITSO_SVC_SPLIT:superuser>detectmdisk

IBM_2145:ITSO_SVC_SPLIT:superuser>lsmdisk
id name                status         mode    mdisk_grp_id mdisk_grp_name capacity
ctrl_LUN_#      controller_name    UID
tier
0  ITSO_V7K_SITEA_SSD0 online unmanaged 0            V7000SITEA     277.3GB
0000000000000000 6005076802890002680000000000000800000000000000000000000000000000
generic_ssd
1  ITSO_V7K_SITEA_SAS0 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000A 6005076802890002680000000000000400000000000000000000000000000000
generic_hdd
2  ITSO_V7K_SITEA_SAS1 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000B 6005076802890002680000000000000500000000000000000000000000000000
generic_hdd
3  ITSO_V7K_SITEA_SAS2 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000C 6005076802890002680000000000000600000000000000000000000000000000
generic_hdd
4  ITSO_V7K_SITEA_SAS3 online         unmanaged 0            V7000SITEA     500.0GB
000000000000000D 6005076802890002680000000000000700000000000000000000000000000000
generic_hdd
```

14. Add the MDisks to the storage pool by using the SAN Volume Controller CLI commands as shown in Example 5-27 and re-create the MDisk to storage pool relationship that existed before the critical event.

    Important: You should remove the previous MDisk (or MDisks) that are still defined in each storage pool but do not physically exist after the critical event (they might appear in an offline or degraded state to the SAN Volume Controller), before you add the newly discovered MDisk (or MDisks).

    Example 5-27 Adding new MDisk to Storage pool

```
IBM_2145:ITSO_SVC_SPLIT:superuser>addmdisk -mdisk ITSO_V7K_SITEA_SAS0 V7000SITEA
```

    Important: After you have added your newly discovered MDisks to the storage pool, the three quorum tandem automatically is fixed. We can check this with the SAN Volume Controller CLI command, as shown in Example 5-28.

15. Check the Quorum Disk status as shown in Example 5-28.

Example 5-28 Quorum status

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsquorum
quorum_index status id name                controller_id controller_name     active
object_type override
0            online 10 ITSO_V7K_SITEC_Q    5             ITSO_V7K_SITEC_Q_N2 yes
mdisk       yes
1            online 6  ITSO_V7K_SITEB_SAS0 0             ITSO_V7K_SITEB_N2   no
mdisk       yes
2            online 4  ITSO_V7K_SITEA_SAS3 1             ITSO_V7K_SITEA_N2   no
mdisk       yes
```

16. Reactivate Volume Mirroring for each volume in accordance with your Volume Mirroring requirements to re-create the same business continuity infrastructure as before the critical event by using the SAN Volume Controller CLI command, as shown in Example 5-29.

Example 5-29 addvdiskcopy example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>addvdiskcopy -mdiskgrp V7000SITEA
ESXi_Cluster_DATASTORE_A
Vdisk [6] copy [0] successfully created
```

17. Check the status of your Volume Mirroring synchronization progress with the SAN Volume Controller CLI command, as shown in Example 5-30.

Example 5-30 vdisksyncprogress example

```
IBM_2145:ITSO_SVC_SPLIT:superuser>lsvdisksyncprogress
```

It is possible to speed up the synchronization progress with the **chvdisk** command, but the more speed you give to the synchronization process, the more effect on the overall performance you might have.

18. With the following SAN Volume Controller CLI command that is shown in Example 5-31, you might consider rebalancing your Split I/O Group configuration to have the Volume Mirroring Primary copy related with the storage pool and preferred node as it was before the critical event, even if they are now in the same site. Doing that helps you with an eventual future stretch of your configuration when a new remote site becomes available. You can do that by using SAN Volume Controller CLI command, as shown in Example 5-31.

Example 5-31 Change Volume primary copy ID

```
IBM_2145:ITSO_SVC_SPLIT:superuser>chvdisk -primary 0 ESXi_Cluster_DATASTORE_A
```

You now recovered a Stretched Cluster configuration after a critical event.

At this point, all of your volumes are accessible from your hosts point of view and the recovery action has not affected your applications.

Some operations were run by using the CLI, CLI and GUI, or just the GUI to show the different possibilities you have.

5.7 PowerVM diagnosis and recovery guidelines

In this section, we provide a brief overview of some of the guidelines that are specific to PowerVM from a problem diagnosis and recovery perspective.

5.7.1 PowerVM diagnosis guidelines

A complete disaster recovery strategy and prerequisite to problem determination for the Virtual I/O Server should include backing up the following components such that you can recover the virtual devices and their physical backing devices. When to back up the Virtual I/O Server, if necessary, is followed by our server backup strategy where we rebuild the AIX operating system-based logical partitions. Also, `mksysb` or `savevg` can be used. The following components make up a virtualized environment; a change to any one of these requires a new backup:

- External device configuration; for example, SAN and storage subsystem configuration.
- Memory, CPU, virtual, and physical devices.
- The Virtual I/O Server operating system.
- User-defined virtual devices that couple the virtual and physical environments. This can be considered virtual device mappings, or metadata.

For more information, see section 2.1.2 of *IBM System p Advanced POWER Virtualization (PowerVM) Best Practices*, REDP-4194, which can be found at this website:

http://www.redbooks.ibm.com/abstracts/redp4194.html

5.7.2 PowerVM recovery guidelines

For more information about recovery guidelines, see section 2.1.3 of *IBM System p Advanced POWER Virtualization (PowerVM) Best Practices*, REDP-4194, which is available at this website:

http://www.redbooks.ibm.com/abstracts/redp4194.html

5.8 PowerHA SystemMirror diagnosis and recovery guidelines

In this section, we provide a brief overview of some of the guidelines that are specific to PowerHA

5.8.1 PowerHA SystemMirror diagnosis guidelines

A periodic backup of nodes should be taken by using `mksysb` or `savevg`. A cluster snapshot also can be taken.

5.8.2 PowerHA SystemMirror recovery guidelines

NIM can be used to restore `mksysb` and `restorevg` for `savevg`. Cluster snapshot also can be restored by using PowerHA SystemMirror smit menus.

Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide more information about the topic in this book. Note that some publications that are referenced in this list might be available in softcopy only:

- *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933
- *Implementing the IBM Storwize V7000 V6.3*, SG24-7938
- *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859
- *SAN Volume Controller: Best Practices and Performance Guidelines*, SG24-7521
- *Implementing an IBM/Brocade SAN with 8 Gbps Directors and Switches*, SG24-6116
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *High Availability and Disaster Recovery Planning: Next-Generation Solutions for Multiserver IBM Power Systems Environments*, REDP-4669
- *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859
- *Configuring SDDPCM for High Availability*, REDP-4928
- *IBM System p5 Approaches to 24x7 Availability Including AIX 5L*, SG24-7196
- *Virtualization and Clustering Best Practices Using IBM System p Servers*, SG24-7349
- *High Availability and Disaster Recovery Options for DB2 for Linux, UNIX, and Windows*, SG24-7363
- *IBM PowerVM Live Partition Mobility*, SG24-7460
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *Exploiting IBM PowerHA SystemMirror Enterprise Edition*, SG24-7841
- *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030
- *IBM PowerVM Best Practices*, SG24-8062
- *IBM SAN and SVC Split Cluster and VMware Solution Implementation*, SG24-8072
- *IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX*, SG24-8106
- *Simple Configuration Example for Storwize V7000 FlashCopy and PowerHA SystemMirror for i*, REDP-4923
- *PowerHA for AIX Cookbook*, SG24-7739

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft, and other materials at the following website:

http://ibm.com/redbooks

© Copyright IBM Corp. 2013. All rights reserved.

VMware online resource

The following VMware resource is available:

http://www.vmware.com/support/pubs/

Other publications

The following publications also are relevant as further information sources:

- *IBM System Storage Open Software Family SAN Volume Controller: Planning Guide*, GA22-1052
- *IBM System Storage Master Console: Installation and User's Guide*, GC30-4090
- *IBM System Storage Open Software Family SAN Volume Controller: Installation Guide*, SC26-7541
- *IBM System Storage Open Software Family SAN Volume Controller: Service Guide*, SC26-7542
- *IBM System Storage Open Software Family SAN Volume Controller: Configuration Guide*, SC26-7543
- *IBM System Storage Open Software Family SAN Volume Controller: Command-Line Interface User's Guide*, SC26-7544
- *IBM System Storage Open Software Family SAN Volume Controller: CIM Agent Developers Reference*, SC26-7545
- *IBM TotalStorage Multipath Subsystem Device Driver User's Guide*, SC30-4096
- *IBM System Storage Open Software Family SAN Volume Controller: Host Attachment Guide*, SC26-7563

Web sites

The following web sites also are relevant as further information sources:

- IBM System Storage home page:

  http://www.storage.ibm.com
- SAN Volume Controller supported platform:

  http://www-1.ibm.com/servers/storage/support/software/sanvc/index.html
- Download site for Windows SSH freeware:

  http://www.chiark.greenend.org.uk/~sgtatham/putty
- IBM site to download SSH for AIX:

  http://oss.software.ibm.com/developerworks/projects/openssh
- Open source site for SSH for Windows and Mac:

  http://www.openssh.com/windows.html
- Cygwin Linux-like environment for Windows:

  http://www.cygwin.com

- IBM Tivoli Storage Area Network Manager site:

  http://www-306.ibm.com/software/sysmgmt/products/support/IBMTivoliStorageAreaNetworkManager.html
- Microsoft Knowledge Base Article 131658:

  http://support.microsoft.com/support/kb/articles/Q131/6/58.asp
- Microsoft Knowledge Base Article 149927:

  http://support.microsoft.com/support/kb/articles/Q149/9/27.asp
- Sysinternals home page:

  http://www.sysinternals.com
- Subsystem Device Driver download site:

  http://www-1.ibm.com/servers/storage/support/software/sdd/index.html
- IBM TotalStorage Virtualization home page:

  http://www-1.ibm.com/servers/storage/software/virtualization/index.html

Help from IBM

IBM Support and downloads:

http://ibm.com/support

IBM Global Services:

http://ibm.com/services

IBM SAN Volume Controller Stretched Cluster with PowerVM and PowerHA

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

IBM SAN Volume Controller Stretched Cluster with PowerVM and PowerHA

Learn about the IBM Stretched Cluster solution and architecture

See how to implement it with Power Systems

Learn about diagnosis and recovery guidelines

This IBM Redbooks publication describes the IBM Storage Area Network and IBM SAN Volume Controller Stretched Cluster solution when combined with PowerVM and PowerHA. We describe guidelines, settings, and the implementation steps that are necessary to achieve a successful implementation.

This book is for administrators who are familiar with the SAN, IBM SAN Volume Controller, and IBM PowerVM and PowerHA Systems.

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

SG24-8142-00 ISBN 0738438502