IBM

Fault Tolerant Storage

Multipathing and Clustering Solutions for Open Systems for the IBM ESS

All you need to know about multipathing software and more!

Integrate multipathing with clustering

Improve your storage availability

Barry Mellish
Bernd Albrecht
Omar Hajime Fidelis
Maciej Struzinski

Redbooks

ibm.com/redbooks

International Technical Support Organization

Fault Tolerant Storage Multipathing and Clustering Solutions for Open Systems for the IBM ESS

April 2002

SG24-6295-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 231. All detailed settings are on an “as is” basis and you are advised to check the IBM Storage websites for the latest information.

First Edition (April 2002)

This edition applies to the Enterprise Storage Server model F.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Contents . . . iii

Figures . . . ix

Tables . . . xi

Preface . . . xiii
The team that wrote this redbook. . . . xiii
Special notice. . . . xiv
IBM trademarks . . . xiv
Comments welcome. . . . xv

Chapter 1. Introduction. . . . 1
1.1 Introduction to the IBM Enterprise Storage System . . . 2
1.1.1 The Seascape architecture . . . 2
1.1.2 Enterprise Storage System overview . . . 2
1.1.3 ESS Copy Services components . . . 3
1.2 Introduction to high availability concepts. . . . 4
1.2.1 Multipathing. . . . 4
1.2.2 Mirroring . . . 7
1.2.3 Multipathing versus mirroring . . . 8
1.2.4 Clustering . . . 8

Chapter 2. Basic concepts of ESS with multipathing and clustering . . . 11
2.1 Concepts of high availability . . . 12
2.1.1 What is high availability?. . . . 12
2.1.2 Single system availability . . . 13
2.1.3 Increasing availability by clustering. . . . 13
2.2 How disks are seen on ESS . . . 15
2.2.1 Overview . . . 15
2.2.2 SCSI . . . 16
2.2.3 Fibre Channel . . . 17
2.2.4 Fibre Channel Storage Area Networks (SANs). . . . 19
2.3 Managing the number of paths to a LUN . . . 19
2.3.1 How many paths are seen . . . 19
2.3.2 How to reduce paths . . . 21
2.3.3 Paths over inter switch links . . . 23

Chapter 3. Multipathing software. . . . 25
3.1 IBM Subsystem Device Driver . . . 26
3.1.1 Path algorithms . . . 27
3.1.2 How many paths to use with SDD. . . . 31
3.1.3 Usage of datapath command . . . 31
3.2 HP-UX LVM built-in support for Physical Volume Links . . . 34
3.3 Veritas VxVM built-in Dynamic MultiPathing software (DMP) . . . 40
3.3.1 Supported disk arrays . . . 43
3.3.2 Path failover and load balancing. . . . 43

Chapter 4. Clustering concepts . . . 45
4.1 What is a cluster? . . . 46

4.2 The benefits of clustering 46
4.2.1 Why consider a cluster? 47
4.2.2 What is high availability? 48
4.2.3 Server consolidation 50
4.3 Types of clusters 50
4.3.1 Software for clusters 50
4.3.2 Hardware for clusters 51
4.3.3 Active and passive servers 52
4.4 IBM cluster strategy 53
4.5 Linux clustering 53
4.5.1 Implementing Linux clustering. 55
4.5.2 Failover service 56
4.5.3 Load balancing 58
4.5.4 Supported services 63
4.5.5 Sharing the data between nodes 63
4.5.6 Putting it all together 67
4.6 RS/6000 Cluster Technology (RSCT) overview 68
4.6.1 Topology Services (TS). 69
4.6.2 Group Services (GS). 71
4.6.3 Event Management (EM) 72

Chapter 5. SDD installation and configuration on AIX 75
5.1 Pre-installation checks 76
5.2 Hardware and software requirements 77
5.2.1 SCSI requirements 78
5.2.2 Fibre Channel requirements 78
5.2.3 Non supported environments 79
5.3 Installing and configuring Fibre Channel device drivers 79
5.3.1 Installing Fibre Channel device drivers 80
5.3.2 Gigabit Fibre Channel Adapter for PCI bus FC 6227 (type 4-S) features 82
5.3.3 2-Gigabit Fibre Channel Adapter for PCI bus FC 6228 (type 4-W) features 84
5.3.4 Problem determination 86
5.4 Installing and configuring the IBM Subsystem Device Driver 91
5.4.1 Installing the IBM Subsystem Device Driver 92
5.4.2 Configuring the Subsystem Device Driver 95
5.4.3 Unconfiguring the Subsystem Device Driver devices 100
5.4.4 Removing the Subsystem Device Driver. 100
5.5 Using IBM SDD on AIX host 101
5.5.1 Providing load-balancing and failover protection 101
5.5.2 Using ESS devices directly 111
5.5.3 Using ESS devices through AIX LVM 111
5.5.4 Migrating non-SDD volume group to ESS SDD multipath volume group 112
5.5.5 SDD utility programs 114
5.5.6 SDD error log messages. 115
5.6 How to use HACMP with SDD 116
5.6.1 Understanding the SDD support for HACMP/6000 117
5.6.2 What's new in SDD for HACMP/6000 119
5.6.3 Special requirements for HACMP/6000 120
5.6.4 Models of the ESS supported in HACMP/6000 environment 122
5.7 Upgrading SDD for AIX 4.2.1, AIX 4.3.2 and AIX 4.3.3 122
5.7.1 Upgrading to SDD 1.3.0.x through a non-disruptive installation 123
5.7.2 Upgrading to SDD 1.3.0.x through a disruptive installation. 123
5.8 Using concurrent download of licensed internal code 124

Chapter 6. DMP installation and configuration on Sun servers 125
6.1 Concepts of multipathing and clustering on Sun platform 126
6.2 Hardware and software requirements 126
6.2.1 Hardware 126
6.2.2 Software 126
6.2.3 Non supported environments 126
6.3 Preparing the environment 127
6.3.1 Installing the operating system 127
6.3.2 Configuring all HBAs 127
6.3.3 Emulex 128
6.3.4 JNI 129
6.3.5 Setting up the ESS 130
6.3.6 Attaching an ESS to a Sun Solaris Enterprise Servers 130
6.4 Using the SDD 130
6.4.1 When to use SDD 130
6.4.2 Installing the Subsystem Device Driver 131
6.4.3 Uninstalling the Subsystem Device Driver 133
6.5 Installing and configuring Veritas Volume Manager 134
6.5.1 Creating a filesystem under Veritas 134
6.5.2 Sun Veritas and ESS logical volumes 144
6.5.3 ESS identification under Veritas 144
6.6 Installing and configuring cluster software 145

Chapter 7. SDD installation and configuration on HP-UX servers 147
7.1 Concepts of multipathing and clustering on HP-UX platform 148
7.2 Hardware and software requirements 148
7.2.1 Hardware 148
7.2.2 Software 148
7.2.3 Non supported environments 148
7.3 Preparing the environment 149
7.3.1 Pre-installation planning 149
7.3.2 Installing the operating system 149
7.3.3 Installing and confirming HP-UX installation patches 150
7.3.4 Confirming correct installation of the host bus adapter 150
7.3.5 Setting up the ESS 151
7.3.6 Attaching an ESS to a Hewlett Packard 9000 151
7.3.7 Confirming storage connectivity 152
7.3.8 Tuning recommendations 153
7.3.9 Installing and configuring cluster software 154
7.4 Installing and uninstalling the SDD 154
7.4.1 What is PV-Link? 154
7.4.2 How to use PV-Links 155
7.4.3 PV-Link versus SDD 155
7.4.4 Installing SDD 156
7.4.5 Uninstalling SDD 161

Chapter 8. Installing Fibre Channel and configuring multipathing on SGI IRIX 6.5 . 167
8.1 Supported configurations 168
8.2 Installation of Fibre Channel on SGI IRIX 6.5 168
8.2.1 Confirming IRIX installation 168
8.2.2 Confirming QLA2200F installation 168
8.2.3 Setting up ESS 170
8.2.4 Installing optical cable 171
8.2.5 Confirming switch connectivity 171

8.2.6 Confirming storage connectivity 173
8.3 Configuring host path failover 177
8.3.1 General considerations 177
8.3.2 Switched fabric 178
8.3.3 Fibre Channel Arbitrated Loop 178
8.3.4 Manually switch IO between the primary and alternate path 179
8.4 Working with ESS volumes 179
8.4.1 Configuring storage. 179
8.4.2 Important SGI disk devices naming convention 182
8.4.3 Tuning recommendations 182
8.4.4 Unsupported utilities 183
8.4.5 Useful SGI information 183

Chapter 9. Installing and configuring IBM SDD on Windows 2000 185
9.1 Pre-installation 186
9.2 Hardware and software requirements. 187
9.2.1 SCSI requirements 188
9.2.2 Fibre Channel requirements 188
9.2.3 Non supported environments 188
9.3 Connecting and configuring SCSI adapters 189
9.4 Connecting and configuring Fibre Channel adapters 192
9.4.1 Fibre Channel adapters, cables and node-to-node distances. 193
9.4.2 LUN affinity, targets and LUNs 194
9.4.3 Installing and configuring the QLogic QLA2100F adapter card 194
9.4.4 Installing and configuring the QLogic QLA2200F adapter card 195
9.4.5 Installing and configuring the Emulex LP8000 adapter card. 197
9.5 A step for Installing Windows 2000 on remote ESS disks 202
9.6 Availability and recoverability for Windows2000 203
9.7 Installing IBM SDD on Windows 2000 host. 204
9.8 Uninstalling or upgrading IBM SDD 207
9.9 Displaying the current version of the SDD 207
9.10 Managing and troubleshooting IBM SDD on Windows 2000 208
9.10.1 Use of datapath query adapter command. 208
9.10.2 Use of datapath query adaptstats command 210
9.10.3 Use of datapath query device command. 210
9.10.4 Use of datapath query devstats command 212
9.10.5 Use of datapath set adapter command. 214
9.10.6 Use of datapath set device command. 215
9.10.7 Adding additional paths to Subsystem Device Driver devices 215
9.11 Using ESS with Veritas VxVM and DMP. 216
9.11.1 Installing Veritas VxVm 216
9.11.2 Using Veritas Volume Manager and configuring DMP 218

Appendix A. Installing adapters on a Windows 2000 host 221
Installing and configuring Adaptec AHA-2944UW adapter card 222
Installing and configuring Symbios 8751D adapter card 223
Installing and configuring the QLogic adapter card 224

Related publications 227
IBM Redbooks 227
Other resources 227
Referenced Web sites 228
How to get IBM Redbooks 229
IBM Redbooks collections. 230

Special notices 231

Index 233

Figures

1-1 Multipathing overview 5
1-2 Operating system sees three instances of the same volume 6
1-3 Consolidated view of an volume 7
2-1 Single server multipathing 13
2-2 Local cluster basics 14
2-3 Stretched clusters 15
2-4 How LUNs are seen. 16
2-5 SCSI connections 20
2-6 Fiber Channel connections 21
3-1 Example of configuration supported by IBM SDD 27
3-2 Example of path status and corresponding adapter status 28
3-3 SDD path failover algorithm. 30
3-4 PV-Links path failover algorithm 35
3-5 Example scenario for PV-Links 36
3-6 Multiple paths connections with PV-Links and two zones configured on the switch . 37
3-7 Multiple paths connections with PV-Links and one zone configured on the switch . . 39
3-8 Example of primary PV-Link change 40
3-9 VxVM concepts 42
4-1 Scalable clusters versus SMP machines. 47
4-2 Failover service in normal operation 56
4-3 Failover service showing actual failover operation 57
4-4 Failover service, resumption of normal operation 57
4-5 Simple Linux Virtual Server setup 58
4-6 Direct routing of returned packets 59
4-7 More sophisticated setup using an internal cluster network 60
4-8 Network address translation 61
4-9 IP tunneling 62
4-10 Using rsync for local server data synchronization 64
4-11 Using NFS for central data storing. 64
4-12 Global FileSystem 65
4-13 Sample Intermezzo setup 66
4-14 Front end clusters example 67
4-15 RSCT infrastructure. 68
4-16 Topology Services process flow 70
4-17 Group Services structure 71
4-18 Event Management functional flow 72
5-1 Where IBM SDD fits in the protocol stack on AIX 77
5-2 Example of device driver software selection on AIX platform 81
5-3 Layout of Fibre Channel FC 6227 (type 4-S) adapter 83
5-4 Layout of Fibre Channel FC 6228 (type 4-W) adapter. 85
5-5 How to uncompress Fibre Channel adapter firmware file 90
5-6 SDD installation on AIX - installp command status 93
5-7 SDD installation on AIX - installp verification. 94
6-1 Volume Manager Storage Administrator 135
6-2 VMSA controller view. 136
6-3 VMSA disk properties 136
6-4 VMSA right-click on rootdg to get Disk Group menu 137
6-5 VMSA New Volume view 138

6-6 VMSA Assign Disks view 139
6-7 VMSA New Volume with disk information 140
6-8 VMSA Add File System view 141
6-9 VMSA Mount Details view 142
6-10 VMSA New Volume with disk and filesystem information 142
6-11 VMSA filesystem/volume creation in progress. 143
6-12 VMSA filesystem/volume creation complete 143
6-13 View of /etc/vfstab entries 144
6-14 VMSA Enclosure view 145
7-1 Installing SDD window 156
7-2 Uninstalling SDD window. 166
9-1 Where the IBM SDD fits in the protocol stack on Windows 2000 187
9-2 Examples of ESS SCSI host interconnections. 190
9-3 Example of multiple SCSI connections 191
9-4 Example of initial window for Emulex configuration tool 200
9-5 Recommended configuration of Emulex LP8000 adapter 201
9-6 Adding WWPN to SCSI ID mapping - Step 1. 201
9-7 Adding WWPN to SCSI ID mapping - Step 2. 201
9-8 Reviewing or adding Fibre Channel LUN to SCSI LUN mapping 202
9-9 Example of list of LUN mapping. 202
9-10 SDD installation - user information 204
9-11 SDD installation - Choose Destination Location 205
9-12 SDD installation - setup type 206
9-13 Disk devices status with SDD installed 206
9-14 Usage of IBM SDD datapath command. 208
9-15 Example of datapath query adapter command output. 209
9-16 Example of degraded adapter 209
9-17 Example of datapath query adaptstats command output. 210
9-18 Example of datapath query device status 211
9-19 Example of datapath query devstats command output 213
9-20 Example of datapath set adapter command 215
9-21 Veritas VxVM personalization and licensing window. 217
9-22 Veritas VxVM features to install. 217
9-23 Multiplied disk instances with DMP installed and not enabled. 218
9-24 Enabling DMP with VxVM - Step 1 219
9-25 Enabling DMP with VxVM - Step 2 220
9-26 Proper configuration of DMP 220

Tables

1-1 Performance bottlenecks where multipathing can help 5
2-1 Host system limitations 16
3-1 Successful I/O operations before attempting to reopen the path. 28
4-1 System availability classification 49
4-2 Downtime costs by application. 49
5-1 List of fixes required for AIX. 78
5-2 Specifications for Fibre Channel adapter FC 6227 83
5-3 Use of FC 6227 adapter LEDs. 83
5-4 Specifications for Fibre Channel adapter FC 6228 85
5-5 Use of FC 6228 adapter LEDs. 85
5-6 SDD installation package filesets. 91
5-7 List of major files installed with IBM SDD 94
5-8 List of all SDD specific SMIT screens 110
5-9 Description of most important utility programs. 114
5-10 List of error messages logged by SDD 115
5-11 IBM Subsystem Device Driver support for HACMP/6000 117
5-12 Supported and unsupported SDD features in HACMP/6000 environment 119
5-13 Models of the ESS supported in HACMP/6000 environment. 122
8-1 Useful IRIX commands 183
8-2 SGI information on the Web. 184
9-1 Fibre Channel cables for the ESS 193
9-2 Recommended settings for Emulex LP8000 adapter 198

Preface

This IBM Redbook explains how to maximize your business benefit with multipathing and the clustering of hosts that are attached to the IBM Enterprise Storage Server (ESS). The ESS is a highly available storage subsystem and can form the basis of a fault tolerant storage subsystem. The role of the Subsystem Device Driver and other multipathing software is explained with details of how to install and configure these systems.

This redbook answers some important questions:

- Are your connections to the ESS reliable enough?
- Did you eliminate all single points of failure in your environment?
- Are some of your connectivity channels to the ESS overloaded, while others are idle?
- Do you need to improve your data paths, but you don't know how to do it?

You will find how disks are seen in a multiple path environment and how they are treated by the operating system. You can learn how to load-balance your channels and establish multiple paths to a single disk, while still maintaining data consistency on this disk. You'll discover all of this using the ESS storage server, on many operating systems, including IBM AIX, Microsoft Windows 2000, HP-UX, Sun Solaris and others.

This book is the result of a seven week project at the ITSO in San Jose during September and October of 2001. It reflects the work that was carried out by the team and was correct at the time of writing. The work has not been submitted for formal testing and all results and settings are on an "as is" basis. Product specifications and microcode levels are continually changing and we advise you to consult the latest information available at IBM Web sites, prior to carrying out work on your systems.

The team that wrote this redbook

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

Barry Mellish is a Project Leader at the International Technical Support Organization, San Jose Center. He has coauthored eight previous redbooks and has taught many classes on storage subsystems. He joined IBM UK 18 years ago, and before joining the ITSO, he worked as a Senior Storage Specialist on the Disk Expert team in EMEA.

Bernd Albrecht is a Systems Engineer in Germany. He has nine years of experience in UNIX and Storage. He holds a master's degree in Computer Science from the Technical University of Dresden. His areas of expertise include ESS, SAN, SSA, AIX and network performance. He has written extensively for two redbook projects.

Omar Hajime Fidelis is a Technical Support Specialist in the Support Distributed Center of IBM Global Services in São Paulo, Brazil. He has two years of experience with UNIX systems, and he joined IBM in 2000. Omar is working at the outsourcing group where he interacts with the Database and SAP groups. His areas of expertise include UNIX platforms and Windows, PSSP, TSM, HACMP and storage systems.

Maciej Struzinski is a Computer Systems Architect at the Technical Department of PROKOM Software SA in Poland. He has seven years of experience in designing and managing computer hardware systems and has been involved in many of PROKOM's projects. His primary areas of expertise focuses on IBM RS/6000, pSeries and AIX solutions. He holds a master's degree in Computer Science from the Technical University of Gdansk, and he is an IBM Certified Advanced Technical Expert for RS/6000 and AIX. PROKOM Software SA is the largest IT solution provider and software house in Poland, as well as an IBM Advanced Business Partner. PROKOM focuses on industry, financial and insurance customers. This is Maciej's first redbook project.

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

Jack Flynn, Richard Heffel, Timothy C Pepper, Srinivasulu Erva, Jean-luc Degrenand, Franck Excoffier, Alejandro B Halili, Donald Herrick, Mike Janini, Jeffry Larson, Robert Moon, Dominick Nguyen, Victoria Perris, Arnel R Rallet, Richard E Ravich, John Tolnay, Rainer Wolafka, Vijayavenkatesh Yelanji, Brian J Smith, Dick Johnson
IBM San Jose

Glauco Jose Pinheiro
IBM Brazil

Special notice

This publication is intended to help storage administrators and system administrators make effective use of their ESS storage subsystems and to obtain optimal performance and utilization of their storage and SAN. It also helps in the positioning of the various multipathing techniques so that correct choices can be made from all of the options. The information in this publication is not intended as the specification of any programming interfaces. See the PUBLICATIONS section of the IBM Programming Announcement for IBM ESS for more information about what publications are considered to be product documentation.

IBM trademarks

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

e (logo)®
IBM ®
AIX®
AS/400®
DB2®
DYNIX®
DYNIX/ptx®
Enterprise Storage Server™
ESCON®
FICON™
FlashCopy™
HACMP/6000™
IBM®
Informix™
iSeries™
Lotus®
Lotus Notes®
Netfinity®
Notes®
Redbooks™
Redbooks (logo)™
NUMA-Q®
OS/390®
Perform™
pSeries™
RETAIN®
RS/6000®
SAA®
S/390®
Seascape®
SP™
SP2®
StorWatch™
Tivoli®
TotalStorage™
xSeries™
zSeries™

Comments welcome

Your comments are important to us!

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

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

  `ibm.com/redbooks`

- Send your comments in an Internet note to:

  `redbook@us.ibm.com`

- Mail your comments to the address on page ii.

1

Introduction

These complex systems are by nature made up of multiple parts, and in a coordinated fashion, each takes advantage of the others' strengths to be successful. In this part of the book we give an overview of the major parts and concepts of high availability with the Enterprise Storage Subsystem (ESS) using multipathing and clustering.

1.1 Introduction to the IBM Enterprise Storage System

The first of IBM's Enterprise Storage Systems (ESS) were introduced in 1999 to meet the need for high performance, scalable, flexible, and available storage systems with advanced management capabilities. The most recent product in IBM's Seascape architecture family, the ESS, sometimes referred to as Shark, is a SAN-ready disk storage system providing universal access across all major server platforms. It employs advanced hardware and software technologies to deliver breakthrough levels of performance and maximize data sharing across the enterprise.

An enhancement to the ESS, ESS Copy Services, provides three varieties of replication of mission critical data:

- **FlashCopy** delivers near-instantaneous, non disruptive, point-in-time copies within the ESS.
- **Peer-to-Peer Remote Copy** (PPRC) implements dynamic synchronous mirroring to a remote ESS.
- Using **Extended Remote Copy** (XRC), asynchronous copying to a remote site is possible in IBM @server zSeries environments.

1.1.1 The Seascape architecture

Seascape is a blueprint for comprehensive storage solutions optimized for a connected world. The Seascape architecture integrates leading technologies from IBM — including disk storage, tape storage, optical storage, powerful processors, and rich software function — to provide highly reliable, scalable, versatile, application-based storage solutions that span the range of servers from PCs to supercomputers.

At its heart, Seascape architecture uses an open, industry-standard, storage server that can scale up exponentially in both power and performance. Since the storage servers can integrate snap-in building blocks and software upgrades, you can quickly deploy new or improved applications and rapidly accommodate new data and media types. In this way, Seascape storage servers become an essential element for storing, manipulating, and sharing data across the network.

1.1.2 Enterprise Storage System overview

The IBM Enterprise Storage System can be configured in a variety of ways to provide scalability in capacity and performance. It employs integrated data caching (both volatile and non-volatile), hardware-level RAID5 support, and redundant systems at all levels. It provides easily configurable shared or secure access to user-variable quantities of storage via SCSI, Fibre Channel, ESCON or FICON I/O interfaces.

Scalability—definition: Scalability is the ability of a computer application or product (hardware or software) to continue to function well as it (or its context) is changed in size or volume in order to meet a user need. And it is the ability not only to function well in the rescaled situation, but to actually take full advantage of it.

A single ESS unit can house up to 11 TB of usable protected storage. Up to 32 GB of cache can be installed in 8 GB increments.

ESS RAID5 storage is configurable by the customer via a convenient Web interface. It can be subdivided into logical disks which are then automatically configured to emulate disk device types that are compatible with the attached host computer systems. Because RAID5 storage is always striped in Enterprise Storage Servers, the I/O benefits of multiple active spindles and parallelism are immediately available; and RAID5 means all data is protected against device failure.

Multiple data paths, both internally and externally, can be specified to each logical disk created on an Enterprise Storage System. This multiplicity both enhances availability and increases I/O bandwidth. It also means that multiple host systems can be attached to either the same (shared) device, or to non shared devices.

Availability—definition: Availability is the ratio between the time during which the system is operational and elapsed time.

The combination of redundancy and hot-swappable components in the ESS translates to continuous availability. Whether it be a line cord or a critical data-filled disk, component failure is automatically handled without interruption of service. Enterprise Storage Systems have phone-home and error notification capabilities built-in.

Additional Web-based performance monitoring software is available for Enterprise Storage Systems; and a variety of standard and custom reports can be specified and scheduled for automated data collection.

1.1.3 ESS Copy Services components

Copy Services is a separately sold feature of the Enterprise Storage Server. It brings powerful data copying and mirroring technologies to Open Systems environments previously available only for mainframe storage. ESS Copy Services has two components: FlashCopy, and Peer-to-Peer Remote Copy (PPRC). Both help to off load backup tasks from your host systems.

Peer-to-Peer Remote Copy (PPRC) and FlashCopy are typically used as data backup tools for creation of test data and for data migration. They can also be used in disaster recovery scenarios.

Peer-to-Peer Remote Copy is a synchronous protocol that allows real-time mirroring of data from one ESS to another. In a disaster recovery scenario, this secondary ESS could be located at another site several kilometers away. PPRC is application independent. Because the copying function occurs at the disk subsystem level, the application has no knowledge of its existence. No host system processor is involved.

Copy Services provides both a Command Line Interface (CLI) and a Web-based interface for setting up and managing its facilities. The CLI allows administrators to execute Java-based Copy Services commands from a command line. The Web-based interface, a part of ESS Specialist, allows storage administrators to manage Copy Services from a browser-equipped computer.

Copy Services is of great use to customers with large IT systems, big data volumes, and a requirement for around-the-clock IS availability.

Copy Services will provide the most benefit to the customer who:

- Needs to have disaster tolerant IT centers
- Is planning to migrate data between systems
- Is often migrating workload

- Has to backup large amounts of data
- Needs to reduce the time the server has to be taken off-line for backup
- Plans to test new applications
- Needs a copy of production data for data warehousing or data mining

Copy Services can be integrated with technologies such as Tivoli Storage Manager (TSM), formerly ADSM, Logical Volume Manager (LVM) mirroring, or SAN Data Gateway mirroring to solve a wide variety of business issues. IBM, with its broad portfolio of products in this industry, has many experts available to discuss the right solution for your business and to help you design and implement a solution that will give you the maximum business benefit.

FlashCopy

Because Enterprise Storage Servers are highly intelligent subsystems, they are capable of performing storage-related activities independently of host computer systems. FlashCopy provides a point in time (PIT) copy of your data. This is known as T(0) copy. With FlashCopy, you can schedule and execute near-instantaneous copies of your data entirely within the ESS itself. Not only does this free your host processors for other activities, it also eliminates host-to-disk I/O normally associated with mirroring or other types of backup. If you look closely, you may notice some very slight performance degradation at the level of the logical disk being copied, but the rest of the system remains unaffected.

Peer-to-Peer Remote Copy (PPRC)

PPRC is essentially a synchronous mirroring activity utilizing two or more enterprise storage servers. A disk write to the source ESS is not complete until its peers have also acknowledged the write. However, since Enterprise Storage Servers can acknowledge writes once the data has been written safely to cache and nonvolatile RAM, PPRC does not require waits on remote drives. This means that PPRC is not burdened by the synchronous write latencies found in other mirroring systems.

Although PPRC is used primarily in high availability or disaster recovery scenarios (remote systems can be located several kilometers away), it can also be used effectively for backups. Once a PPRC relationship has been established, the slave (also called a “clone”) will always track the master ESS. However, this tracking mechanism can be suspended long enough for a backup to be made, then resumed. This is called a “split-mirror” backup.

1.2 Introduction to high availability concepts

This section gives an overview of concepts that are used to acquire high availability systems. Multipathing, mirroring, and clustering are discussed.

> **High availability—definition:** High availability refers to a system or component that is continuously operational for a desirably long length of time. Availability can be measured relative to "100% operational" or "never failing." A widely-held but difficult-to-achieve standard of availability for a system or product is known as "five 9s" (99.999 percent) availability.

1.2.1 Multipathing

This section gives an overview of multipathing and what it is used for.

What is multipathing?

Multipathing is the use of different ways or paths to get to the same physical device (Figure 1-1).

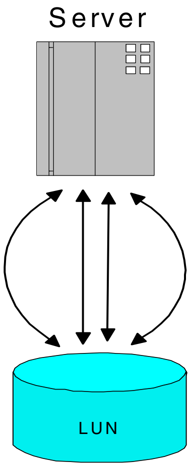


Figure 1-1 Multipathing overview

This can be happen by using:

- Several adapters in a server assigned to the same volume
- Using multiple ports of an ESS
- Using switches or a SAN
- Or a mix of the above

When is it used?

Multipathing is used in two cases: increase the availability of the application and increase the performance of the application.

Increase the availability of the application

Multipathing protects the application and server availability for the following cases:

- Adapter failure at the server
- Adapter failure at ESS
- Cable failure
- Switch / Hub failure
- Temporary unavailability of components through updates of microcodes

Increase the performance of the application

Performance of an application depends on several components like CPU, memory, disks or disk subsystem, adapters, connections speed of disks and network, and many parameters of the components and the application.

The way to increase performance is to remove bottlenecks. Multipathing can help you to remove the following bottlenecks (Table 1-1).

Table 1-1 Performance bottlenecks where multipathing can help

Bottleneck	Action
IO Rate of the server adapter is on the limit	Add an additional adapter on the server
IO Rate of the ESS host adapter is on the limit	Use/Add an additional host adapter on the ESS

Bottleneck	Action
The throughput rate of the server adapter is on the limit	Add an additional adapter on the server
The throughput rate of the ESS host adapter is on the limit	Use/Add an additional host adapter on the ESS
Some times too bad performance, though merging traffic on the SAN	using dynamic multipathing over a couple of paths on much servers as possible, to reduce peeks on connections

Static and dynamic load balancing

Several *path selection policies* are available:

- Load Balancing (lb): The path to use is chosen by estimating the load on the adapter that each path is attached to.
- Round Robin (rr): The path to used in rotation from those paths not used in the last I/O.
- Failover Only (fo): All I/Os use the same (preferred) path unless the path fails because of I/O errors.

Note: Not all policies are possible on each platform and each type of multipathing software.

Why you need multipathing software

- There are no failover or dynamic load balancing features in many operating systems.
- In most cases they cannot handle multiple instances of the same volume. If there are two paths from an HBA to a disk then the operating system will see two instances of the one actual disk unless multipathing is used or is built into the operating system.

In Figure 1-2 you see an system with three adapters attached to the same volume. The operating system sees three instances of the same volume.

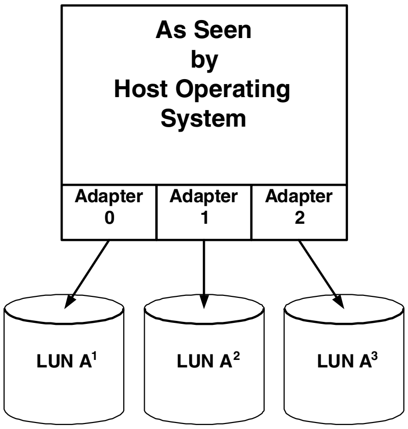


Figure 1-2 Operating system sees three instances of the same volume

To handle this you need a multipathing software which consolidates the view to the volume

- By masking the volume. The operating system sees only one volume.
- By creating an additional volume, that should be used.

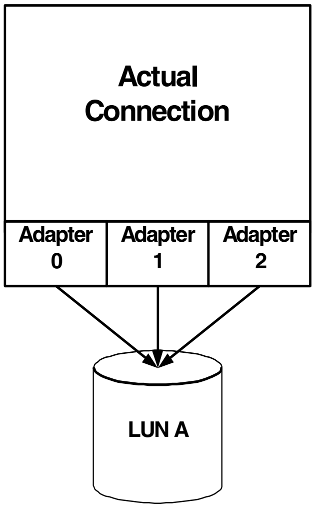


Figure 1-3 Consolidated view of an volume

1.2.2 Mirroring

This section gives an overview of mirroring and what it is used for.

What is mirroring?

Disk mirroring uses data in two or more copies simultaneously. It is always connected with at least 100% data redundancy or more. It is also known RAID1.

The data are exits and will be changed in two or more different places at the approximately the same time. This is done by hardware (controller, subsystems) or software.

Software mirroring

Most of currently available operating systems are already prepared for disk mirroring through their own built-in or separately available mechanism.

Here are some examples of software that allows mirroring of logical volumes and disk drives:

- AIX Logical Volume Manager (LVM), which is built-in part of AIX operating system.
- Mirror Disk/UX, which is separately available and a licensed product for HP-UX operating system.
- Solstice Disk Suite, which is a separately available product for Solaris operating system. No additional license is required for Solstice Disk Suite.
- Windows 2000 Disk Management, which is a built-in operating system tool.
- Veritas Volume Manager VxVM, which is a separate product available for most operating systems including AIX, HP-UX, Solaris and Windows 2000.

Hardware mirroring

In this case the mirroring will be done by the hardware and there is no additional CPU load for mirroring.

Hardware mirroring can be done in three ways, by:

1. Raid controller in the server

In this case the Controller is able to mirror the data (RAID1). The distance of the copies depends on the type of connection and the configuration.

2. Raid controller in a storage subsystem

   In this case the Controller is located in the subsystem. There are mirrors of the data, but all in the same box.

3. Storage subsystem with the possibility of remote copy

   In this case the subsystem is able to mirror the data in one or more subsystems. Long distances are possible. The server normally knows nothing about the mirror and is not able to use the mirror without respective system changes.

Note: For performance reasons mirroring will be used in combination with striping.

Why is it used?

- In all cases mirroring protects against data lost by disk failure.
- A distance between the mirrors can also protect against disasters like fire or others.
- It always increases the availability of a server by continuous data usage in the event of a drive failure.
- If the controller of the software is able to read from several copies at the same time, it increases the read performance, compared with a single disk, in a multitasking or OLTP environment.

1.2.3 Multipathing versus mirroring

Mirroring should be used to protect against data loss, by disk failure, subsystem failure or other disasters like fire and so on. It also increases the availability of servers. But if there is only one path to the data, it is an single point of failure. This single point of failure can be eliminated by using multipathing.

It also makes sense to use multipathing in an non mirrored environment. It increases the availability of the server and allows for example microcode updates for each redundant component.

Tip: A combination of mirroring and multipathing is a very good solution for high availability.

1.2.4 Clustering

This section gives an overview of clustering and what it is used for.

What is it?

A cluster is a group of servers and other resources that provide a set of network resources to a client, like a single system, and enables high availability and, in some cases, load balancing and parallel processing.

Clustering is the use of multiple computers, typically PCs or UNIX workstations and servers, multiple storage devices, and redundant interconnections, to form what appears to users as a single highly available system. Clustering can be used for load balancing. Advocates of clustering suggest that the approach can help an enterprise achieve 99.999% availability in some cases. One of the main ideas of clustering is that, to the outside world, the cluster appears to be a single system.

Load balancing—definition: Load balancing is dividing the amount of work that a computer has to do between two or more computers so that more work gets done in the same amount of time and, in general, all users get served faster. Load balancing can be implemented with hardware, software, or a combination of both. Typically, load balancing is the main reason for computer server clustering.

Parallel processing—definition: Parallel processing is the processing of program instructions by dividing them among multiple processors with the objective of running a program in less time.

99.999—definition: 99.999 (often called "five 9s") refers to a desired percentage of availability of a given computer system. Such a system would probably have what some refer to as high availability. As Evan Marcus, Principal Engineer at Veritas Software, observes, 99.999 availability works out to 5.39 minutes of total downtime - planned or unplanned - in a given year.

Why is it used?

Clustering means that the client is isolated and protected from changes to the physical hardware, which brings a number of benefits. Perhaps the most important of these benefits is high availability. Resources on clustered servers act as highly available versions of unclustered resources.

2

Basic concepts of ESS with multipathing and clustering

The primary reason to use multipathing, which is multiple routes from the host to storage, is to maintain access to data when one path fails. Clustering has two main purposes, first to provide an alternate host in the event of host failure, and second to be able to share the workload if it is too great for a single host. In this chapter we introduce and discuss the concepts for using multipathing and clustering, and how they can be implemented using the IBM ESS and Subsystem Device Driver (SDD). For a more detailed discussion on clustering see Chapter 4, “Clustering concepts” on page 45.

2.1 Concepts of high availability

This section discusses concepts of how to increase the availability of applications. High availability, particularly 24x7 availability is increasing in necessity for today’s business environment.

2.1.1 What is high availability?

High availability is always targeted on the availability of the application for the users. It is often important for the success and growth of your business.

Here are some basic terms:

- **Outage**: A period when the system is not available to users. During a scheduled outage, you deliberately make your system unavailable to users. You might use a scheduled outage to run batch work, save your system, or apply program temporary fixes (PTFs). An unscheduled outage is usually caused by a failure of some type.
- **High availability**: The system has no unscheduled outages.
- **Continuous operations**: The system has no scheduled outages.
- **Continuous availability**: The system has no scheduled or unscheduled outages.

Downtime issues

In the past, most disaster recovery focused on unscheduled downtime. This type of disaster includes fire, storm, flood, plane crashes, etc. The natural consequences were that the failure occurred, and the business stopped, and then moved to a remote recovery site. The business interruption could be measured in many hours or even days. This disaster scenario is well known and well documented. There are also many unscheduled downtimes that are not because of disaster. These can vary from simple acts like someone inadvertently pushing the Emergency Power Off (EPO) button in the machine room, to deliberate acts of sabotage. Both of these actions may crash applications and possibly the systems. Other examples of unscheduled downtimes are:

- No more free space on a file system
- An application error
- Power or environmental system loss
- Hardware or network failure

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

- Hardware upgrades
- Software upgrades
- Fix applies

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

The paradigm has changed. The cost of scheduled downtime is so great that businesses demand zero downtime.

2.1.2 Single system availability

Most systems in the market place today offer single system availability. The availability of a single system depends on many components like processors, memory, DASD or drives, power supplies, fans, individual I/O adapters, connection to the network, connections to storage, the storage subsystem itself and others. Some of these components have redundancy and failover capability to increase their availability, like dual power supply, dual fans, ECC checks, hardware mirroring or others.

Multipathing helps you to guarantee the availability of your server and applications in case of failure of these:

- Individual IO host bus adapter (HBA) in the server
- Connection between IO adapter and ESS
- Host adapter failure on ESS
- Temporary unavailability of adapters or SAN components through microcode updates or configuration changes

Figure 2-1 shows the increasing availability of the application of a single server by using multipathing. A failure of one of these marked components or a cable, does not disturb the availability of the application.

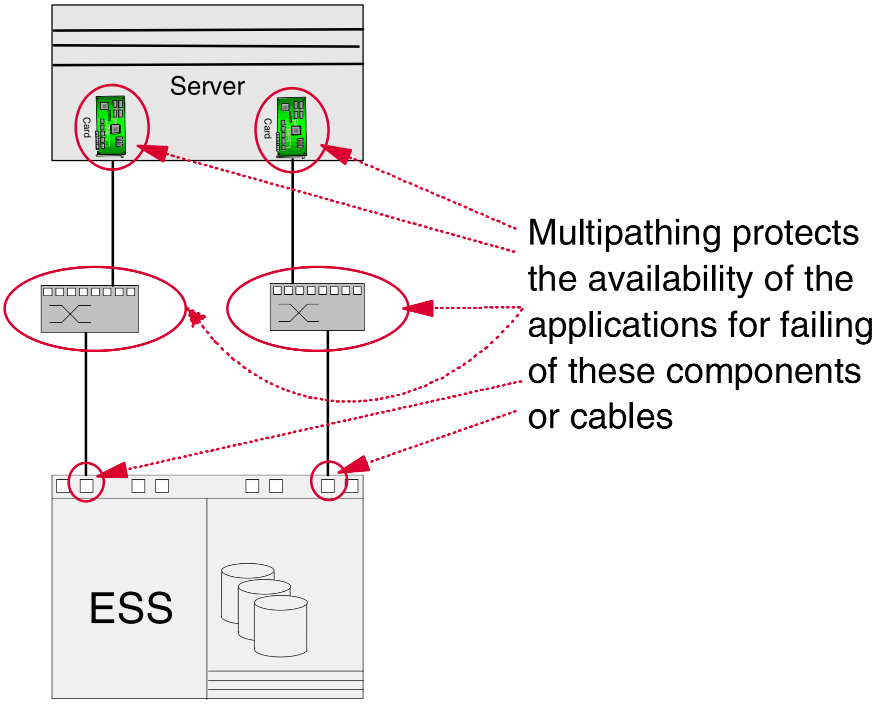

Server
Card
Card
Multipathing protects the availability of the applications for failing of these components or cables
ESS

Figure 2-1 Single server multipathing

2.1.3 Increasing availability by clustering

The target for many applications is continuous operation. A hardware failure or planned maintenance will cause an outage, some large servers can take as long as 30 minutes to reboot. A major upgrade or service repair to a server can take several hours. Also in the case of major disaster, such as a fire, a single server cannot provide availability. A second or more servers, perhaps on a remote site are required to serve the application so that it is still available for the users.

Clustering concept basics

There are many clustering solution on the market. Some are specific for a platform, Veritas Cluster, HACMP, Microsoft Cluster and others, and are specific for an application or included in an application (Lotus Notes cluster, Microsoft Exchange Cluster, Oracle cluster, Informix cluster and others). All of these solutions have one thing in common, sharing data between servers. This data should be stored on a highly available storage subsystem. Highly redundant storage servers such as the ESS helps you to realize this high availability.

Figure 2-2 shows the structure of a local cluster. The servers are redundant. In many cases the IP connection is also redundant as there will be dual LANs. Generally there is a non IP connection between the servers for the heartbeat, that is the servers monitor each other to check that they are still online. This heartbeat is run on connections, such as serial links, SAA or SCSI target modes.

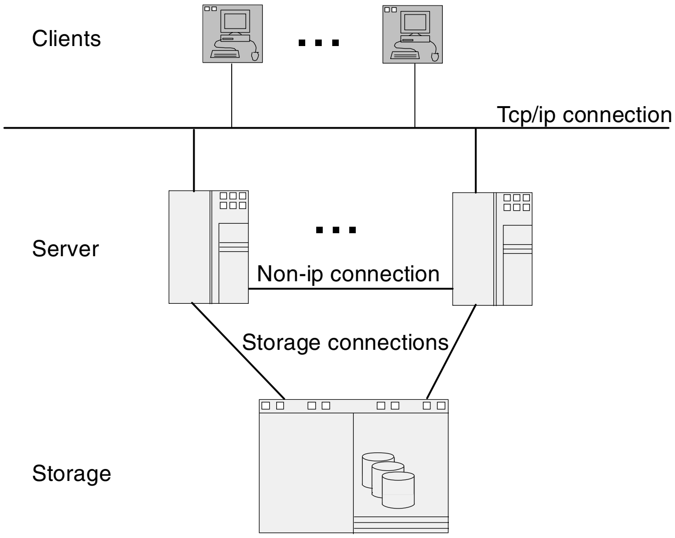


Figure 2-2 Local cluster basics

Local clusters

The servers can work in either of two modes: active and passive (standby); or active and active. Target of the system is to have no single point of failure (SPOF).

Notes:

- Any switching of the applications between the servers can include an outage, so that's why it is important to ensure that each single server is highly available. Multipathing to the storage unit and a redundant and high availability storage unit like ESS is essential.
- A local cluster increases the availability of an application, however it does not protect against disasters like fire, water or other that damage the facility.

Stretched clusters

To provide protection against disasters it is necessary to locate the servers and storage / storage mirror on different locations like other fire barriers sectors, other buildings, other cities or other places.

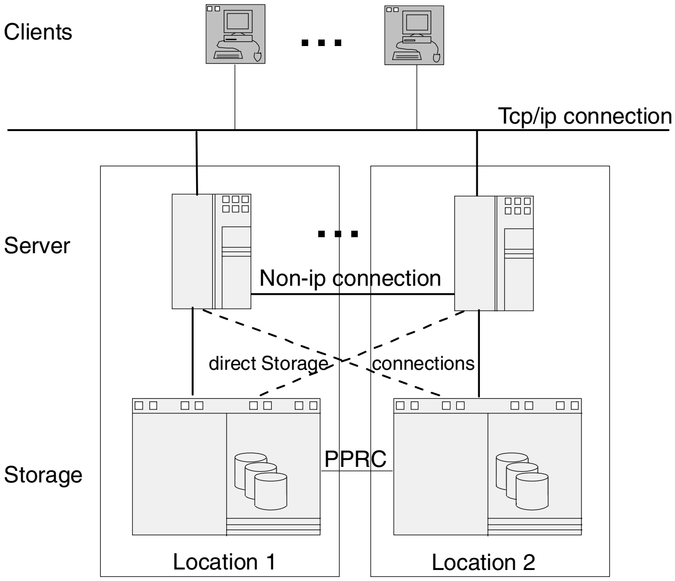


Figure 2-3 Stretched clusters

One of the basic requirements for stretched clusters is the mirroring of data. This can be done by the server operating system (LVM mirroring, for example) or by the server HBA. But in the case of longer distances, for example more than 10 km, you need a product, such as peer to peer remote copy (PPRC) for synchronous data transfer, to ensure and protect the data integrity. However, it also is necessary to keep the individual server highly available so there is still a requirement for multipathing to the ESS.

Note: For short distances it makes sense to connect both mirrors of the data to both servers. If one of the mirrors has problems, the other mirror is still available to the server. The application is still running without any outages.

2.2 How disks are seen on ESS

This section describes how disks are seen by hosts using SCSI and Fibre Channel and the effect that multipathing has on this.

2.2.1 Overview

There are fundamental differences between SCSI and Fibre Channel (FC) attachment and how the assignment of logical unit numbers (LUNs) is made to one or more hosts. LUN masking, that is preventing unauthorized hosts from having access to storage is also implemented differently for SCSI and FC attachment.

Figure 2-4 shows an example of LUN assignments. The SCSI hosts can only see the disks that are assigned to the port that they are attached to. FC attached hosts have the ability to see any open systems LUNs that are defined on the ESS. LUN masking is used to prevent unauthorized access.

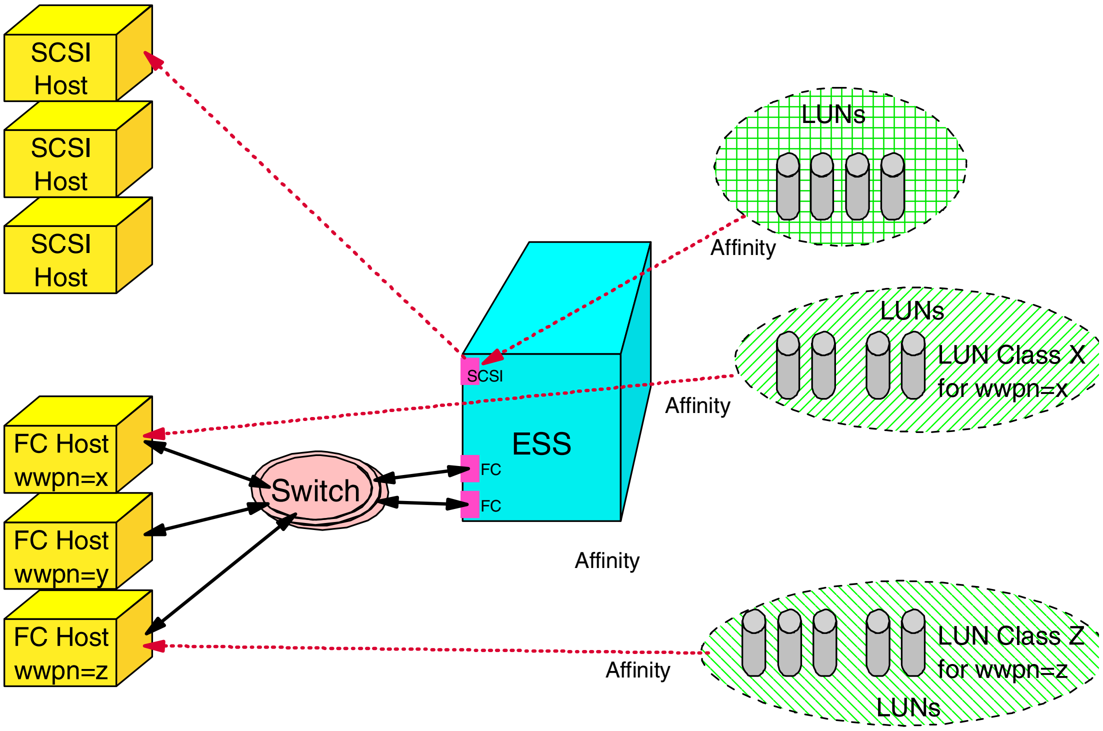


Figure 2-4 How LUNs are seen

A detailed description and updates can be found in *Host Systems Attachment Guide 2105 Models E10,E20,F10, and F20,* SC26-7296. The following is based on Version 4. See also the redbook *Implementing Fibre Channel Attachment on ESS,* SG24-6113.

2.2.2 SCSI

For SCSI attachment, LUNs have an affinity to SCSI ports, independent of which hosts might be attached to the ports. If you attach multiple hosts to a single SCSI port, each host has the exact same access to all the LUNs available on that port.

Targets and LUNs

For SCSI attachment, each SCSI bus can attach a combined total of 16 initiators and targets. Because at least one of these attachments must be a host initiator, that leaves a maximum of 15 that can be targets. The ESS is capable of defining all 15 targets on each of its SCSI ports. Each can support up to 64 LUNs. The software in many hosts is only capable of supporting 8 or 32 LUNs per target, but the architecture allows for 64. Therefore, the ESS can support 960 LUNs per SCSI port (15 targets x 64 LUNs =960).

SCSI host system limitations

Table 2-1 shows the configuration limitations for the host systems. These limitations can be caused by the device drivers, hardware or different adapters that the host systems support.

Table 2-1 Host system limitations

Host system	LUN Assignments per target	Configuration notes
Data General	0 -7	None
HP 9000	0 -7	None

Host system	LUN Assignments per target	Configuration notes
IBM AS/400 [a]	0 -7	The target SCSI ID is always 6. Sixteen LUNs are supported for each feature code 6501. For ESS, the two ports on the feature code 6501 each supports eight drives at full capacity for RAID. Real 9337s running RAID-5 must account for parity. Therefore, the eight drives provide the equivalent of a 7-drive capacity.
IBM @server iSeries (Fibre Channel) [b]	0 -31	There is one target per AS/400 and iSeries adapter.
IBM Personal Computer Server	0 -7	None
IBM RS/6000 and IBM @server pSeries	0 -31	AIX 4.3.3 supports 64 LUNs per target.
Sun Ultra A	0 -7	None
Sun Ultra B	0 -31	Use Sun Solaris 2.6, 2.7 or 2.8. (Solaris 2.6 and 2.7 require a Solaris patch to enable 32 LUNs per target).
Windows NT	0 -7	None
Windows 2000	0 -7	None
Novell NetWare	0 -31	None
Compaq AlphaServer	0 -7	Open VMS[c] - all versions
Compaq AlphaServer	0-7	Tru-64 Unix 4.0f and 4.0g
Compaq AlphaServer	0-15	Tru-64 Unix 5.0a and 5.1
NUMA-Q (UNIX)	0 -7	Use a minimum operating system level of DYNIX/ptx V4.4.7

[a] The naming convention for the AS/400 now defines a machine connected through a 6501 bus using SCSI cables.

[b] You can use the model 270 and 8xx for a Fibre Channel connection.

[c] Virtual Memory System (VMS) is an operating system from the Digital Equipment Corporation (DEC) that runs in its computers. VMS originated in 1979 as a new operating system for DEC's new VAX computer, the successor to DEC's PDP-11. VMS is a 32-bit system that exploits the concept of virtual memory.

2.2.3 Fibre Channel

For Fibre Channel attachment, LUNs have an affinity to the host’s Fibre Channel adapter through the worldwide port name (WWPN) for the host adapter. In a switched fabric configuration, a single Fibre Channel host could have physical access to multiple Fibre Channel ports on the ESS. In this case, you can configure the ESS to allow the host to use either:

- All physically accessible Fibre Channel ports on the ESS
- Only a subset of the physically accessible Fibre Channel ports on the ESS

In either case, the set of LUNs that are accessed by the Fibre Channel host are the same on each of the ESS ports that can be used by that host.

Targets and LUNs

For Fibre Channel attachment, each Fibre Channel host adapter has architecturally one WWPN of 2^{64} possible addresses. Each fabric can use 2^{24} addresses. On an Fibre Channel Arbitrated Loop only 128 devices are possible.The ESS supports a maximum of 4096 LUNs divided into a maximum of 16 logical subsystems each with up to 256 LUNs. If the software in the Fibre Channel host supports the SCSI command Report LUNs, then you can configure all 4096 LUNs on the ESS to be accessible by a host on a single adapter. Otherwise, you can configure no more than 256 of the LUNs in the ESS to be accessible by that host adapter.

Fibre Channel access modes

The Fibre Channel architecture allows any Fibre Channel initiator to access any Fibre Channel device, without access restrictions. However, in some environments this kind of flexibility can represent a security exposure. Therefore, the Enterprise Storage Server allows you to restrict this type of access when IBM sets the access mode for your ESS during initial configuration.

Note: Changing the access mode is a disruptive process, and requires that you shut down and restart both clusters of the ESS.

There are two types of LUN access modes: Access-any mode and Access-restricted mode.

Access-any mode

The access-any mode allows all Fibre Channel attached host systems that do not have an access profile to access all non-AS/400 and non-iSeries open system logical volumes that you have defined in the ESS.

Note: If you connect the ESS to more than one host system with multiple platforms and use the access-any mode without setting up an access profile for the hosts, the data in the LUN used by one open-systems host might be inadvertently corrupted by a second open-systems host. Certain host operating systems insist on overwriting specific LUN tracks during the LUN discovery phase of the operating system start process.

Access-restricted mode

The access-restricted mode prevents all Fibre Channel attached host systems that do not have an access profile from accessing any volumes that you have defined in the ESS. This is the default mode.

Your IBM service support representative (SSR) can change the LUN access mode. However, changing the access mode is a disruptive process, and requires that you shut down and restart both clusters of the ESS.

Access profiles

Whichever access mode you choose, any Fibre Channel attached host system that has an access profile can access only those volumes that are defined in the profile.

Depending on the capability of the particular host system, an access profile can contain up to 256 or up to 4096 volumes. The setup of an access profile is transparent to you when you use the ESS Specialist Web interface to configure the hosts and volumes in the ESS.

Configuration actions that affect the access profile are as follows:

- When you define a new Fibre Channel attached host system in the ESS Specialist by specifying its worldwide port name (WWPN) using the Modify Host Systems panel, the access profile for that host system is automatically created. Initially the profile is empty. That is, it contains no volumes. In this state, the host cannot access any logical volumes that are already defined in the ESS.
- When you add new logical volumes to the ESS using the Add Fixed Block Volumes panel, the new volumes are assigned to the host that you select. The new volumes are created and automatically added to the access profile of the selected host.
- When you assign volumes to Fibre Channel attached hosts using the Modify Volume Assignments panel, the selected volumes are automatically added to the access profile of the selected host.
- When you remove a Fibre Channel attached host system from the ESS Specialist using the Modify Host Systems panel, you delete the host and its access profile.

The anonymous host

When you run the ESS in access-any mode, the ESS Specialist Web interface displays a dynamically created pseudo-host called anonymous. This is not a real host system connected to the storage server, but is intended to represent all Fibre Channel attached host systems (if any) that are connected to the ESS that do not have an access profile defined. This is a visual reminder to the user that certain logical volumes defined in the ESS can be accessed by hosts which have not been specifically identified to the ESS.

2.2.4 Fibre Channel Storage Area Networks (SANs)

A SAN is a specialized, high-speed network that attaches servers and storage devices. A SAN is also called the network behind the servers. With a SAN, you can perform an any-to-any connection across the network using interconnect elements such as routers, gateways, hubs and switches. With a SAN, you can eliminate the connection between a server and storage and the concept that the server effectively owns and manages the storage devices. The SAN also eliminates any restriction to the amount of data that a server can access, which is limited by the number of storage devices, that can be attached to the individual server. Instead, a SAN introduces the flexibility of networking to enable one server or many heterogeneous servers to share a common storage utility, which might comprise many storage devices, including disk, tape, and optical storage. You can locate the storage utility far from the servers that use it.

2.3 Managing the number of paths to a LUN

This section discusses how to manage multiple routes from a host to a LUN. While SDD will support up to 32 paths to a LUN, in our opinion, no more than four paths to a device are a good choice, as this offers good redundancy and there is no performance gain by using more paths.

2.3.1 How many paths are seen

Here we discuss the differences between the SCSI and FC connection types and the way that the disks are seen.

SCSI connection

On SCSI, the number of paths results in the number of host bus adapters to host adapter connections. Figure 2-5 shows some possible configurations for disks.

.

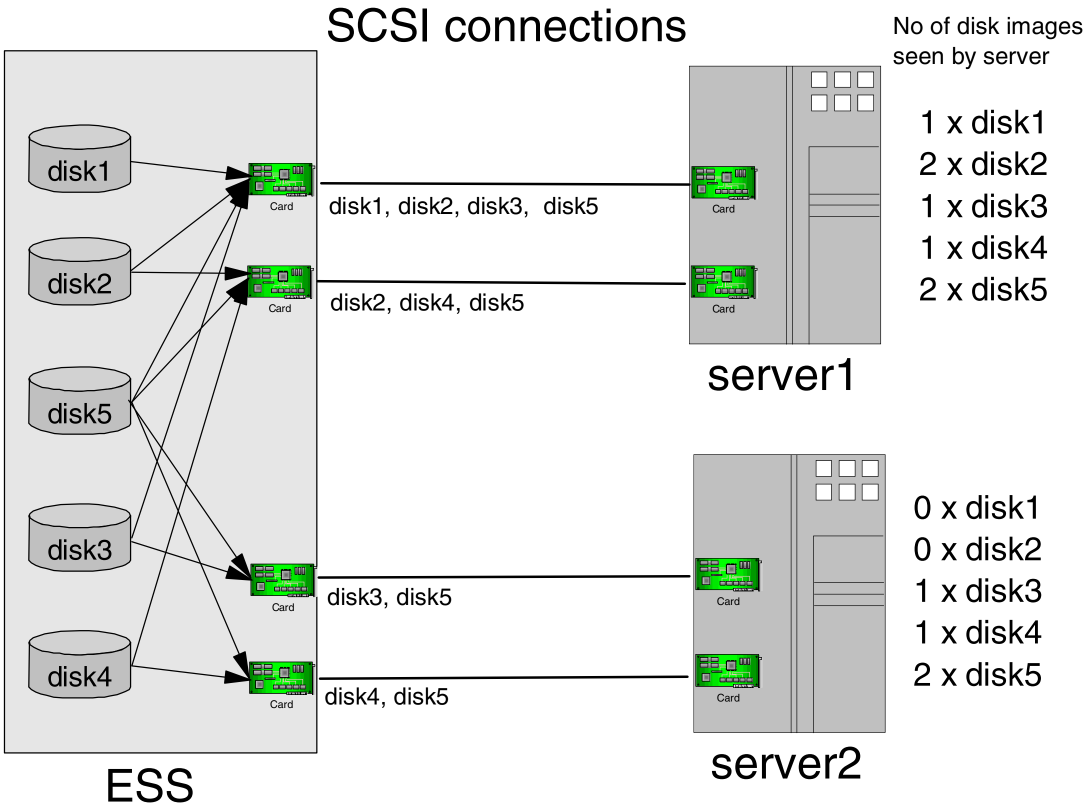


Figure 2-5 SCSI connections

Disk1 is only assigned to one port of the ESS connected with one port on server1. There is one path, no redundancy.

Disk2 is assigned to two ports of the ESS. Each port is connected with a port on server1. There are two paths to server1. Multipathing support must been used.

Disk3 is assigned to two ports of the ESS. One of the ports is connected with server1 the other with server2. There is only one path to each server. No redundancy on the individual server. There is no multipathing support necessary. There is disk sharing between servers for clustering.

Disk4 is the same configuration like Disk3. The combination of Disk3 and Disk4 allow a static load distribution in a cluster environment.

Disk5 is assigned to four ports of the ESS. Two of the ports are connected with server1 the other two with server2. There are two paths to each server. There is path redundancy on each server. Multipathing support is necessary. There is also disk sharing between servers for clustering.

Fibre Channel connections

On Fibre Channel, if there are no restrictions on the number of paths by Zoning or other methods then any physical connection between Host and ESS will be used. In Figure 2-6 we show examples of connections and the visible paths to one device.

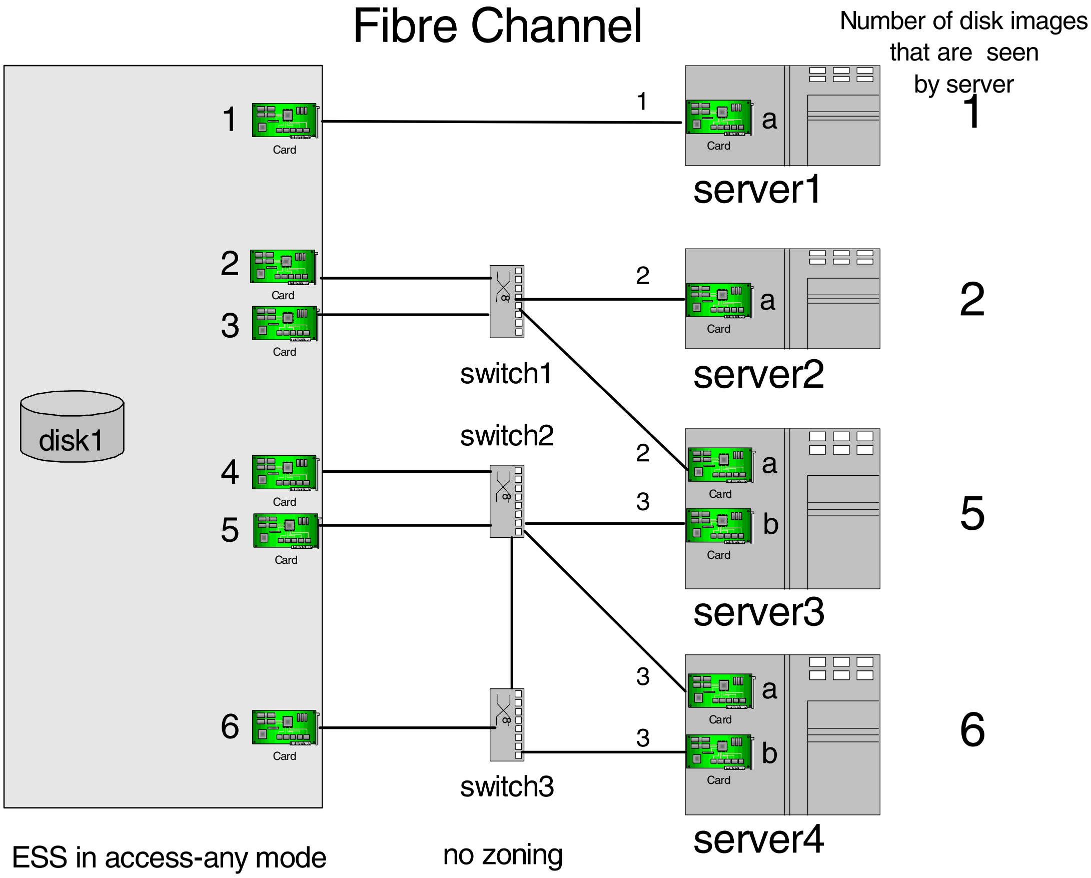


Figure 2-6 Fiber Channel connections

Server1 has one connection to one adapter. There is no redundancy. The disk is seen one time. If there is a second direct connection to the ESS, the disk is seen two times and multipathing support is necessary.

Server2 has one connection to the switch. The switch has two connections to the ESS. The server adapter sees both ESS adapters, so there are two paths to the ESS and the disk will be seen two times. Multipathing support is necessary.

Server3 has two adapter cards, card (a) is connected to switch1 and card (b) to switch2. Switch1 has two paths to the disk. Server3 also has three paths over the adapter card (b). Two over switch2 and one over switch2 to switch3 and then to ESS. In sum the disk will be seen five times. Multipathing support is necessary.

Server4 has over each adapter three paths to the ESS. In sum the disk will be seen six times. This kind of connection can be a good solution for high available clusters in a SAN environment. Multipathing support is necessary.

2.3.2 How to reduce paths

There are several ways to reduce the number of paths to a disk. All examples in this part are based on the configuration in Figure 2-6 on page 21.

Reduce by ESS configuration

The ESS provides two ways to reduce the paths to a disk.

LUN definition (LUN masking)

The standard mode of an ESS is the access-restricted mode. Each Fibre Channel adapter on the server is defined as a separate Host with a WWPN.

Here are some examples based on the configuration in Figure 2-6 on page 21:

- If disk1 is only assigned to adapter (b) of server4, only server4 sees the disk1 three times over adapter (b). No other server or adapter of server4 can see or access disk1.
- If disk1 is assigned to server2 adapter (a) and server3 adapter (b), server 2 sees disk 1 two times, and server3 three times. The paths of the server to the ESS are totally independent. So it is a good configuration for a cluster. If using adapter (a) of server3 instead of adapter (b) of server3, there is a single point of failure (switch1).

Host definition

In the access-restricted mode of the ESS it is necessary to define each Fibre Channel adapter as a separate Host.

On the Host definition panel of the IBM StorWatch Enterprise Storage Server Specialist, it is possible to reduce the Fibre Channel adapter of ESS used by this Fibre Channel adapter assigned to this hostname. The standard uses all installed ports. To reduce the used adapter in the ESS, it is necessary to know the bay number and the card number in the bay you plan to use.

Be sure there is a connection between the server and these Fibre Channel ports of the ESS. Here are some examples based on the configuration in Figure 2-6 on page 21. To make it easier, we did not use the bays and card numbers, only the numbers in the figure.

- If server2 is configured to use only ESS adapter 2, disk1 will be seen once. No multipathing support is necessary.
- To prevent traffic in the inter switch link (ISL) between switch2 and switch3, it is necessary to configure adapter card (a) on server3 to use only ESS ports 4 and 5. Adapter card (b) on server3 should only use ESS port 6.

Reduce by zoning

Zoning is done by switch management. On the switch you can define, which port or WWPN can see which other ports or WWPNs.

Here are some examples based on the configuration in Figure 2-6 on page 21:

- To prevent seeing disk1 two times, adapter card (a) from server2 should be in only one zone together with ESS port 2 **or** port 3.
- To prevent traffic in the inter switch link (ISL) between switch2 and switch3 there should be two zones. The first with adapter card (a) of server4 and ESS ports 4 and/or 5. The second one with adapter card (b) of server4 and ESS port 6.

Reduce by adapter profiles

Some server Fibre Channel adapters, especially in the windows environment, are able to make LUN masking on the adapter, so that on a Windows level not all the disks are seen.

Here are some examples based on the configuration in Figure 2-6 on page 21:

- To prevent seeing disk1 two times, adapter card (a) from server2 should be seen in only one ESS port 2 **or** 3.
- To prevent traffic in the inter switch link (ISL) between switch2 and switch3 adapter card (a) of server4 should only seen ESS ports 4 and/or 5. Adapter card (b) of server4 should only see ESS port 6.

Reduce by SAN management software

There are several software products that are available to manage SANs. Some of these, like Tivoli Storage Network Manager (TSNM), are provide LUN masking for the SAN. So it also is possible to reduce paths to devices using this feature. Further information for Tivoli Storage Network Manager can be found at:

`http://www.tivoli.com/products/index/storage_net_mgr`

2.3.3 Paths over inter switch links

In some situations there are many paths to a device. Normally no more than four active paths make sense. Some paths go over inter switch links and may have a bottleneck or are not the shortest way to the ESS. In these situations it makes sense to have these paths, but sets these paths offline. So the inter switch links are not used in the normal operation, however, if there no other paths available, these paths will be activated and our application is still available.

3

Multipathing software

In this chapter we discuss software which can be used for multipathing solutions based on IBM Enterprise Storage Server (ESS) 2105. In particular we discuss the following:

- IBM Subsystem Device Driver (SDD)
- HP-UX Logical Volume Manager built-in software Physical Volume Links (PV-Links)
- Veritas Volume Manager built-in Dynamic MultiPathing software (DMP)

3.1 IBM Subsystem Device Driver

The IBM Subsystem Device Driver (SDD) is a multipathing software designed by IBM especially to use with the IBM Enterprise Storage Server 2105 (ESS). It cannot be used with any other storage servers or storage devices. IBM SDD is not a disk driver itself. It resides on the host server above the native disk device driver that is attached to the IBM 2105. The purpose of SDD is to present a single image of a disk or ESS LUN to the host operating system when multiple paths would ordinarily present multiple views. SDD enables redundant connections between the host server and disk storage in the ESS to provide enhanced performance and data availability.

IBM Subsystem Device Driver allows you to dynamically manage multiple paths and recover to another path in case of disaster. It has the following features:

- Dynamic load balancing between multiple paths
- Dynamic path failover in case of disaster
- Dynamic path recovery when the failed path becomes operational
- Enables concurrent download of licensed internal code (LIC)

Note: Concurrent download of licensed internal code is the capability to download and install licensed internal code on an ESS while applications continue to run. During the time when new licensed internal code is being installed in an ESS, the upper-interface adapters inside the ESS may not respond to host I/O requests for approximately 30 seconds. The IBM Subsystem Device Drivers makes this transparent to the host through its path selection and retry algorithms. Path algorithms are discussed later in this chapter.

Figure 3-1 shows an example of configuration supported by SDD. Both types of adapters: SCSI and Fibre Channel are supported by SDD and can be used for multipathing, but they should not be mixed at the same time for the same LUNs. In the other words, if any LUN is assigned and accessed through a SCSI HBA within the ESS, for multipathing purposes only another SCSI HBA can be used or the LUN has to be moved to Fibre Channel HBA and vice-versa. If the LUN is accessed through a Fibre Channel adapter, only other Fibre Channel adapters can be used for multipathing.

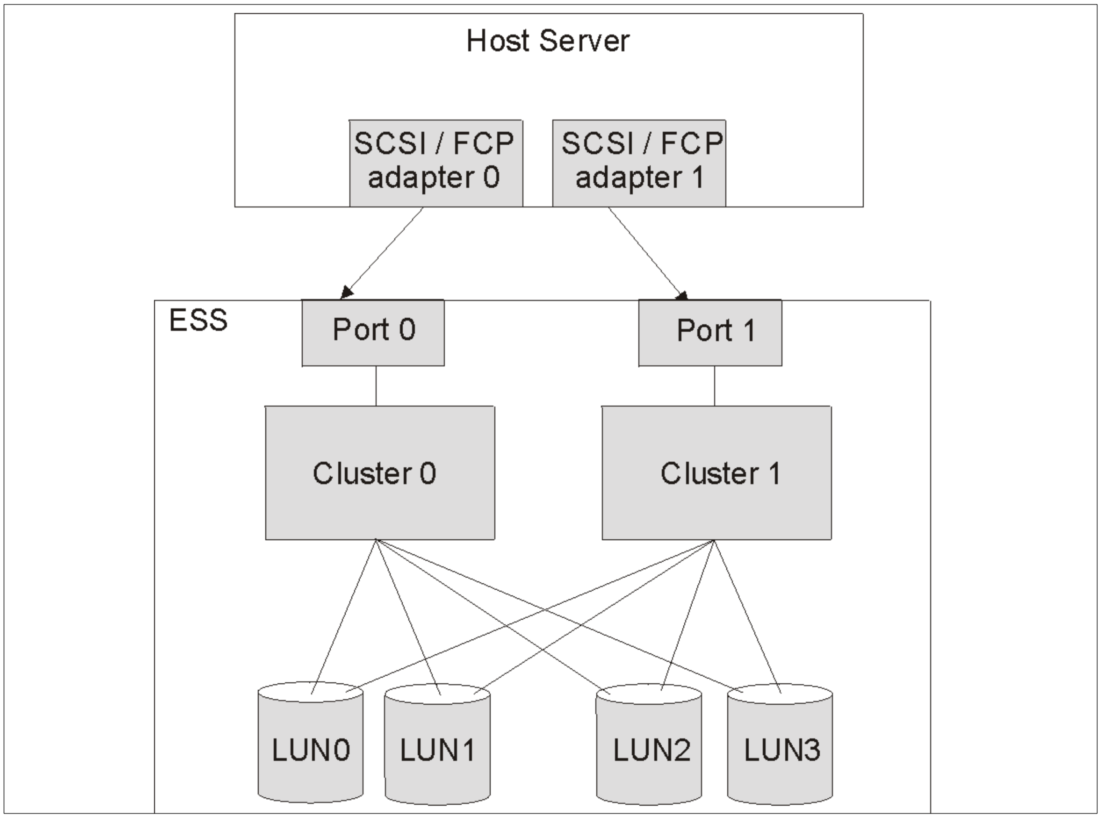


Figure 3-1 Example of configuration supported by IBM SDD

3.1.1 Path algorithms

The path algorithms basically work the same for all the platforms that the Subsystem Device Driver runs on. There are two modes of operation: *single-path mode* and *multiple-path mode*. Both of them are described.

Single-path mode

The host server has only one path that is configured to an ESS logical unit number (LUN). The Subsystem Device Driver in single-path mode has the following characteristics:

- When an I/O error occurs, the I/O is retried a sufficient number of times to bypass the interval when the ESS upper-interface adapters are not available. This behavior is required by concurrent download of licensed internal code.
- This path is never put into the *Dead* state.

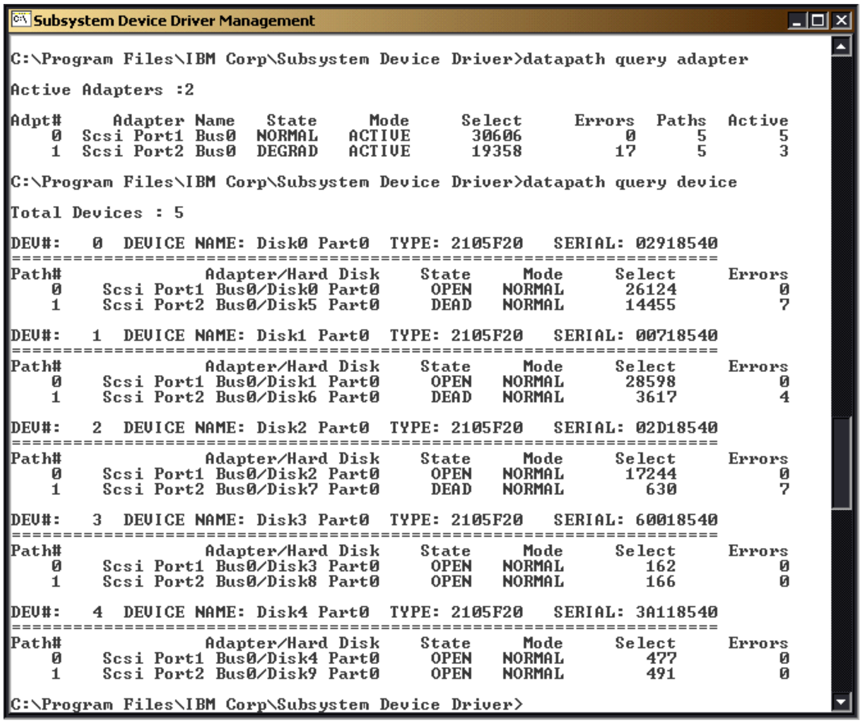
Multiple-path mode

The host server has multiple paths that are configured to an ESS LUN(s). The Subsystem Device Driver in multiple-path mode has the following characteristics:

- LUN(s) with only one operational path are in single-path mode.
- If an I/O error occurs on a path, the Subsystem Device Driver does not attempt to use the path again until 2,000 successful I/O operations have been performed on a remaining operational path. This process is known as bypassing a path. The Subsystem Device Driver bypasses a failing path twice (until the I/O error count reaches three), and then the path is changed to the *Dead* state. After the path is put into the *Dead* state, the Subsystem Device Driver uses this same bypass algorithm an additional two times.

> **Note:** You can always bring the path online by using the `datapath` command.

- When a path to a LUN is changed to the *Dead* state, the corresponding adapter state is changed from *Normal* to *Degraded* state. This adapter remains in the *Degraded* state as long as at least one path to its subsequent LUNs remains in the *Dead* state. Figure 3-2 shows an example of the corresponding path and adapter statuses.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adapter

Active Adapters :2

Adpt#     Adapter Name   State     Mode     Select     Errors  Paths  Active
    0  Scsi Port1 Bus0  NORMAL   ACTIVE      30606          0      5       5
    1  Scsi Port2 Bus0  DEGRAD   ACTIVE      19358         17      5       3

C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query device

Total Devices : 5

DEV#:    0  DEVICE NAME: Disk0 Part0  TYPE: 2105F20   SERIAL: 02918540
======================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk0 Part0     OPEN   NORMAL      26124          0
    1    Scsi Port2 Bus0/Disk5 Part0     DEAD   NORMAL      14455          7

DEV#:    1  DEVICE NAME: Disk1 Part0  TYPE: 2105F20   SERIAL: 00718540
======================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk1 Part0     OPEN   NORMAL      28598          0
    1    Scsi Port2 Bus0/Disk6 Part0     DEAD   NORMAL       3617          4

DEV#:    2  DEVICE NAME: Disk2 Part0  TYPE: 2105F20   SERIAL: 02D18540
======================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk2 Part0     OPEN   NORMAL      17244          0
    1    Scsi Port2 Bus0/Disk7 Part0     DEAD   NORMAL        630          7

DEV#:    3  DEVICE NAME: Disk3 Part0  TYPE: 2105F20   SERIAL: 60018540
======================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk3 Part0     OPEN   NORMAL        162          0
    1    Scsi Port2 Bus0/Disk8 Part0     OPEN   NORMAL        166          0

DEV#:    4  DEVICE NAME: Disk4 Part0  TYPE: 2105F20   SERIAL: 3A118540
======================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk4 Part0     OPEN   NORMAL        477          0
    1    Scsi Port2 Bus0/Disk9 Part0     OPEN   NORMAL        491          0

C:\Program Files\IBM Corp\Subsystem Device Driver>
```

Figure 3-2 Example of path status and corresponding adapter status

- After the Subsystem Device Driver puts a path into the *Dead* state, it will attempt to reuse it and put it back into the *Open* state after a certain number of successful I/O operations have completed on a remaining operational path. This number is operating system specific. Table 3-1 lists the number of successful I/O operations that must be completed on an operational path before a previously failed path is changed from the *Dead* to *Open* state.
- If an I/O error occurs on the last operational path to a device, the Subsystem Device Driver immediately attempts to reuse (or fail back to) one of the previously failed paths.

Table 3-1 Successful I/O operations before attempting to reopen the path

Operating system	**Number of successful I/O operations**
AIX	50 000
Windows NT/2000	50 000
HP-UX	200 000
Solaris	200 000

- If the first I/O operation fails after the path is put back into the *Open* state, the Subsystem Device Driver puts the path into the *Dead* state immediately and permanently. You must manually bring the path online by using the datapath command.

Note: The Subsystem Device Driver never puts the last operational path to a LUN into the Dead state. This is true even if I/O errors have occurred on the path.

- If an I/O error occurs on all the paths to a LUN, the Subsystem Device Driver returns an I/O error back to the application.

As we can see in described above path failover algorithm, every change of path state is triggered on I/O error basis and the number of I/O operations on remaining operational paths to the LUN. As long as there is no I/O activity on the path to the LUN, this path will never change its state to *Dead* even if it will fail. In turn, as long as there in no sufficient activity on remaining and operational paths to that LUN, the path will never change its state from *Dead* to *Open*. In this case we can manually bring the path online by using the `datapath` command.

The path failover algorithm is shown in the flowchart form in Figure 3-3.

Important: The algorithm on Figure 3-3 shows clearly, that SDD attempts to automatically fail back the path to the LUN from its `Dead` state to `Open` *only based on number of I/O requests completed on remaining and operational paths to that LUN*. This process is *not time-dependent*. In different environments, depending on I/O load, it may take from several seconds up to many hours. Remember, that you can always (at any time) manually bring the path online by using the `datapath` command.

Note: For updated and additional information not included in this redbook, see the README file on the IBM Subsystem Device Driver compact disc or visit the Subsystem Device Driver Web site at:

http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/storsw

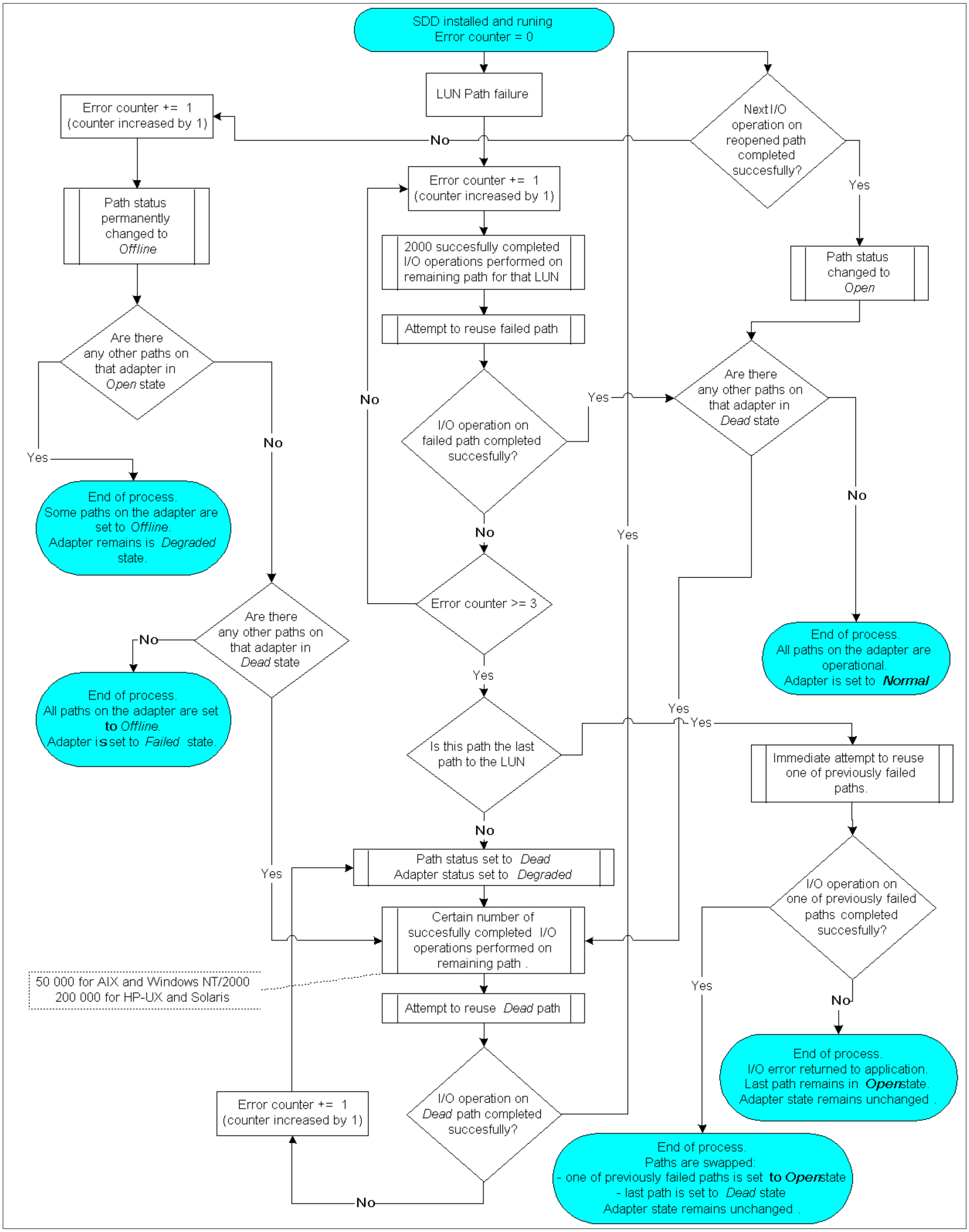

SDD installed and runing
Error counter = 0
LUN Path failure
Error counter += 1
(counter increased by 1)
No
Next I/O operation on reopened path completed succesfully?
Yes
Path status permanently changed to Offline
Error counter += 1
(counter increased by 1)
2000 succesfully completed I/O operations performed on remaining path for that LUN
Path status changed to Open
Attempt to reuse failed path
Are there any other paths on that adapter in Open state
No
Yes
I/O operation on failed path completed succesfully?
Are there any other paths on that adapter in Dead state
Yes
No
End of process.
Some paths on the adapter are set to Offline.
Adapter remains is Degraded state.
No
Yes
No
Error counter >= 3
End of process.
All paths on the adapter are operational.
Adapter is set to Normal
Are there any other paths on that adapter in Dead state
No
Yes
End of process.
All paths on the adapter are set to Offline.
Adapter is set to Failed state.
Is this path the last path to the LUN
Yes
Yes
Immediate attempt to reuse one of previously failed paths.
No
Yes
Path status set to Dead
Adapter status set to Degraded
I/O operation on one of previously failed paths completed succesfully?
Certain number of succesfully completed I/O operations performed on remaining path .
50 000 for AIX and Windows NT/2000
200 000 for HP-UX and Solaris
Attempt to reuse Dead path
Yes
No
End of process.
I/O error returned to application.
Last path remains in Open state.
Adapter state remains unchanged .
Error counter += 1
(counter increased by 1)
I/O operation on Dead path completed succesfully?
End of process.
Paths are swapped:
- one of previously failed paths is set to Open state
- last path is set to Dead state
Adapter state remains unchanged .
No

Figure 3-3 SDD path failover algorithm

3.1.2 How many paths to use with SDD

Depending on the environment there is no simple answer of how many paths should be used to a single device with SDD installed. The best performance and high availability results are always achieved, when separate paths use separate host adapters on both sides: within the ESS and within the host system. It is not required to use an equal number of HBAs on both sides, however, the system administrators and storage administrators have to consider if one-to-many HBAs connections will not produce a bottleneck on that side, where only one HBA is involved. Use of IBM ESS StorWatch Expert and host platform-dependent performance monitor software is highly recommended in more complex environments to check overall disk subsystem performance and find potential bottlenecks.

> **Tip:** IBM laboratory test results show that there are no additional performance benefits when using more than four paths to a LUN. As we mentioned earlier, the best result is achieved, when individual paths use different HBAs on both sides. When more than four paths are configured to a single LUN, the path load balancing algorithm becomes more complex due to increased number of calculations. The potential performance boost is decreased by delays needed to calculate which path should be used for I/O operation. In some cases this may not increase, but decrease an overall disk subsystem performance.

3.1.3 Usage of datapath command

In this section we describe in detail the syntax and usage of the IBM Subsystem Device Driver `datapath` command.

Usage of *`datapath query adapter`* command

The `datapath query adapter` command displays basic information about the status of all adapters or a single adapter in a host system. Returned information contains basic statistic information. The syntax for that command is as shown in Example 3-1.

Example 3-1 Syntax for datapath query adapter command

```
datapath query adapter [adapter_number]

   Parameters:
      adapter_number - the number of adapter for which you want the information to be
      displayed. If you do not enter an adapter number, information about all adapters is
      displayed.
```

The meaning of individual columns in the output of the command is as follows:

- `Adpt#` - the number of the adapters in the system.
- `Adapter Name` - the name of the adapter.
- `State` - the condition of the named adapter. It can be either:
  - `Normal` - adapter is in use.
  - `Degraded` - one or more paths are not functioning.
  - `Failed` - the adapter is no longer being used by Subsystem Device Driver.
- `Mode` - the mode of the named adapter, which is either Active or Offline.
- `Select` - the number of times this adapter was selected for input or output.
- `Errors` - the number of errors on all paths that are attached to this adapter.
- `Paths` - the number of paths that are attached to this adapter.
- Active - the number of functional paths that are attached to this adapter. The number of functional paths is equal to the number of paths minus any that are identified as failed or offline.

Usage of *datapath query adaptstats* command

The `datapath query adaptstats` command displays performance information for all SCSI and Fibre Channel adapters that are attached to Subsystem Device Driver devices. The syntax for that command is shown in Example 3-2.

Example 3-2 Syntax for datapath query adaptstats command

```
datapath query adaptstats [adapter_number]

   Parameters:
      adapter_number - the number of adapter for which you want the information to be
      displayed. If you do not enter an adapter number, information about all adapters is
      displayed.
```

The meaning of individual columns in the output of the command is as follows:

- `Total Read`
  - `I/O` - total number of completed read requests
  - `SECTOR` - total number of sectors that have been read
- `Total Write`
  - `I/O` - total number of completed write requests
  - `SECTOR` - total number of sectors that have been written
- `Active Read`
  - `I/O` - total number of read requests in process
  - `SECTOR` - total number of sectors to read in process
- `Active Write`
  - `I/O` - total number of write requests in process
  - `SECTOR` - total number of sectors to write in process
- `Maximum`
  - `I/O` - the maximum number of queued I/O requests
  - `SECTOR` - the maximum number of queued sectors to read/write

Usage of *datapath query device* command

The `datapath query device` command displays basic information about the status of all disk devices or a single disk device that are under control of IBM Subsystem Device Driver. Returned information contains basic statistic information. The syntax for that command is shown in Example 3-3.

Example 3-3 Syntax for datapath query device command

```
datapath query device [device_number]

   Parameters:
      device_number - the number of device for which you want the information to be
      displayed. If you do not enter a device number, information about all devices is
      displayed.
```

The meaning of individual columns in the output of the command is as follows:

- `Dev#` - the number of this device
- `Name` - the name of this device
- `Type` - the device product ID from inquiry data
- `Serial` - the logical unit number (LUN) for this device
- `Path` - the path number
- `Adapter` - the name of the adapter that the path is attached to
- `Hard Disk` - the name of the logical device that the path is bound to
- `State` - the condition of the named device:
  - `Open` - path is in use

  - `Close` - path is not being used
  - `Dead` - path is no longer being used. It was either removed by the IBM SDD due to errors or manually removed using the `datapath set device n path m offline` or `datapath set adapter n offline` command.
  - `Invalid` - path verification failed. The path was not opened.
- `Mode` - the mode of the named device. It is either Normal or Offline.
- `Select` - the number of times this path was selected for input or output
- `Errors` - the number of errors on a path that is attached to this device

Usage of *datapath query devstats* command

The `datapath query devstats` command displays the performance information for all disk devices or a single disk device that are under the control of IBM Subsystem Device Driver. The syntax for that command is shown in Example 3-4.

Example 3-4 Syntax for datapath query devstats command

```
datapath query devstats [device_number]

   Parameters:
      device_number - the number of device for which you want the information to be
      displayed. If you do not enter a device number, information about all devices is
      displayed.
```

The meaning of individual columns in the output of the command is as follows:

- `Total Read`
  - `I/O` - total number of completed read requests
  - `SECTOR` - total number of sectors that have been read
- `Total Write`
  - `I/O` - total number of completed write requests
  - `SECTOR` - total number of sectors that have been written
- `Active Read`
  - `I/O` - total number of read requests in process
  - `SECTOR` - total number of sectors to read in process
- `Active Write`
  - `I/O` - total number of write requests in process
  - `SECTOR` - total number of sectors to write in process
- `Maximum`
  - `I/O` - the maximum number of queued I/O requests
  - `SECTOR` - the maximum number of queued sectors to read/write
- `Transfer size`
  - <= 512: the number of I/O requests received, whose transfer size is 512 bytes or less
  - <= 4k: the number of I/O requests received, whose transfer size is 4 KB or less, but greater then 512 bytes
  - <= 16k: the number of I/O requests received, whose transfer size is 16 KB or less, but greater then 4 KB
  - <= 64k: the number of I/O requests received, whose transfer size is 64 KB or less, but greater then 16 KB
  - > 64k: the number of I/O requests received, whose transfer size is greater than 64 KB

Usage of *datapath set adapter* command

The `datapath set adapter` command sets all device paths attached to the adapter either to `Online` or `Offline` state. The syntax for that command is shown in Example 3-5.

Example 3-5 Syntax for datapath set adapter command

```
datapath set adapter adapter_number online/offline
```

```
   Parameters:
      adapter_number - the number of the adapter for which you want to change the status,
      online - sets the adapter online,
      offline - sets the adapter offline
```

Restrictions: The following restrictions apply when issuing `datapath set adapter` command (see the chapter “Path algorithms” on page 27 for details):

- This command will not remove the last path to a device.
- The `datapath set adapter offline` command fails if there is any device having the last path attached to this adapter.
- This command can be issued even when the devices are closed.
- If all paths are attached to a single Fibre Channel adapter, that connects to multiple ESS ports through a switch, the `datapath set adapter 0 offline` command fails and all the paths are not set offline.

Usage of *`datapath set device`* command

The `datapath set device` command sets the path to the device either to `Online` or `Offline` state. The syntax for that command is shown in Example 3-6.

Example 3-6 Syntax for datapath set device command

```
datapath set device device_number path path_number online/offline

   Parameters:
      device_number - the index number for a device for which you want to change the
      status,
      path_number - the number of the path to that device, for which you want to change
      the status
      online - sets the path online,
      offline - sets the path offline
```

Restrictions: The following restrictions apply when issuing the `datapath set device` command (see “Path algorithms” on page 27 for details):

- This command will not remove the last path to a device.
- This command can be issued even when the devices are closed.

3.2 HP-UX LVM built-in support for Physical Volume Links

PV-Links are built into a Logical Volume Manager algorithm to support multipathing solutions on HP-UX platform. PV-Links are available in 10.20 or later releases of HP-UX operating system or operating environment. PV-Links allow multiple paths to be connected between host server and its storage devices and uses them for failover purposes in case of primary path failure. When the primary path fails, after a certain period of time, HP-UX Logical Volume Manager will automatically switch I/O traffic to one of remaining available paths. This time period value depends on certain logical volume and physical volume timeout settings. The simplified version of PV-Links path failover algorithm is shown in Figure 3-4. Since no dynamic load balancing is available with PV-Links, some static load balancing may be established through careful implementation of PV-Links.

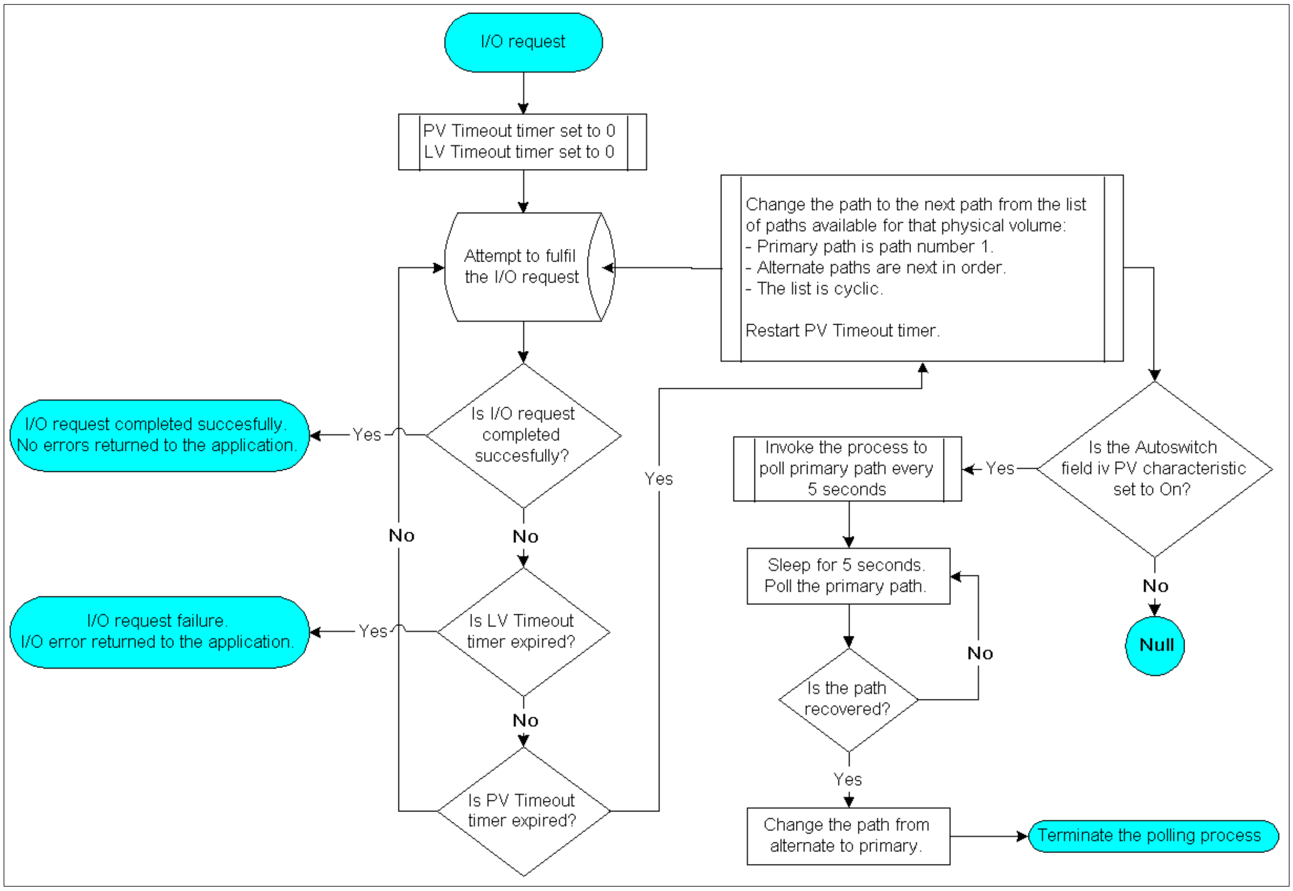

I/O request
PV Timeout timer set to 0
LV Timeout timer set to 0
Attempt to fulfil the I/O request
Change the path to the next path from the list of paths available for that physical volume:
- Primary path is path number 1.
- Alternate paths are next in order.
- The list is cyclic.
Restart PV Timeout timer.
I/O request completed succesfully. No errors returned to the application.
Yes
Is I/O request completed succesfully?
No
No
Yes
Invoke the process to poll primary path every 5 seconds
Yes
Is the Autoswitch field iv PV characteristic set to On?
Sleep for 5 seconds. Poll the primary path.
No
Null
I/O request failure. I/O error returned to the application.
Yes
Is LV Timeout timer expired?
No
No
Is the path recovered?
Yes
Is PV Timeout timer expired?
Change the path from alternate to primary.
Terminate the polling process

Figure 3-4 PV-Links path failover algorithm

As we can see in Figure 3-4, there are some dependencies between PV Timeout value and LV Timeout value. The most important conclusions are as follows:

- If LV Timeout is lower than (or equal to) PV Timeout, LVM will never attempt to switch to the alternate link. The LV Timeout timer will expire earlier (or at the same time) than the PV Timeout timer and an attempt to switch to the alternate path will never be made. An I/O error is returned to the application.
- The LV Timeout value should be an integer multiplication of PV Timeout value. If we want the Logical Volume Manager to test all alternate paths, the LV Timeout value should be set (at least) to a value of PV Timeout multiplied by number of paths available for that physical volume (including primary path).

Integrating PV-Links requires some manipulation of LVM's Volume Groups (VG). When planning a multipath topology, careful attention should be paid to the applications and data to be transferred over the SAN. By analyzing the data to be transferred, it is often possible to establish effective, static load balancing through the use of PV-Links. When creating a VG, the first path established to a device is always used as the primary path during normal operation. Any subsequent definition to the same device is used as an alternate path. Therefore, in a balanced I/O configuration with two paths to the storage, assigning half of the storage volumes to one path as primary, and the other half to the second path, we may get an average of fifty percent of the I/O for each of two paths. This is of course a theoretical scenario, but it demonstrates the possibilities.

PV-Links may be established by the same means used to extend volume groups with single paths. Either the System Administration and Management tool (SAM) or the command line may be used. To add an alternate path to the storage, simply extend the VG with the alternate paths. HP-UX LVM takes care of the details. It recognizes the following cases:

- The disk is already a member of that volume group - in this case an alternate path is added to a disk device
- The disk is already a member of another volume group - Logical Volume Manager will not allow to extend the volume group by the disk already in use
- The disk is not a member of any volume group - volume group is extended with a new disk device

After extending the VG with the alternate paths, a `vgdisplay -v` will show all primary and alternate paths. HP-UX versions 10.20 through 11.i support up to eight alternate paths to each primary. Each path may be a primary in a particular volume group, and an alternate in another. This can be very useful if multiple paths are used, and static load balancing is implemented. By using all possible paths as alternates to other volume groups, availability is maximized.

To show the possibilities of static load balancing, we discuss the following scenario (Figure 3-5):

- One system hosts attaching disk devices, equipped with two Fibre Channel adapters
- One Storage Area Network switch used
- One IBM Enterprise Storage Server with four LUNs configured as JBOD disks, all of them assigned to two Fibre Channel adapters: HBA1 and HBA2
- Disks used to store database. LVM mirror copies between disks involved.

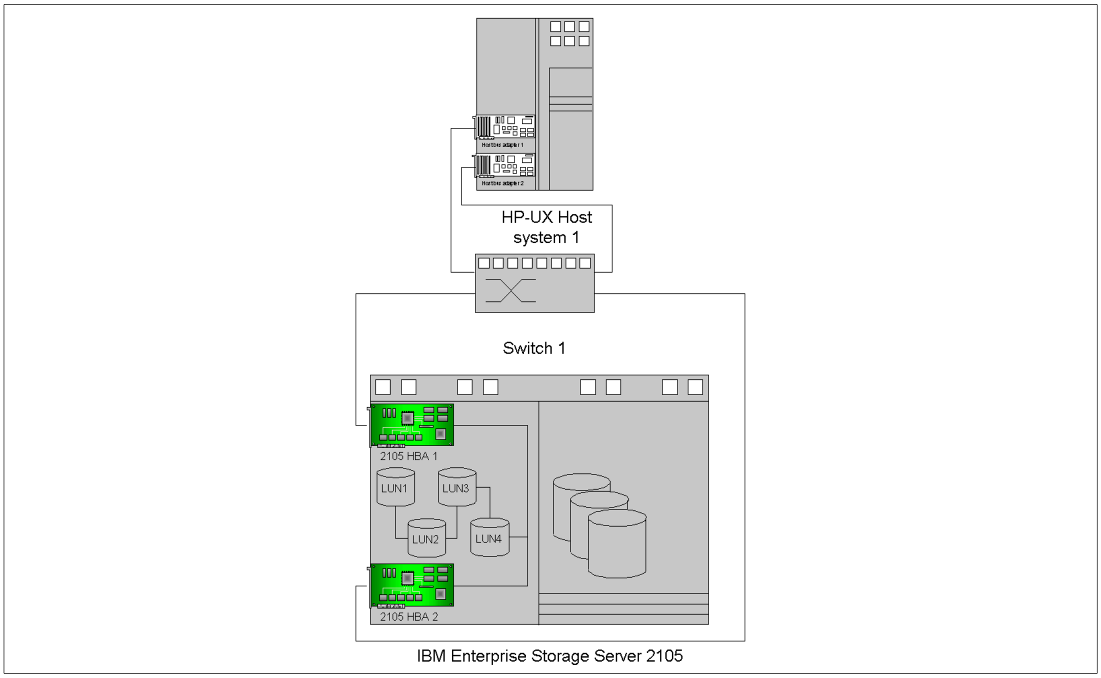


Figure 3-5 Example scenario for PV-Links

Depending on how zoning is configured on the switch, we may get the following two results (as shown in Figure 3-6 and Figure 3-7):

- If one zone is configured on the switch and it contains all four ports which are used on the switch, the operating system on the HP-UX host will see 16 LUNs available:
  - LUN1, LUN2, ..., LUN8 available on the first adapter

  - LUN9, LUN10, ..., LAN16 available on the second adapter
- If two zones are configured on the switch (one containing HBA1 from the ESS and HBA1 from hosts system, the second containing HBA2 form the ESS and HBA2 from the hosts system), the operating system on HP-UX host will see eight LUNs available:
  - LUN1, LUN2, LUN3 and LUN4 available on the first adapter
  - LUN5, LUN6, LAN7 and LUN8 available on the second adapter

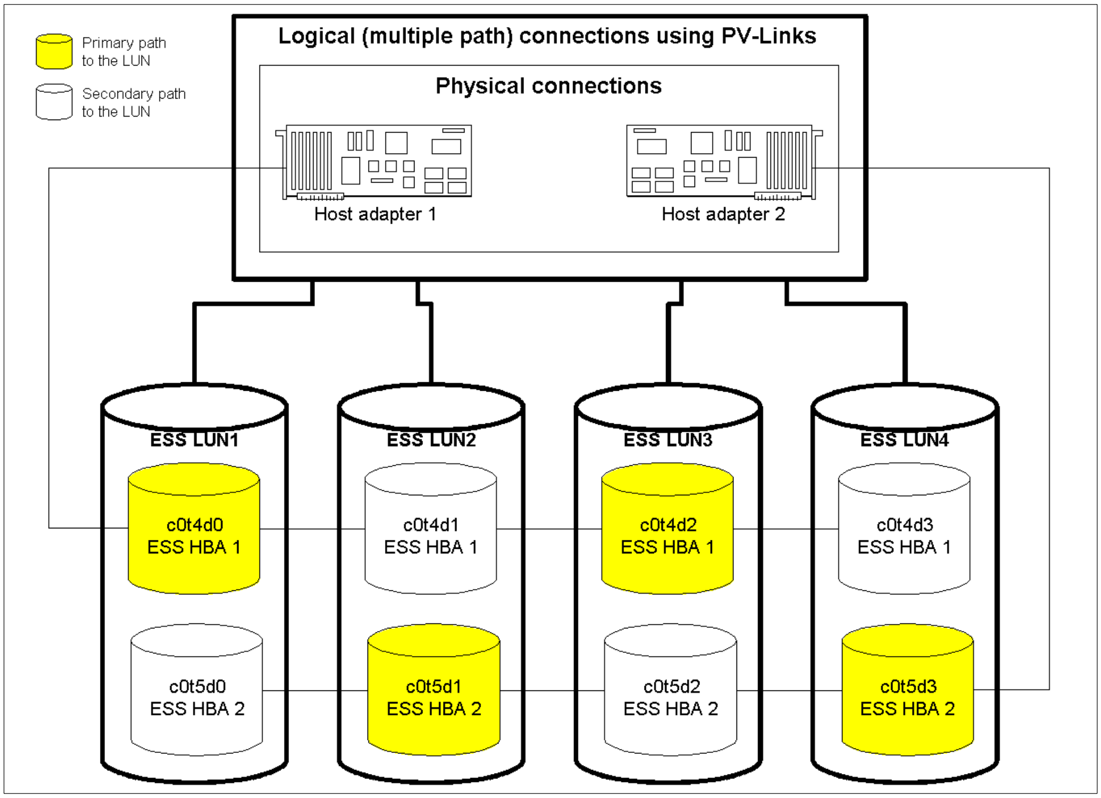


Figure 3-6 Multiple paths connections with PV-Links and two zones configured on the switch

In both cases described above, careful implementation of PV-Links allows to obtain a static load-balancing between available paths. Since PV-Links uses only primary path for I/O traffic, while secondary path is used only in case of primary path failure, we can implement mirror pairs as follows:

- In the case of two zones configured on the switch (see Figure 3-6):
  - Use device c0t4d0 as primary path to ESS LUN1 and device c0t5d0 as secondary path to ESS LUN1
  - Use device c0t5d1 as primary path to ESS LUN2 and device c0t4d1 as secondary path to ESS LUN2
  - Use device c0t4d2 as primary path to ESS LUN3 and device c0t5d2 as secondary path to ESS LUN3
  - Use device c0t5d3 as primary path to ESS LUN4 and device c0t4d3 as secondary path to ESS LUN4
  - Create first mirror pair between ESS LUN1 (first mirror copy) and ESS LUN2 (second mirror copy)
  - Create second mirror pair between ESS LUN4 (first mirror copy) and ESS LUN3 (second mirror copy)

  In this case, every adapter in the hosts system operates as primary for two paths and as backup adapter for two other paths. Interleaving primary paths between two copies of the same mirror pair allows us to direct I/O traffic for primary and secondary mirror copies to two different adapters. Because I/O characteristics for write requests are different from I/O characteristics for read requests (when mirroring is used, the same number of writes occurs to both copies, while read requests are done mainly from the primary mirror copy) we strongly recommend to interleave adapters between primary copies for different mirror pairs. That’s why ESS LUN4 (and *not* ESS LUN3) was used as the primary mirror copy for the second mirror pair. During write requests the same I/O traffic occurs for both adapters, but during read requests higher I/O traffic occurs to the primary copy. Using ESS LUN4 for the primary mirror copy of the second mirror pair prevents host adapter 1 from overloading with read requests - for reading from the first mirror pair host adapter 1 is used, while reads from the second mirror pair go through host adapter 2.

- In the case of one zone configured on the switch (see Figure 3-7):
  - Use device c0t4d0 as primary path to ESS LUN1, device c0t5d0 as secondary path to ESS LUN1, device c0t4d4 as third path and device c0t5d4 as fourth path
  - Use device c0t5d1 as primary path to ESS LUN2, device c0t4d1 as secondary path to ESS LUN2, device c0t5d5 as third path and device c0t4d5 as fourth path
  - Use device c0t4d6 as primary path to ESS LUN3, device c0t5d6 as secondary path to ESS LUN3, device c0t4d2 as third path and device c0t5d2 as fourth path
  - Use device c0t5d7 as primary path to ESS LUN4, device c0t4d7 as secondary path to ESS LUN4, device c0t5d3 as third path and device c0t4d3 as fourth path
  - Create first mirror pair between ESS LUN1 (first mirror copy) and ESS LUN2 (second mirror copy)
  - Create second mirror pair between ESS LUN4 (first mirror copy) and ESS LUN3 (second mirror copy)

    In this case, the same recommendations apply as in the case when two zones are configured on the switch. Every adapter in the host system operates as a primary for two paths and as backup adapter for two other paths. Interleaving primary paths between two copies of the same mirror pair allows it to direct I/O traffic for primary and secondary mirror copies to two different adapters. Because I/O characteristics for write requests are different from the I/O characteristic for read requests (when mirroring is used, the same number of writes occurs to both copies, while read requests are mainly from the primary mirror copy) we strongly recommend to interleave adapters also between primary copies for different mirror pairs. That’s why ESS LUN4 (and *not* ESS LUN3) were used as the primary mirror copy for the second mirror pair. During write requests the same I/O traffic occurs for both adapters, but during read requests higher I/O traffic occurs to primary copy. Using ESS LUN4 for the primary mirror copy of the second mirror pair prevents host adapter 1 from overloading in read requests - for reading from first mirror pair host adapter 1 is used, while reads from second mirror pair goes through host adapter 2.

Statistically, in both cases we get 50% of total system disk load for each of two adapters in host system. This may differ of course in particular installations and depends on the applications that generates the I/O traffic. Because only static load balancing is possible with PV-Links (the HP-UV operating system is unable to use PV-Links for dynamic load balancing), when significant differences occur between demands of applications generating disk I/O, we recommend for you to redesign as shown in Figure 3-6 and Figure 3-7. As examples of symmetrical implementation of PV-Links to asymmetric configuration, consider real I/O rate and number of I/O requests (both read and write) for each of the applications that are in use. Changing the configuration of PV-Links from primary to alternate is very easy, as its managed via the HP-UX LVM:

- To remove a primary or alternate path from a volume group, simply reduce the physical volume associated with that path (special device file) from configuration of the volume group. To do that, issue the command **`vgreduce vg_name pv_path`**. If the primary path is reduced from a volume group, LVM automatically switches to the alternate path, which is treated since then as a primary path. This means, that the volume group never exists without the primary paths to all subsequent physical volumes configured.

- To add an alternate path to a volume group, simply extend the volume group with a physical volume associated with that path (special device file). To do that, issue the command **`vgextend vg_name pv_path`**. If the LVM recognizes that `pv_path` as an alternate path to one of already configured physical volumes in this volume group, it marks it automatically as an `alternate link`.
- To switch the primary path to one of the alternate paths, simply issue the command **`pvchange -s pv_path`**. This will change the primary path to the specified one. Note, that you don't need to specify the volume group name. An example of changing the primary path for a physical volume is shown in Figure 3-8.

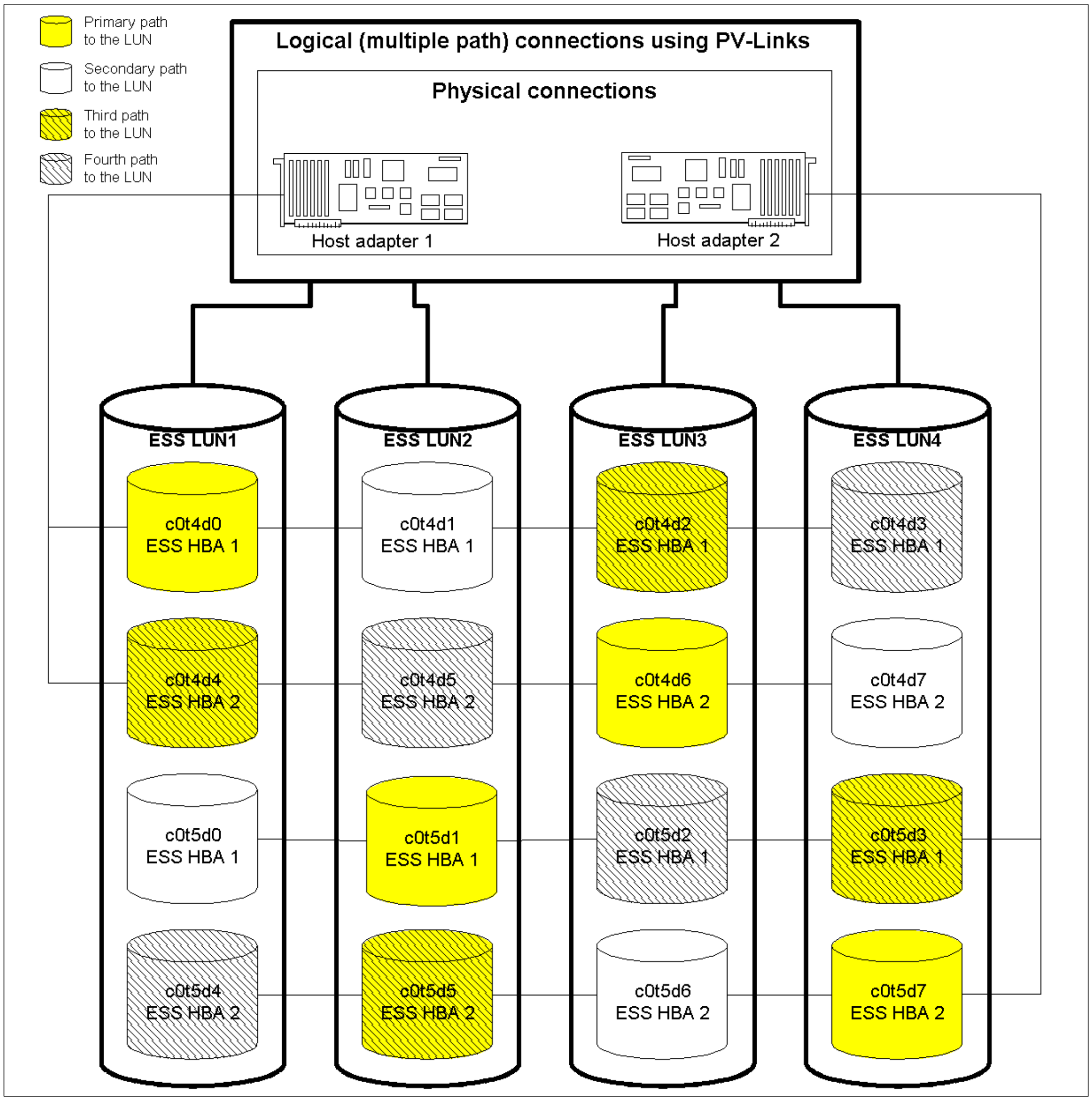
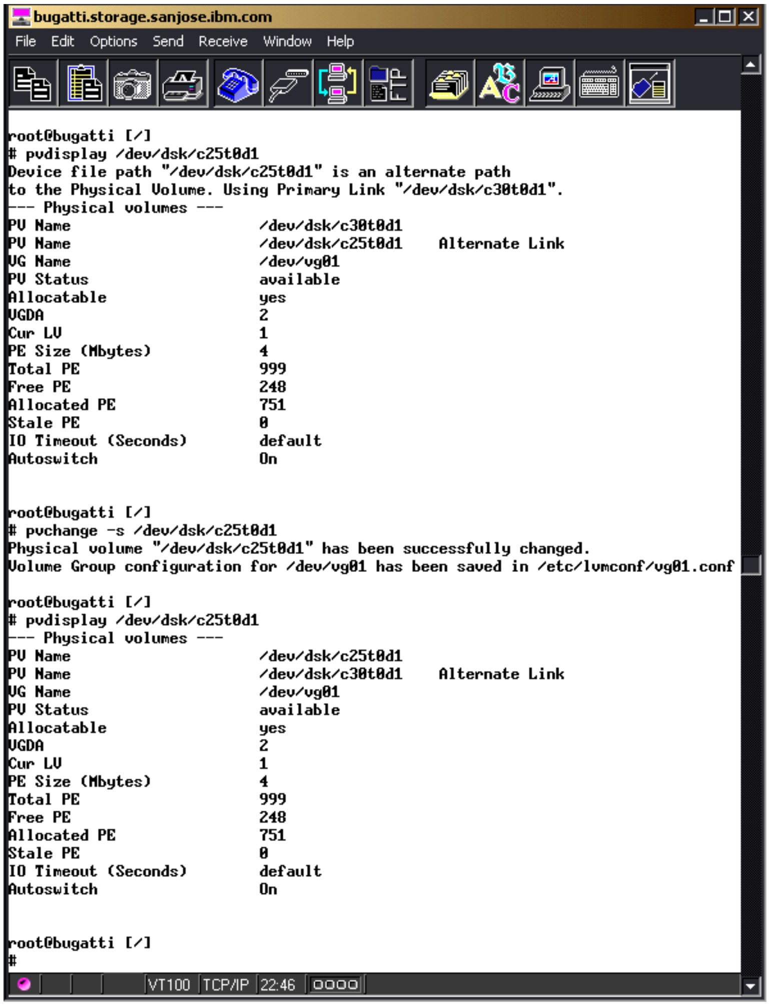


Figure 3-7 Multiple paths connections with PV-Links and one zone configured on the switch

As we can see, designing a proper PV-Links implementation should be done carefully, and in some environments is a time-consuming effort, but can be extremely useful when attaching HP systems to Storage Area Networks (SAN).


bugatti.storage.sanjose.ibm.com
File Edit Options Send Receive Window Help
VT100 TCP/IP 22:46 OOOO

```
root@bugatti [/]
# pvdisplay /dev/dsk/c25t0d1
Device file path "/dev/dsk/c25t0d1" is an alternate path
to the Physical Volume. Using Primary Link "/dev/dsk/c30t0d1".
--- Physical volumes ---
PV Name                     /dev/dsk/c30t0d1
PV Name                     /dev/dsk/c25t0d1    Alternate Link
VG Name                     /dev/vg01
PV Status                   available
Allocatable                 yes
VGDA                        2
Cur LV                      1
PE Size (Mbytes)            4
Total PE                    999
Free PE                     248
Allocated PE                751
Stale PE                    0
IO Timeout (Seconds)        default
Autoswitch                  On


root@bugatti [/]
# pvchange -s /dev/dsk/c25t0d1
Physical volume "/dev/dsk/c25t0d1" has been successfully changed.
Volume Group configuration for /dev/vg01 has been saved in /etc/lvmconf/vg01.conf

root@bugatti [/]
# pvdisplay /dev/dsk/c25t0d1
--- Physical volumes ---
PV Name                     /dev/dsk/c25t0d1
PV Name                     /dev/dsk/c30t0d1    Alternate Link
VG Name                     /dev/vg01
PV Status                   available
Allocatable                 yes
VGDA                        2
Cur LV                      1
PE Size (Mbytes)            4
Total PE                    999
Free PE                     248
Allocated PE                751
Stale PE                    0
IO Timeout (Seconds)        default
Autoswitch                  On


root@bugatti [/]
#
```

Figure 3-8 Example of primary PV-Link change

The `Autoswitch` fields have the following meaning:

- When set to `On`, the physical volume will automatically switch back to the original primary path after recovery from its failure. In this case the Logical Volume Manager will poll the primary path every five seconds. When the path recovers, LVM will switch all I/O requests to the primary path.
- When set to `Off`, the physical volume will remain on the alternate path even after recovery from primary path failure.

3.3 Veritas VxVM built-in Dynamic MultiPathing software (DMP)

Dynamic Multipathing is a built-in feature of Veritas Volume Manager to manage and operate multiple paths to the same disk device. In some cases Veritas DMP requires an additional licence to activate. Veritas Volume Manager uses two mechanisms to detect devices attached to the host operating system through multiple paths:

- If the disk device attached to the host system provides unambiguous identification through the *Universal World-Wide Device Identifier* (WWD ID), this identifier is used to determine if the device is already connected through a different path.
- If the disk device is unable to identify itself using WWD ID, the volume manager recognizes the metadata identifiers, which are stored on the disk.

VxVM creates metanodes for all physical disk devices which are accessible at the operating system level. Each metanode represents a metadevice and contains mapping to one or more of the operating system disk devices (if multiple paths to the same physical disk are detected) and is configured with the appropriate multipathing policy. In particular, metanodes are created for each multiple path disk device which VxVM has detected.

Note, that the metanode is not created for each available path to a device, but to a whole set of all available paths. Usually each path is represented at the operating system level as a separate special device file. Figure 3-9 shows how VxVM concepts fit into the operating system protocol stack.

Veritas Volume Manager DMP allows you to manage two types of multiple paths to a LUN:

- Active/Active - this kind of multiple target is used, when a target device is a disk array which concurrently allows to use more than one path to a LUN.
- Active/Passive - this kind of multiple target is used, when a target device is a disk array which allows to use at the same time only one active path to a LUN. All other connected paths are treated as alternate paths and are used only in case of active path failure.

As we can see, an Active/Passive configuration is not able to provide load balancing mechanism due to only one path available as active at the same time. Active/Active configuration allows to balance the load across multiple paths to the disk device. Usually Active/Passive configuration does not require an additional licensing, while Active/Active configuration requires an update of a base VxVM licence.

All information related to DMP configuration are stored in the DMP Database. It contains information about the following items (only major items are listed here):

- System disk controllers
- Disk devices
- Disk arrays
- Paths

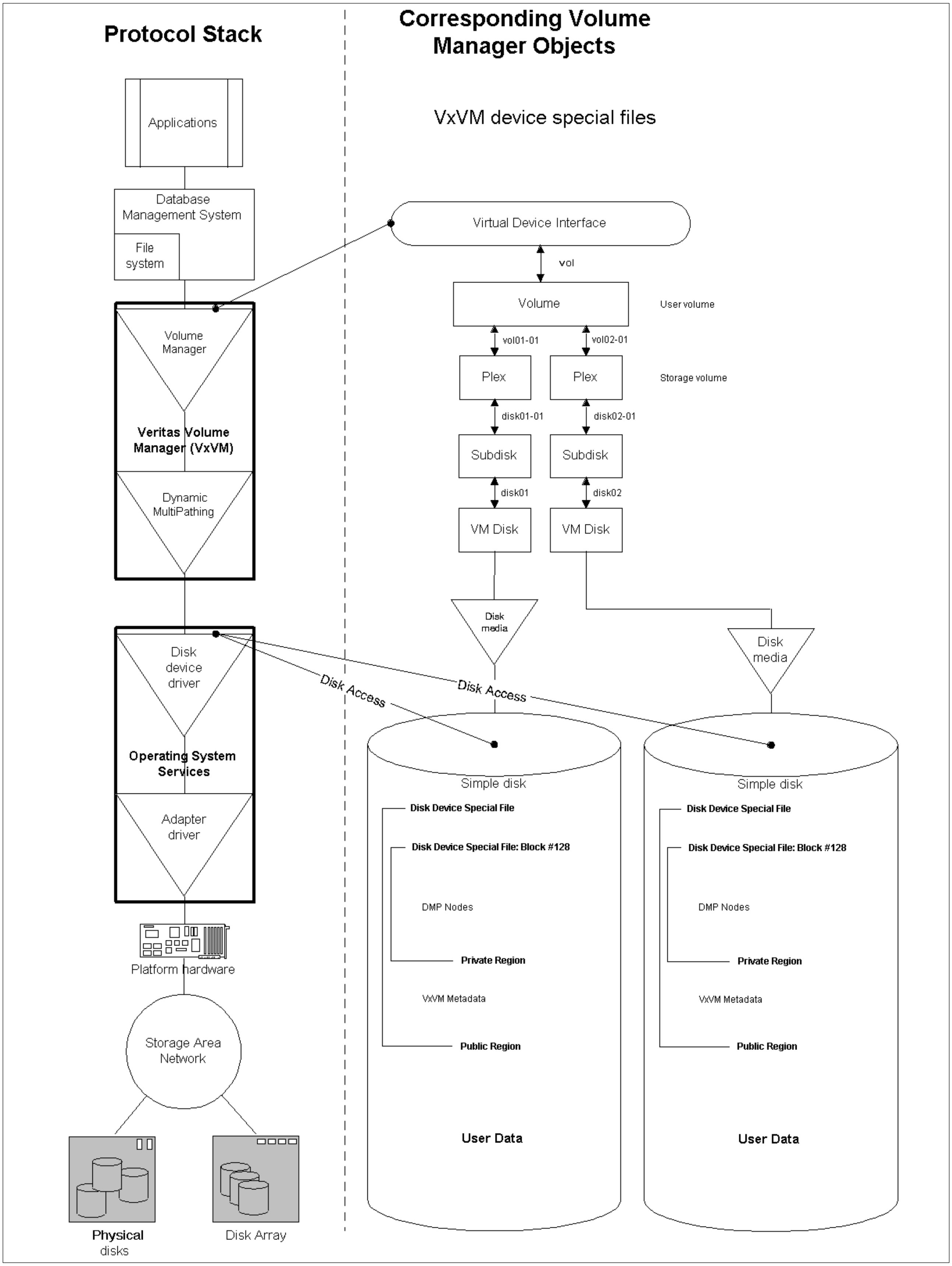


Figure 3-9 VxVM concepts

3.3.1 Supported disk arrays

Disk arrays supported for Veritas VxVM and Dynamic Multipathing are as follows.

- IBM Enterprise Storage Server™ 2105 (ESS)
- EMC Symmetrix™
- HP SureStore™ E Disk Array XP256 and XP512
- Hitachi Data Systems™ 5700E Disk Array Subsystem™
- Hitachi Data Systems 5800E/7700E Disk Array Subsystem™
- Sun StorEdge A5x00 Array™
- Sun StorEdge T3 Array™
- JBOD (Just a Bunch of Disks)
- SEAGATE disks that return unique serial numbers in standard SCSI inquiry data
- Storage Computer™ OmniRaid™ disk array
- ECCS™ Synchronix™ Array

3.3.2 Path failover and load balancing

In the case of active path failure, Veritas DMP automatically selects the next available path to complete the I/O request. Once the primary path fails, the I/O requests are switched over to the other available path. DMP allows the system administrator to indicate whether the primary path is recovered from the failure. This mechanism is called DMP reconfiguration. DMP reconfiguration also allows to detect new devices added to the system or devices removed from the system configuration. Because Veritas Volume Manager uses operating system subroutines to access and manage disk devices, changes can be recognized properly only if underlaying levels of protocol stack detects the changes.

Load balancing for Active/Passive configurations is not supported to avoid continuous transfer of ownership of LUNs from one controller to another. This always results in I/O performance degradation. This is particularly important in clustered environments to make sure, that in Active/Passive configurations, all hosts in the cluster access disk devices through the same physical path. Otherwise, for this same reason, simultaneous use of multiple paths will decrease I/O performance.

DMP supports load balancing for Active/Active configurations through *balanced path mechanism*. Usually this increases an overall system I/O throughput by utilizing the full bandwidth of all paths. Because many physical disk devices (or LUNs created within a storage array) support track caching, sequential I/O read requests are sent through the same path to utilize the effect of data caching, however, large sequential I/O reads are distributed across many paths to take advantage of I/O balancing.

DMP allows also to manually turn off/on an individual adapter for maintenance or administrative purposes. This is not so important in Active/Active configurations, because I/O requests may go through any available path. For Active/Passive configurations, Volume Manager schedules all I/O requests to the primary path, until it fails. Therefore, to take any administrative or maintenance actions in Active/Passive configuration it is required first to disable the controller. This will automatically switch all I/O requests to another available path on another active adapter.

4

Clustering concepts

This chapter briefly describes cluster concepts and its funcionalities. We introduce what a cluster is, what clusters can do for you (and what they cannot do), the different ways that clusters can be implemented, and cluster installation and configuration with multipathing.

The following topics are described:

- What is a cluster?
- The benefits of clustering
- Types of clusters
- IBM cluster strategy
- Linux clustering
- RS/6000 Cluster Technology (RSCT) overview

For additional information about clustering, see the following redbooks:

- *Exploiting HACMP 4.4: Enhancing the Capabilities of Cluster Multi-Processing*, SG24-5979
- *RS/6000 SP/Cluster: The Path to Universal Clustering*, SG24-5374
- *IBM @server xSeries Clustering Planning Guide*, SG24-5845
- *Universal Clustering Problem Determination Guide*, SG24-6602
- *Linux HPC Cluster Installation*, SG24-6041
- *Installing and Managing Microsoft Exchange Clusters*, SG24-6265

4.1 What is a cluster?

A cluster is a group of servers and other resources that provide a set of network resources to a client like a single system and enable high availability and, in some cases, load balancing and parallel processing.

Clustering is the use of multiple computers, typically PCs or UNIX workstations and servers, multiple storage devices, and redundant interconnections, to form what appears to users as a single highly available system. Clustering can be used for load balancing as well as for high availability. Advocates of clustering suggest that the approach can help an enterprise achieve 99.999 percent availability in some cases. One of the main ideas of clustering is that, to the outside world, the cluster appears to be a single system.

A common use of clustering is to load balance traffic on high-traffic Web sites. A Web page request is sent to a "manager" server, which then determines which of several identical or very similar Web servers to forward the request to for handling. Having a server farm (as such a configuration is sometimes called) allows traffic to be handled more quickly.

Clustering has been available since the 1980s when it was used in DEC's VMS systems. IBM, Microsoft, Sun Microsystems, Hewllet-Packard and other leading hardware and software companies offer clustering packages that are said to offer scalability as well as availability. As traffic or availability assurance increases, all or some parts of the cluster can be increased in size or number.

Clustering can also be used as a relatively low-cost form of parallel processing for scientific and other applications that lend themselves to parallel operations.

4.2 The benefits of clustering

Clustering means that the client is isolated and protected from changes to the physical hardware, which brings a number of benefits. Perhaps the most important of these benefits is high availability. Resources on clustered servers act as highly available versions of unclustered resources.

If a node (an individual computer) in the cluster is unavailable or too busy to respond to a request for a resource, the request is transparently passed to another node capable of processing it. Clients are therefore unaware of the exact locations of the resources they are using. For example, a client can request the use of an application without being concerned about either where the application resides or which physical server is processing the request. The user simply gains access to the application in a timely and reliable manner.

Another benefit is scalability. If you need to add users or applications to your system and want performance to be maintained at existing levels, additional systems can be incorporated into the cluster. A typical example would be a Web site that shows rapid growth in the number of demands for Web pages from browser clients. Running the site on a cluster would allow the growth in demand to be easily accommodated by adding servers to the cluster as needed.

Buying a large symmetric multiprocessing (SMP) machine and just adding central processing units (CPUs) and memory as demand increases is not a viable long-term solution for scalability. An SMP machine scales very poorly when the number of CPUs increases beyond a certain point that depends on the SMP implementation. The primary bottleneck is the bandwidth available to access the system’s memory. As the CPU count increases, so does the amount of traffic on the memory bus, which eventually limits system throughput. In contrast, a well-implemented cluster can scale almost linearly. This can be seen in Figure 4-1.

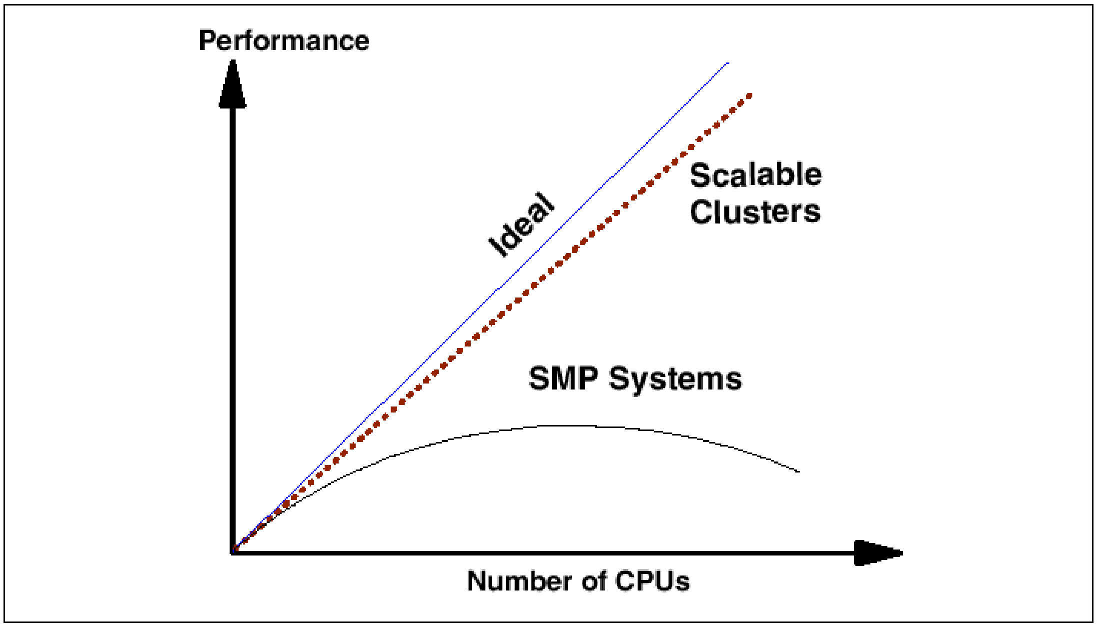


Figure 4-1 Scalable clusters versus SMP machines

In an ideal cluster, users would never notice node failures and administrators could add or change nodes at will. Unfortunately, this is not the case today. Current Intel-based clusters provide many of the features and functions of an idealized cluster but fall short in some areas as we will see in the coming chapters. IBM and others in the industry are working to get closer to the ideal.

4.2.1 Why consider a cluster?

There are several different reasons you might want to implement a cluster. We have already touched upon high availability and scalability and they are reiterated below along with other desirable characteristics of clusters.

High availability

If one node in a cluster fails, its workload is passed to one or more of the other servers in the cluster. Failure of a unclustered server means work comes to a halt.

Scalability

As the demands placed on a system grow, it will begin to suffer from overload. With clustering, if you outgrow a configuration, a new node can be added to the cluster with minimal or no downtime.

Performance

Using a cluster for load balancing could allow you to support a larger number of simultaneous users.

Price/performance

A clustered system can get leading-edge performance by linking inexpensive industry-standard systems together.

Manageability

Administrators can use a graphical console to move resources between the different nodes. This is used to manually balance workloads and to unload computers for planned maintenance without downtime (rolling upgrades).

Administrative clustering

For ease of administration, servers can be consolidated and clustered.

4.2.2 What is high availability?

Most companies are concerned about system availability or uptime.

Mission-critical applications, such as e-business servers, cannot afford to suffer downtime due to unplanned outages. However, because computer systems are constructed using components that can wear out or fail, including software, system design must expect such failures and minimize their impact.

Traditionally, companies have used very reliable mainframes to host critical applications. Users and business managers have become used to this level of availability, which is not usually achieved in low-end PC systems. However, the cost/performance characteristics of Intel-based systems are compelling. Therefore, many system administrators now have a task to significantly improve the availability of their PC servers.

Before we discuss ways to increase system availability, we will try to define it. Simply stated, availability is the percentage of time that a system is running and available for access by its users. Availability is calculated only for the hours during which a system is supposed to be available. For example, if your business requires a system to be up from 6:00 a.m. to 11:00 p.m. each day, then downtime for system maintenance from 11:01 p.m. to 5:59 a.m., the next day does not affect your system availability. However, if you host an online store that is open 24 hours a day, seven days a week, each second of downtime counts.

High availability is a relative characteristic: a highly available system will be operational for a higher percentage of the total time it is required to be available than it would be if no special system features or operational procedures were in place. As a reference, normal system availability in a mainframe environment has typically been measured at around 99.5%. For highly available systems, this improves to, perhaps, 99.99% or better. You can reach this level of availability only by eliminating or masking unplanned outages during scheduled periods of operations. To accomplish this, an advanced system design incorporating fault tolerance has to be used. Advanced system design with fault tolerance enables a system to continue to deliver acceptable service in the event of a component failure. To achieve this, the proper configuration of system features and operational procedures have to be in place. The most common method of providing fault tolerance is to provide redundancy of critical resources, either in the same machine or elsewhere on the network, so that the backup can be made available in the event of a failing primary resource.

Some components are able to predict failures and employ preventive measures to avoid them or at least prevent these failures from affecting normal operation of the system. For instance, even in unclustered systems, hard drives using predictive failure analysis (PFA) can alert the system of an impending disk failure, allowing the disk controller to move the drive offline and to replace it with a hot spare without any manual intervention or downtime at all.

Clustering goes one step further. In clustered solutions, major components or subsystems belonging to a node may fail without users being affected. The clustering software detects the failure and makes another instance of the resource available from another system in the cluster. Users, at worst, see a brief interruption in availability of the resource. In many cases, they may be completely unaware that a problem has occurred.

When implementing a cluster solution for high availability, the classifications in Table 4-1 are often used. These levels are sometimes referred to by the number of nines in the Percent Available column. For example, a four 9s solution means you will only suffer a little under an hour's downtime per year; a five 9s solution reduces this to about five minutes. The more 9s you want, the more you will have to initially invest in your system. You will have to make a business judgment, balancing the cost of downtime against this investment.

Table 4-1 System availability classification

Percent Available	Downtime/year	Classification
99.5	3.7 days	Conventional
99.9	8.8 hours	Available
99.99	52.6 minutes	Highly Available
99.999	5.3 minutes	Fault Resilient
99.9999	32 seconds	Fault Tolerant

High availability is important for almost every industry in today's business world. Failure of a critical IT system can quickly bring business operations to a grinding halt, and every minute of downtime means lost revenue, productivity, or profit. While avoiding downtime is not a new requirement for businesses, its importance is emphasized by business strategies that are either based on or incorporate enterprise resource planning (ERP) and e-business applications.

There is a growing demand for solutions with increased availability that allow businesses to be up and running, 24 hours a day, 365 days a year without interruption. Without high availability, these businesses do not operate at their full potential. For the worst case, the cost of downtime can be enough to put a company out of business. Table 4-2 indicates some estimated costs of downtime for different types of applications.

Table 4-2 Downtime costs by application

Application	Cost per minute
Call location	$29,300
ERP	$14,300
Supply chain management	$12,000
E-commerce	$11,000
Customer service center	$4,200
ATM/POS/EFT	$3,800

As an example, take the price of downtime per minute for an e-commerce application from Table 4-2, which is $11,000 and total that against a 99.9% availability figure; 8.8 hours is 528 minutes of downtime. At $11,000 per minute, this is a total of 5.81 million dollars, expensive for any company and potentially catastrophic.

IBM offers a 99.9% Guarantee Program. This program, which requires specific hardware configurations and specific installation and maintenance services, offers a very high level of availability of the physical and logical layers of the solution. See the following Web page for more information:

`http://www.pc.ibm.com/ww/netfinity/999guarantee.html`

4.2.3 Server consolidation

Server consolidation can be approached in three ways according to the Gartner Group. These are: logical consolidation, physical consolidation, and re-centralization.

Logical consolidation

Logical server consolidation normalizes the operational server environment for procedures such as backup and user maintenance. The benefits from this are a reduction in administrative staff or the local administrator's workload and, at the same time, having the lowest overall associated risk while providing a reasonable return on investment (ROI).

Physical consolidation

Servers are relocated into a centralized data center to be racked and stacked, allowing improved physical security and capacity planning across the servers and better sharing of peripherals. This in turn means reduced hardware, packaging, and cabling costs.

Re-centralization

A number of servers are collapsed into a single, more powerful and larger server. This process can be iterated to reduce the total number of servers in an organization by a significant factor.

An obvious benefit of this is the possibility of reducing the total unused capacity of the replaced servers, but it also has a number of spin-offs:

- Operating system consolidation
- Reductions in the number and complexity of software licenses
- Application instance consolidation
- Reduction in the number of application versions supported

4.3 Types of clusters

There are several different ways you can categorize a cluster:

- Is the cluster technology software or hardware based?
- Does the cluster operate generally (as part of an operating system) or is it for a specific application?
- What kind of hardware approach to data clustering is used?

Today’s Intel-based clusters utilize a number of different approaches. To help you understand the products available, this section discusses some useful ways to classify clustering technologies.

4.3.1 Software for clusters

Depending on what you are trying to accomplish and the availability of suitable products, there are different methods of implementing your cluster.

From a software perspective, the primary types of clustering available are:

- At the operating system (OS) level

  Clustering software is either directly built into the operating system, or it is a middle ware product that adds the function to a base operating system.

  Although these clusters often include an application programming interface to allow applications to take advantage of clustering features, an important aspect of these products is that many existing applications can also gain the benefits of clustering without any modification.

- At the application level

  Most applications are written to run on a single machine. Some, however, particularly those intended for execution on a server, are written to take advantage of the multiple processors available in a symmetric multiprocessing (SMP) machine. An SMP-aware application divides its tasks into separate threads that can be executed in parallel with each other. The SMP machine's operating system then distributes the application threads among the system's processors.

  The problem is, SMP machines eventually run into performance bottlenecks that prevent them from scaling as processors are added. Clustering is regarded as the way to improve performance beyond that attainable through SMP. However, today's Intel-based servers offer little in the way of scalability when clusters are implemented in the operating system.

  As a way of providing the advantages of clustering in this marketplace, several server application vendors have implemented proprietary forms of clustering within their applications.

- A combination of OS and application clustering

4.3.2 Hardware for clusters

Hardware approaches to providing storage within a cluster also give us a way to classify clusters. The two most common cluster types are the shared disk cluster and the shared nothing (sometimes referred to as partitioned) cluster.

Shared disk

Disk storage is provided by a common disk subsystem that can be accessed by all cluster members. The clustering software manages disk accesses to prevent multiple systems from attempting to make changes to the same data simultaneously.

Shared nothing

Each cluster node has its own disk storage space. When a node in the cluster needs to access data owned by another cluster member, it must ask the owner. The owner performs the request and passes the result back to the requesting node. If a node fails, the data it owns is assigned to another node or another set of nodes in the cluster.

Symmetric multiprocessing (SMP) systems have overhead associated with managing communication between the individual CPUs in the system that eventually means that SMP machines do not scale well. In a similar way, adding nodes to a cluster produces overhead in managing resources within the cluster. Cluster management data has to be transferred between members of the cluster to maintain system integrity. Typically, cluster nodes are linked by a high-speed interconnect that carries a heartbeat signal for node failure detection and cluster-related data. However, careful design of clustering software, coupled with efficient intracluster communication, can minimize these overheads so that the linear scalability of an ideal cluster can be approached.

As already suggested, the disk subsystem and intracluster connections are two important elements of clustering. To date, these have generally been provided by extensions of mature technology. For example, a typical disk subsystem for clustering can be formed by having a common SCSI bus between two systems. Both systems are able to access disks on the common bus.

Similarly, the interconnect is typically implemented with a dedicated 100 Mbps Ethernet link. As the development of faster and more flexible systems continues, and the demand for clusters supporting more than two nodes grows, high-speed centralized disk subsystems (storage area networks or SANs) and switched interconnects will become increasingly common.

4.3.3 Active and passive servers

Nodes in a cluster can operate in different ways, depending on how they are set up. In an ideal two-node cluster, both servers are active concurrently. That is, you run applications on both nodes at the same time. In the event of a node failure, the applications that were running on the failed node are transferred over to the surviving system. This does, of course, have implications on server performance, since the work of two nodes now is handled by a single machine.

A solution for this is to have one node passive during normal operation, stepping into action only when the active node fails. However, this is not a particularly cost-effective solution, since you have to buy two servers to do the work of one. Although performance in the failure mode is as good as before the failure, the price/performance ratio in normal operation is comparatively high.

We, therefore, have another way we can usefully classify clusters (particularly two-node clusters).

Active / active

This is the most common clustering model. It provides high availability and acceptable performance when only one node is online. The model also allows maximum utilization of your hardware resources.

Each of the two nodes makes its resources available through the network to the network's clients. The capacity of each node is chosen so that its resources run at optimum performance, and so that either node can temporarily take on the added workload of the other node when failover occurs.

All client services remain available after a failover, but performance is usually degraded.

Active / passive

Though providing maximum availability and minimum performance impact on your resources, the active/passive model requires a fully equipped node that performs no useful work during normal operation.

The primary (active) node handles all client requests while the secondary (passive) node is idle. When the primary node fails, the secondary node restarts all resources and continues to service clients without any noticeable impact on performance (providing the nodes are themselves comparable in performance).

Hybrid

A hybrid model is a combination of the two previous models. By enabling failover only for critical applications, you can maintain high availability for those applications while having less critical, non clustered applications conveniently running on the same server in normal operation.

In a failover situation, the less critical applications that were running on the failed server become unavailable and do not have any adverse impact on the performance of the surviving applications. You can therefore balance performance against the fault tolerance of your entire application suite.

4.4 IBM cluster strategy

There are technical challenges in implementing effective Intel CPU-based clusters. Hardware manufacturers have to develop high-speed interconnect methods, efficient storage subsystems and powerful processor complexes. Software designers need to provide clustering versions of operating systems, middleware layers (such as, databases, online transaction processing (OLTP), and decision support), and applications. Importantly, this has to be achieved while conforming to industry standards and price points.

To address these challenges, IBM has developed a three-pronged clustering strategy:

- Migration of established technologies from IBM's high-end clustering portfolio onto the Intel platform to drive the industry in the development and exploitation of the necessary technology.
- Help establish and lead industry efforts to provide open, industry-standard cluster solutions.
- Provide solutions to customers across major operating system and application platforms.

IBM range of Intel-based server clusters offer key advantages to customers:

- High availability systems from cost-effective mainstream servers to high-performance enterprise class systems
- Support for Windows 2000 Advanced Server, Windows NT 4.0 Enterprise Edition, NetWare, Windows 2000 Datacenter Server and Linux (as a new operating system)
- A wide choice of disk subsystems and connectivity options
- Industry-standard implementations
- Enhanced cluster system management capability
- Worldwide service and support

Server cluster—definition: A server cluster is a group of computers acting as server and housed together in a single location. A server cluster is sometimes called a server farm.

4.5 Linux clustering

With the adoption of Linux as a mature server operating system by IBM in early 2000, what had been a relatively obscure “hacking” project, suddenly became the talk of the IT world. The approach taken by IBM towards Linux has “legitimized” Linux in the eyes of IBM’s more traditional customers, and has caused these customers to think seriously about Linux for the first time.

Linux now offers an alternative server operating system and is an ideal match for the IBM range of Intel-based servers — the xSeries and Netfinity systems.

In this section we study specific Linux solutions for creating clusters of machines to provide high availability configurations: software solutions using Linux to provide higher availability with multiple machines than with single server solutions. The combination of the Linux operating system, sophisticated and reliable software clustering, and xSeries and Netfinity hardware offers high availability at a low price. Even in an enterprise environment, where an obvious choice for a highly reliable back-end database server would be the zSeries Parallel Sysplex environment, for example, Linux high availability clustering solutions can provide a reliable front-end Web server.

The two primary benefits of a Linux high availability cluster are:

- Fault tolerance — if a single Linux server in a cluster should fail, then the server function of the total cluster solution is not impacted.
- Scalability — as workload demands grow it should be possible to add machines to an existing cluster to handle the load. This compares with a single-box solution in which at some point a total hardware replacement is required to upgrade the server function.

The three typical ways clustering is used in a Linux environment are: high performance computing or scientific computing; load balancing or scalability; and high availability and failover.

High performance computing or scientific computing

The most commonly known implementation of Linux clustering is Beowulf. The Beowulf Project implements high performance computing (HPC) using message-passing parallel programs. To really use the Beowulf concept, your application has to be written (or rewritten) using parallel virtual machine (PVM) or message passing interface (MPI). At least you should be able to run the processing in parallel using shell script front ends, so that each node works at a specific range of the whole task.

Beowulf is not a single software package — instead, it consists of different parts (PVM, MPI, Linux kernel, some kernel patches, etc.). You can get more information about Beowulf at:

http://www.beowulf.org/

Load balancing or scalability

This is a very important topic for any fast growing business, which most of today’s e-business sites are. Most of these sites start small with only a few Web servers and a back-end database. So when they grow, they have to change their hardware more often, as their number of customers and the number or level of services they provide increases. Changing or upgrading your hardware means outages, downtimes, and lost money and it does not look professional nor does it provide the kind of service your business needs to grow.

With a load-balancing cluster, you can just add another box into the cluster if the demand or load you get increases. If one server fails, just change the cluster configuration automatically and take the broken server out for service. Later you can reintegrate this server or a replacement box back into the cluster again.

Most of today’s Linux load-balancing cluster solutions are based on the Linux Virtual Server (LVS) project and one of the major products implementing LVS is TurboLinux Cluster Server. For more information on LVS, see:

http://www.linuxvirtualserver.org/

Another approach to load sharing and distributed computing is called MOSIX (the Multicomputer OS for UNIX). It allows you to run processes distributed on a collection of clustered nodes transparently. MOSIX migrates processes from very loaded nodes to other less loaded nodes dynamically and scales very well. No special support from the application is necessary. It simply looks like normal SMP but with more than one physical box.

Actually, MOSIX does not exactly fit in any of these categories. It's something between HPC and load sharing, but currently does not provide improved additional availability. For more information see:

http://openmosix.sourceforge.net

High availability and failover

High availability is also part of load balancing as discussed above and is known as an active/active configuration (where all nodes are doing real or active work). However, high availability can also be configured as active/passive — in Linux, this concept is called the Fail Over Service (FOS).

With two-node FOS systems, you have one master and one standby system. In normal operation the master is running your service or application and the second system is just watching the master. If the master fails, the second system takes over the service immediately and shuts the master down (if this has not already happened). This provides you with a highly available system.

FOS is also provided by the Linux Virtual Server project. One currently available commercial product is Red Hat HA Server and the next release of TurboLinux Cluster Server will provide FOS as well.

4.5.1 Implementing Linux clustering

In the following sections, we examine the latter two aspects of Linux clustering as they are implemented in the Linux Virtual Server:

- High availability and failover
- Load balancing

Note: While both Fail Over Service (FOS) and load balancing are provided by the Linux Virtual Server project, most people refer to LVS when talking about load balancing, and FOS when they mean real high availability, and so do we.

As with other clustering implementations, such as Microsoft Cluster Server, a Linux cluster has the following components and characteristics:

- A heartbeat connection between the nodes.

  With a two-node cluster this can be a simple crossover Ethernet connection, but with three or more nodes, a private switched Ethernet network is recommended, either 100-BaseT or Gigabit Ethernet.

- Separate network connections to the network for normal data traffic.

- The cluster as a whole is virtualized as one single server complete with one or more virtual IP addresses.

4.5.2 Failover service

Failover means that we have two servers, one primary or master node and one secondary or backup node. Both know about each other via heartbeat and are attached to the client network as well, as shown in Figure 4-2.

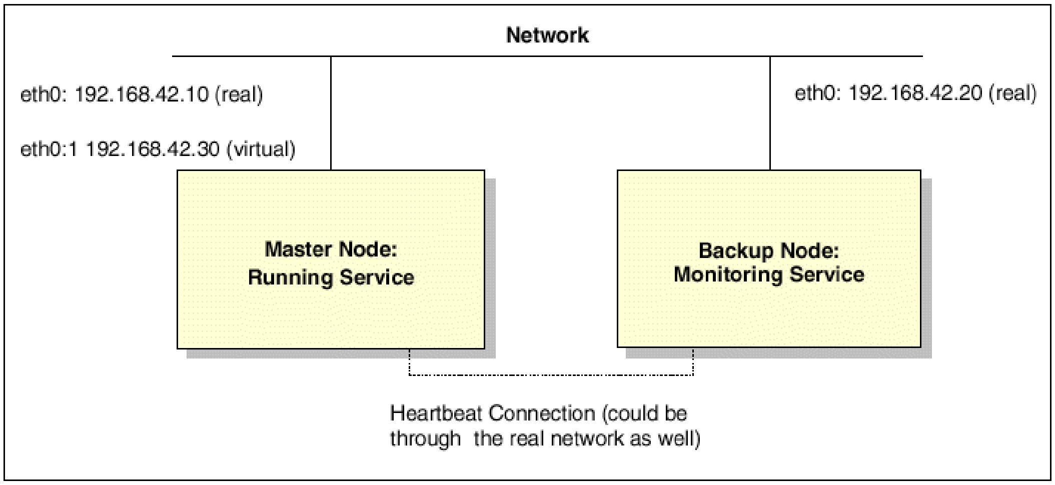

Network
eth0: 192.168.42.10 (real)
eth0:1 192.168.42.30 (virtual)
eth0: 192.168.42.20 (real)
Master Node:
Running Service
Backup Node:
Monitoring Service
Heartbeat Connection (could be
through the real network as well)

Figure 4-2 Failover service in normal operation

In normal operation, the master server is running and providing the service, a Web server, for example. The backup node monitors the master such as by trying to connect to the master server's HTTP port (80) every 10 seconds and retrieve a Web page. Heartbeats are exchanged by both servers. As the picture implies, both servers have a real IP address assigned (192.168.42.10 for the master and 192.168.42.20 for the backup, in this case) to their real interfaces (eth0). As the master node is active, it gets a second IP address, the virtual cluster IP address. In Linux terms, this IP address is an alias address (eth0:1) defined on top of the real network interface (eth0).

Both real and virtual interfaces can be seen via Address Resolution Protocol (ARP), responsible for the IP to MAC address mapping. Actually, both eth0 and eth0:1 share the same MAC address, which is why eth0:1 is called an alias.

What happens if the service on the master server becomes unavailable or the server itself goes down? This situation will be noticed via monitoring (if only the service fails) or via heartbeat (if the complete machine goes down).

Figure 4-3 shows what will happen then.

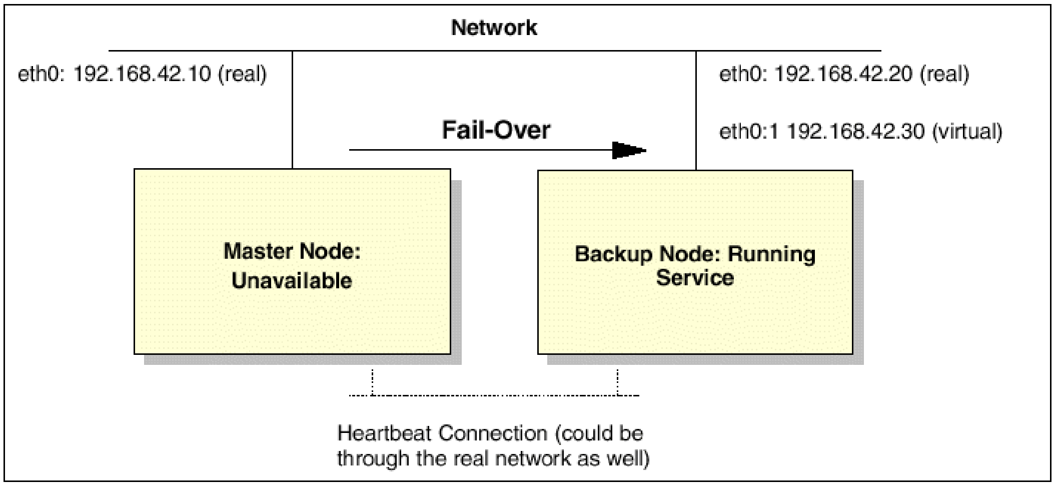

Network
eth0: 192.168.42.10 (real)
eth0: 192.168.42.20 (real)
Fail-Over
eth0:1 192.168.42.30 (virtual)
Master Node: Unavailable
Backup Node: Running Service
Heartbeat Connection (could be through the real network as well)

Figure 4-3 Failover service showing actual failover operation

The backup node takes over the virtual cluster IP address from the master node and gets its aliased eth1:0 up and running. After that it starts the service that was originally available on the master node and everything is fine again. This process is called failover.

As the virtual IP address is transferred to another real network interface, its associated MAC address changes too. To get this change reflected to all other computers on the network, the new active (and former backup) node broadcasts an ARP message for the IP address of the cluster containing the new MAC address. This process is known as gratuitous ARP or courtesy ARP and enables the other machines on the network to update their ARP tables with the new MAC address of the cluster.

If now the master becomes available again (observed via heartbeat), also called failback or fallback can take place (see Figure 4-4). The backup node stops running the service, the master node takes over the virtual IP address, issues the gratuitous ARP broadcast and starts its service. At this time everything looks like no failover had happened at all.

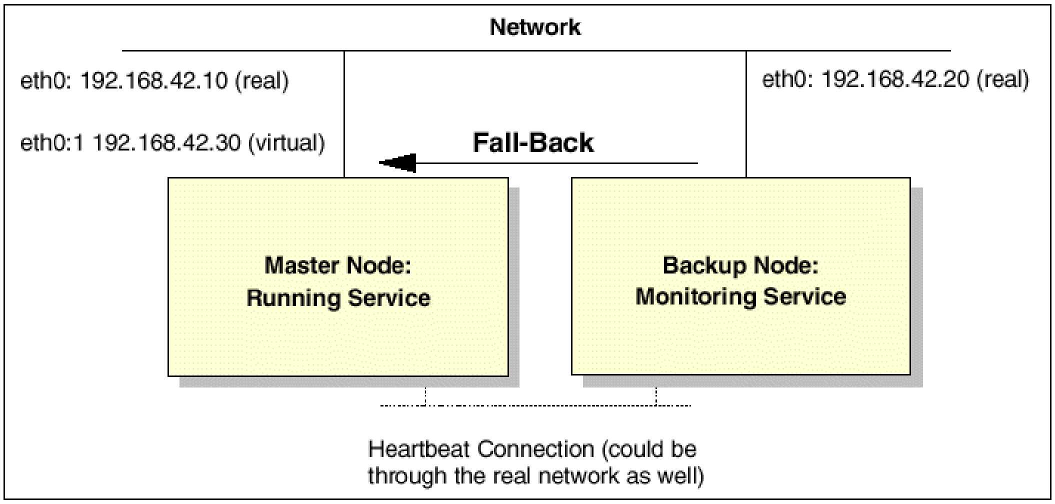

Network
eth0: 192.168.42.10 (real)
eth0: 192.168.42.20 (real)
eth0:1 192.168.42.30 (virtual)
Fall-Back
Master Node: Running Service
Backup Node: Monitoring Service
Heartbeat Connection (could be through the real network as well)

Figure 4-4 Failover service, resumption of normal operation

With the current Linux product implementations, some restrictions apply:

- Only two node FOS configurations are supported.

- No selective failover (for individual services) is possible; all services are monitored and failover as a group.

4.5.3 Load balancing

Load balancing works similar to Fail Over Service, but aims at scalability and reducing system outages. It spreads incoming traffic to more than one server and lets all these servers look like one large server. It uses heartbeats like FOS, but implements another concept, unique to load balancing: traffic monitors or managers. A very simple LVS setup is shown in Figure 4-5. There is no dedicated, internal cluster network; all machines are connected to the same physical network.

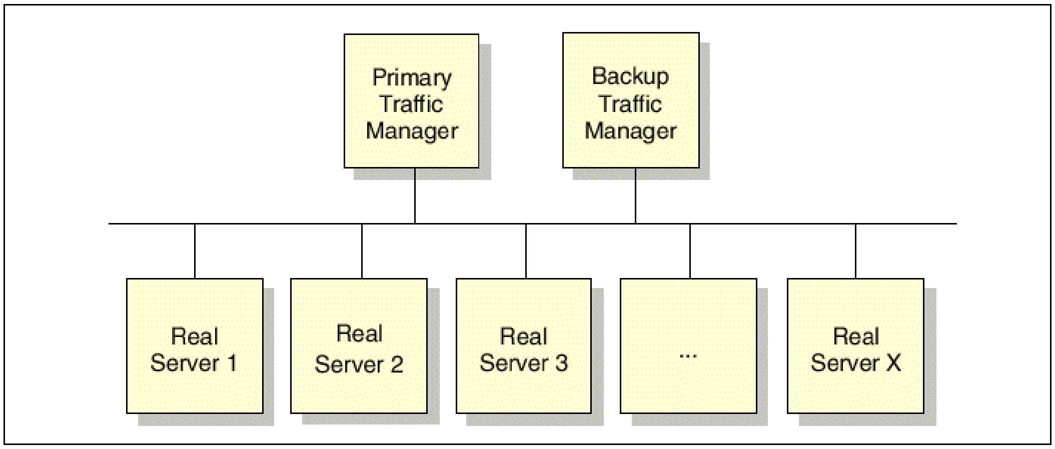


Figure 4-5 Simple Linux Virtual Server setup

As with FOS, there's a virtual server formed out of individual boxes. The primary and backup traffic manager behave like a FOS cluster concerning network connection and heartbeat service. The active traffic manager gets the virtual IP address assigned and redirects the incoming traffic to the real servers, based on the chosen load balancing and routing scheme. The traffic manager monitors the real servers for heartbeat, service, and load (if supported).

Scheduling mechanisms for distributing the incoming traffic can be one of the following, depending on the product:

- Round robin — all traffic is equally distributed to all real servers.
- Least connections — more traffic is distributed to real servers with fewer active connections.
- Weighted round robin — more traffic gets distributed to the more powerful servers (as specified by the user) and dynamic load information is taken into account.
- Weighted least connections — more traffic is spread to the servers with fewer active connections (based on a user-configured capacity) and dynamic load information is taken into account.

The next steps are to get requests from the traffic manager to the cluster nodes and then respond to the clients.There are three options, depending on the product you use:

- Direct routing
- Network address translation
- Tunneling

Direct routing

In Figure 4-6, the client accesses the virtual server (192.168.42.30). Its traffic gets routed to the traffic manager, which redirects it to the real server by simply changing the MAC address of the data frame and retransmitting on the LAN. The real server itself has a physical network interface (eth0) for the incoming traffic and one aliased, ARP-hidden network interface (lo:0) for the outgoing traffic.

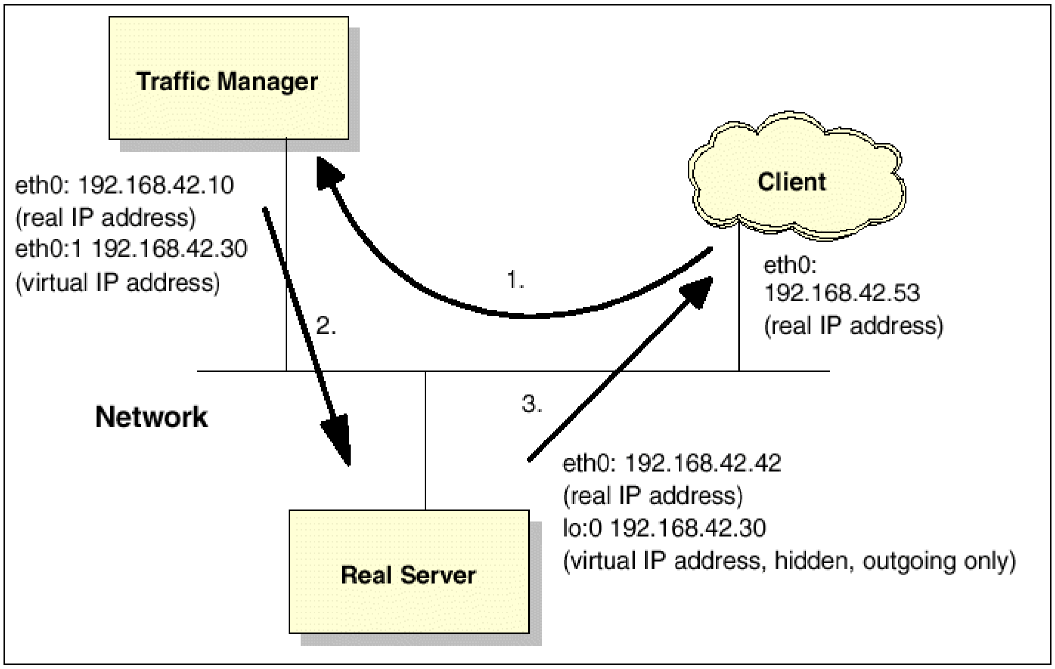


Figure 4-6 Direct routing of returned packets

So the real server sends the response back directly to the requesting Client 1 using lo:0 as its source address, therefore using the virtual IP address. From the perspective of the client, an IP packet has been sent to the virtual server's address and a response has been received from the same address. The client never sees any response to its request as coming from the server's "real" eth0 address. It only sees the virtual IP address.

The lo:0 address in Figure 4-6 is called a "hidden" address because it must be configured in such a way that the server owning this network interface will not respond to ARP requests for the IP address. The only network device that should respond to ARP requests is the traffic manager. The traffic manager determines which actual server is to be used for the received packet and forwards the packet to the server by re-transmitting the received packet onto the network, but with the destination Layer 2 MAC address of the packet now being the MAC address of the desired server.

The server will receive the packet, because it is now destined to its hardware MAC address, and will examine the packet and discover that it contains an IP packet destined for an IP address known to the server as its internal "hidden" IP address. It will then pass the packet to the IP application (such as a Sockets application) bound to this IP address. The application will respond and the same IP address will be used as the source address in the response, and the response packet will be sent out over the network directly to the client. The response does not pass through the traffic manager.

In our cluster implementations we examined all-Linux environments, in which the traffic managers and the servers themselves are running the same distribution of Linux code; this certainly eases implementation of the clusters but it should be noted that other operating system environments such as Windows 2000 or even OS/390 can be used for the server environments in a Linux cluster. The only requirement is that the servers themselves must be configured with both "real" and "hidden" IP addresses in a similar manner to Linux servers.

Because only the traffic manager responds to ARP requests for the IP address of the cluster, a full implementation of a load-balancing cluster environment will include a backup traffic manager as shown in Figure 4-5. There will now be an additional requirement for the backup traffic manager to maintain cluster state information such as information on the state of open TCP connections into the cluster, and this information will allow the backup traffic manager to take over operation of the cluster without disrupting existing connections.

Although not shown explicitly in Figure 4-5, the traffic manager function can also reside on the same physical server as one of the "real" servers. The function can be "co-located" with the server itself. This reduces the total number of machines required to implement a load-balancing cluster if this is an issue. A cluster could be implemented on only two machines, with one machine acting as the primary traffic manager and the other as the backup traffic manager and with the server functions themselves residing on the same machines.

This basic configuration is the easiest and fastest solution to implement, but has one major disadvantage: the traffic manager and the real servers must have interfaces to the same physical LAN segment. As traffic to the cluster increases, this may lead to congestion. Each packet inbound to the cluster appears on the network twice (once to the traffic manager from outside and once from the traffic manager to the actual server) and then each response packet also crosses the same network.

It's a good idea to have a separate internal cluster network where possible like the one shown in Figure 4-7. In this network traffic between the traffic managers and the servers flows over the private network, and this network could also be used for the flow of heartbeat information, meaning that all intracluster network traffic is isolated from the external client network environment.

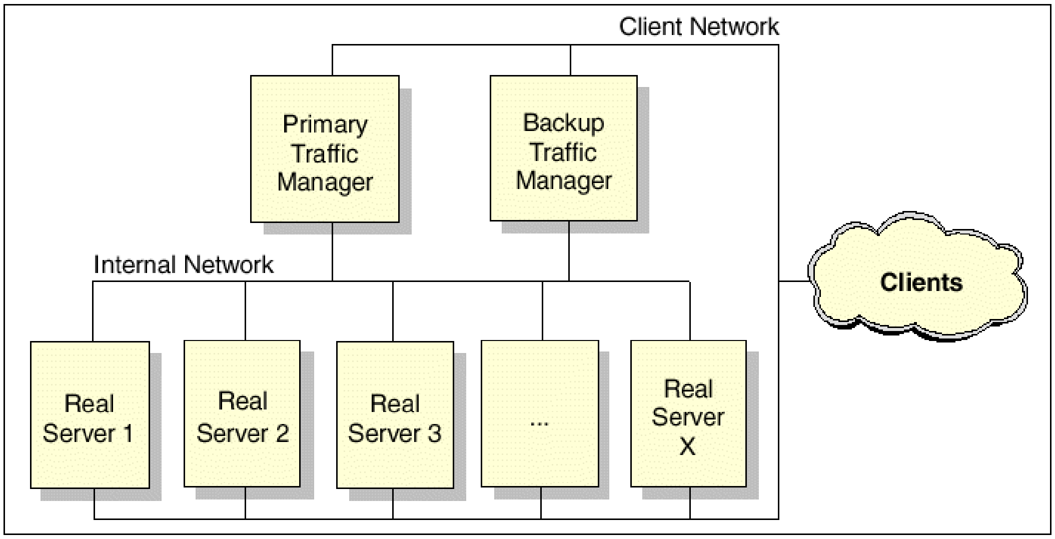


Figure 4-7 More sophisticated setup using an internal cluster network

Network Address Translation

Another option for hiding the internal cluster network is called Network Address Translation (NAT). NAT requires the traffic managers to take on one more job role; they have to translate the IP addresses of incoming traffic to direct it to one of the real servers and on the way back they have to re-translate the IP addresses of the outgoing traffic. Unlike the previous configurations, this requires that both inbound and outbound traffic have to flow through the traffic manager. Figure 4-8 shows this process.

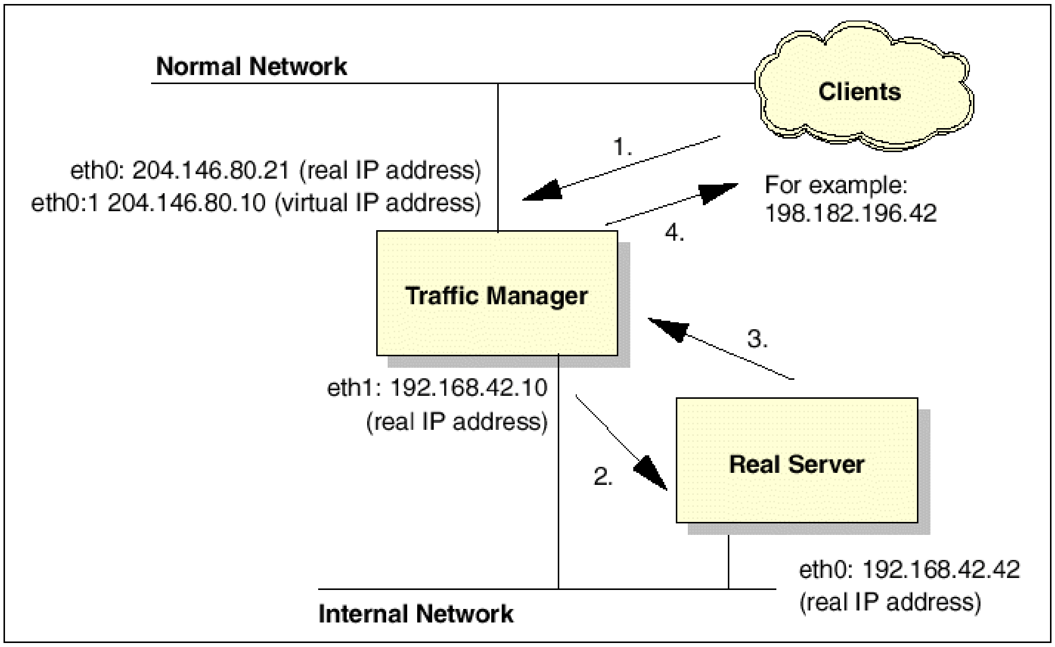


Figure 4-8 Network address translation

When the client talks to the virtual server represented by the traffic manager, its traffic looks like:

Source = 198.182.196.56
Destination = 204.146.80.10

Now the traffic manager selects a real server for this traffic and after translating the addresses passes on the traffic:

Source = 198.182.196.56
Destination = 192.168.42.42

After the real server does its job, it sends back the response using:

Source = 192.168.42.42
Destination = 198.182.196.56

Finally the traffic manager forwards the traffic to the outside world after a retranslation:

Source = 204.146.80.10
Destination = 198.182.196.56

The translation is done inside the traffic manager using a hash table and IP address-port mappings.

This is a very convenient way to implement a cluster because it only requires a single external IP address and all the destination servers can be defined on an internal private IP network. It does have one really significant disadvantage, mentioned above: all outgoing traffic has to pass through the traffic manager now. One of the justifications for permitting inbound traffic to pass through the traffic manager in a basic cluster is that outgoing traffic is usually much more significant in volume than the incoming traffic. This is because incoming requests such as HTTP requests are small in comparison to the volume of traffic sent back in response. Your traffic manager finally becomes the bottleneck of your cluster, and so NAT is suitable for a smaller cluster environment with not too much expected traffic.

And, in any case, what's to stop the servers in Figure 4-6 from being configured with "real" IP addresses in a private IP network? There is no reason why even a simple cluster environment with a single network address cannot implement multiple IP networks over the same physical infrastructure. So NAT may not be required even in cases where multiple external IP addresses are not possible.

However, NAT has one other major attraction in that the destination servers themselves do not need to be configured with a "hidden" IP address at all. In the early days of clustering this was a problem on certain operating systems, and certainly adds complexity to the server configuration process even today. The NAT solution means that absolutely any IP server platform can be used as the target servers in a cluster without having to consider the quirks of IP addressing using "hidden" IP addresses configured on loopback interfaces.

Tunneling

Another interesting option for building up a LVS cluster is to use IP tunneling. It allows you to cluster real servers spread around the world, being part of different networks. But it needs the support of IP tunneling on each server of the cluster. Figure 4-9 shows the setup.

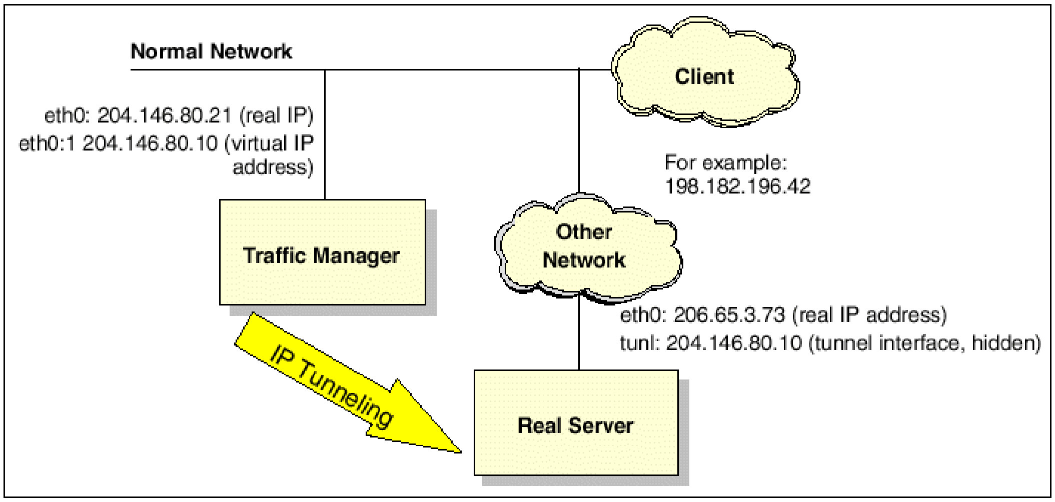

Normal Network
Client
eth0: 204.146.80.21 (real IP)
eth0:1 204.146.80.10 (virtual IP address)
For example:
198.182.196.42
Traffic Manager
Other Network
eth0: 206.65.3.73 (real IP address)
tunl: 204.146.80.10 (tunnel interface, hidden)
IP Tunneling
Real Server

Figure 4-9 IP tunneling

Here, when a client accesses the virtual server, the client sends a packet to the traffic manager, which advertises the IP address of the cluster and responds to ARP requests for it. Having received a packet from a client, the traffic manager encapsulates the packet into an IP datagram addressed to the real server, forwards it and stores this connection information in its hash table. All subsequent IP packets belonging to this connection end up at the same real server over the same IP tunnel. The real server itself de-encapsulates the packet and responds to the client directly using the virtual IP address as its source address.

IP tunneling is a very flexible way to build up a widespread cluster solution, but depends on the IP encapsulation protocol support of all participating cluster servers/nodes. In current implementations, this requires that all the servers be Linux servers, whereas the other solutions we discussed can use a mix of server operating systems in a single cluster.

4.5.4 Supported services

All IP services using a direct socket connection can be implemented with the current Linux clustering solutions. Here are some examples:

- HTTP
- FTP (INETD)
- SMTP
- POP
- IMAP
- LDAP
- NNTP
- SSH
- Telnet

Services depending on a secondary port connection besides the listening port are not supported.

4.5.5 Sharing the data between nodes

One of the most important aspects of the cluster is the availability to the nodes and the consistency of that data between the nodes. There are, as always, multiple solutions, mostly depending on the frequency of changes to your data and the amount of data involved. We cover the following, starting with the simplest:

- rsync
- Network File System
- Global File System
- Intermezzo
- Back-end databases

rsync

If your content is primarily static Web pages (contact information, for example) or a reasonably small FTP site, you can store all the data locally on each of the actual servers. Then to keep the data synchronized you can simply use a mirroring tool such as rsync that runs periodically, say twice an hour. With this solution you get good availability, since all data is stored on each server individually. It doesn’t matter if one server goes down for some reason, nor do you rely on a central storage server. Figure 4-10 shows this solution.

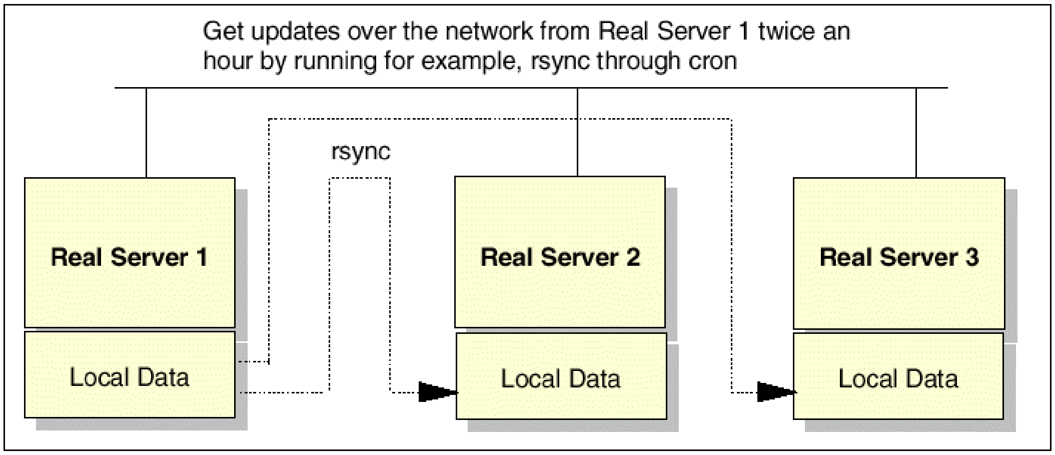

Get updates over the network from Real Server 1 twice an hour by running for example, rsync through cron
rsync
Real Server 1
Real Server 2
Real Server 3
Local Data
Local Data
Local Data

Figure 4-10 Using rsync for local server data synchronization

But this solution will not be suitable if you have really large amounts of data (more than a few gigabytes) changing more often (more than a few times in a week) and you have to keep it synchronized. Now network or distributed file systems come into the picture.

Network File System (NFS)

Again, starting with the simplest approach, we can use NFS, the widely used, commonly known and stable network file system. It's easy to use and requires a central NFS server that exports the shared data. This data is "mounted" by the real servers across the network. This approach looks like the one shown in Figure 4-11.

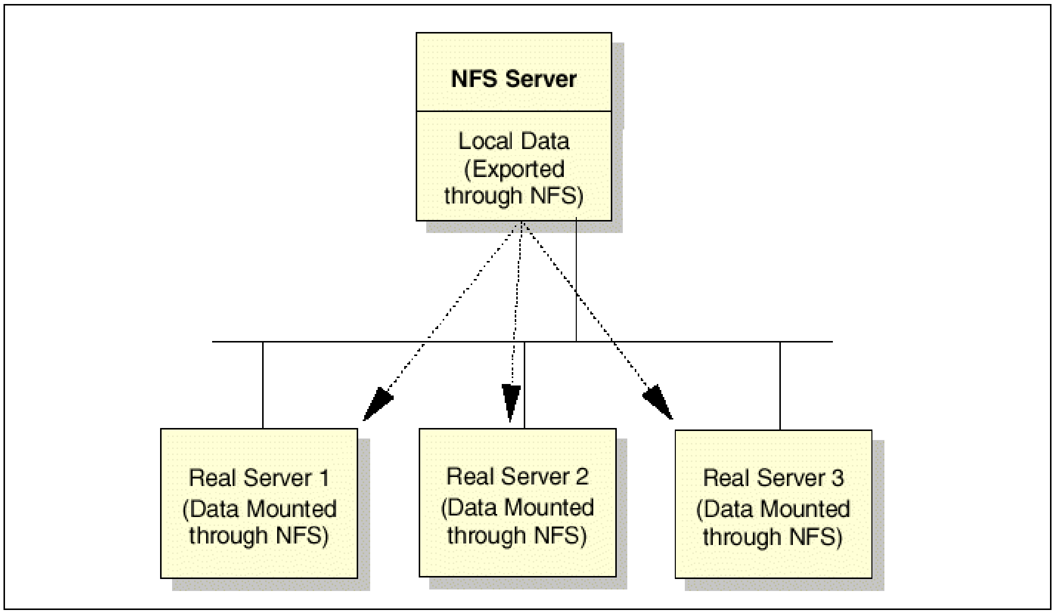

NFS Server
Local Data (Exported through NFS)
Real Server 1 (Data Mounted through NFS)
Real Server 2 (Data Mounted through NFS)
Real Server 3 (Data Mounted through NFS)

Figure 4-11 Using NFS for central data storing

Although NFS is simple to implement and use (on Linux), it has two major drawbacks:

- Slow performance
- Single point of failure

Although the performance may be acceptable for a small cluster solution, you should always be aware that if the NFS server dies then the real servers will no longer be able to get to your data, and therefore will not be able to provide their service. This might make you think about setting up a redundant, highly available NFS server, but that's no trivial thing to attempt; you have to take care of the clients' file handles, keep the clustered NFS servers synchronized, and there is no "out of the box" solution here. So after all, NFS is no real solution for a cluster environment.

That's why there are real cluster-capable file systems, such as the Global FileSystem (GFS) and Intermezzo, which offers different approaches to a cluster file system.

Global File System (GFS)

GFS implements the sharing of storage devices over a network. This includes Shared SCSI, Fibre Channel (FC), and Network Block Device (NBD). The Global File System sitting on top of these storage devices appears as a local file system for each box (Figure 4-12).

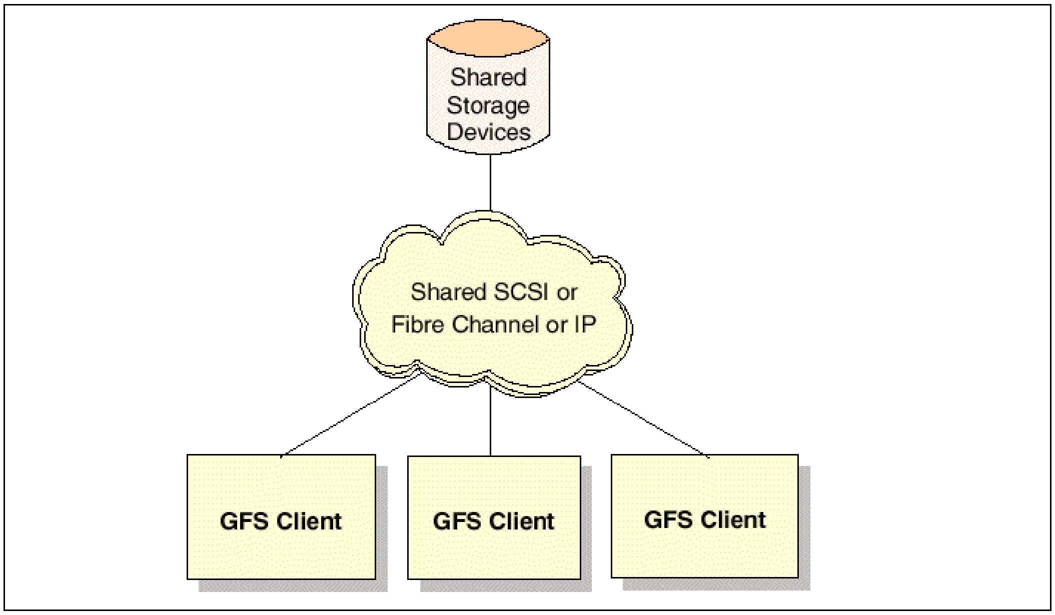


Figure 4-12 Global FileSystem

The Global File System is a 64-bit, shared disk file system focusing on:

- Availability — if one of the clients goes offline, the data can still be accessed by all the other GFS clients.
- Scalability — which means it doesn't suffer from concepts based on a central file server, as NFS does.

Furthermore GFS is able to pool separate storage devices into one large volume and to load balance between the workload generated by all the clients.

To set up GFS, you need to decide on the transport medium to use:

- Shared SCSI (although typically you are limited to clusters of two nodes)
- Fibre Channel
- IP (akin to using tunneling to attach to your client over a traditional network), not yet a widely used option and limited by the network bandwidth but allows you to attach any client without direct FC or SCSI connection to your storage pool

GFS itself implements the storage pooling, the file locking, and the real file system. It's still under development, but should be quite usable already.

Intermezzo

Unlike GFS, which is a shared file system, Intermezzo is an implementation of a distributed file system. This means that there's no central storage pool, but each machine has its own kind of storage locally. The storage gets synchronized via traditional TCP/IP networks.

Intermezzo features a client/server model. The server holds the authoritative data, while the clients only have a locally cached version of the data, which is kept synchronized. Intermezzo even supports disconnected operation and is able to reintegrate when connected again. Figure 4-13 shows a simple Intermezzo configuration

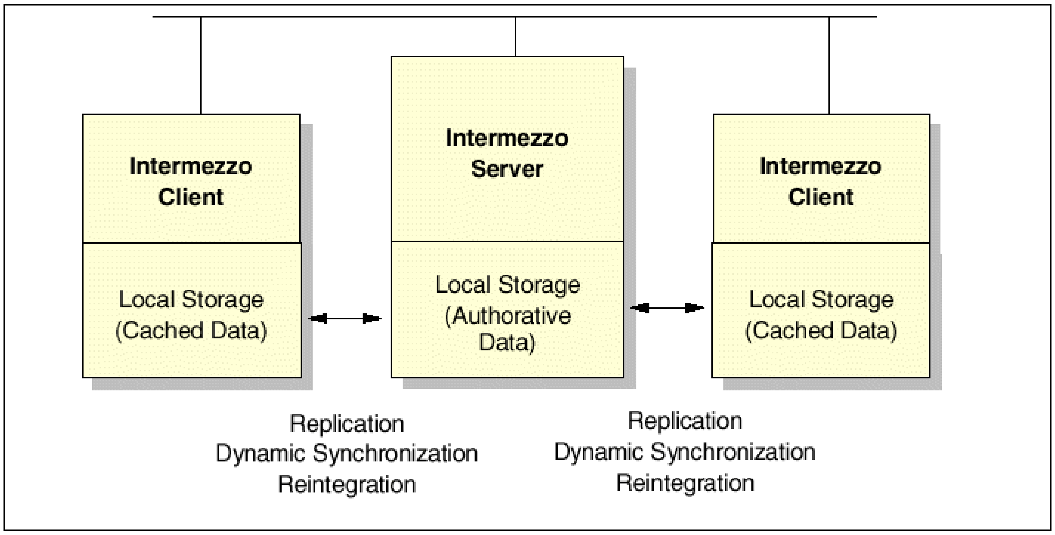


Figure 4-13 Sample Intermezzo setup

Intermezzo uses a traditional file system such as ext2 to hold its data and puts a layer in between that is responsible for journaling updates and keeping the data synchronized. Intermezzo, like GFS, is still under development, but already usable (it requires a Linux kernel recompilation to implement it today).

Back-end database

Another option for storing and accessing your data, and one that you might already have in place, is a back-end database, such as DB2. This database itself can be highly available, but that's not part of the Linux clustering solution. Your real servers simply have to be capable of connecting to this database and putting data into it or getting data from it, using, for example, remote SQL queries inside PHP featuring dynamic Web pages. This is a very convenient and widely used option. An example of such a configuration is shown in Figure 4-14. Consider the back-end database as the existing enterprise database server running on S/390 or other UNIX platforms, for example.

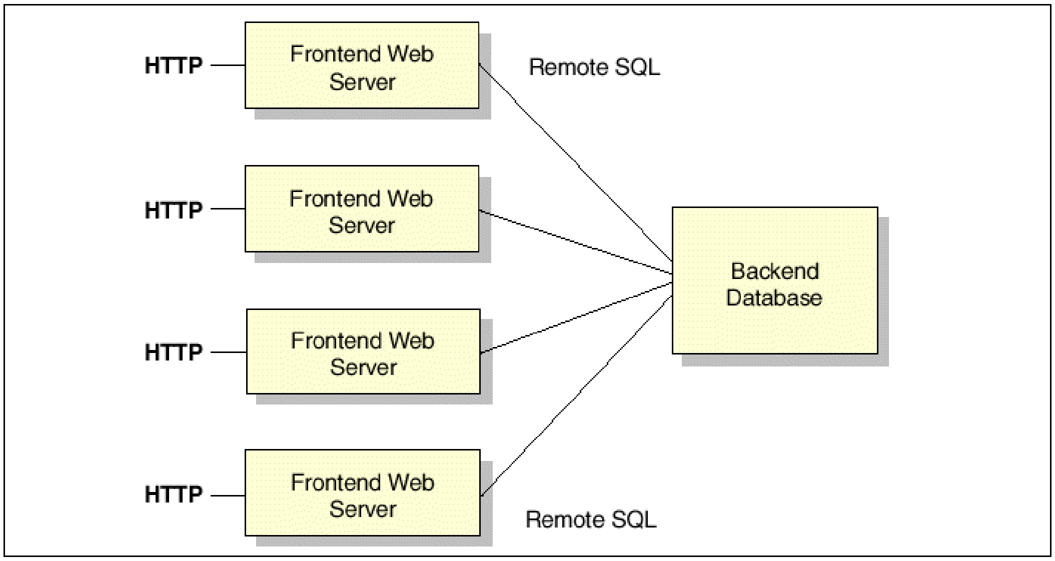


Figure 4-14 Front end clusters example

4.5.6 Putting it all together

After discussing the different aspects of clustering, we can now put all these things together to get a complete picture. The first question to ask is:

Why do we want to do clustering? The possible answers are:

- We want a scalable solution. So we go for *load balancing*.
- We want a highly available solution. So we go for *Fail Over Service*.

The important things about load balancing are:

- Make sure your services are able to run in a cluster environment (single listening TCP/IP port) and can be balanced (servers can act in parallel).
- Keep your system scalable from the point of network technology and cluster implementation.
- Think about a backup traffic manager. Otherwise, all your real servers are useless if the traffic manager dies.
- Based on the amount of data and change frequency, select an appropriate method to access and store your data.

Regarding high availability, consider the following thoughts:

- Make sure you can monitor the services accordingly.
- Think about storing and accessing your data safely and think about availability. There's little sense in building up a high availability Web server if the database it connects to does not offer comparable high availability.
- Be cautious (up to a certain point). For example, consider a second, backup Internet provider if you want to offer Internet services. Otherwise you may end up with a really highly available Internet service locally that is not accessible if your provider goes offline.
- Think about high availability from the hardware side — uninterruptable power supply (UPS).
- Think about disaster prevention, such as putting nodes in separate buildings.

- Finally, don't forget to test your setup on a regular basis.

For more information, see the following:

- Linux HA project: http://linux-ha.org/
- Linux Virtual Server (LVS) project: http://www.linuxvirtualserver.org/
- Red Hat HA server project: http://ha.redhat.com/
- Global File System (GFS): http://www.sistina.com/products_gfs.htm
- Intermezzo: http://www.inter-mezzo.org/

4.6 RS/6000 Cluster Technology (RSCT) overview

RSCT is a distributed group of subsystems, running across multiple nodes or machines, that communicate with each other through multiple networks to provide high availability, online monitoring and automatic recovery actions. This distributed group of subsystems, known as a stack, runs in a single partition on the RS/6000 SP. There may be more than one partition per SP but each RSCT stack is separate from the other.

The three principal components of the RSCT stack are:

- Topology Services (TS)
- Group Services (GS)
- Event Management (EM)

This infrastructure is represented pictorially, as shown in Figure 4-15.

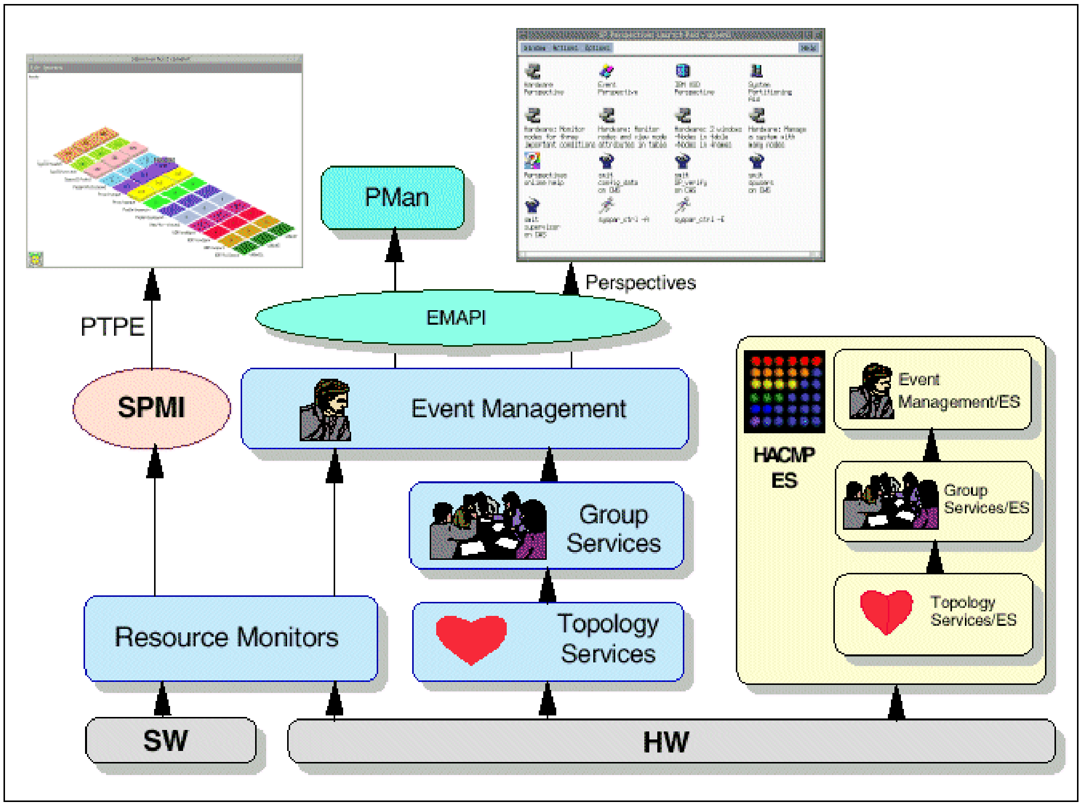


Figure 4-15 RSCT infrastructure

A HACMP/ES domain is also shown in this figure, which you can see contains another RSCT stack of subsystems. These stacks are independent of each other and although the problem determination of the RSCT stack in the HACMP environment is similar, it is not the intent of this book to detail the differences.

Refer to the following IBM Redbooks for more information about HACMP/ES:

- *HACMP Enhanced Scalability Handbook*, SG24-5328
- *HACMP/ES Customization Examples*, SG24-4498
- *HACMP Enhanced Scalability: User-Defined Events*, SG24-5327

4.6.1 Topology Services (TS)

Topology Services (TS) is the lowest level of the RSCT subsystems. It provides and maintains connectivity and availability information about the nodes and network adapters. Most problems are automatically recovered without intervention; however, often to understand or isolate a problem in a higher level of the RSCT system you need to examine the state of Topology Services.

TS overview

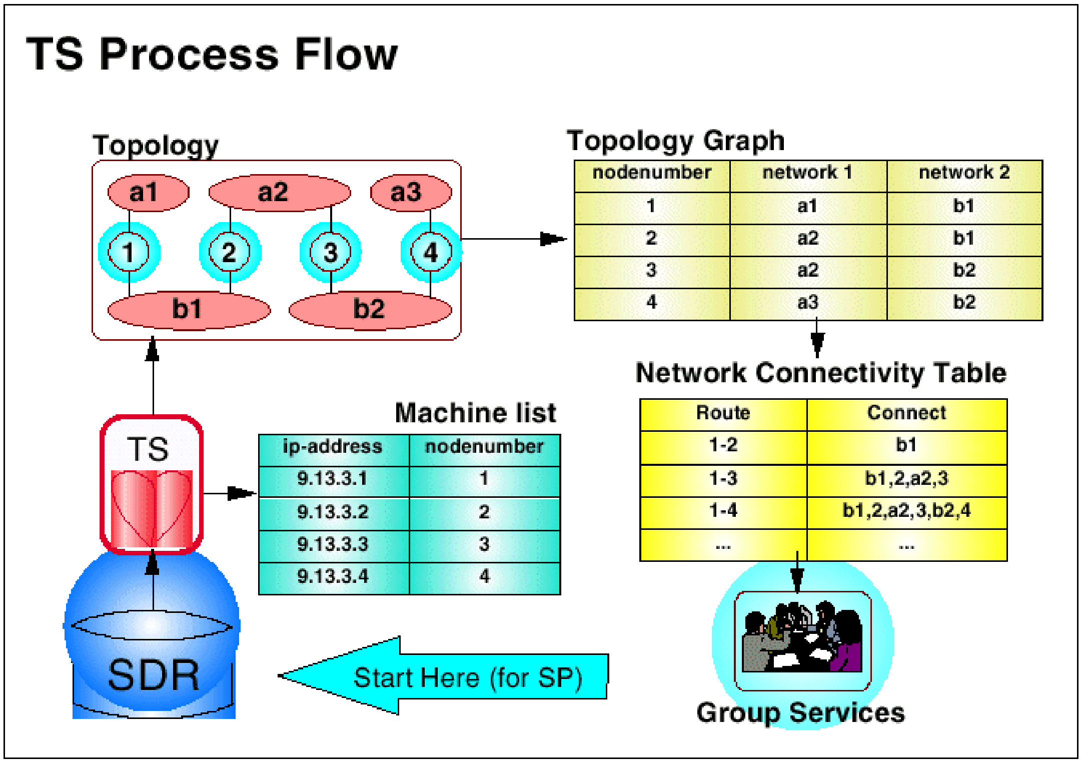
Topology Services is a distributed subsystem of the IBM RS/6000 Cluster Technology (RSCT) software on RS/6000 systems. The RSCT software provides a set of services that support high availability on your SP system. Other services in the RSCT software are the Event Management and Group Services distributed subsystems. These three distributed subsystems operate within a domain. A domain is a set of RS/6000 machines upon which the RSCT components execute and, exclusively of other machines, provide their services. On an SP system, a domain is a system partition. Note that a machine might be in more than one RSCT domain; the control workstation is a member of each system partition, and, therefore, a member of each RSCT domain. When a machine is a member of more than one domain, there is an executing copy of each RSCT component per domain.

Topology Services provides other high availability subsystems with network adapter status, node connectivity information, and a reliable messaging service. The adapter status and node connectivity information is provided to the Group Services subsystem upon request, Group Services then makes it available to its client subsystems. The Reliable Messaging Service, which takes advantage of node connectivity information to reliably deliver a message to a destination node, is available to the other high availability subsystems.

This adapter status and node connectivity information is discovered by an instance of the subsystem on one node, participating in concert with instances of the subsystem on other nodes, to form a ring of cooperating subsystem instances. This ring is known as a heartbeat ring, because each node sends a heartbeat message to one of its neighbors and expects to receive a heartbeat from its other neighbor. Actually each subsystem instance can form multiple rings, one for each network it is monitoring. Usually, each subsystem monitors two rings; the SP Ethernet and the SP switch. This system of heartbeat messages enables each member to monitor one of its neighbors and to report to the heartbeat ring leader, called the Group Leader, if it stops responding. The Group Leader, in turn, forms a new heartbeat ring based on such reports and requests for new adapters to join the membership. Every time a new group is formed, it lists which adapters are present and which adapters are absent, making up the adapter status notification that is sent to Group Services.

In addition to the heartbeat messages, connectivity messages are sent around all rings. Connectivity messages for each ring will forward its messages to other rings, so that all nodes can construct a connectivity graph. It is this graph that determines node connectivity and defines a route that Reliable Messaging would use to send a message between any pair of nodes that have connectivity.

Upon the startup of the Topology Services daemon, the initial configuration information is supplied from the SDR. This is used to build a Machine List file, and adapter groups are established and a topology table (connectivity and availability table) is built. This in turn is used to build the Network Connectivity Table (NCT) in shared memory, and this information is passed via the Reliable Messaging subsystem to Group Services (GS) as a client of Topology Services. This process flow is shown in Figure 4-16.



TS Process Flow

Topology

Topology Graph

nodenumber	network 1	network 2
1	a1	b1
2	a2	b1
3	a2	b2
4	a3	b2

Network Connectivity Table

Route	Connect
1-2	b1
1-3	b1,2,a2,3
1-4	b1,2,a2,3,b2,4
...	...

Machine list

ip-address	nodenumber
9.13.3.1	1
9.13.3.2	2
9.13.3.3	3
9.13.3.4	4

Figure 4-16 Topology Services process flow

To manage the changes in heartbeat rings, the following roles are defined within Topology Services:

Group Leader (GL): The node whose adapter has the highest IP address; it proclaims the group and handles join requests and death notifications, coordinates with group members, and distributes connectivity information. The GL node will not necessarily be the same for the different heartbeat rings.

Crown Prince: The second highest IP address; detects the death of the Group Leader and takes over the GL role.

Mayor: Picked by the Group Leader to broadcast messages to the group members in a given subnet.

Generic: Any other member of the group, who monitors the heartbeat message from its neighbor and informs the Group Leader if there is a problem.

All of these roles are dynamic; they are continuously re-evaluated and reassigned.

4.6.2 Group Services (GS)

Group Services (GS) is a client of Topology Services, and is the next level in the RSCT structure. GS provides coordination and synchronization services to client subsystems, such as Event Management (EM) and Recoverable Virtual Shared Disk (RVSD).

Refer to the following IBM publications for more information about Group Services:

- Chapter 25, “Group Services Subsystem in the Parallel System Support” in *Parallel System Support Programs for AIX: Diagnosis Guide*, GA22-7350
- *Programs for AIX: Administration Guide*, SA22-7348
- *RSCT Group Services: Programming Cluster Applications*, SG24-5523

GS overview

GS runs as a distributed daemon (hagsd) running on all nodes in a system partition, and communication between the nodes is through the Reliable Messaging Library. On the CWS there will be one instance of the hagsd daemon running for each system partition. The GS structure is shown in Figure 4-17.

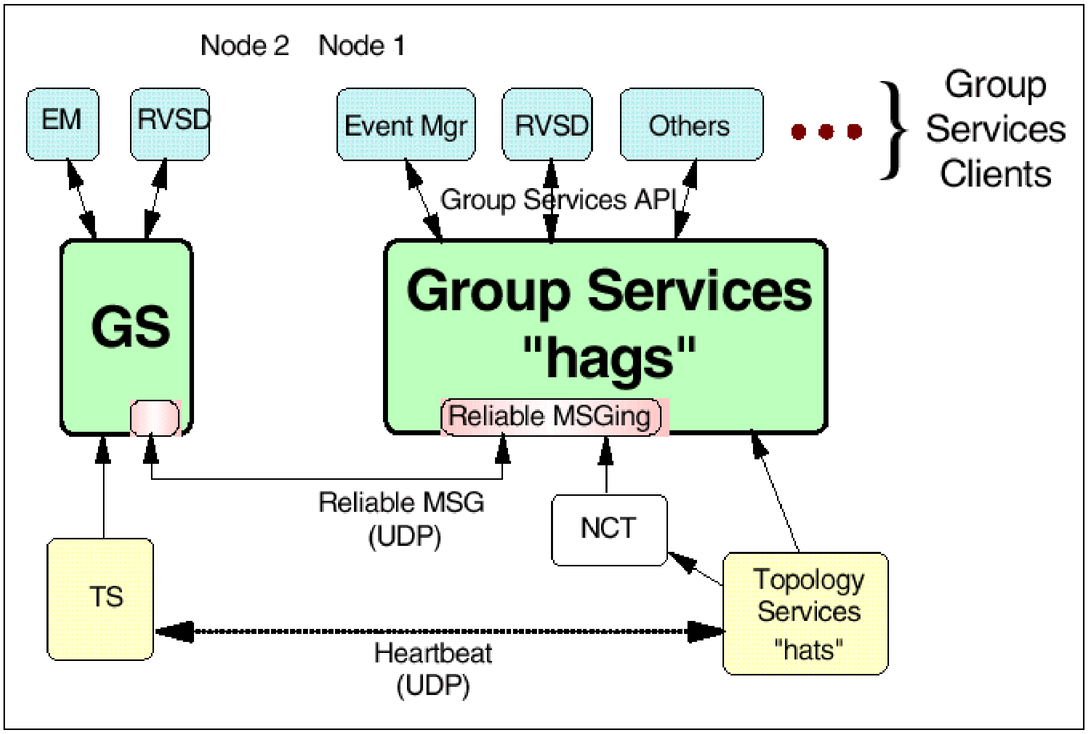


Figure 4-17 Group Services structure

GS clients are either providers, processes that join the group, or subscribers, that monitor the group. The group state, is maintained by the GS subsystem consisting of a group membership list, a list of the providers, and a group state value, this is controlled by the providers. Subscribers do not appear in the group membership lists, they are known to the GS subsystem but not by the providers. The client subsystems connect to GS and form groups by using the Group Services API (GSAPI).

Any provider in a group can initiate a change to the group state, by either joining or leaving the group. Changes to the group state are serialized, that is the change must complete before another change can start. GS establishes a single group namespace across a SP partition. Each SP partition would be a separate namespace and can have more than one group within that namespace. To keep track of the changes to the client groups GS nominates a nameserver (NS), this is not the same as a DNS nameserver. The nameserver, if all nodes are booted at once, will be the node with the lowest IP address. If there is a GS daemon already running within a namespace, then it will be the NS and will remain so taking responsibility for tracking group state changes.

4.6.3 Event Management (EM)

Event Management (EM) is the top level of the RSCT subsystems, it is a client of Group Services. It provides a monitoring service of client requested system resources, such as file systems, processes, CPUs, and notifies those clients when certain conditions are met. It runs as a daemon, haemd. The functional flow is shown inFigure 4-18.

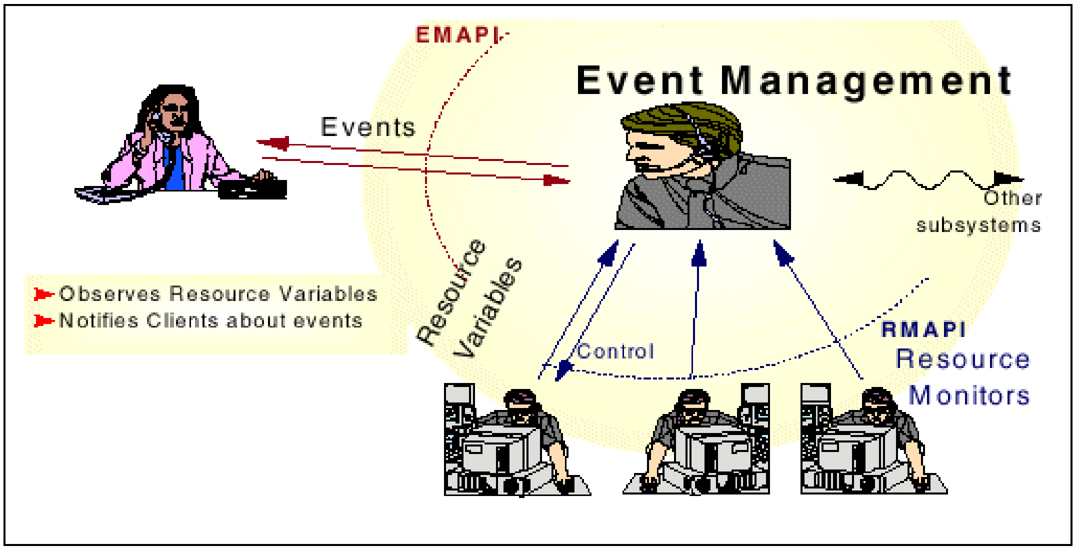


Figure 4-18 Event Management functional flow

EM overview

By monitoring the state of the resource conditions against the client system resources, the client is notified in advance of any event that can cause a possible system failure. Therefore, using this information is useful in trying to recover from any events that can possibly cause system failures in advance of the problem. An example would be detecting a file system on a node starting to fill up, communicating this to the client, and the client (such as pmand) then taking action to make available space in the monitored file system.

There are three components of EM:

- Resource Monitors, which keep track of information related to system attributes, transform this information to resource variables and communicate them to the EM subsystem.
- The EM subsystem communicates between the Resource Monitors and the EM clients. It receives and keeps track of information from the Resource Monitors, as well as tracking information for which the EM clients have expressed an interest in.
- The EM client acts upon information regarding system resources. An EM client can be an application or a subsystem.

The Event Management Configuration Database (EMCDB) holds all the definitions of the resource monitors and the resource variables which are written to the SDR. It is a binary file that is created from the EM SDR classes.

5

SDD installation and configuration on AIX

In this chapter we describe how to install and set up the Subsystem Device Driver on an AIX platform attached to an IBM Enterprise Storage Server. For updated and additional information not included in this chapter, see the `README` file on the compact disc included with your ESS or visit the Subsystem Device Driver Web site at:

http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/storsw

5.1 Pre-installation checks

Before installing the IBM Subsystem Device Driver, you must first configure the ESS for single-port or multiple-port access for each LUN. The Subsystem Device Driver requires a minimum of two independent paths that share the same logical unit to use the load balancing and failover features.

For information about configuring your ESS, see the following publications:

- *IBM TotalStorage ESS Introduction and Planning Guide*, GC26-7294; see:

  `http://ssddom02.storage.ibm.com/disk/ess/documentation.html`

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420; see:

  `http://publib-b.boulder.ibm.com/Redbooks.nsf/RedbookAbstracts/sg245420.html`

- *Implementing Fibre Channel Attachment on the ESS*, SG24-6113; see:

  `http://publib-b.boulder.ibm.com/Redbooks.nsf/RedbookAbstracts/sg246113.html`

Figure 5-1 shows where IBM Data Path Optimizer (DPO) fits in the protocol stack on AIX platforms. As we can see in that figure, each IBM SDD device can later be used by AIX LVM or as a raw device. Each IBM SDD device points also to a single physical disk device and can support up to 32 different paths to that device. Devices handled by IBM SDD on an AIX platform behave like ordinary hdisk devices except, that support for multipathing and load balancing is added.

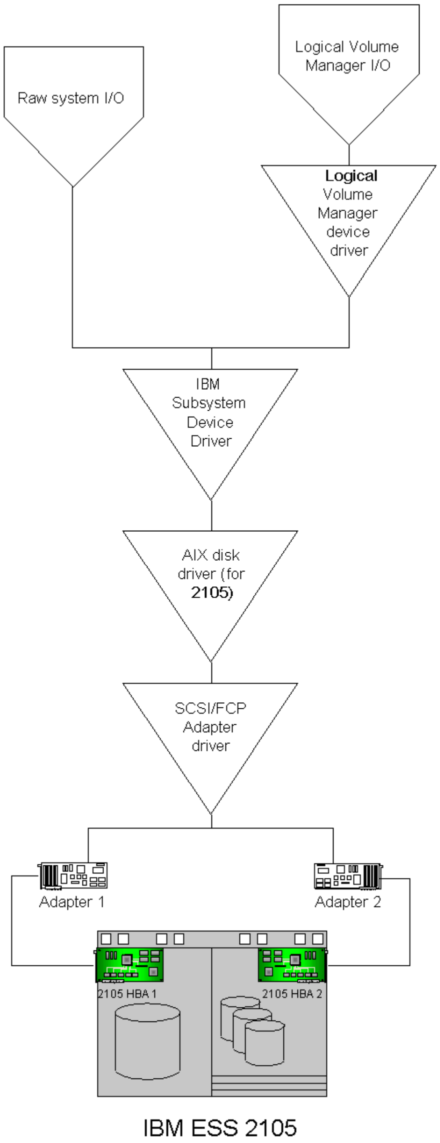

Logical Volume Manager I/O
Raw system I/O
Logical Volume Manager device driver
IBM Subsystem Device Driver
AIX disk driver (for 2105)
SCSI/FCP Adapter driver
Adapter 1
Adapter 2
2105 HBA 1
2105 HBA 2
IBM ESS 2105

Figure 5-1 Where IBM SDD fits in the protocol stack on AIX

5.2 Hardware and software requirements

The IBM Subsystem Device Driver has following hardware and software requirements:

- Hardware:
  - The IBM Enterprise Storage Server
  - IBM RS/6000 or pSeries host system
  - SCSI and/or Fibre Channel adapters and cables
- Software:
  - AIX 4.2.1, AIX 4.3.2, AIX 4.3.3 or AIX 5.1.0 with appropriate fixes installed. See Table 5-1 for the list of required fixes.
  - ESS package `ibm2105.rte` installed
  - SCSI and Fibre Channel device driver installed

Table 5-1 List of fixes required for AIX

AIX level	PTF number	Component name	Component level
4.2.1	IX62304		
	U451711	perfagent.tools	2.2.1.4
	U453402	bos.rte.libc	4.2.1.9
	U453481	bos.adt.prof	4.2.1.11
	U458416	bos.mp	4.2.1.15
	U458478	bos.rte.tty	4.2.1.14
	U458496	bos.up	4.2.1.15
	U458505	bos.net.tcp.client	4.2.1.19
	U462492	bos.rte.lvm	4.2.1.16
4.3.2	U461953	bos.rte.lvm	4.3.2.4

Attention: The presented list of fixes is valid at the date of this book’s publishing. For the latest APARs, maintenance level fixes, and microcode updates, go to the following Web site:

http://techsupport.services.ibm.com/server/support

5.2.1 SCSI requirements

To use the Subsystem Device Driver SCSI support, ensure your host system meets the following requirements:

- The maximum number of SCSI adapters that is supported is 32.
- A SCSI cable is required to connect each SCSI host adapter to an ESS port.
- The `bos.adt` package must be installed on the host operating system. The host system can be a uniprocessor or a multiprocessor system, such as SMP.
- The Subsystem Device Driver I/O load-balancing and failover features require a minimum of two SCSI adapters.

Note: The Subsystem Device Driver also supports one SCSI adapter on the host system. With single-path access, the concurrent download of licensed internal code is supported with SCSI devices. However, the load-balancing and failover features are not available.

For current information about the SCSI adapters that can attach to your AIX host system, go to the Web site at:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

5.2.2 Fibre Channel requirements

To use the Subsystem Device Driver Fibre Channel support, ensure your host system meets the following requirements:

- The AIX host system is an IBM RS/6000 or pSeries with AIX 4.3.3 or AIX 5.1.0 installed.

- The AIX host system has the Fibre Channel device drivers installed along with APARs `IY10201`, `IY10994`, `IY11245`, `IY13736`, `IY17902`, and `IY18070`.
- The `bos.adt` package must be installed on the host operating system. The host system can be a uniprocessor or a multiprocessor system, such as SMP.
- A fiber-optic cable is required to connect each Fibre Channel adapter to an ESS port or Fibre Channel switch.
- The Subsystem Device Driver I/O load-balancing and failover features require a minimum of two Fibre Channel adapters.

For current information about the Fibre Channel adapters that can attach to your AIX host system go to the Web site at:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

5.2.3 Non supported environments

The following environments are not supported by the Subsystem Device Driver:

- A host server with a single-path Fibre Channel connection to an ESS is not supported. There is no reason for install SDD when only one path is available.

Note: A host server with a single fibre adapter that connects through a switch to multiple ESS ports is considered a multipath Fibre Channel connection and therefore is a supported environment.

- A host server with SCSI channel connections and a single-path Fibre Channel connection to an ESS is not supported.
- A host server with both a SCSI channel and Fibre Channel connection to a shared LUN is not supported.

5.3 Installing and configuring Fibre Channel device drivers

In this section we describe in detail procedures for how to install Fibre Channel adapters within RS/6000 or pSeries system, how to install appropriate device drivers, and configure Fibre Channel devices.

The Fibre Channel adapters (F/C 6227 or F/C 6228) suitable for AIX are manufactured by Emulex. Device drivers for those adapters are developed by IBM.

Important: If more than one adapter is attached to a *Peripheral Component Interconnect* (PCI) bus, all adapter devices on that bus will be configured. Sometimes, though, one adapter saturates the entire PCI bus and causes command timeouts.

The Emulex LP7000E adapter should be attached to its own PCI bus and the bus should not be shared with other PCI adapters. See the *PCI Adapter Placement Reference Guide*, SA38-0538 for general information about the number of Fibre Channel adapters suitable for your host system and the locations of the adapters. You can find this reference guide at:

http://www-1.ibm.com/servers/eserver/pseries/library/hardware_docs/

As mentioned in 5.2.2, "Fibre Channel requirements" on page 78, for the Fibre Channel support, the AIX host system must be an IBM RS/6000 or pSeries system with AIX 4.3.3 or AIX 5.1.0. The AIX host system should have the Fibre Channel device driver installed along with the following APARs: `IY10201`, `IY10994`, `IY11245`, `IY13736`, `IY17902`, and `IY18070`.

There are two supported in AIX Fibre Channel adapters: FC 6227 and FC 6228. These are described in 5.3.2, "Gigabit Fibre Channel Adapter for PCI bus FC 6227 (type 4-S) features" on page 82 and 5.3.3, "2-Gigabit Fibre Channel Adapter for PCI bus FC 6228 (type 4-W) features" on page 84.

5.3.1 Installing Fibre Channel device drivers

If the AIX operating system is not yet installed on the host server, we only need to install the hardware adapter. The device driver software will be installed automatically during the installation of AIX, when the adapter hardware is detected. Depending on the release of the AIX operating system installation media, required AIX APARs may or may not be installed. We need to check this after the installation of AIX is complete. To check if all APARs required for Fibre Channel proper operation are installed, issue the commands listed in Example 5-1.

Tip: You can download all required AIX fixes from:

http://techsupport.services.ibm.com/rs6k/fixdb.html

Example 5-1 Checking installation of all required APARs on AIX

```
instfix -ik IY10201
instfix -ik IY10994
instfix -ik IY11245
instfix -ik IY13736
instfix -ik IY17902
instfix -ik IY18070

For each of APARs listed above, the instfix command should return the message:
   All filesets for XXXXXXX were found.
      where XXXXXXX is a number of APAR

If the instfix command returns one of messages listed below, the APAR is not installed (or
is installed incompletely) and needs to be installed (re-installed):
   There was no data for XXXXXXX in the fix database.
   Not all filesets for XXXXXXX were found.
```

If AIX is already installed, we need to install the hardware adapter, device driver software and all required APARs. To do this perform the following steps:

1. Install the hardware adapter in appropriate PCI slot within your host. To obtain a host model-dependent list of PCI slots suitable for the Fibre Channel adapters, see the *PCI*

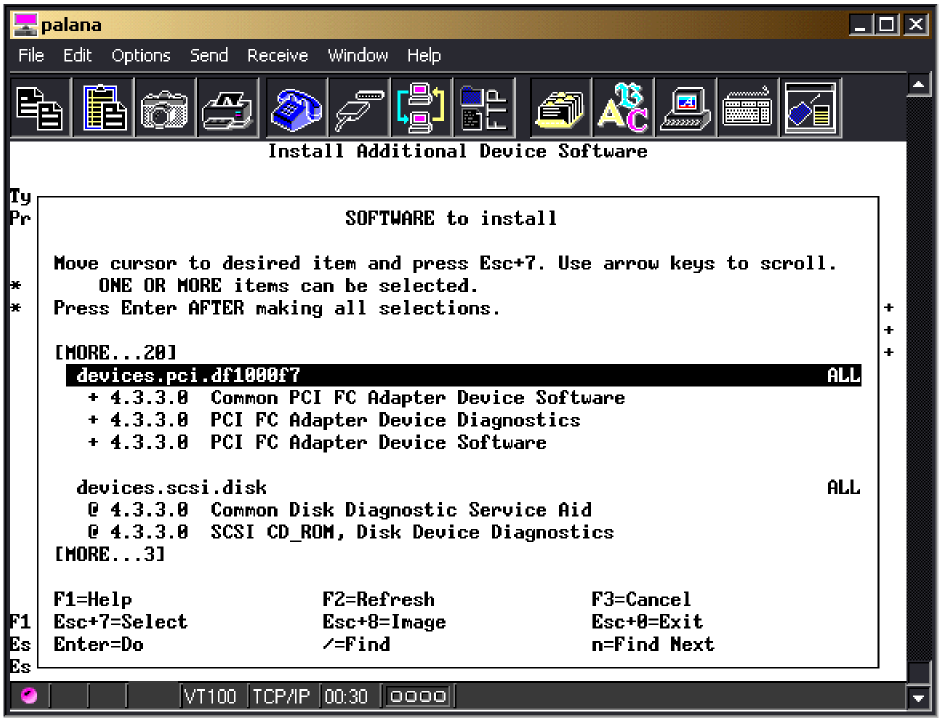
Adapter Placement Reference Guide, SA38-0538. You can find this Reference Guide at: http://www-1.ibm.com/servers/eserver/pseries/library/hardware_docs/

2. Boot the system and log in as user *root*.
3. Insert the media containing the device driver software (in most cases this is the AIX CD-ROM installation disk) into the appropriate media device. Type the following: `smitty devinst`, and press **Enter**.
4. The Install Additional Device Software menu is displayed and the **INPUT device/directory for software** option is highlighted. Select or type your input device as described below:

   a. To select the input device from the list press **F4** (or **ESC** + **4**) for a list of devices and select the appropriate device. Press **Enter**.

   b. If you want to manually enter the input device, type the full path of the special file associated with the input device you are using (or the full path to the directory where installation images are stored) in the entry field and press **Enter** (for example, `/dev/cd0`).

5. The Install Additional Device Software menu is expanded and the **SOFTWARE to install** option is highlighted. Press **F4** (or **ESC** + **4**) for a list of software items available to install. When a list of software items expands, press "/" (slash) key to display the Find window.
6. Type the following: `devices.pci.df1000f7`, and press **Enter**. The system will find and highlight the device driver software as shown in Figure 5-2.



```
                         Install Additional Device Software

Ty
Pr                               SOFTWARE to install

   Move cursor to desired item and press Esc+7. Use arrow keys to scroll.
*      ONE OR MORE items can be selected.
*  Press Enter AFTER making all selections.                                   +
                                                                              +
   [MORE...20]                                                                +
     devices.pci.df1000f7                                                 ALL
      + 4.3.3.0  Common PCI FC Adapter Device Software
      + 4.3.3.0  PCI FC Adapter Device Diagnostics
      + 4.3.3.0  PCI FC Adapter Device Software

     devices.scsi.disk                                                    ALL
      @ 4.3.3.0  Common Disk Diagnostic Service Aid
      @ 4.3.3.0  SCSI CD_ROM, Disk Device Diagnostics
   [MORE...3]

   F1=Help                 F2=Refresh              F3=Cancel
F1 Esc+7=Select            Esc+8=Image             Esc+0=Exit
Es Enter=Do                /=Find                  n=Find Next
Es
```

Figure 5-2 Example of device driver software selection on AIX platform

Press **F7** to select the highlighted device driver software. Invoke the Find window again (pressing "/" button) to find and select `devices.fcp.disk` fileset and after that to find and select `devices.common.IBM.fc` fileset.

7. When all three filesets: `devices.pci.df1000f7`, `devices.fcp.disk` and `devices.common.IBM.fc` are selected, press **Enter**. The Install Additional Device Software menu displays again. The entry data fields are automatically updated. Press **Enter** to accept the values.

8. The ARE YOU SURE pop-up window displays. Press **Enter** to continue with the installation.
9. The COMMAND STATUS is displayed. After the installation process has completed, **OK** will be displayed. Scroll to the bottom to view the results to ensure that the installation was successful.
10. Remove the installation media from the drive. Press **F10** (or **ESC** + **0**) to exit SMIT.

Verifying the installation

The installation can be verified, by performing one or both of the following procedures:

- **To verify the hardware installation (`lsdev`)**, log in as `root` user and type: `lsdev -C | grep fcs`. Press **Enter**. If the Gigabit Fibre Channel PCI Adapter is properly installed and configured, an output similar to that shown in Example 5-2 should display on your screen. The adapter should be in `Available` state.

Example 5-2 Example of properly configured Fibre Channel PCI adapter

```
      fcs0 Available 20-60
```

  If no adapter information is displayed, or if it is shown as `Defined`, refer to 5.3.4, “Problem determination” on page 86, to determine the cause of the problem.

- **To verify the software installation (`lslpp`)**, log in as `root` user and type `lslpp -h | grep -p df1000f7`. Press **Enter**. If the Gigabit Fibre Channel PCI Adapter Device driver software is properly installed, an output similar to that shown in Example 5-3 should display on your screen.

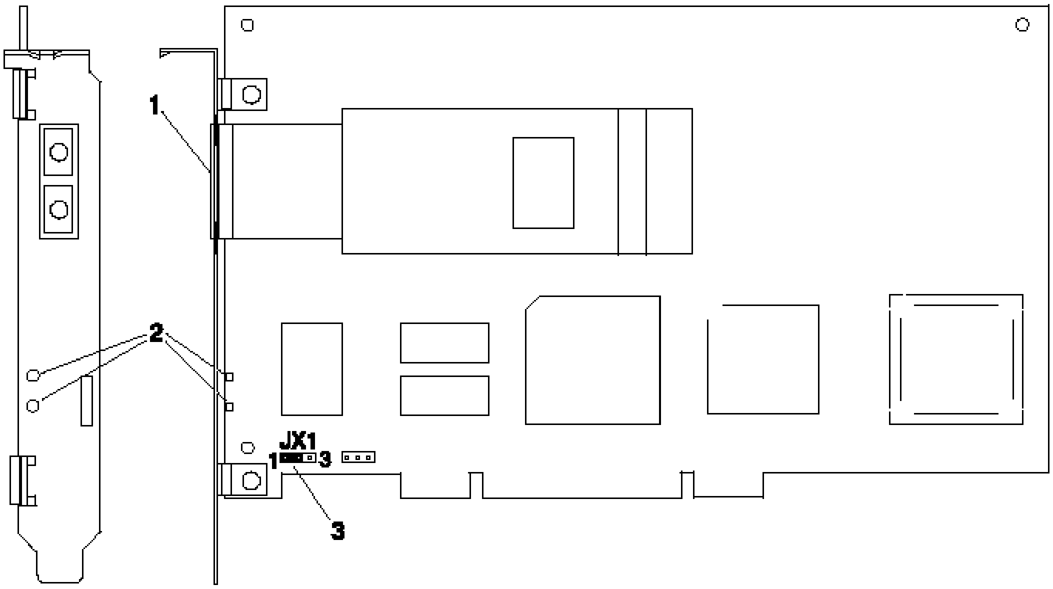
Example 5-3 Example of properly installed Fibre Channel PCI adapter driver software

```
  Fileset         Level     Action       Status       Date         Time
  ----------------------------------------------------------------------------
Path: /usr/lib/objrepos
  devices.pci.df1000f7.rte
                  4.3.3.0   COMMIT       COMPLETE     10/12/01     19:36:44

  devices.pci.df1000f7.com
                  4.3.3.0   COMMIT       COMPLETE     10/12/01     19:36:44

  devices.pci.df1000f7.diag
                  4.3.3.0   COMMIT       COMPLETE     10/12/01     19:36:44

  devices.fcp.disk.rte
                  4.3.3.0   COMMIT       COMPLETE     10/12/01     19:36:51

  devices.common.IBM.fc.rte
                  4.3.3.0   COMMIT       COMPLETE     10/12/01     19:36:57
```

  If no device driver information is displayed, or some information is missing, refer to 5.3.4, “Problem determination” on page 86, to determine the cause of the problem.

5.3.2 Gigabit Fibre Channel Adapter for PCI bus FC 6227 (type 4-S) features

The Gigabit Fibre Channel Adapter for PCI bus FC 6227 provides the attachment of external storage using the Fibre Channel Arbitrated Loop protocol. The protocol is sent over a shortwave (multimode) fiber optic cable. This adapter have on-board FC-AL protocol engine and buffers and is FC-PH and PCI 2.1 compliant. Figure 5-3 shows the layout of a 6227 adapter, and Table 5-2 gives the specifications for a 6227 adapter.



1. Multimode Fiber SC Connector
2. Data Link Status LEDs
3. Jumper JX1, Pins 1 to 2 only

Figure 5-3 Layout of Fibre Channel FC 6227 (type 4-S) adapter

Table 5-2 Specifications for Fibre Channel adapter FC 6227

Item	Description
FRU number	09P1173
BUS architecture	PCI 2.1
Card type	Half
Adapter slots	For system-specific adapter placement, see the *PCI Adapter Placement Reference Guide*, SA38-0538. You can find this Reference Guide at: `http://www-1.ibm.com/servers/eserver/pseries/library/hardware_docs/`
Connector	ANSI specified SC duplex
Wrap plug	Shipped with assembly or 16G5609
Cables	50 or 62.5 micron multimode fiber-optic, customer provided

As shown in Figure 5-3, the FC 6227 PCI Fibre Channel adapter is equipped with two LEDs: green and yellow located near the SC connector. The information displayed on the LEDs is very useful for adapter and connection problem solving. The meaning of the adapter LEDs is explained in Table 5-3.

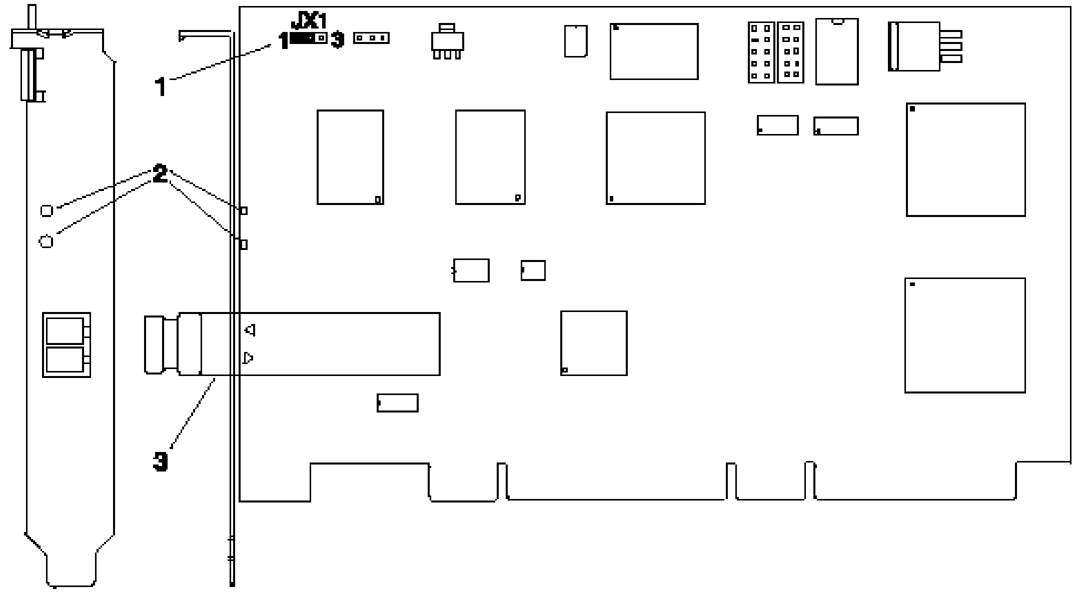
Table 5-3 Use of FC 6227 adapter LEDs

Green LED	Yellow LED	Adapter state
OFF	OFF	Wakeup failure (adapter is defective)
OFF	ON	POST failure (adapter is defective)
OFF	Slow blink (1 Hz)	Wakeup failure

Green LED	Yellow LED	Adapter state
OFF	Fast blink (4 Hz)	Failure in POST
OFF	Flashing irregularly	POST processing in progress
ON	OFF	Failure while functioning
ON	ON	Failure while functioning
ON	Slow blink (1 Hz)	Normal - inactive
ON	Fast blink (4 Hz)	Normal - busy
ON	Flashing irregularly	Normal - active
Slow blink	OFF	Normal - link down or not yet started
Slow blink	ON	Off-line for download
Slow blink	Slow blink (1 Hz)	Restricted off-line mode (waiting for restart)

5.3.3 2-Gigabit Fibre Channel Adapter for PCI bus FC 6228 (type 4-W) features

The 2-Gigabit Fibre Channel Adapter for PCI bus FC 6228 provides attachment of external storage using the Fibre Channel Arbitrated Loop protocol. The protocol is sent over a shortwave (multimode) fiber optic cable. This adapter have on-board FC-AL protocol engine and buffers and is FC and PCI 2.2 compliant. Figure 5-4 shows layout of the 6228 adapter and Table 5-4 gives the specifications of the 6228 adapter.



1. Jumper JX1, Pins 1 to 2 only
2. Data Link Status LEDs
3. Multimode Fiber LC Connector

Figure 5-4 Layout of Fibre Channel FC 6228 (type 4-W) adapter

Table 5-4 Specifications for Fibre Channel adapter FC 6228

Item	Description
FRU number	09P0102
BUS architecture	PCI 2.2
Card type	Half
Adapter slots	For system-specific adapter placement, see the *PCI Adapter Placement Reference Guide*, SA38-0538. You can find this Reference Guide under at: `http://www-1.ibm.com/servers/eserver/pseries/library/hardware_docs/`
Connector	ANSI specified LC duplex
Wrap plug	Shipped with assembly or 05N6768
Cables	50 or 62.5 micron multimode fiber-optic, customer provided

As shown in Figure 5-4, the FC 6228 PCI Fibre Channel adapter is equipped with two LEDs: green and yellow located near the LC connector. The information displayed on the LEDs is very useful for adapter and connection problem solving. The meaning of the adapter LEDs is explained in Table 5-5.

Table 5-5 Use of FC 6228 adapter LEDs

Green LED	Yellow LED	Adapter state
OFF	OFF	Wakeup failure (adapter is defective)
OFF	ON	POST failure (adapter is defective)
OFF	Slow blink (1 Hz)	Wakeup failure

Green LED	Yellow LED	Adapter state
OFF	Fast blink (4 Hz)	Failure in POST
OFF	Flashing irregularly	POST processing in progress
ON	OFF	Failure while functioning
ON	ON	Failure while functioning
ON	slow blink (1 Hz)	Normal - inactive
ON	Fast blink (4 Hz)	Normal - busy
ON	Flashing irregularly	Normal - active
Slow blink	OFF	Normal - link down or not yet started
Slow blink	Slow blink (1 Hz)	Off-line for download
Slow blink	Fast blink (4 Hz)	Restricted off-line mode (waiting for restart)

5.3.4 Problem determination

In this section we provide some basic information about problem determination procedures in a Fibre Channel environment. This section is *not* intended to describe detailed problem determination procedures, but to give an idea of where to look for potential problem causes and how to proceed with Fibre Channel problems. If a problem occurs in the Fibre Channel environment, you will need a number of pieces of information to successfully correct the problem. Here we discuss Fibre Channel environment-specific problems. If problems are experienced with the AIX system, see your AIX documentation.

The Fibre Channel environment can be complex, and because of the potential distances between components of the system, and the diverse nature of these components, additional information will be required to aid in problem determination. The information is available from several sources:

- Gigabit Fibre Channel PCI Adapter Service LEDs:

  The Gigabit Fibre Channel PCI Adapter has two LEDs located near the connectors. These can be used to determine the state of the adapter. For details on these LEDs, see Table 5-3 on page 83 and Table 5-5 on page 85.

- AIX system problem determination information:

  The AIX system provides problem determination information from its operator display codes, error logging facilities, and application messages. For more information on AIX error logs, see your AIX documentation.

- Fibre Channel Director problem determination information:

  The Fibre Channel Director provides problem determination information from its operator panel, LED indicators on the port cards, and the enterprise fabric connectivity management terminal.

- Problem determination information from other devices:

  Other Fibre Channel devices, including disk storage subsystems, provide problem determination information in various ways, such as status LEDs, operator panels, and logout information.

Complexity of the Fibre Channel Environment

The Fibre Channel environment can be difficult to troubleshoot. A typical Fibre Channel configuration, such as a Storage Area Network (SAN), may contain some or all of the following:

- One or more system hosts, perhaps running any of several different operating systems.
- One or more Disk Storage Subsystems with a number (perhaps a very large number) of Disk Devices (LUNs) in RAID or non-RAID configurations.
- One or more Tape Subsystems connected by their native Fibre Channel interfaces.
- One or more hubs connecting system hosts and Disk Storage Subsystems in loop configurations.
- One or more Fibre Channel switches connecting the various devices and system hosts in a fabric environment.
- One or more SAN Data Gateways allowing the introduction of SCSI attachable Disk Storage Subsystems or Magnetic Tape Subsystems into the Fibre Channel environment.
- A large number of Fibre Channel jumper cables interconnecting the various system hosts and devices.
- Fiber trunks carrying data between floors and between buildings.
- Patch panels connecting the various jumper cables and trunk cables.

Troubleshooting the Fibre Channel environment is further complicated by the fact that the various hosts and devices may be physically separated by considerable distance, and located in different rooms, on different floors, and even in different buildings.

Nature of Fibre Channel environment problems

In the complex and diverse Fibre Channel environment, a wide variety of problems can be encountered. These problems may include, but are by no means limited to:

- A Gigabit Fibre Channel PCI Adapter in an AIX system host has a hardware defect.
- A Gigabit Fibre Channel PCI Adapter in an AIX system host is not at required firmware level.
- A Gigabit Fibre Channel PCI Adapter has been incorrectly configured.
- The device driver for a Gigabit Fibre Channel PCI Adapter has been incorrectly installed or is exhibiting incorrect behavior.
- A Fibre Channel SCSI I/O Controller Protocol Device is not properly configured.
- A logical hard disk in the AIX system is not properly configured.
- A port adapter in a Fibre Channel switch has a hardware defect.
- A port in a Fibre Channel switch is incorrectly zoned or blocked.
- Ports in a Fibre Channel switch have been soft rezoned or reblocked and the `cfgmgr` command has not been run to set up the new configuration parameters.
- Host-to-switch cabling has been changed or swapped and the `cfgmgr` AIX command has not been run to update the configuration attributes. In this case, results of commands such as `lsattr -El` will not yield the correct information for attributes such as the `scsi_id` field.
- A port adapter in a Fibre Channel hub has a hardware defect.
- A Fibre Channel port adapter in a SAN Data Gateway has a hardware defect.
- A SCSI port adapter in a SAN Data Gateway has a hardware defect.
- A port adapter in a disk storage subsystem has a hardware defect.

- A disk drive in a disk storage subsystem has a hardware defect.
- A Fibre Channel jumper cable is defective.
- A Fibre Channel cable connector is not properly seated, or is dirty.
- A Fibre Channel trunk has a defective fiber.
- A patch panel connection is defective or incorrectly plugged.
- A host or device has defective logic, memory, or control circuitry, or a defective power or cooling system.
- Optical components somewhere in the environment are defective and are causing intermittent failures.

As we can see in the previous list, problems can be encountered anywhere throughout the Fibre Channel configuration. Sometimes the problem is distinctly reported by, and at the failing component. Often however, the AIX system host, as the initiator, will detect and report the error condition. As a result, Fibre Channel errors reported by the AIX system must be analyzed carefully to determine the true origin of the failure.

As demonstrated above, the Fibre Channel environment is very complex and no procedure can provide 100% problem determination coverage. It should be noted that because of the complexity of the environment, a single Fibre Channel problem can produce a large volume of error log entries in the AIX system. In such a case, it is necessary to carefully analyze these logged errors to find the one which represents the original, root cause. In addition, while Fibre Channel environment problems are often reported by the AIX system, indiscriminate replacement of the Gigabit Fibre Channel PCI Adapter is not the recommended problem determination procedure.

Upgrading the Fibre Channel adapter firmware

Sometimes Fibre Channel problems on AIX machines are caused by incorrect firmware level of your Fibre Channel Gigabit PCI Adapter. In this section we describe in details, how to find your adapter firmware level and upgrade them to the latest available level.

Determining the Emulex adapter firmware level

To determine if adapter firmware level is possible cause of problem you *must* check the firmware level that is currently installed on the adapter. Latest firmware levels for Fibre Channel Gigabit PCI Adapters, as of date of this book publishing, are as follows:

- For Gigabit Fibre Channel Adapter for PCI bus FC 6227 (type 4-S) firmware level is `sf322A0`.
- For 2-Gigabit Fibre Channel Adapter for PCI bus FC 6228 (type 4-W) firmware level is `sf382A0`.

> **Important:** You are required to install new adapter firmware only if your current adapter firmware is not at latest available level. You can check out and download the latest level of adapter firmware from the following Web site:
>
> `http://www.rs6000.ibm.com/support/micro/download.html#adapter`

Perform the following steps to obtain your current Emulex adapter firmware:

1. List all Fibre Channel adapters installed in your host system. To do that issue the command `lsdev -Cc adapter | grep fcn`.
2. For all adapters determine the firmware level that is currently installed. Issue the `lscfg -vl fcsX` command, where `X` is the number of adapter instance listed in previous step. The adapter's vital product data is displayed.

3. Look at the ZB field. The ZB field should look similar to shown below:

   ```
   (ZB)............S2F3.22A0
   ```

4. To determine the firmware level, ignore the second character in the ZB field. In this example, the firmware level is sf322A0.

If the adapter firmware level is at the required level, there is no need to upgrade. Otherwise, the adapter firmware level must be upgraded.

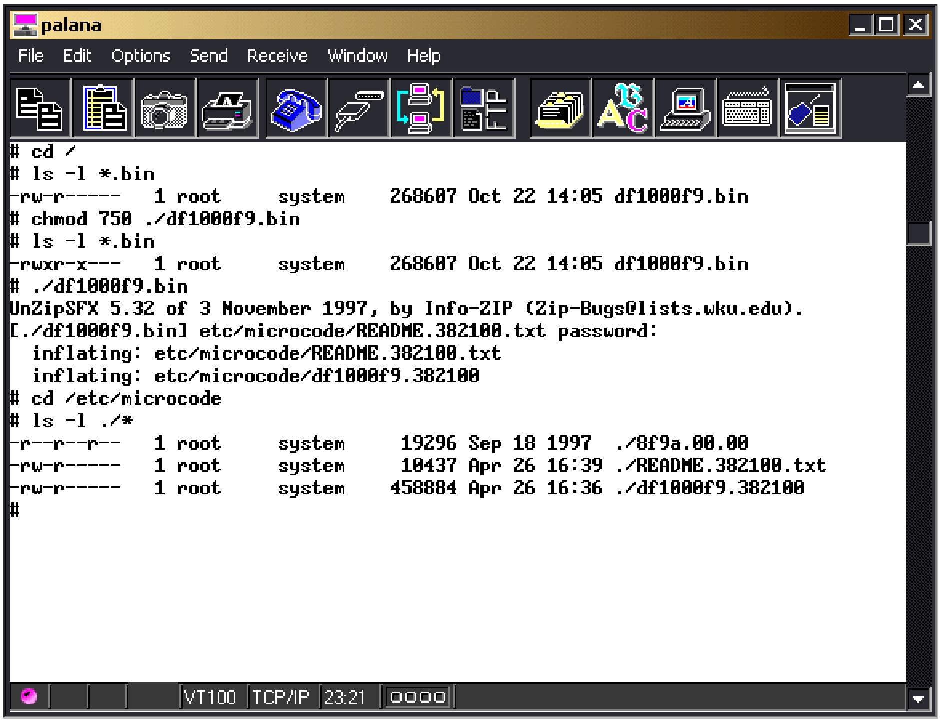
Update the AIX Diagnostic Software to Current Level

The Diagnostic Microcode Download software has been updated to support a new naming convention for the microcode binary files. If you are running AIX 4.3.2 or 4.3.3, please install PTF U473607 or APAR IY14502 from FIXDIST to upgrade the operating system diagnostic to support new naming convention. This is required before upgrading the adapter microcode. If you don't upgrade the operating system diagnostic, the diag command will be unable to find firmware binary files stored with new naming convention.

Upgrading the Emulex adapter firmware level

Upgrading the firmware level consists of downloading the firmware (microcode) from your AIX host system to the adapter. All Fibre Channel attached devices, which are accessible through the adapter being upgraded, must be closed before continuing. Perform the following steps to download the firmware:

1. Go to the Web site http://www.rs6000.ibm.com/support/micro/flicense.html. Read the "IBM eServer pSeries & RS/6000 License Agreement for Machine Code". Click on the link "I have read and understood this license agreement and I agree to abide by its terms". You will obtain a password which is required to uncompress the file containing latest firmware release.
2. Download the latest firmware level available for your adapter type from the following URL: http://www.rs6000.ibm.com/support/micro/download.html#adapter and save it *directly under the / (root)* directory. Change your current directory to the *root* directory.
3. Change the mode of the file to make it executable. To do this, issue the command **chmod 750 /filename.bin**, where filename.bin is the name of the file you have just downloaded. This is shown in Figure 5-5, where the firmware file for Fibre Channel adapter type 4-W (df1000f9.bin) is used.



```
# cd /
# ls -l *.bin
-rw-r-----   1 root     system    268607 Oct 22 14:05 df1000f9.bin
# chmod 750 ./df1000f9.bin
# ls -l *.bin
-rwxr-x---   1 root     system    268607 Oct 22 14:05 df1000f9.bin
# ./df1000f9.bin
UnZipSFX 5.32 of 3 November 1997, by Info-ZIP (Zip-Bugs@lists.wku.edu).
[./df1000f9.bin] etc/microcode/README.382100.txt password:
  inflating: etc/microcode/README.382100.txt
  inflating: etc/microcode/df1000f9.382100
# cd /etc/microcode
# ls -l ./*
-r--r--r--   1 root     system     19296 Sep 18 1997  ./8f9a.00.00
-rw-r-----   1 root     system     10437 Apr 26 16:39 ./README.382100.txt
-rw-r-----   1 root     system    458884 Apr 26 16:36 ./df1000f9.382100
#
```

Figure 5-5 How to uncompress Fibre Channel adapter firmware file

> **Note:** The self-extracting zip file will unzip the `Readme` and the microcode file into the `etc/microcode` directory under the current directory. Since on AIX all microcode files must be stored in `/etc/microcode` directory, you must ensure that your current directory is the *root* directory, before uncompressing the file. This allows you to uncompress the files directly to `/etc/microcode` directory.

4. Please ensure that all Fibre Channel attached devices, which are accessible through the adapter being upgraded are closed. Vary off all volume groups which are accessible through that adapter.
5. From the AIX command prompt, type `diag` and press **Enter**. The window will change to DIAGNOSTIC OPERATING INSTRUCTIONS window. Press **Enter** again. Select the **Task Selection (Diagnostics, Advanced Diagnostics, Service Aids, etc.)** -> **Download Microcode**.
6. Select all the Fibre Channel adapters to which you want to download firmware, by pressing Enter key when appropriate adapter is highlighted. When finished, press **F7** (or **ESC** + **7**). The Download window is displayed with one of the selected adapters highlighted. Press **Enter** to continue.
7. Type the filename for the firmware that is contained in the `/etc/microcode` directory and press **Enter**, or use the **Tab** key to toggle to `Latest`.
8. Follow the instructions that are displayed to download the firmware, one adapter at a time.
9. After the download is complete, issue the **`lscfg -vl fcsX`** command (where `X` is the adapter number found from the "**`lsdev -Cc adapter | grep fcn`**" command) to verify the firmware level for each Fibre Channel adapter.
10. Vary on all previously varied off volume groups and mount the filesystems.

Note: You can also use the command line to download the microcode to the adapter. Type `diag -c -d fcsX -T "download -s /etc/microcode -f -l latest"`, where `X` is the adapter number found from the "`lsdev -Cc adapter | grep fcn`" command. Repeat that command for all adapters that need to be updated.

5.4 Installing and configuring the IBM Subsystem Device Driver

The following skills and information are required to install, configure and verify the installation of the IBM Subsystem Device Driver on your AIX host:

- An AIX system administrator skills
- An AIX system operator with root user authority
- A chart showing the Fibre Channel cabling scheme
- A list of hardware, microcode, and device driver levels for the Gigabit Fibre Channel PCI Adapter and all devices in the Fibre Channel configuration.

To install SDD, use the installation package that is appropriate for your environment. Table 5-6 lists and describes the SDD installation package file names (filesets).

Table 5-6 SDD installation package filesets

Package file name	Description
ibmSdd_421.rte	AIX 4.2.1
ibmSdd_432.rte	AIX 4.3.2 or AIX 4.3.3 (also use when running HACMP with AIX 4.3.3 in concurrent mode)
ibmSdd_433.rte	AIX 4.3.3 (only use when running HACMP with AIX 4.3.3 in non-concurrent mode)
ibmSdd_510.rte	AIX 5.1.0 (also use when running HACMP with AIX 5.1.0 in concurrent mode)
ibmSdd_510nchacmp.rte	AIX 5.1.0 (only use when running HACMP with AIX 5.1.0 in non-concurrent mode)

The following restrictions apply when running IBM SDD on an AIX host:

- SDD 1.3.0.x does not support AIX 5.1.B.
- SDD 1.3.0.x installed from either the `ibmSdd_432.rte` or `ibmSdd_433.rte` fileset is a 32-bit device driver. This version supports 32-bit and 64-bit mode applications on AIX 4.3.2 and AIX 4.3.3 host systems. A 64-bit mode application can access a SDD device directly or through the logical volume manager (LVM).
- SDD 1.3.0.x installed from the `ibmSdd_433.rte` fileset is supported on AIX 4.3.3 and is for High Availability Cluster Multi-Processing (HACMP) environments only. It supports non-concurrent and concurrent modes. However, in order to make the best use of the manner in which the device reserves are made, IBM recommends that you:
  - Use the `ibmSdd_432.rte` fileset for SDD 1.3.0.x when running HACMP with AIX 4.3.3 in concurrent mode.
  - Use the `ibmSdd_433.rte` fileset for SDD 1.3.0.x when running HACMP with AIX 4.3.3 in non-concurrent mode.

- The SDD 1.3.0.x installed from either `ibmSdd_510.rte` or `ibmSdd_510nchacmp.rte` filesets is supported on AIX 5.1.0. It contains both 32-bit and 64-bit drivers. Based on the kernel mode currently running on the system, the AIX loader will load the correct mode of the SDD into the kernel.
- SDD 1.3.0.x contained in the `ibmSdd_510nchacmp.rte` fileset supports HACMP in both concurrent and non-concurrent modes. IBM recommends that you:
  - Install SDD 1.3.0.x from the `ibmSdd_510.rte` fileset if you run HACMP with AIX 5.1.0 in concurrent code only.
  - Install SDD 1.3.0.x from the `ibmSdd_510nchacmp.rte` fileset if you run HACMP with AIX 5.1.0 in non-concurrent mode.
- SDD does not support a system restart from a SDD pseudo device.
- SDD does not support placing system paging devices (for example, `/dev/hd6`) on a SDD pseudo device.
- SDD 1.3.0.x installed from the `ibmSdd_421.rte`, `ibmSdd_432.rte` and `ibmSdd_510.rte` filesets do not support any application that depends on a reserve/release device on AIX 4.2.1, AIX 4.3.2, AIX 4.3.3, and AIX 5.10.
- The published AIX limitation on one system is 10,000 disk devices. The combined number of `hdisk` and `vpath` devices should not exceed the number of supported devices by AIX. In a multipath environment, since each path to a disk creates an hdisk, this limit applies to the total number of disk devices multiplied by a number of paths available for them.

5.4.1 Installing the IBM Subsystem Device Driver

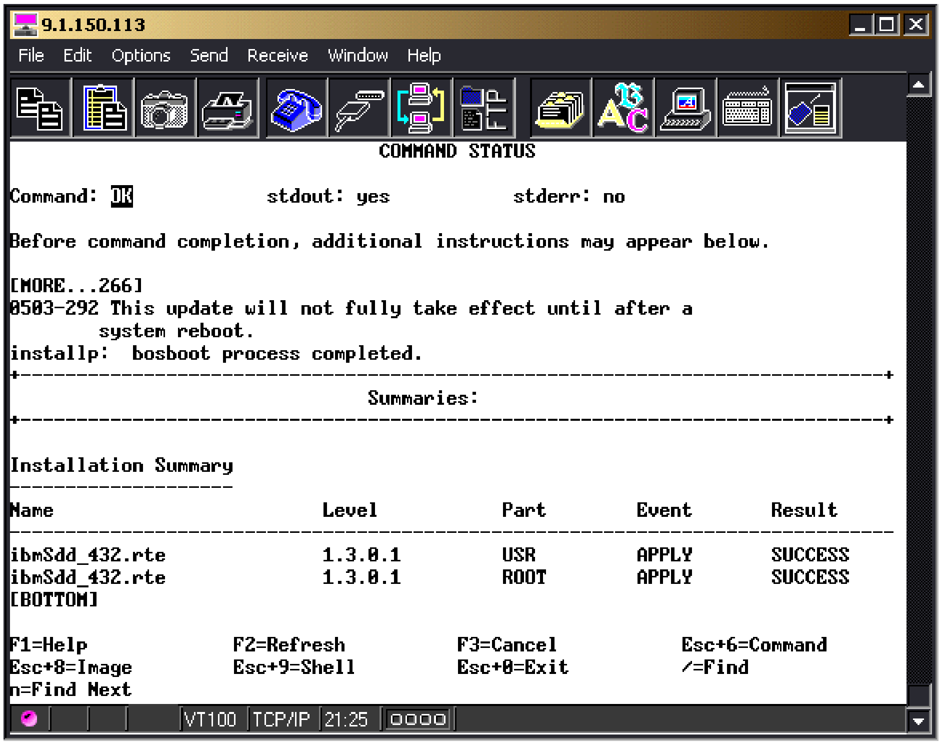
You can use the System Management Interface Tool (SMIT) or command line to install the IBM Subsystem Device Driver into your AIX operating system. For this procedure we assume that a non graphical (text-based) interface of SMIT will be used.

Perform the following SMIT steps to install the SDD package on your system:

1. Log in as the root user.
2. Load your installation media into the appropriate device drive. Usually you will use the installation CD-ROM supplied with your IBM 2105 ESS server or install SDD form hard disk, if you are using IBM SDD installation image downloaded from the Internet. You can download the latest version of SDD from the following Web site:
   `http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/storsw`
3. From your desktop window, type `smitty install_update` and press **Enter** to go directly to the Install and Update Software screen of SMIT. Choose **Install and Update from LATEST Available Software** and press **Enter**.
4. The Install and Update from LATEST Available Software screen is displayed and the **INPUT device/directory for software** option is highlighted. Select or type your input device as described below:
   a. To select the input device from the list press **F4** (or **ESC** + **4**) for a list of devices and select the appropriate device. Press **Enter**.
   b. If you want to manually enter the input device, type the full path of the special file associated with the input device you are using (or the full path to the directory where installation images are stored) in the entry field and press **Enter** (for example, `/dev/cd0`).
5. The Install and Update from LATEST Available Software menu is expanded and the **SOFTWARE to install** option is highlighted. Press **F4** (or **ESC** + **4**) for a list of software items available to install. When a list of software items expands, select the installation

package that is appropriate for your environment. Table 5-6 on page 91 lists and describes the SDD installation package file names (filesets). Press **Enter**. The Install and Update from LATEST Available Software menu displays again. The entry data fields are automatically updated. Press **Enter** to accept the values.

6. Check the default option settings to ensure that they are what you need. Press **Enter** to install.
7. The ARE YOU SURE pop-up window displays. Press **Enter** to continue with the installation. The installation process can take several minutes to complete. Successfully completed `installp` process should finish with SMIT screen similar to shown on Figure 5-6. When the installation is complete, press F10 (or ESC + 0) to exit from SMIT. Remove the installation media from the device.
8. As shown in Figure 5-6, to complete an installation process a system reboot is required. Please reboot your system.



```
                              COMMAND STATUS

Command: OK            stdout: yes            stderr: no

Before command completion, additional instructions may appear below.

[MORE...266]
0503-292 This update will not fully take effect until after a
        system reboot.
installp:  bosboot process completed.
+-----------------------------------------------------------------------------+
                                Summaries:
+-----------------------------------------------------------------------------+

Installation Summary
--------------------
Name                        Level           Part        Event       Result
-------------------------------------------------------------------------------
ibmSdd_432.rte              1.3.0.1         USR         APPLY       SUCCESS
ibmSdd_432.rte              1.3.0.1         ROOT        APPLY       SUCCESS
[BOTTOM]

F1=Help             F2=Refresh          F3=Cancel           Esc+6=Command
Esc+8=Image         Esc+9=Shell         Esc+0=Exit          /=Find
n=Find Next
```

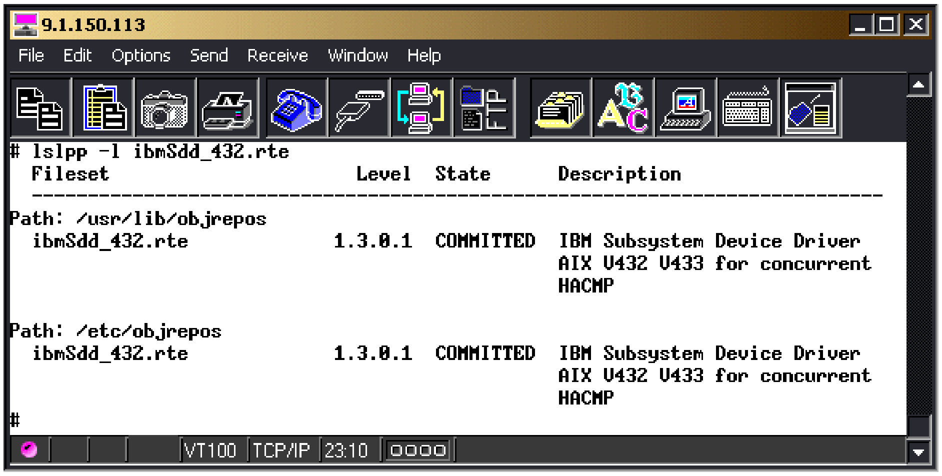
Figure 5-6 SDD installation on AIX - installp command status

Verifying the SDD Installation

To verify that SDD has been successfully installed, issue the `lslpp -l fileset_name` command, where the `fileset_name` is the name of the fileset you have installed, as described in Table 5-6 on page 91.

If you have successfully installed the appropriate IBM Subsystem Device Driver fileset, the output from the `lslpp -l fileset_name` command should look like Figure 5-7, where fileset for AIX 4.3.3 and HACMP in concurrent mode is used. Accordingly to the version of SDD you have installed, the name of fileset in `lslpp -l fileset_name` command output may differ, but the status of installed fileset should remain the same - `COMMITED`. If during installation of the SDD fileset you set the option **COMMIT software updates** to “`no`“ and **SAVE replaced files** to “`yes`”, the status of installed software may be `APPLIED`, which means that the software is installed and operational, but you did not commit the software installation. This is also a correct status for installed software, but until you commit the software installation for that specific fileset, you may roll it back to a previous version, if any previous version was installed. If no previous version of software was installed, the `lslpp` command output will always return `COMMITED` status.

> **Note:** Preserving previous versions of installed software is very useful if you are making a new version software trial installation or you are not sure that new version will not cause any problems on your system, but it always consumes additional disk space. If you have only a limited capacity of free disk space, this may be a cause for software installation failure, due to insufficient free disk space available.



```
# lslpp -l ibmSdd_432.rte
  Fileset                      Level  State      Description
  ----------------------------------------------------------------------------
Path: /usr/lib/objrepos
  ibmSdd_432.rte             1.3.0.1  COMMITTED  IBM Subsystem Device Driver
                                                 AIX V432 V433 for concurrent
                                                 HACMP

Path: /etc/objrepos
  ibmSdd_432.rte             1.3.0.1  COMMITTED  IBM Subsystem Device Driver
                                                 AIX V432 V433 for concurrent
                                                 HACMP
#
```

Figure 5-7 SDD installation on AIX - installp verification

Major files installed with SDD software

Table 5-7 contains the list of major files installed with IBM Subsystem Device Driver and an appropriate description for each of installed files.

Table 5-7 List of major files installed with IBM SDD

File	Description
/usr/lib/methods/defdpo	Define method of the SDD pseudo parent Data Path Optimizer (DPO).
/usr/lib/methods/cfgdpo	Configure method of the SDD pseudo parent DPO.
/usr/lib/methods/define_vp	Define method of the SDD vpath devices.
/usr/sbin/addpaths	The command that dynamically adds more paths to Subsystem Device Driver devices while they are in `Available` state. This command is supported only with SDD for AIX 4.3.2 and higher. It is not available if you have the ibmSdd_421.rte fileset installed.
/usr/lib/methods/cfgvpath	Configure method of SDD vpath devices.
/usr/lib/methods/cfallvpath	Fast-path configure method to configure the SDD pseudo parent dpo and all vpath devices.
/usr/lib/drivers/vpathdd	Subsystem device driver.
/usr/sbin/hd2vp	The SDD script that converts an ESS hdisk device volume group to a Subsystem Device Drive vpath device volume group.

File	Description
/usr/sbin/vp2hd	The SDD script that converts a SDD vpath device volume group to an ESS hdisk device volume group.
/usr/sbin/datapath	The SDD driver console command tool.
/usr/sbin/lsvpcfg	The SDD driver query configuration status command.
/usr/sbin/mkvg4vp	The command that creates a SDD volume group.
/usr/sbin/extendvg4vp	The command that extends SDD devices to a SDD volume group.
/usr/sbin/dpovgfix	The command that fixes a SDD volume group that has mixed vpath and hdisk physical volumes.
/usr/sbin/savevg4vp	The command that backs-up all files belonging to a specified volume group with SDD devices.
/usr/sbin/restvg4vp	The command that restores all files belonging to a specified volume group with SDD devices.

5.4.2 Configuring the Subsystem Device Driver

The following section describes the steps required to properly configure the IBM Subsystem Device Driver in an AIX operating environment.

Preparing to configure the Subsystem Device Driver

Before you configure SDD, ensure that:

- The IBM 2105 Enterprise Storage Server is operational and all ESS LUNs are configured properly for your AIX host system.
- The proper fileset for your environment is installed on the AIX host system (as described in Table 5-6 on page 91).
- The ESS LUNs are seen as `hdisks` and configured correctly on the AIX host system.

Configure the ESS LUNs before you configure the SDD. If you configure multiple paths to an ESS LUN, make sure that all paths (hdisks) are in `Available` state. Otherwise, some SDD devices will lose multiple-path capability.

To check if the ESS LUNs are configured correctly on the AIX operating system level issue the following command: `lsdev -Cc disk | grep 2105`. Look at the command output and check if all hdisks are present and are in `Available` state.

If you have already created some ESS volume groups, vary off (deactivate) all active volume groups with ESS subsystem disks by using the `varyoffvg` (LVM) command.

Attention: Before you vary off a volume group, unmount all file systems in that volume group and close other applications which are directly accessing logical volumes (such as database engines). If some ESS devices (hdisks) are used as physical volumes of an active volume group, and there are file systems of that volume group being mounted, then you must unmount all file systems, and vary off (deactivate) all active volume groups with ESS SDD disks.

Configuring the Subsystem Device Driver

Perform the following steps to configure SDD using SMIT:

1. Log in as `root` user.

2. Type `smitty device` from your desktop window. The Devices screen of SMIT is displayed.
3. Choose **Data Path Device** and press **Enter**. The **Data Path Device** screen of SMIT is displayed.
4. Choose **Define and Configure All Data Path Devices** and press **Enter**. The configuration process begins.
5. Check the SDD configuration status. See "Displaying the ESS vpath device configuration" on page 101 for details.
6. Vary on all previously deactivated ESS volume groups by using the `varyonvg` (LVM) command.
7. If you want to convert the ESS hdisk volume group to SDD vpath devices, you must run the hd2vp utility. See 5.5.5, "SDD utility programs" on page 114 for more detailed information about this utility.
8. Mount the file systems for all volume groups that were previously unmounted.

Tip: The following error might occur if you run the `cfgmgr` command with all `vpath` paths (`hdisks`) in the `Open` state:

```
0514-061 Cannot find a child device
```

Ignore this error if it is returned by the `cfgmgr` command when all `vpath` paths (`hdisks`) are in the `Open` state. You can use the `datapath query device` command to verify the status of all `vpath` paths.

Verifying the SDD configuration

To check the SDD configuration, you can use either the Display Device Configuration screen of SMIT or the `lsvpcfg` console command. Perform the following steps to verify the SDD configuration on an AIX host system:

1. Log in as root. Type smitty device from your desktop window. The Devices screen of SMIT is displayed.
2. Select **Data Path Device** -> **Display Data Path Device Configuration**. Press **Enter** to display the condition (`Defined` or `Available`) of all SDD pseudo devices and the paths to each device.

If any device is listed as `Defined`, the configuration was not successful. Check the configuration procedure again. See "Configuring the Subsystem Device Driver" on page 95 for information about the procedure.

Tip: To verify that multiple paths are properly configured for each LUN configured within the ESS, refer to "Displaying the ESS vpath device configuration" on page 101.

Changing the path-selection policy

IBM SDD supports path-selection policies that increase the performance of a multipathing environment and make path failures transparent to applications. The path selections policies are available on AIX platform only. The following policies are supported:

Load balancing (lb)

The path to use for an I/O operation is chosen by estimating the load on the adapter to which each path is attached. The load is a function of the number of I/O operations currently in process. If multiple paths have the same load, a path is chosen at random from those paths.

Round robin (rr)

The path to use for each I/O operation is chosen at random from those paths not used for the last I/O operation. If a device has only two paths, SDD alternates between the two.

Failover only (fo)

All I/O operations for the device are sent to the same (preferred) path until the path fails because of I/O errors. Then an alternate path is chosen for subsequent I/O operations.

The path-selection policy is set at the SDD device level. The default path-selection policy for a SDD device is load balancing. You can change the policy for a SDD device with the `chdev` command. Before changing the path-selection policy, determine the active attributes for the SDD device. Type the **`lsattr -El vpathX`** command, where `X` represents the `vpath` number. The output should look similar to Example 5-4.

Example 5-4 Example of lsattr -El vpathX command output

```
pvid         0004379001b90b3f0000000000000000 Data Path Optimizer Parent False
policy       df                               Scheduling Policy          True
active_hdisk hdisk1/30C12028                  Active hdisk               False
active_hdisk hdisk5/30C12028
```

The path-selection policy is the only attribute of a SDD device that can be changed. The valid policies are `rr`, `lb`, `fo`, and `df`. Here are the explanations for these policies:

- rr - round robin,
- fo - failover only,
- lb - load balancing,
- df - default policy (load balancing).

> **Attention:** By changing a SDD device's attribute, the `chdev` command unconfigures and then reconfigures the device. You must ensure the device is not in use if you are going to change its attribute. Otherwise, the command fails.

To change the SDD path-selection policy, type the command:

```
chdev -l vpathX -a policy=[rr/fo/lb/df]
```

Adding paths to SDD devices which belongs to a volume group

You can add more paths to SDD devices that belong to a volume group after you have initially configured SDD. This section shows you how to add paths to SDD devices from AIX 4.2.1 and AIX 4.3.2 or higher host systems.

Adding paths from AIX 4.3.2 or higher host systems

If your host system is AIX 4.3.2 or higher, you can use the **`addpaths`** command to add paths to SDD devices of a volume group. The **`addpaths`** command allows you to dynamically add more paths to SDD devices while they are in the `Available` state. It also allows you to add paths to `vpath` devices belonging to active volume groups.

The addpaths command automatically opens a new path (or multiple paths) if the `vpath` is in the `Open` state and if the `vpath` has more than one existing path.

Before you use the **`addpaths`** command, make sure that ESS logical volume sharing is enabled for all applicable devices. You can enable ESS logical volume sharing through the ESS Specialist.

Complete the following steps to add paths to SDD devices with the **`addpaths`** command:

1. Issue the **`lspv`** command to list the physical volumes.

2. Identify the volume group that contain the SDD devices to which you want to add more paths.
3. Verify that *all* the physical volumes belonging to the SDD volume group are SDD devices (**vpathX**). If they are not, you must fix the problem before proceeding to the next step. Otherwise, the entire volume group loses the path-failover protection. You can issue the **dpovgfix vg_name** command to ensure that all physical volumes within the SDD volume group are SDD devices.
4. Terminate all I/O operations in the volume group. The addpaths command is designed to add paths when there are no I/O activities. The command fails if it detects active I/Os.
5. Run the AIX configuration manager in one of the following ways to recognize all new hdisk devices. Ensure that all logical drives on the ESS are identified as hdisks before continuing.
   - Run the **cfgmgr** command *n* times, where *n* represents the maximum used in your environment number of paths to a single ESS LUN(S).
   - Run the **cfgmgr -l [scsiX/fcsX]** command for each relevant SCSI or Fibre Channel adapter.
6. Issue the **addpaths** command from the AIX command line to add more paths to the SDD devices.
7. Type the **lsvpcfg** command from the AIX command line to verify the configuration of the SDD devices in the volume group. SDD devices should show two or more hdisks associated with each SDD device for failover protection to be active.

Adding paths from AIX 4.2.1 host systems

On the AIX 4.2.1 operating system, the addpath command is not available with SDD. To activate additional paths to a SDD device, the related SDD devices must be unconfigured and then reconfigured. The SDD conversion scripts should be run to enable the necessary SDD associations and links between the SDD vpath (pseudo) devices and the ESS hdisk devices.

Important: Ensure that logical volume sharing is enabled at the ESS for all applicable LUN(s). See *IBM TotalStorage Enterprise Storage Server Web Interface User's Guide*, SC26-7346, for information about enabling volume sharing.

Perform the following steps to activate additional paths to SDD devices belonging to a volume group from your AIX 4.2.1 host system:

1. Identify the volume groups containing the SDD devices to which you want to add additional paths. To do this, type the command **lspv**.
2. Verify that *all* the physical volumes belonging to the SDD volume group are SDD devices (vpathX). If they are not, you must fix the problem before proceeding to the next step. Otherwise, the entire volume group loses the path-failover protection. You can issue the **dpovgfix vg_name** command to ensure that all physical volumes within the SDD volume group are SDD devices.
3. Identify the associated file systems for the selected volume group. Type the command:

   ```
   lsvgfs vg_name
   ```

4. Identify the associated mounted file systems for the selected volume group. To do this type the command:

   ```
   mount
   ```

5. Unmount the file systems of the selected volume group listed in step 3. Type the command:

```
umount name_of_mounted_filesystem
```

6. Run the vp2hd volume group conversion script to convert the volume group from SDD devices to ESS hdisk devices. To do this type the command:

   ```
   vp2hd vg_name
   ```

   When the conversion script completes, the volume group is in the Active condition (varied on).

7. Vary off the selected volume group in preparation for SDD reconfiguration. Type the command:

   ```
   varyoffvg vg_name
   ```

8. Run the AIX configuration manager in one of the following ways to recognize all new hdisk devices. Ensure that all logical drives on the ESS are identified as hdisks before continuing.

   – Run the **cfgmgr** command *n* times, where *n* represents the maximum used in your environment number of paths to a single ESS LUN(S), or

   – Run the **cfgmgr -l [scsiX/fcsX**] command for each relevant SCSI or Fibre Channel adapter.

9. Ensure that all logical drives on the ESS are identified as hdisks before continuing.

10. Unconfigure affected SDD devices to the Defined condition by using the **rmdev -l vpathX** command, where X represents the vpath number you want to set to the Defined condition. This command allows you to unconfigure only SDD devices for which you are adding paths.

    > **Note:** Use the rmdev -l dpo -R command if you need to unconfigure *all* Subsystem Device Driver devices. SDD volume groups must be inactive before unconfiguring. This command attempts to unconfigure *all* SDD devices recursively.

11. Reconfigure SDD devices by using either the System Management Interface Tool (SMIT) or the command-line interface.

    a. If you are using SMIT, type smitty device from your desktop window. The Devices screen of SMIT is displayed. Choose **Data Path Devices** -> **Define and Configure All Data Path Devices** and press **Enter**. SMIT executes a script to define and configure all SDD devices that are in the Defined condition.

    b. If you are using the command line interface, type the **mkdev -l vpathX** command for each SDD device or type the cfallvpath command to configure all SDD devices.

12. Verify your vpaths configuration using either SMIT or the command line interface.

    a. If you are using SMIT, type smitty device from your desktop window. The Devices screen of SMIT is displayed. Choose **Data Path Devices** -> **Display Data Path Device Configuration** and press Enter.

    b. If you are using the command line interface, type the lsvpcfg command to display the SDD configuration status.

    SDD devices should show two or more hdisks associated with each SDD device for failover protection to be active.

13. Vary on the volume groups selected in Step 3. Type the command:

    ```
    varyonvg vg_name
    ```

14. Run the hd2vp script to convert the volume group from ESS hdisk devices back to SDD vpath devices. To do this type the command:

    ```
    hd2vp vg_name
    ```

15. Mount all file systems for the volume groups that were previously unmounted.

5.4.3 Unconfiguring the Subsystem Device Driver devices

Before you can unconfigure SDD devices, all the file systems belonging to the SDD volume groups *must* be unmounted. Then, run the `vp2hd` conversion script to convert the volume group from SDD devices (`vpathX`) to ESS subsystem devices (`hdisks`).

> **Important:** If you are running IBM HACMP/6000 with `ibmSdd_510nchacmp.rte` or `ibmSdd_433.rte` for SDD 1.3.0.x fileset installed on your host system, there are special requirements regarding unconfiguring and removing SDD 1.3.0.x. `vpath` devices. See the 5.6.3, "Special requirements for HACMP/6000" on page 120 for details.

Using either the System Management Interface Tool (SMIT) or the command line interface, you can unconfigure the SDD devices in two ways:

- Without deleting the device information from the Object Database Management (ODM) database. All information about the device and its configuration remains in the ODM database as well as the operating system reports that the device is in the `Defined` condition. You can return the device to the `Available` condition using **`mkdev -l vpathX`** command.
- Deleting device information from the ODM database at the same time you delete the device from the operating system. All information related to this device and its configuration are deleted from the ODM database and you are unable to return the device to the `Available` state, unless you follow the procedure described in 5.4.2, "Configuring the Subsystem Device Driver" on page 95 (assuming, that underlying `hdisks` are still available).

Perform the following steps to unconfigure SDD devices:

1. Log in as `root` user.
2. Type `smitty device` from your desktop window. The Devices screen of SMIT is displayed.
3. Select **Devices** -> **Data Path Device** -> **Remove a Data Path Device**. Press **Enter**. A list of all SDD devices and their conditions (either `Defined` or `Available`) is displayed.
4. Select the device that you want to unconfigure and whether or not you want to delete the device information from the ODM database. Press **Enter**. The device is unconfigured to the `Defined` condition or completely removed from the system (depending upon your selection).
5. To unconfigure more SDD devices you have to repeat step 4 for each SDD device.

> **Tip:** The fast-path command to unconfigure all SDD devices from the `Available` to the `Defined` condition is: `rmdev -l dpo -R`
>
> The fast-path command to remove all Subsystem Device Driver devices from your system is: `rmdev -dl dpo -R`

5.4.4 Removing the Subsystem Device Driver

Before you remove the SDD package from your AIX host system, all the SDD devices *must* be removed from your host system. The `rmdev -dl dpo -R` command removes all the SDD devices from your system. After all SDD devices are removed, perform the following steps to remove SDD fileset:

1. Log in as `root` user.
2. Type `smitty deinstall` from your desktop window. The Remove Installed Software screen of SMIT is displayed.
3. Type `ibmSdd_421.rte`, `ibmSdd_432.rte`, `ibmSdd_433.rte`, `ibmSdd_510.rte`, or `ibmSdd_510nchacmp.rte` in the SOFTWARE name field and press **Enter**.
4. Press the **Tab** key in the PREVIEW Only? field to toggle between `Yes` and `No`. Select `No` to remove the software package from your AIX host system. If you select `Yes`, the process previews only what you are removing. The results of your pre-check are displayed without removing the software. If the condition for any SDD device is either `Available` or `Defined`, the process fails.
5. Select `No` for the remaining fields on this panel. Press **Enter**.
6. The ARE YOU SURE pop-up window displays. Press **Enter** to continue with the removal process. This might take a few minutes.
7. When the process is complete, the SDD software package is removed from your system.

5.5 Using IBM SDD on AIX host

This section provides instructions for using the IBM Subsystem Device Driver on AIX host. It explains how to configure SDD to provide I/O load-balancing and path failover protection.

5.5.1 Providing load-balancing and failover protection

IBM Subsystem Device Driver provides load-balancing and failover protection for AIX applications and for the LVM when ESS `vpath` devices are used. These devices *must* have a minimum of two paths to a physical Logical Unit Number (LUN) within an ESS for failover protection to exist.

Displaying the ESS vpath device configuration

To display which ESS vpath devices are available to provide failover protection, choose the **Display Data Path Device Configuration** SMIT screen, or run the `lsvpcfg` command. Perform the following steps if you want to use to use SMIT:

1. Type `smitty device` from your desktop window. The Devices screen of SMIT displayed.
2. Select **Data Path Devices** -> **Display Data Path Device Configuration**. Press **Enter**. You will see an output similar to shown in Example 5-5.

Example 5-5 Example of vpaths configuration

```
vpath0 (Avail pv vpathvg) 018FA067 = hdisk1 (Avail )
vpath1 (Avail ) 019FA067 = hdisk2 (Avail )
vpath2 (Avail ) 01AFA067 = hdisk3 (Avail )
vpath3 (Avail ) 01BFA067 = hdisk4 (Avail ) hdisk27 (Avail )
vpath4 (Avail ) 01CFA067 = hdisk5 (Avail ) hdisk28 (Avail )
vpath5 (Avail ) 01DFA067 = hdisk6 (Avail ) hdisk29 (Avail )
vpath6 (Avail ) 01EFA067 = hdisk7 (Avail ) hdisk30 (Avail )
vpath7 (Avail ) 01FFA067 = hdisk8 (Avail ) hdisk31 (Avail )
vpath8 (Avail ) 020FA067 = hdisk9 (Avail ) hdisk32 (Avail )
vpath9 (Avail pv vpathvg) 02BFA067 = hdisk20 (Avail ) hdisk44 (Avail )
vpath10 (Avail pv vpathvg) 02CFA067 = hdisk21 (Avail ) hdisk45 (Avail )
vpath11 (Avail pv vpathvg) 02DFA067 = hdisk22 (Avail ) hdisk46 (Avail )
vpath12 (Avail pv vpathvg) 02EFA067 = hdisk23 (Avail ) hdisk47 (Avail )
vpath13 (Avail pv vpathvg) 02FFA067 = hdisk24 (Avail ) hdisk48 (Avail )
```

The following information is displayed:

- The name of each SDD vpath device, such as vpath1.
- The configuration condition of the SDD vpath device. It is either in Defined or Available state. There is no failover protection if only one path is in the Available condition. At least two paths to each SDD vpath device must be in the Available condition to have failover protection.
- The name of the volume group to which the device belongs, such as vpathvg.
- The unit serial number of the ESS LUN, such as 018FA067.
- The names of the AIX disk devices that comprise the SDD vpath devices, their configuration conditions, and the physical volume status.

In Example 5-5 vpath devices vpath0, vpath1, and vpath2 have a single path and therefore do not provide failover protection. The other ESS vpath devices each have two paths and therefore can provide failover protection.

Attention: The configuration condition also indicates whether or not the SDD vpath device is defined to AIX as a physical volume (pv flag). If pv flag is displayed for both SDD vpath devices and ESS hdisk devices that comprise that SDD vpath, you might not have failover protection. Run the dpovgfix command to fix this problem.

You can also use the datapath command to display information about a SDD vpath device. This command displays the number of paths to the device. The datapath query device 10 command might produce output similar to Example 5-6, where we can see that vpath10 device has two operational paths to the ESS LUN. Each single path is seen at the operating system level as hdisk21 and hdisk45.

Example 5-6 Example of datapath query device command output

```
DEV#: 10    DEVICE NAME: vpath10    TYPE: 2105B09    SERIAL: 02CFA067
==================================================================
Path#      Adapter/Hard Disk   State     Mode    Select        Errors
    0          scsi6/hdisk21    OPEN   NORMAL        44             0
    1          scsi5/hdisk45    OPEN   NORMAL        43             0
```

Configuring a volume group for failover protection

It is possible to create a volume group that has only a single path (like devices vpath0, vpath1, and vpath2 shown in Example 5-5) and then add paths later by reconfiguring the ESS. However, a SDD volume group does not have the failover protection if any of its physical volumes has only a single path to the ESS LUN.

Perform the following steps to create a new volume group with SDD vpaths:

1. Log in as root.
2. Type smitty from your desktop window and press **Enter**. This will invoke the non-graphical, text-based System Management Interface Tool.
3. Select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Volume Group** and press **Enter**. The Add Volume Group with Data Path Devices screen of SMIT is displayed.
4. Select **Add Volume Group with Data Path Devices** and press **Enter**. Highlight the PHYSICAL VOLUME names field and press **F4** (or **ESC** + **4**) to list all available vpaths.

Important: If you use a script file to create a volume group, you *must* modify your script file and replace the mkvg command with the mkvg4vp command to use SDD vpath devices.

All the functions that apply to a regular volume group also apply to a SDD volume group. You can create a logical volume group (mirrored, striped, or compressed) or a file system (mirrored, striped, or compressed) on a SDD volume group and SDD vpaths.

Once you create the volume group, AIX creates the SDD vpath device as a physical volume (the pv flag is set for that vpath device). In Example 5-5 on page 101, vpath9 through vpath13 are included in a volume group vpathvg and they become physical volumes. Also vpath0 is included in that volume group, but since it comprise only of one path (hdisk1), the whole vpathvg volume group may not be able to provide path failover mechanism.

To list all the physical volumes known to AIX, use the **lspv** command. Any ESS vpath devices that were created into physical volumes are included in the output of that command, which may look similar to Example 5-7, where output of the command is consistent with vpaths configuration shown in Example 5-5 on page 101.

Example 5-7 Example of lspv command output

```
hdisk0        0001926922c706b2    rootvg
hdisk1        none                None
...
hdisk10       none                None
hdisk11       00000000e7f5c88a    None
...
hdisk48       none                None
hdisk49       00000000e7f5c88a    None
vpath0        00019269aa5bc858    vpathvg
vpath1        none                None
vpath2        none                None
vpath3        none                None
vpath4        none                None
vpath5        none                None
vpath6        none                None
vpath7        none                None
vpath8        none                None
vpath9        00019269aa5bbadd    vpathvg
vpath10       00019269aa5bc4dc    vpathvg
vpath11       00019269aa5bc670    vpathvg
vpath12       000192697f9fd2d3    vpathvg
vpath13       000192697f9fde04    vpathvg
```

To display the devices that comprise a volume group, enter the lsvg -p vg_name command. The lsvg -p vpathvg command might produce the output shown in Example 5-8.

Example 5-8 Example of lsvg -p vg_name command output

```
PV_NAME           PV STATE    TOTAL PPs   FREE PPs    FREE DISTRIBUTION
vpath0            active      29          4           00..00..00..00..04
vpath9            active      29          4           00..00..00..00..04
vpath10           active      29          4           00..00..00..00..04
vpath11           active      29          4           00..00..00..00..04
vpath12           active      29          4           00..00..00..00..04
vpath13           active      29          28          06..05..05..06..06
```

Importing a volume group with SDD

You can import a new volume group definition from a set of physical volumes with SDD `vpath` devices using the Volume Groups screen of System Management Interface Tool. To use this command, you must either have `root` user authority or be a member of the `system` group.

> **Attention:** IBM SDD does not automatically create the `pvid` attribute in the `ODM` database for each `vpath` device. The AIX disk driver automatically creates the `pvid` attribute in the `ODM` database, if a `pvid` exists on the physical device. Therefore, the *first time* you import a new SDD volume group to a new cluster node, you *must* import the volume group using `hdisks` as physical volumes. Next, run the `hd2vp` conversion script (see 5.5.5, “SDD utility programs” on page 114) to convert the volume group's physical volumes from ESS `hdisks` to `vpath` devices. This conversion step not only creates `pvid` attributes for *all* `vpath` devices which belong to that imported volume group, but it also deletes the `pvid` attributes for these `vpath` devices underlying hdisks. Later on you can import and vary on the volume group directly from the `vpath` devices. These special requirements apply to both concurrent and non-concurrent volume groups.

Under certain conditions, the state of a `pvid` on a system is not always as we expected. So it is necessary to determine the state of a `pvid` as displayed by the `lsvp` command, in order to select the appropriate action. There are four possible scenarios, which are shown in Example 5-9, Example 5-10, Example 5-11 and Example 5-12.

Example 5-9 Scenario 1, where lspv displays pvid's for both hdisks and vpath

```
>lspv
hdisk1          003dfc10a11904fa          None
hdisk2          003dfc10a11904fa          None
vpath0          003dfc10a11904fa          None
```

Example 5-10 Scenario 2, where lspv displays pvid's for hdisks only

```
>lspv
hdisk1          003dfc10a11904fa          None
hdisk2          003dfc10a11904fa          None
vpath0          None                      None
```

For both Scenario 1 and Scenario 2, the volume group should be imported using the `hdisk` names and then converted using the `hd2vp` command:

```
importvg -y vg_name -V vg_major_number hdisk1
hd2vp vg_name
```

Example 5-11 Scenario 3, where lspv displays pvid's for vpaths only

```
>lspv
hdisk1          None                      None
hdisk2          None                      None
vpath0          003dfc10a11904fa          None
```

For Scenario 3, the volume group should be imported using the `vpath` name. To do this issue the following command:

```
importvg -y vg_name -V vg_major_number vpath0
```

Example 5-12 Scenario 4, where lspv does not display the pvid's nor for vpaths and hdisks

```
>lspv
hdisk1          None                      None
hdisk2          None                      None
```

```
vpath0          None                      None
```

For Scenario 4, the `pvid` will need to be placed in the `ODM` database for the `vpath` devices and then the volume group can be imported using the `vpath` name:

```
chdev -l vpath0 -a pv=yes
importvg -y vg_name -V vg_major_number vpath0
```

You can also use SMIT to import a volume group with SDD devices or hdisk devices:

1. Log in as `root`.
2. Type **`smitty`** from your desktop window. The System Management Interface Tool is displayed.
3. Select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Volume Groups** -> **Import a Volume Group**. Press **Enter**. The Import a Volume Group screen of SMIT is displayed.
4. In the Import a Volume Group panel, perform the following tasks:
   a. Type in the volume group name under which you want to import the volume group.
   b. Type in the physical volumes that you want to import over. You can press the **F4** key (or **ESC** + **4**) for a list of choices.
   c. Press **Enter** after making all desired changes.

Exporting a volume group with SDD

You can export a volume group definition from a set of physical volumes with SDD `vpath` devices using the Volume Groups screen of SMIT.

The `exportvg` command removes the definition of the volume group specified by the Volume Group parameter from the system. Since all system knowledge of the volume group and its contents are removed, an exported volume group is no longer accessible. The `exportvg` command does not modify any user data in the volume group.

A volume group is a non-shared resource within the system. It should not be accessed by another system until it has been explicitly exported from its current system and imported on another. The primary use of the `exportvg` command, coupled with the `importvg` command, is to allow portable volumes to be exchanged between systems. Only a complete volume group can be exported, not individual physical volumes.

Using the **`exportvg`** command and the **`importvg`** command, you can also switch ownership of data on physical volumes shared between two systems. To use this command, you must either have `root` user authority or be a member of the `system` group.

Perform the following steps to export a volume group with SDD devices:

1. Log in as `root`.
2. Type **`smitty`** from your desktop window. The System Management Interface Tool is displayed.
3. Select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Volume Groups** -> **Export a Volume Group**. Press **Enter**. The Export a Volume Group screen of SMIT is displayed.
4. Type in the volume group name you want to export and press **Enter**. You can press the **F4** key (or **ESC** + **4**) to generate a list of choices.

How failover protection can be lost

AIX can only create volume groups from disk (or pseudo) devices that are physical volumes. If a volume group is created using a device that is not a physical volume, AIX makes it a physical volume as part of the procedure of creating the volume group. A physical volume has a physical volume identifier (`pvid`) written on its sector 0 and also has a `pvid` attribute attached to the device attributes in the `CuAt` object of `ODM` database. The `lspv` command lists all the physical volumes known to AIX as shown in Example 5-7 on page 103.

In some cases, access to data is not lost, but failover protection might not be present. Failover protection can be lost in several ways:

- Through the loss of a device path
- By creating a volume group from single-path `vpath` (pseudo) devices
- As a side effect of running the disk change method
- Through running the `mksysb` restore command
- By manually deleting devices and running the configuration manager (`cfgmgr`)

Here is more information about the ways that failover protection can be lost.

Through the loss of a device path

Due to hardware errors, SDD might remove one or more paths to a `vpath` pseudo device. A pseudo device loses failover protection when it only has a single path. You can use the datapath query device command to show the state of paths to a pseudo device. You cannot use any paths in the `Dead` state for I/O operations.

By creating a volume group from single-path vpath (pseudo) devices

A volume group created using any single-path pseudo devices does not have failover protection because there is no alternate path to the ESS LUN(s).

As a side effect of running the disk change method

It is possible to modify attributes for an `hdisk` device by running the `chdev` command. The `chdev` command invokes the `hdisk` configuration method to make the requested change. In addition, the `hdisk` configuration method sets the `pvid` attribute for an `hdisk` if it determines that the hdisk has a `pvid` written on sector 0 of the LUN. This causes the `vpath` pseudo device and one or more of its `hdisks` to have the same `pvid` attribute in the `ODM` database. If the volume group containing the `vpath` pseudo device is activated, the `LVM` uses the first device it finds in the `ODM` with the desired `pvid` to activate the volume group.

Assume, that configuration of all `vpaths` in the operating system is as shown in Example 5-5 on page 101. By issuing the command `chdev -l hdisk46 -a queue_depth=30` (which could also set the `pvid` attribute in the `ODM` database for an `hdisk`), the output of the `lsvpcfg` command would look similar to Example 5-13.

Example 5-13 Example of lsvpcfg command output after setting pvid attribute for an hdisk

```
vpath0 (Avail pv vpathvg) 018FA067 = hdisk1 (Avail )
vpath1 (Avail ) 019FA067 = hdisk2 (Avail )
vpath2 (Avail ) 01AFA067 = hdisk3 (Avail )
vpath3 (Avail ) 01BFA067 = hdisk4 (Avail ) hdisk27 (Avail )
vpath4 (Avail ) 01CFA067 = hdisk5 (Avail ) hdisk28 (Avail )
vpath5 (Avail ) 01DFA067 = hdisk6 (Avail ) hdisk29 (Avail )
vpath6 (Avail ) 01EFA067 = hdisk7 (Avail ) hdisk30 (Avail )
vpath7 (Avail ) 01FFA067 = hdisk8 (Avail ) hdisk31 (Avail )
vpath8 (Avail ) 020FA067 = hdisk9 (Avail ) hdisk32 (Avail )
vpath9 (Avail pv vpathvg) 02BFA067 = hdisk20 (Avail ) hdisk44 (Avail )
vpath10 (Avail pv vpathvg) 02CFA067 = hdisk21 (Avail ) hdisk45 (Avail )
vpath11 (Avail pv vpathvg) 02DFA067 = hdisk22 (Avail ) hdisk46 (Avail pv vpathvg)
vpath12 (Avail pv vpathvg) 02EFA067 = hdisk23 (Avail ) hdisk47 (Avail )
```

```
vpath13 (Avail pv vpathvg) 02FFA067 = hdisk24 (Avail ) hdisk48 (Avail )
```

The output of the `lsvpcfg` command shows that `vpath11` contains `hdisk22` and `hdisk46`. However, `hdisk46` is the one with the `pv` attribute set. If you run the `lsvg -p vpathvg` command again, you might see output similar to Example 5-14.

Example 5-14 Example of lsvg -p vg_name command output for mixed volume group

```
PV_NAME           PV STATE    TOTAL PPs   FREE PPs    FREE DISTRIBUTION
vpath0            active      29          4           00..00..00..00..04
vpath9            active      29          4           00..00..00..00..04
vpath10           active      29          4           00..00..00..00..04
hdisk46           active      29          4           00..00..00..00..04
vpath12           active      29          4           00..00..00..00..04
vpath13           active      29          28          06..05..05..06..06
```

Notice that now device `vpath11` has been replaced by `hdisk46`. That is because `hdisk46` is one of the hdisk devices included in `vpath11`, it has a `pvid` attribute in the `ODM` database and a record for `hdisk46` is stored in the ODM earlier than the record for `vpath11`. AIX LVM always uses the *first* data occurrence in ODM which match the required criteria. In this example, the criteria is `pvid` field and this is the reason why LVM used `hdisk46` instead of `vpath11` when it activated volume group `vpathvg`. The volume group is now in a mixed mode of operation because it partially uses `vpath` pseudo devices and partially uses `hdisk` devices. This is a problem that must be fixed because failover protection is effectively disabled for the `vpath11` physical volume of the `vpathvg` volume group.

Note: The way to fix this problem with the mixed volume group is to run the `dpovgfix vg_name` command after running the chdev command.

Through running the mksysb restore command

If a system is restored from a `mksysb` restore file or tape, the `vpath` pseudo device `pvid` attribute is not set. All logical volumes made up of `vpath` pseudo devices use `hdisk` devices instead of `vpath` devices. You can correct the problem by using the `hd2vp` shell script to convert the volume group back to using `vpath` devices. This is necessary to run `hd2vp` script every time the operating system is restored from the `mksysb` image and multipathing is in place.

By manually deleting devices and running the configuration manager (cfgmgr)

Assume that in the situation shown in Example 5-5 on page 101, `vpath3` is made up of `hdisk4` and `hdisk27` and additionally `vpath3` is currently a physical volume. If the `vpath3`, `hdisk4`, and `hdisk27` devices are all deleted by using the `rmdev` command and then `cfgmgr` is invoked at the command line, only one path of the original `vpath3` is configured by AIX. The following set of commands would produce this situation:

```
rmdev -dl vpath3; rmdev -dl hdisk4; rmdev -dl hdisk27
cfgmgr
```

The `datapath query device` command displays the vpath3 configuration status.

Next, all paths to the vpath must be restored. You can restore the paths in one of the following ways:

- Run `cfgmgr` once for each installed SCSI or Fibre Channel adapter.
- Run `cfgmgr` *n* times, where *n* represents the number of paths per SDD device.

Running the AIX configuration manager (`cfgmgr`) *n* times for *n*-path configurations of ESS devices is not always required. It depends on whether the ESS device has belonged as a physical volume to a volume group or not. If it has, it is necessary to run `cfgmgr` *n* times for a *n*-path configuration. Since the ESS device has been used as a physical volume in a volume group, it has a `pvid` value written on its sector 0. When the first SCSI or Fibre Channel adapter is configured by `cfgmgr`, the AIX disk driver configuration method creates a `pvid` attribute in the AIX `ODM` database with the `pvid` value it read from the device. It then creates a logical name (`hdiskX`), and puts the `hdiskX` in the `Defined` condition.

When the second adapter is configured, the AIX disk driver configuration method reads the `pvid` from the same device again, and searches the `ODM` database to see if there is already a device with the same `pvid` in the `ODM`. If there is a match, and that hdiskX is in a `Defined` condition, the AIX disk driver configuration method does not create another `hdisk` logical name for the same device. That is why only one set of `hdisks` gets configured the first time `cfgmgr` runs. When `cfgmgr` runs for the second time, the first set of `hdisks` are in the `Available` condition, so a new set of `hdisks` are `Defined` and configured to the `Available` condition. That is why you must run `cfgmgr` *n* times to get *n* paths configured.

If the ESS device has never belonged to a volume group, that means there is no `pvid` written on its sector 0. In that case, you only need to run `cfgmgr` command once to get all multiple paths configured.

> **Note:** The `addpaths` command allows you to dynamically add more paths to Subsystem Device Driver devices while they are in `Available` state. In addition, this command allows you to add paths to `vpath` devices (which are then opened) belonging to active volume groups.
>
> This command will open a new path (or multiple paths) automatically if the `vpath` is in the `Open` state, and the original number of path of the `vpath` is more than one. You can use either the Add Paths to Available Data Path Devices SMIT screen, or run the `addpaths` command from the AIX command line. See “Adding paths to SDD devices which belongs to a volume group” on page 97 for more information about the addpaths command.
>
> SDD does not support the addpaths command for AIX 4.2.1.

If you have the `ibmSdd_421.rte` fileset installed, you can run the `cfgmgr` command instead of restarting the system. After all the ESS `hdisk` devices are restored, you must unconfigure all SDD devices to the `Defined` condition. Then reconfigure the SDD devices to the `Available` condition in order to restore all paths to the SDD (`vpath`) devices.

Below are examples of commands which can be used to unconfigure or configure SDD devices:

- To unconfigure a *single* SDD device to the `Defined` condition type the command `rmdev -l vpathX`.
- To unconfigure *all* SDD devices to the `Defined` condition type the command `rmdev -l dpo -R`.
- To configure a *single* `vpath` device to the `Available` condition type the command `mkdev -l vpathX`.
- To configure *all* `vpath` devices to the `Available` condition type the command `cfallvpath`.

Recovering from mixed volume groups

Run the dpovgfix shell script to recover a mixed volume group. The syntax for that command is dpovgfix vg_name. The script tries to find a pseudo device corresponding to each hdisk in the volume group and replaces the hdisk with the vpath pseudo device. In order for the shell script to be executed, all mounted file systems of this volume group have to be unmounted. After successful completion of the dpovgfix shell script, mount the file systems again.

Extending an existing SDD volume group

You can extend a volume group with SDD vpath devices using the Logical Volume Groups screen of SMIT. The SDD vpath devices to be added to the volume group should be chosen from those that can provide failover protection. It is possible to add a SDD vpath device to a SDD volume group that has only a single path (like vpath0 in Example 5-5 on page 101) and then add paths later by reconfiguring the ESS. However, with a single path, failover protection is not provided for that specific vpath. See “Adding paths to SDD devices which belongs to a volume group” on page 97 for more information about adding paths to a SDD device.

Perform the following steps to extend a volume group with SDD devices:

1. Log in as root.
2. Type **smitty** from your desktop window. The System Management Interface Tool is displayed.
3. Select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Volume Group** -> **Add Volume Group with Data Path Devices**.
4. Type in the volume group name and physical volume name and press **Enter**. You can also use the **F4** key (or **ESC** + **4**) to list all the available SDD devices, and then select the devices you want to add to the volume group.

Important: If you use a script file to extend an existing SDD volume group, you *must* modify your script file and replace the extendvg command with the extendvg4vp command.

Backing-up all files belonging to a SDD volume group

You can back up all files belonging to a specified volume group with Subsystem Device Driver vpath devices using the Volume Groups screen of SMIT.

Perform the following steps to back up all files belonging to a SDD volume group:

1. Log in as root.
2. Type smitty from your desktop window. The System Management Interface Tool is displayed.
3. Select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Volume Groups** -> **Back Up a Volume Group with Data Path Devices**. The Back Up a Volume Group with Data Path Devices screen of SMIT is displayed.
4. In the Back Up a Volume Group with Data Path Devices screen, perform the following steps:
   a. Type in the Backup DEVICE or FILE name.
   b. Type in the Volume Group to back up.
   c. Press **Enter** after making all desired changes.

   You can also use the **F4** key (or **ESC** + **4**) to list all available backup devices as well as to generate a list of all configured volume groups.

Important: If you use a script file to back up all files belonging to a specified SDD volume group, you *must* modify your script file and replace the `savevg` command with the `savevg4vp` command.

Backing-up files (running the **`savevg4vp`** command) will result in the loss of all data previously stored on the selected output medium. Data integrity of the archive may be compromised if a file is modified during system backup. Keep system activity at a minimum during the system backup procedure.

Restoring all files belonging to a SDD volume group

You can restore all files belonging to a specified volume group with Subsystem Device Driver `vpath` devices using the Volume Groups screen of SMIT.

Perform the following steps to restore all files belonging to a specified SDD volume group:

1. Log in as `root`.
2. Type `smitty` from your desktop window. The System Management Interface Tool is displayed.
3. Select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Volume Groups** -> **Remake a Volume Group with Data Path Devices**. The Remake a Volume Group with Data Path Devices screen of SMIT is displayed.
4. In the Remake a Volume Group with Data Path Devices screen, type in the Restore DEVICE or FILE name and press **Enter**. You can also use the **F4** key (or **ESC** + **4**) to generate a list of all available restore devices.

Important: If you use a script file to restore all files belonging to a specified SDD volume group, you *must* modify your script file and replace the `restvg` command with the `restvg4vp` command.

SDD-specific SMIT panels

SDD supports several specialized SMIT screens. Some SMIT screens provide SDD-specific functions, while other provide AIX functions, but requires SDD specific commands. Table 5-8 lists all SDD specific SMIT screens.

Table 5-8 List of all SDD specific SMIT screens

SMIT screen title	Description
Display Data Path Device Configuration	Displays configuration of all or several vpaths
Display Data Path Device Status	Displays the status of all or several vpaths
Display Data Path Device Adapter Status	Displays the status of all or several adapters
Define and Configure all Data Path Devices	Configures all vpaths
Add Paths to Available Data Path Devices Please note, that this screen of SMIT is available only for AIX 4.3.2 or higher.	Adds path or paths to already configured vpath which are in Available state
Configure a Defined Data Path Device	Configures a vpath which is in a Defined state
Remove a Data Path Device	Removes a specified vpath device
Add a Volume Group with Data Path Devices	Adds a new volume group using vpath devices

SMIT screen title	Description
Add a Data Path Volume to a Volume Group	Extends an existing volume group with new vpath device
Remove a copy from a datapath Logical Volume	Removes a mirror copy for specified logical volume
Back Up a Volume Group with Data Path Devices	Backs up a specified volume group
Remake a Volume Group with Data Path Devices	Restores a volume group from a specified restore device

5.5.2 Using ESS devices directly

After you configure the SDD, it creates SDD devices (`vpath` devices) for all available ESS LUNs. ESS LUNs are accessible through the connection between the AIX host server SCSI or Fibre Channel adapter and the ESS ports. The AIX disk driver creates the original ESS devices (`hdisks`). Therefore, with SDD, an application has two ways to access ESS devices:

- Using `hdisk` devices, which means that you are not using the SDD load-balancing and failover features
- Using `vpath` devices, which *must* be used if you want to use the SDD load-balancing and failover features.

You can still access vpath devices in raw mode (raw device) or through the AIX logical volume manager (LVM). For applications which are accessing vpath devices through the LVM you must create a volume group with the SDD vpath devices.

If your application used ESS hdisk device special files directly before installing SDD, convert it to using the SDD vpath device special files. To do this, perform the following steps, after installing SDD:

1. Log in as `root`.
2. Type `smitty` from your desktop window. The System Management Interface Tool is displayed.
3. Select **Devices** -> **Data Path Devices** -> **Display Data Path Device Configuration**. The system displays all SDD `vpaths` with their attached multiple paths (`hdisks`).
4. Search the list of `hdisks` to locate the `hdisks` your application is using.
5. In your application replace each `hdisk` with its corresponding SDD `vpath` device. Depending upon your application, the manner in which you replace these files is different. In this book we do not cover how to replace special storage files for particular applications. Refer to your application’s administration guide for more information.

Tip: Alternately, you can type `lsvpcfg` from the command-line interface rather than using SMIT. This displays all configured SDD `vpath` devices and their underlying paths (`hdisks`).

5.5.3 Using ESS devices through AIX LVM

If your application accesses ESS devices through LVM, determine the volume group that it uses before you convert volume groups. To avoid any potential problems we strongly recommend that you use the System Management Interface Tool instead of ordinary LVM command-line commands (like `mkvg`). Otherwise, the path failover capacity could be lost.

To convert the volume group from the original ESS device hdisks to the SDD `vpaths` perform the following steps:

1. Log in as `root`.
2. Determine the logical volumes that your application accesses and write down their mount points:
   a. Type `smitty` from your desktop window. The System Management Interface Tool is displayed.
   b. To determine the logical volumes select **System Storage Management (Physical & Logical Storage)** -> **Logical Volume Manager** -> **Logical Volumes** -> **List All Logical Volumes by Volume Group** to determine the logical volumes that belong to this volume group and their logical volume mount points. Press **Enter**. The logical volumes are listed by volume group.
3. Unmount these of the selected file systems, which are mounted.
4. Close the applications that are directly accessing other logical volumes in this volume group (such as database engines).
5. Enter the command `hd2vp vg_name` to convert the volume group from the original ESS hdisks to SDD vpaths.
6. When the conversion is complete, mount all file systems that you previously unmounted.

When the conversion is complete, your application now accesses ESS physical LUNs through SDD vpath devices. This provides load balancing and failover protection for your application.

5.5.4 Migrating non-SDD volume group to ESS SDD multipath volume group

Before you migrate your non-SDD volume group to a SDD volume group, make sure that you have completed the following tasks:

1. The SDD for the AIX host system is installed and configured. See 5.4.1, “Installing the IBM Subsystem Device Driver” on page 92 for details.
2. The ESS subsystem devices to which you want to migrate have multiple paths configured for each LUN. See “Displaying the ESS vpath device configuration” on page 101 for details.
3. Make sure the SDD `vpath` devices you are going to migrate to do not belong to any other volume group, and that the corresponding physical device (ESS LUN) does not have a `pvid` written on it. Use the `lsvpcfg` command output to check the SDD `vpath` devices that you are going to use for migration. Make sure there is no `pv` displayed for this `vpath` and its paths (`hdisks`). If a LUN has never belonged to any volume group, there is no `pvid` written on it. In case there is a `pvid` written on the LUN and the LUN does not belong to any volume group, you need to clear the `pvid` from the LUN before using it to migrate a volume group. The commands to clear the `pvid` are:

   ```
   chdev -l hdiskX -a pv=clear
   chdev -l vpathX -a pv=clear
   ```

   See “How failover protection can be lost” on page 106 and “Recovering from mixed volume groups” on page 109 for additional information related to this problem.

> **Note:** Exercise care when clearing a pvid from a device with this command. Issuing this command to a device that does belong to an existing volume group can cause system failures.

You should complete the following steps to migrate a non-SDD volume group to a multipath SDD volume group in concurrent mode:

1. Add new SDD `vpath` devices to an existing non-SDD volume group:
   a. Type **`smitty`** and press Enter from your desktop window. The System Management Interface Tool screen is displayed.
   b. Select **System Storage Management (Physical & Logical)** -> **Logical Volume Manager** -> **Volume Group** -> **Add a Data Path Volume to a Volume Group**. Press **Enter**.
   c. Type the volume group name and `vpath` physical volume name and press **Enter**. Alternately, you can use the **F4** key (or **ESC** + **4**) to generate a list of all available SDD `vpath` devices and use the **F7** key (or **ESC** + **7**) to select the vpath you want to add.
2. Mirror *all* logical volumes from the original volume to a Subsystem Device Driver ESS volume. From the command line use the command `smitty mklvcopy` or choose **System Storage Management (Physical & Logical)** -> **Logical Volume Manager** -> **Logical Volumes** -> **Set Characteristic of a Logical Volume** -> **Add a Copy to a Logical Volume**. The Add Copies to a Logical Volume screen of SMIT is displayed and the **LOGICAL VOLUME name** option is highlighted.
   a. Type the logical volume name for which you want to add a mirror copy as described below:
      i. To select the logical volume name from the list press **F4** (or **ESC** + **4**) for a list of all logical volumes and select the desired one. Press **Enter**.
      ii. If you want to manually enter the logical volume name, simply type in the input field where the cursor blinks.
   b. Use the new Subsystem Device Driver `vpath` devices for copying all logical volumes. Do not forget to include JFS log volumes. Do *not* force the synchronization at this moment for a single logical volume. We can do it later for the whole volume group, which is a much faster way.
3. Repeat step 2 until *all* logical volumes in that volume group are mirrored.
4. Synchronize logical volumes (LVs). Use the command `syncvg -v vg_name [&]` to synchronize all the logical volumes in a volume group at once.
5. Remove the mirror and delete the original LVs. Use the `rmlvcopy lv_name 1 pv_name` command to remove the original copy of all the logical volumes from all original non-SDD physical volumes. The above command `lv_name` is a name of a logical volume for which you want to remove a mirror copy, and number “1” is a new summarized number of mirror copies for that logical volume and `pv_name` is a name of original non-SDD physical volume from which you want to remove a copy.

   An example of above command is `rmlvcopy lv01 1 hdisk13`, which removes a mirror copy of logical volume `lv01` from non-SDD physical volume `hdisk13`.
6. Remove the original non-SDD devices from the volume group. To do this use the **`reducevg vg_name pv_name`** command, where `vg_name` is the name of volume group you want to reduce and `pv_name` is the name of original non-SDD physical volume you want to remove from that volume group.

   An example of the above command is `reducevg vg1 hdisk13`, which removes a non-SDD physical volume `hdisk13` from a volume group `vg1`.

5.5.5 SDD utility programs

In this section we briefly describe how to use most important utility programs provided with IBM SDD installation on AIX. Please note, that some of utility programs listed in Table 5-9 may not be available on platforms other than AIX platforms.

Table 5-9 Description of most important utility programs

Utility program	Description
addpath	The command that dynamically adds more paths to Subsystem Device Driver devices while they are in `Available` state. This command is supported only with SDD for AIX 4.3.2 and higher. It is not available if you have the ibmSdd_421.rte fileset installed. You can use the `addpaths` command to dynamically add more paths to SDD devices while they are in the `Available` state. In addition, this command allows you to add paths to `vpath` devices (which are then opened) belonging to active volume groups. This command will open a new path (or multiple paths) automatically if the `vpath` is in `Open` state, and the original number of path of the vpath is more than one. You can use either the “Add Paths to Available Data Path Devices” SMIT screen, or run the `addpaths` command from the AIX command line.

hd2vp and vp2hd	SDD provides two conversion scripts, `hd2vp` and `vp2hd`. The `hd2vp` script converts a volume group from ESS hdisks into SDD `vpaths`, and the `vp2hd` script converts a volume group from SDD `vpaths` into ESS hdisks. Use the `vp2hd` program when you want to configure your applications back to original ESS hdisks, or when you want to remove the SDD from your AIX host system.
	You *must* convert all your applications and volume groups to the original ESS hdisk device special files before removing SDD software from your system.
	The syntax for these conversion scripts is as follows:
	`hd2vp vg_name`
	`vp2hd vg_name`
	These two conversion programs require that a volume group contain either all original ESS hdisks or all SDD vpaths. The program fails if a volume group contains both kinds of device special files (mixed volume group).
	Always use SMIT screen to create a volume group of SDD devices. This avoids the problem of a mixed volume group.

Utility program	Description
dpovgfix	You can use the dpovgfix script tool to recover mixed volume groups. Performing AIX system management operations on adapters and ESS hdisk devices might cause original ESS hdisks to be contained within a SDD volume group. This is known as a mixed volume group. Mixed volume groups happen when a SDD volume group is inactivated (varied off), and certain AIX commands to the hdisk put the pvid attribute of hdisk back into the ODM database. The following is an example of a command that does this: `chdev -l hdiskX -a queue_depth=30` If this disk is an active hdisk of a vpath that belongs to a SDD volume group, and you run the varyonvg command to activate this SDD volume group, LVM might pick up the hdisk device rather than the vpath device. The result is that a SDD volume group partially uses SDD vpath devices, and partially uses ESS hdisk devices. In this case the volume group loses path failover capability for that physical volume. The dpovgfix script tool fixes this problem. The command syntax is dpovgfix vg_name
lsvpcfg	You can use the lsvpcfg script tool to display the configuration status of SDD devices. This displays the configuration status for all SDD devices. The lsvpcfg command can be issued in two ways: - Without any parameters it displays information about all vpaths (SDD devices) configured. - With the vpath device name as a parameter it displays information about that particular SDD device(s). The command syntax is: lsvpcfg vpathX_0 vpathX_1 vpathX_2 ...
mkvg4vp	You can use the mkvg4vp command to create a SDD volume group. For more information about this command, go to Configuring a volume group for failover protection.
extendvg4vp	You can use the extendvg4vp command to extend an existing SDD volume group. For more information about this command, go to Extending an existing SDD volume group.

Usage of datapath command

For more detailed information about usage of the datapath command refer to 3.1.3, “Usage of datapath command” on page 31.

5.5.6 SDD error log messages

IBM Subsystem Device Driver logs error conditions into the AIX error log system. To check if SDD has generated an error log message, type the command `errpt -a | grep VPATH`. Table 5-10 the lists the SDD error log messages with appropriate description.

Table 5-10 List of error messages logged by SDD

Message	Description
VPATH_XBUF_NOMEM	An attempt was made to open a SDD vpath file and to allocate kernel-pinned memory. The system returned a null pointer to the calling program and kernel-pinned memory was not available. The attempt to open the file failed.

Message	Description
VPATH_PATH_OPEN	SDD device file failed to open one of its paths (hdisks). An attempt to open a `vpath` device is successful if at least one attached path opens. The attempt to open a `vpath` device fails only when *all* the `vpath` device paths fail to open.
VPATH_DEVICE_OFFLINE	Several attempts to retry an I/O request for a `vpath` device on a path have failed. The path state is set to `Dead` and the path is taken offline. Use the `datapath` command to set the offline path to online.
VPATH_DEVICE_ONLINE	SDD supports `Dead` path auto_failback and `Dead` path reclamation. A `Dead` path is put Online, and its state changes to Open after it has been bypassed by 50 000 I/O requests on an operational path. See 3.1.1, “Path algorithms” on page 27 for detailed information about path automatic failback and path reclamation.
New and modified error log messages by SDD for HACMP The following list shows the new and modified error log messages generated by SDD installed from the ibmSdd_433.rte or ibmSdd_510nchacmp.rte fileset. This SDD release is for HACMP environments only.	
VPATH_DEVICE_OPEN	The SDD device file failed to open one of its paths (hdisks). An attempt to open a `vpath` device is successful if at least one attached path opens. The attempt to open a `vpath` device fails only when *all* the `vpath` device paths fail to open. In addition, this error log message is posted when the `vpath` device fails to register its underlying paths or fails to read the persistent reserve key for the device.
VPATH_OUT_SERVICE	There is no path available to retry an I/O request that failed for a `vpath` device. The I/O error is returned to the calling program and this error log is posted.
VPATH_FAIL_RELPRESERVE	An attempt was made to close a `vpath` device that was not opened with the `RETAIN_RESERVE` option on the persistent reserve. The attempt to close the `vpath` device was successful, however, the persistent reserve was not released. The user is notified that the persistent reserve is still in effect, and this error log is posted.
VPATH_RESV_CFLICT	An attempt was made to open a `vpath` device, but the reservation key of the `vpath` device is different from the reservation key currently in effect. The attempt to open the device fails and this error log is posted. The device could not be opened because it is currently reserved by someone else.

5.6 How to use HACMP with SDD

In this section we focus on the differences and special requirements while using the IBM Subsystem Device Driver with the High Availability Cluster Multi-Processing (HACMP/6000).

5.6.1 Understanding the SDD support for HACMP/6000

You can run the Subsystem Device Driver in both: concurrent and non-concurrent multiple host environments in which more than one host is attached to the same LUNs on the ESS. RS/6000 (or pSeries) servers running HACMP/6000 in concurrent or non-concurrent mode are supported. Different SDD releases support different kinds of environments. See Table 5-6 on page 91 and Table 5-11 on page 117 to find out which fileset of SDD is proper for your particular environment and which APARs are you required to install.

Important: The list of required APARs contained in Table 5-11 is valid at the date of this book's publishing. For the most up-to-date list of required APARs go to the following Web site: `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

HACMP/6000 provides a reliable way for clustered IBM RS/6000 and pSeries servers which share disk resources to recover from server and disk failures. In an HACMP/6000 environment, each server in a cluster is a node. Each node has access to shared disk resources that are accessed by other nodes. When there is a failure, HACMP/6000 transfers ownership of shared disks and other resources based on how you define the relationship among nodes in a cluster. This process is known as node failover or node failback. HACMP supports two modes of operation:

- Non-concurrent — only one node in a cluster is actively accessing shared disk resources while other nodes are standby for those particular disk resources. Very often people misunderstand what is the core on non-concurrent environment and think, that use of non-concurrent environment forces only one node in a cluster to become an active node, while all other nodes are stand-by. This statement is not true and we want to explain, that non-concurrent environment does *not mean*, that only *one node* in a cluster can be an active node. In non-concurrent environment *all nodes* in a cluster can be active nodes, but only one node at the same time can take the ownership of particular disk resource group. For the other nodes different disk resources can be defined and actively accessed at the same time.
- Concurrent — multiple nodes in a cluster are actively accessing shared disk resources at the same time.

SDD supports both SCSI adapters and Fibre Channel adapters in HACMP/6000 environment. The kind of attachment support depends on the version of SDD that you have installed. Table 5-11 and Table 5-12 on page 119 summarizes the software requirements to support HACMP/6000.

Table 5-11 IBM Subsystem Device Driver support for HACMP/6000

SDD version and release level	HACMP 4.3.1 + APARs	HACMP 4.4 + APARs
Support for HACMP in concurrent mode		
SDD 1.1.4.0 (SCSI only)	IY07392 IY03438 IY11560 IY08933 IY11564 IY12021 IY12056 F models require IY11110	IY11563 IY11565 IY12022 IY12057 F models require IY11480

SDD version and release level	HACMP 4.3.1 + APARs	HACMP 4.4 + APARs
SDD 1.2.0.0 (SCSI/Fibre Channel)	IY07392 IY13474 IY03438 IY08933 IY11560 IY11564 IY12021 IY12056 F models require IY11110	IY13432 IY11563 IY11565 IY12022 IY12057 F models require IY11480
SDD 1.2.2.x (SCSI/Fibre Channel)	IY07392 IY13474 IY03438 IY08933 IY11560 IY11564 IY12021 IY12056 F models require IY11110	IY13432 IY11563 IY11565 IY12022 IY12057 F models require IY11480
SDD 1.3.0.x (SCSI/Fibre Channel)	IY07392 IY13474 IY03438 IY08933 IY11560 IY11564 IY12021 IY12056 F models require IY11110	IY13432 IY11563 IY11565 IY12022 IY12057 F models require IY11480
Support for HACMP in non-concurrent mode		
SDD 1.2.2.x (SCSI/Fibre Channel)	IY07392 IY13474 IY03438 IY08933 IY11560 IY11564 IY12021 IY12056 IY14682 F models require IY11110	IY13432 IY11563 IY11565 IY12022 IY12057 IY14683 F models require IY11480
ibmSdd_433.rte fileset for SDD 1.3.0.x (SCSI/Fibre Channel)	IY07392 IY13474 IY03438 IY08933 IY11560 IY11564 IY12021 IY12056 IY14682 F models require IY11110	IY13432 IY11563 IY11565 IY12022 IY12057 IY14683 F models require IY11480
Support for HACMP in concurrent mode on AIX 5.1.0 (32-bit kernel only)		

SDD version and release level	HACMP 4.3.1 + APARs	HACMP 4.4 + APARs
ibmSdd_510.rte fileset for SDD 1.3.0.x (SCSI/Fibre Channel)	Version 4.3.1 of HACMP/6000 not available for AIX 5.1.0	IY11563 IY11565 IY12022 IY12057 IY13432 IY14683 IY17684 IY19089 IY19156 F models require IY11480
Support for HACMP in non-concurrent mode on AIX 5.1.0 (32-bit kernel only)		
ibmSdd_510nchacmp.rte fileset for SDD 1.3.0.x (SCSI/Fibre Channel)	Version 4.3.1 of HACMP/6000 not available for AIX 5.1.0	IY11563 IY11565 IY12022 IY12057 IY13432 IY14683 IY17684 IY19089 IY19156 F models require IY11480

Even though SDD supports HACMP/6000, certain combinations of features are not supported. Table 5-12 lists those combinations.

Table 5-12 Supported and unsupported SDD features in HACMP/6000 environment

Feature	Support for HACMP
ESS concurrent download of licensed internal code	Yes
Subsystem Device Driver load balancing	Yes
Support for SCSI adapters	Yes
Support for Fibre Channel adapters	Yes
Single path Fibre Channel connection to an ESS LUN(s)	No
SCSI and Fibre Channel connections to the same LUN from one host (mixed environment)	No

5.6.2 What's new in SDD for HACMP/6000

The `ibmSdd_433.rte` and `ibmSdd_510nchacmp.rte` filesets for SDD 1.3.0.x have different features compared with `ibmSdd_432.rte` and `ibmSdd_510.rte` filesets for SDD 1.3.0.x. The `ibmSdd_433.rte` and `ibmSdd_510nchacmp.rte` filesets implement the SCSI-3 Persistent Reserve command set, in order to support HACMP in non-concurrent mode with single-point failure protection. The `ibmSdd_433.rte` and `ibmSdd_510nchacmp.rte` filesets require the ESS G3 level microcode on the ESS to support the SCSI-3 Persistent Reserve command set. If the ESS G3 level microcode is not installed, the `ibmSdd_433.rte` and `ibmSdd_510nchacmp.rte` filesets will switch the multi-path configuration to a single-path configuration. There is no single-point failure protection for single-path configurations.

The ibmSdd_433.rte and ibmSdd_510nchacmp.rte filesets have a new attribute under its pseudo parent (dpo), that reflects whether the ESS supports the Persistent Reserve Command set or not. The attribute name is persistent_resv. If SDD detects that G3 level microcode is installed, the persistent_resv attribute is created in the CuAt ODM and its value is set to yes. Otherwise this attribute only exists in the PdAt ODM and its value is set to no (this is the default value, since it exists in PdAt object of ODM, which is object for Predefined Attributes). You can use the command odmget -q name=dpo CuAt to check if the persistent_resv attribute is set to yes, *after* the SDD device configuration is complete. The output of that command is shown in Example 5-15, where we can see, that G3 level microcode is installed in the ESS.

Example 5-15 Example of odmget -q name=dpo CuAt command output

```
name = "dpo"
attribute = "persistent_resv"
value = "yes"
generic = "D"
rep = "sl"
nls_index = 0
```

In order to implement the Persistent Reserve command set, each host server needs a unique 8-byte reservation key. There are two ways to get a unique reservation key. In HACMP/6000 environments, HACMP/6000 generates a unique key for each node in the ODM database. When SDD cannot find that key in the ODM database, it generates a unique reservation key by using the middle eight bytes of the output from the uname -m command. To check the Persistent Reserve Key for an HACMP node, issue the command:

```
odmget -q name=ioaccess CuAt
```

The output should look similar to Example 5-16.

Example 5-16 Example of odmget -q name=ioaccess CuAt command output

```
name = "ioaccess"
attribute = "perservekey"
value = "01043792"
type = "R"
generic = ""
rep = "s"
nls_index = 0
```

5.6.3 Special requirements for HACMP/6000

There is a special requirement regarding unconfiguring and removing the ibmSdd_433.rte and ibmSdd_510nchacmp.rte filesets for SDD 1.3.0.x vpath devices. You *must* unconfigure and remove the vpath devices *before* you unconfigure and remove the underlying ESS hdisks. Otherwise if the ESS hdisks are unconfigured and removed first, the persistent reserve will *not* be released, even though the vpath devices have been successfully unconfigured and removed.

SDD does *not* automatically create the pvid attribute in the ODM database for each vpath device. The AIX disk driver automatically creates it if a pvid exists on the physical device. Therefore, the first time you import a new SDD volume group to a new cluster node, you must import the volume group using hdisks as physical volumes. Next, run the hd2vp conversion script to convert the volume group from ESS hdisks to SDD vpath devices. This conversion step not only creates pvid attributes for all vpath devices which belong to that imported volume group, but it also deletes the pvid attributes from ODM for underlying hdisks. Later on you can import and vary on the volume group directly from the vpath devices. These special requirements apply to both concurrent and non-concurrent volume groups. See “Importing a volume group with SDD” on page 104 for more detailed information of how to import a volume group with SDD installed.

How to recover paths that are lost during HACMP/6000 node failover

Normally, when there is a node failure, HACMP/6000 transfers ownership of shared disks and other resources, through a process known as node failover. Certain situations, such as a loose or disconnected SCSI or Fibre Channel adapter card, can cause your vpath devices to lose one or more underlying paths during node failover. Perform the following steps to recover these paths:

1. Check to ensure that *all* the underlying paths (hdisks) are in the Available state.
2. Run the addpaths command to add the lost paths back to the SDD devices. See “Adding paths to SDD devices which belongs to a volume group” on page 97 for details.

> **Note:** Simply running the cfgmgr command while the vpath devices are in the Available state will *not* recover the lost paths. That is why you need to run the addpaths command to recover the lost paths.

IBM SDD does not support the addpaths command for AIX 4.2.1 (it is not available if you have the ibmSdd_421.rte fileset installed and only supports SDD for AIX 4.3.2 and higher). If you have the ibmSdd_421.rte fileset installed, and if your vpath devices have lost one or more underlying paths and they belong to an active volume group.

> **Tip:** When there is a node failure, HACMP/6000 transfers ownership of shared disks and other resources, through a process known as node failover. To recover these paths, you need to first check to ensure that *all* the underlying paths (hdisks) are in the Available state. Next, you need to unconfigure and reconfigure your SDD vpath devices.

Perform the following steps to recover the lost paths:

1. Run the lspv command to find the volume group name for the vpath devices that have lost paths.
2. Run the lsvgfs vg_name command to find out the file systems for the volume group.
3. Run the mount command to find out if any file systems of the volume group were mounted. If yes, run the umount filesystem_name command to unmount any file systems that were mounted.
4. Close any other applications that are using other logical volumes in this volume group (such as database engines).
5. Run the vp2hd vg_name command to convert the volume group from vpath devices to the ESS hdisks.
6. Vary off the volume group. This puts the physical volumes (hdisks) in the Close state.
7. Run the rmdev -l vpathX command on each vpath device that has lost a path. Run the mkdev -l vpathX command on the same vpath devices to recover the paths.
8. Run the lsvpcfg or lsvpcfg vpathX_0 vpathX_1 ... vpathX_N command to ensure that all paths are configured.
9. Vary on the volume group:
   a. Use the varyonvg vg_name command for non-concurrent volume groups.

   b. Use the `/usr/sbin/cluster/events/utils/convaryonvg vg_name` or `varyonvg -u vg_name` command for concurrent volume groups.

10. Run the `hd2vp vg_name` command to convert the volume group back to SDD `vpath` devices.

11. Mount all the file systems you previously unmount and run other applications that are using logical volumes in this volume group (such as database engines).

5.6.4 Models of the ESS supported in HACMP/6000 environment

HACMP/6000 is not supported on all models of the ESS. Table 5-13 shows support matrix for the ESS in HACMP environment.

Table 5-13 Models of the ESS supported in HACMP/6000 environment

ESS Model	AIX 4.2.1 HACMP 4.2.2	AIX 4.3.3 HACMP 4.2.2	AIX 4.3.3 HACMP 4.3.1	AIX 4.3.3 & 5.1 HACMP 4.4.0	AIX 4.3.3 & 5.1 HACMP 4.4.1
ESS E10-E20	SCSI only	SCSI only	SCSI + FC	SCSI + FC	SCSI + FC
ESS F10-F20	Not supported	Not Supported	SCSI + FC	SCSI + FC	SCSI + FC

Note: For latest information about supported ESS models and required ESS microcode levels, go to the following Web site and download the Supported Server List:
http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

5.7 Upgrading SDD for AIX 4.2.1, AIX 4.3.2 and AIX 4.3.3

SDD 1.3.0.x allows for a non-disruptive installation if you are upgrading from any one of the following filesets:

- ibmSdd_421.rte, ibmSdd.rte.421
- ibmSdd_432.rte, ibmSdd.rte.432
- ibmSdd_433.rte, ibmSdd.rte.433

If you have previously installed from any of the listed filesets, SDD 1.3.0.x allows you to upgrade while:

- All of the Subsystem Device Driver file systems are mounted
- All of the Subsystem Device Driver volume groups are varied-on

If you are upgrading from a previous version of the SDD that you installed from other filesets, you cannot do the non-disruptive installation. To upgrade SDD to a newer version, all the SDD filesets must be uninstalled.

You can verify what version of SDD you have previously installed by issuing the command `lslpp -L | grep sdd`. If the SDD is installed from one of the filesets listed above, proceed to “5.7.1, “Upgrading to SDD 1.3.0.x through a non-disruptive installation” on page 123”. Otherwise, proceed to “5.7.2, “Upgrading to SDD 1.3.0.x through a disruptive installation” on page 123”.

5.7.1 Upgrading to SDD 1.3.0.x through a non-disruptive installation

SDD 1.3.0.x allows for a non-disruptive installation if you are upgrading from any of the filesets listed in “5.7, “Upgrading SDD for AIX 4.2.1, AIX 4.3.2 and AIX 4.3.3” on page 122“. Perform the following steps to upgrade to SDD 1.3.0.x with a non-disruptive installation:

1. Terminate all I/O operations to the SDD volume groups. You don’t need to unmount filesystems and vary off volume groups.
2. Complete the installation instructions provided in 5.4, “Installing and configuring the IBM Subsystem Device Driver” on page 91.
3. Restart your system by typing the `shutdown -Fr` command.
4. Verify your currently installed version of the SDD by completing the instructions provided in “Verifying the SDD Installation” on page 93.
5. Verify the SDD configuration by typing the `lsvpcfg` command. Refer to “Verifying the SDD configuration” on page 96 for more detailed information.

Attention: If a SDD volume group are mixed with `hdisk` devices and `vpath` devices, you must run the `dpovgfix` utility to fix this problem. Otherwise, SDD will not function properly. Use the `dpovgfix vg_name` command to fix this problem. Refer to “How failover protection can be lost” on page 106 for more details related to this problem.

5.7.2 Upgrading to SDD 1.3.0.x through a disruptive installation

If you are upgrading from a previous version of the SDD that you installed with a fileset not listed in “5.7, “Upgrading SDD for AIX 4.2.1, AIX 4.3.2 and AIX 4.3.3” on page 122“, you cannot do the non-disruptive installation. Perform the following steps to upgrade to SDD 1.3.0.x:

1. Remove any `.toc` files generated during previous SDD or DPO installations. Type the command `rm .toc` to delete any `.toc` file found in the `/usr/sys/inst.images directory`. Ensure that this file is removed because it contains information about the previous version of SDD or DPO.
2. Run the `lspv` command to find out all the Subsystem Device Driver volume groups.
3. Run the `lsvgfs` command for each SDD volume group, to find out its mounted file systems. To do this issue the command `lsvgfs vg_name`.
4. Run the `umount` command to unmount all file systems belonging to SDD volume groups. To do this type the command `umount filesystem_name`.
5. Run the `vp2hd` script to convert the volume group from SDD devices to ESS `hdisk` devices. Type the command `vp2hd vg_name` for all volume groups using SDD devices.
6. Run the `varyoffvg` command to vary off the volume groups. Type the command `varyoffvg vg_name` for all required volume groups.
7. Remove all SDD devices. Type the command `rmdev -dl dpo -R`.
8. Uninstall SDD fileset. See 5.4.4, “Removing the Subsystem Device Driver” on page 100 for a step-by-step procedure on uninstalling SDD.
9. Install the newer version of SDD. See 5.4, “Installing and configuring the IBM Subsystem Device Driver” on page 91 for more detailed information how to install IBM SDD.
10. Configure all the SDD devices to the Available condition. See 5.4.2, “Configuring the Subsystem Device Driver” on page 95 for a step-by-step procedure.
11. Verify the SDD configuration by typing the `lsvpcfg` command. Refer to “Verifying the SDD configuration” on page 96 for more detailed information.

12. Run the `varyonvg vg_name` command for each volume group that was previously varied offline.

13. Run the `hd2vp` script for each SDD volume group, to convert the physical volumes from ESS `hdisk` devices back to SDD `vpath` devices. To do this type the command `hd2vp vg_name`.

14. Run the `lspv` command to verify that all physical volumes in the SDD volume groups are SDD vpath devices and that `hdisk` devices are *not* used.

> **Attention:** If a SDD volume group are mixed with `hdisk` devices and `vpath` devices, you must run the `dpovgfix` utility to fix this problem. Otherwise, SDD will not function properly. Use the `dpovgfix vg_name` command to fix this problem. Refer to “How failover protection can be lost” on page 106 for more details related to this problem.

5.8 Using concurrent download of licensed internal code

Concurrent download of licensed internal code is the capability to download and install licensed internal code on an ESS while applications continue to run. This capability is supported for single-path (SCSI only) and multiple-path (SCSI or Fibre Channel) access to an ESS.

During the download of licensed internal code, the AIX error log might overflow and excessive system paging space could be consumed. When the system paging space drops too low it could cause your AIX system to hang. To avoid this problem, you can perform the following steps prior to doing the download:

1. Save the existing error report by typing the command `errpt > file.save` from the AIX command-line interface.
2. Delete the error log from the error log buffer by typing the command `errclear 0`.
3. Enlarge the system paging space by using the SMIT tool.
4. Stop the AIX error log daemon by typing the command `/usr/lib/errstop`.

Once you have completed this procedure, you can perform the download of the ESS licensed internal code. After the download completes, type `/usr/lib/errdemon` from the command-line interface to restart the AIX error log daemon.

6

DMP installation and configuration on Sun servers

This chapter describes the concepts to use multipathing and cluster software on servers running the Sun operating system (OS). We show you how to install and configure the volume manager, multipathing and cluster software.

This chapter describes:

- Concepts of multipathing and clustering on Sun platform
- Hardware and software requirements
- Preparing the environment
- Using the SDD
- Installing and configuring Veritas Volume Manager
- Installing and configuring cluster software

6.1 Concepts of multipathing and clustering on Sun platform

The Enterprise Storage Server has several features that are desirable for Sun's Enterprise Servers — both in high availability (HA) and distributed environments.

However, several system variables need to be configured properly to successfully employ the ESS and the Sun Enterprise Server. While it is possible to determine these values through the process of "trial and error", such a process is, to say the least, costly and time consuming.

The following sections will provide the integrator or system administrator with the changes necessary to bring the IBM ESS and Sun Enterprise Server(s) together and allow them to operate successfully.

6.2 Hardware and software requirements

The IBM Subsystem Device Driver has the following hardware and software requirements:

6.2.1 Hardware

- The IBM Enterprise Storage Server (ESS)

  Important: The TotalStorage Enterprise Storage System (ESS) is supported in many environments, and customers are responsible for ensuring that the specific host system configuration used (that is, server model, operating system level and host adapter combination) is a valid and supported configuration by the server manufacturer.

  To confirm the list of supported servers, go to "ESS Supported Servers" at:

  `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

- Sun Enterprise Servers

  The listed hardware or software has been thoroughly tested and certified for use. If you use unsupported hardware or software you can have unpredictable results. For the latest IBM supported hardware and software, please see the supported server Web sites:

  `http://www.storage.ibm.com/hardsoft/products/ess/supserver_summary_open.htm`
  `http://www.storage.ibm.com/hardsoft/products/ess/pdf/1012-01.pdf`

- SCSI and/or Fibre Channel adapters and cables
- Fibre Channel Switch, if using Fibre Channel adapters (not necessary)

6.2.2 Software

- Sun Solaris 2.6, 2.7 and 2.8 with appropriate packages installed
- SCSI and Fibre Channel device driver installed
- Sun Veritas Dynamic Multi Pathing
- Sun Veritas Clustering Software
- Sun Veritas Fast Filesystem
- Sun Veritas Volume Manager

6.2.3 Non supported environments

The following environments are not supported by the Subsystem Device Driver:

- A host server with a single-path Fibre Channel connection to an ESS is not supported. There is no reason to install SDD when only one path is available.

Note: A host server with a single fibre adapter that connects through a switch to multiple ESS ports is considered a multipath Fibre Channel connection and therefore is a supported environment.

- A host server with SCSI channel connections and a single-path Fibre Channel connection to an ESS.
- A host server with both a SCSI channel and Fibre Channel connection to a shared LUN is not supported.

Note: The Subsystem Device Driver also supports one SCSI adapter on the host system. With single-path access, concurrent download of licensed internal code is supported with SCSI devices. However, the load balancing and failover features are not available.

For current information about the SCSI/Fibre Channel adapters which can attach to your HP host system go to the Web site at:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

6.3 Preparing the environment

In this section we discuss the initial preparation of the environment.

6.3.1 Installing the operating system

For the steps to install and configure the Sun Solaris 2.6, 2.7 and 2.8 operating system, see the Sun Solaris documentation (Installation Guide Collection) at: http://docs.sun.com/

6.3.2 Configuring all HBAs

Three variables are required in the /etc/system file. These modifications should be inserted above any forceload statements.

sd_max_throttle

```
set sd:sd_max_throttle =“calculated”
```

The sd_max_throttle variable assigns the default value lpfc will use to limit the number of outstanding commands per sd device. This value is global, affecting each sd device recognized by the driver. The maximum sd_max_throttle setting supported is 256. To determine the correct setting, perform the following calculation for each HBA:

```
256 / (number of LUN's per adapter) = sd_max_throttle value
```

Here, for example, is a server with two HBAs installed, 20 LUNs defined to HBA1, and 26 LUNs defined to HBA2.

```
HBA1= 256 / 20 = 12.8 and HBA2 = 256 / 26 = 9.8
```

Rounding down yields 12 for HBA1 and 9 for HBA2. In this example, the correct sd_max_throttle setting would be the lowest value obtained or 9.

sd_io_time

```
set sd:sd_io_time = 120
```

The sd_io_time variable determines how long a queued job will wait for any sd device I/O to fail. Originally, sd_io_time is set to 60. This is too low for most configurations. Setting it to 120 provides the host more time to complete I/O operations.

set maxphys

```
set maxphys = 8388608
```

The maxphys value determines the maximum number of bytes that can be transferred with a SCSI transaction. The original value is too small to allow the Fibre Channel HBA(s) to run efficiently. Set this to 8388608.

6.3.3 Emulex

The Emulex HBAs require modifications in the /kernel/drv/lpfc.conf file. The following variables must be modified as specified below for all supported Emulex HBAs:

automap=1;

The automap variable is used to turn on or off the retention of SCSI IDs on the fibre. If automap is set, SCSI IDs for all FCP nodes without persistent bindings will be automatically generated. If new FCP devices are added to the network when the system is down, there is no guarantee that these SCSI IDs will remain the same when the system is booted again. If one of the above FCP binding methods is specified, then automap devices will use the same mapping method to preserve SCSI IDs between link down and link up. If automap is 0, only devices with persistent bindings will be recognized by the system. Set this to 1.

fcp-on=1;

The fcp-on variable controls whether or not Fibre Channel port access is enabled or not. Set this to 1 to enable FCP access.

lun-queue-depth

```
lun-queue-depth=“sd_max_throttle from /etc/system”;
```

The lun-queue-depth variable determines how many requests can be accepted for each of the LUNs the host has access to. This value is global in nature as it affects all LUNs on the host. Set this to be equal to the

sd_max_throttle value obtained from the /etc/system file. See “sd_max_throttle” on page 127.

network-on=0;

The network-on variable determines whether networking is enabled for the HBA. Networking will be enabled if set to 1, and disabled if set to 0. This variable will be set during the installation of the driver via pkgadd. Verify it is set to 0.

topology=2;

The topology variable is used to let the lpfc driver know how to attempt to start the HBA. It can be set to start only one mode or to attempt one mode, and then failover to the other mode should the first mode fail to connect.

- 0x00 = attempt loop mode, if it fails attempt point-to-point mode.
- 0x02 = attempt point-to-point mode only
- 0x04 = attempt loop mode only

- 0x06 = attempt point-to-point mode, if it fails attempt loop mode

Set the variable to point-to-point mode to run as an N_Port or FC-SW. Set the variable to loop mode to run as an NL_Port or FC-AL. The above setting reflects FC-SW only.

zone-rscn=1;

The zone-rscn variable allows the driver to check with the NameServer to see if an N_Port ID received from an RSCN applies. Setting zone-rscn to 1 causes the driver to check with the NameServer. If Soft Zoning is used, with Brocade Fabrics, this should be set to 1. Set this to 1.

Tip: To obtain more information about EMULEX adapter, see: `http://www.emulex.com/`

6.3.4 JNI

Under the /kernel/drv directory, modifications will be necessary in the appropriate configuration file.

fca_nport

```
fca_nport = 0;
```

Or

```
fca_nport = 1;
```

The fca_nport variable is used to setup either FC-AL or FC-SW. If false (0), then fca initializes on a loop. If true (1), then fca initializes as an N_Port and fabric operation is enabled. This variable can be overridden by public_loop (see below). For fabric, set this to 1.

public_loop = 0;

The public_loop variable can override the fca_nport variable. If public_loop is false (0), then fca initializes according to what fca_nport is set to. If true (1), then fca initializes as an NL_Port on a public loop and fabric operation is enabled via the FLPort of the switch. Also, if public_loop = 1, then fca_nport is overridden to be 0. Set this to 0.

ip_disable = 1;

The ip_disable variable allows the IP side of the driver to be enabled or disabled. If false (0), then the IP side of the driver is enabled. If true (1), then the IP side of the driver is completely disabled. Set this to 1.

scsi_probe_delay = 5000;

The scsi_probe_delay variable uses a 10 millisecond resolution to set the delay before SCSI probes are allowed to occur during boot. This allows time for the driver to build a network port list for target binding. Set this to 5000.

failover = 60;

The failover variable represents the number of seconds after a target is declared offline before the target is declared as failed and all pending commands are flushed back to the application. Using the IBM 2109 or the Brocade switch, set this to 60. If using a McData switch, set to 300.

Tip: To obtain more information about JNI adapter, see: `http://www.jni.com/`

6.3.5 Setting up the ESS

The ESS comes with network-based software called the ESS Specialist to manage and monitor the ESS. ESS Specialist must be accessed from a PC located on a secure network with the ESS. Any browser supporting Java 1.1.8 JDK may be used. See these IBM Redbooks for a detailed discussion of configuring ESS with ESS Specialist:

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *IBM Enterprise Storage Server*, SG24-5465
- *Implementing ESS Copy Services on UNIX and Windows NT/2000*, SG25-5757

On the ESS Specialist Open System Storage window it is possible to add and remove host connections, configure Host Adapter ports, set up disk groups, add and remove volumes and modify volume assignments.

6.3.6 Attaching an ESS to a Sun Solaris Enterprise Servers

This section describes the host system requirements and provides procedures to attach an ESS to a Sun Solaris Enterprise Server host system with Fibre Channel adapters. *Host Systems Attachment Guide 2105 Models E10, E20, F10, and F20*, SC26-7296, at this Web site shows all the steps to connect the servers to an ESS:

http://www.storage.ibm.com/hardsoft/products/ess/pubs/f2ahs04.pdf

The reader is expected to be able to administer a Sun Solaris system, as well as configure and administer an ESS on an IBM SAN. Detailed information on these topics will not be covered in this redbook. For a detailed discussion of ESS configuration and SAN topics, please see the following IBM Redbooks:

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *Implementing Fibre Channel Attachment on the ESS*, SG23-6113
- *Implementing an Open IBM SAN*, SG24-6116

Additionally, please refer to the following Sun documentation available from:

http://docs.sun.com

6.4 Using the SDD

This section discusses when to use SDD and when to use DMP.

6.4.1 When to use SDD

At the date of writing this book, the last version of SDD (Version 1.2.0.5) on the IBM Web site is not supported for IBM when using in a cluster environment (Veritas Cluster Manager or Sun Cluster Software) or when DMP is enabled in the Veritas Volume Manager.

If your system already has a volume manager, software application, or DBMS installed that communicates directly with the Solaris disk device drivers, you need to insert the new SDD device layer between the program and the Solaris disk device layer. You also need to customize the volume manager, software application, or DBMS in order to have it communicate with the SDD devices instead of the Solaris devices.

In addition, many software applications and DBMSs need to control certain device attributes such as ownership and permissions. Therefore, you must ensure that the new SDD devices and these software applications or DBMSs have the same attributes as the Solaris sd devices that they replace. You may need to customize the software application or DBMS to accomplish this.

See http://ftp.software.ibm.com/storage/subsystem/tools/f2asdd00.htm, to know more about using application with SDD in the following environments:

- Standard UNIX applications
- Network File System file systems
- Oracle
- Veritas Volume Manager

Important: Before starting the process to install the last SDD version, verify with an IBM representative or at the following Web site, to confirm if the product is supported in a cluster environment:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

6.4.2 Installing the Subsystem Device Driver

You need to complete the following procedure if you are installing SDD for the first time on your Sun host.

Important: To install the SDD using Veritas Volume Manager, Version 3.1 and above, it is necessary to disable DMP. The new versions of VxVM do not allow to disable the DMP and it is not supported for IBM.

The steps described in the following Web site show how to disable DMP:

http://seer.support.veritas.com/docs/180452.htm

1. Make sure the SDD compact disc is available.
2. Insert the compact disc into your CD-ROM drive.
3. Change to the install directory:

   ```
   # cd /cdrom/cdrom0/sun32bit or
   # cd /cdrom/cdrom0/sun64bit
   ```

4. Run pkgadd, and point the -d option of pkgadd to the directory containing IBMdpo. See Example 6-1.

Example 6-1 Installing SDD

```
pkgadd -d /cdrom/cdrom0/sun32bit IBMdpo or
pkgadd -d /cdrom/cdrom0/sun64bit IBMdpo
```

5. You should see messages similar to Example 6-2.

Example 6-2 Installing output

```
+-------------------------------------------------------------------------------+
|Processing package instance <IBMdpo> from <var/spool/pkg>                      |
|                                                                               |
|                                                                               |
|IBM DPO driver                                                                 |
|(sparc) 1                                                                      |
|## Processing package information.                                             |
```

```
|## Processing system information.                                               |
|## Verifying disk space requirements.                                           |
|## Checking for conflicts with packages already installed.                      |
|## Checking for setuid/setgid programs.                                         |
|                                                                                |
|This package contains scripts which will be executed with super-user            |
|permission during the process of installing this package.                       |
|                                                                                |
|Do you want to continue with the installation of <IBMdpo> [y,n,?]               |
+--------------------------------------------------------------------------------+
Type Y and press Enter to proceed.
You should see messages similar to this:
+--------------------------------------------------------------------------------+
|Installing IBM DPO driver as <IBMdpo>                                           |
|                                                                                |
|## Installing part 1 of 1.                                                      |
|/etc/defvpath                                                                   |
|/etc/rc2.d/S00vpath-config                                                      |
|/etc/rcS.d/S20vpath-config                                                      |
|/kernel/drv/vpathdd                                                             |
|/kernel/drv/vpathdd.conf                                                        |
|/opt/IBMdpo/cfgvpath                                                            |
|/opt/IBMdpo/datapath                                                            |
|/opt/IBMdpo/devlink.vpath.tab                                                   |
|/opt/IBMdpo/etc.system                                                          |
|/opt/IBMdpo/pathtest                                                            |
|/opt/IBMdpo/showvpath                                                           |
|/usr/sbin/vpathmkdev                                                            |
|[ verifying class <none>  ]                                                     |
|## Executing postinstall script.                                                |
|                                                                                |
|DPO: Configuring 24 devices (3 disks * 8 slices)                                |
|                                                                                |
|Installation of <IBMdpo> was successful.                                        |
|                                                                                |
|The following packages are available:                                           |
|1 IBMcli ibm2105cli                                                             |
|         (sparc) 1.1.0.0                                                        |
|2 IBMdpo IBM DPO driver Version: May-10-2000 16:51                              |
|         (sparc) 1                                                              |
|Select package(s) you wish to process (or 'all' to process                      |
|all packages). (default: all) [?,??,q]:                                         |
+--------------------------------------------------------------------------------+
Type q and press Enter to proceed.
You should see messages similar to this:
+--------------------------------------------------------------------------------+
|*** IMPORTANT NOTICE ***                                                        |
|This machine must now be rebooted in order to ensure                            |
|sane operation. Execute                                                         |
|       shutdown -y -i6 -g0                                                      |
|and wait for the "Console Login:" prompt.                                       |
|                                                                                |
|DPO is now installed. Proceed to Post-Installation.                             |
+--------------------------------------------------------------------------------+
```

Note: You can verify that SDD has been successfully installed by issuing the datapath query device command. If the command executes, SDD is installed.

Post-installation

After the installation is complete, manually unmount the compact disc. Run the umount /cdrom command from the root directory. Go to the CD-ROM drive and press the Eject button.

After SDD is installed, your system must be rebooted to ensure proper operation. Type the command:

```
# shutdown -i6 -g0 -y
```

6.4.3 Uninstalling the Subsystem Device Driver

Attention: You must uninstall the current level of SDD before upgrading to a newer level.

Upgrading SDD consists of uninstalling and reinstalling the IBMdpo package. Perform the following steps to uninstall SDD:

1. Reboot or umount all SDD file systems.
2. If you are using SDD with a database, such as Oracle, edit the appropriate database configuration files (database partition) to remove all of the SDD devices.
3. If you are using a database, restart the database.
4. Type `# pkgrm IBMdpo` and press **ENTER**.

Attention: A number of different installed packages is displayed. Make sure you specify the correct package to uninstall.

A message similar to Example 6-3 is displayed.

Example 6-3 Uninstalling SDD

```
+--------------------------------------------------------------------------------+
|The following packages are available:                                           |
|1 IBMcli ibm2105cli                                                             |
|         (sparc) 1.1.0.0                                                        |
|2 IBMdpo IBM DPO driver Version: May-10-2000 16:51                              |
|         (sparc) 1                                                              |
|                                                                                |
+--------------------------------------------------------------------------------+
```

5. Type **Y** and press **ENTER**. A message similar to Example 6-4 is displayed.

Example 6-4 Removing window

```
+--------------------------------------------------------------------------------+
|## Removing installed package instance <IBMdpo>                                 |
|                                                                                |
|This package contains scripts that will be executed with super-user             |
|permission during the process of removing this package.                         |
|                                                                                |
|Do you want to continue with the removal of this package [y,n,?,q] y            |
|                                                                                |
+--------------------------------------------------------------------------------+
```

6. Type **Y** and press **ENTER**. A message similar to Example 6-5 is displayed.

Example 6-5 Command output

```
+--------------------------------------------------------------------------------+
|## Verifying package dependencies.                                             |
|## Processing package information.                                             |
|## Executing preremove script.                                                 |
|Device busy                                                                    |
|Cannot unload module: vpathdd                                                  |
|Will be unloaded upon reboot.                                                  |
|## Removing pathnames in class <none>                                          |
|/usr/sbin/vpathmkdev                                                           |
|/opt/IBMdpo                                                                    |
|/kernel/drv/vpathdd.conf                                                       |
|/kernel/drv/vpathdd                                                            |
|/etc/rcS.d/S20vpath-config                                                     |
|/etc/rc2.d/S00vpath-config                                                     |
|/etc/defvpath                                                                  |
|## Updating system information.                                                |
|                                                                               |
|Removal of <IBMdpo> was successful.                                            |
|                                                                               |
+--------------------------------------------------------------------------------+
```

Attention: Do not reboot at this time.

6.5 Installing and configuring Veritas Volume Manager

ESS logical volumes appear and can be used just like any fibre-attached disk drive. The logical volumes can be formatted, partitioned, and encapsulated under Sun Veritas Volume Manager as simple or sliced disks, or used as raw disks.

When using a logical volume as a file system — whether under Veritas control or just a disk partition — be sure to keep in mind the following:

- Large files — Solaris has the ability to “help” other functions and applications deal more intelligently with files over 2 GB in size. This is done during the mount phase with the large files option. Otherwise, some applications may not behave nicely when files grow larger than 2 GB in size.
- Logging — Sun’s Veritas Fast Filesystem allows for faster disk access, faster crash recovery, almost instantaneous filesystem creation, and a better method of handling large numbers of files on a filesystem. However, if the VxFS product is not installed, it is still possible to get the faster recovery and psuedo-journaling for the filesystem using the logging option during the mount phase of a UFS filesystem. Logging takes approximately 1 MB of disk space for each 1 GB capacity up to a maximum of 64 MB.

This is definitely an option to turn on to reduce the `fsck` phase after a crash — especially with large filesystems.

6.5.1 Creating a filesystem under Veritas

The Sun Veritas Volume Manager and the Sun Veritas Volume Manager Storage Administrator provide the building blocks to administer any amount of disk space from any type of disk array attached to a Sun server — including the ESS.

The Volume Manager Storage Administrator (VMSA) should be run as the root account. Otherwise, changes to the disks, subdisks, and volumes will not be possible. Launching VMSA is done by entering /opt/VRTSvmsa/bin/vmsa, assuming the application was installed in the default directory, and the application bin directory is not in the search path. Next, enter the host, account name, and password for this session. This will display the standard VMSA view as seen in Figure 6-1.

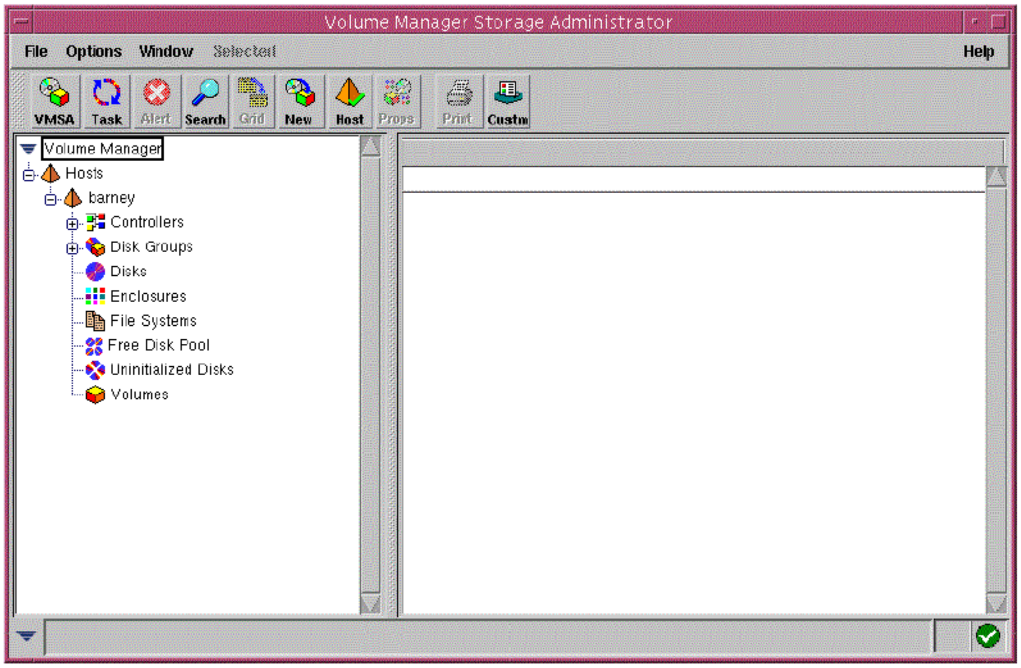
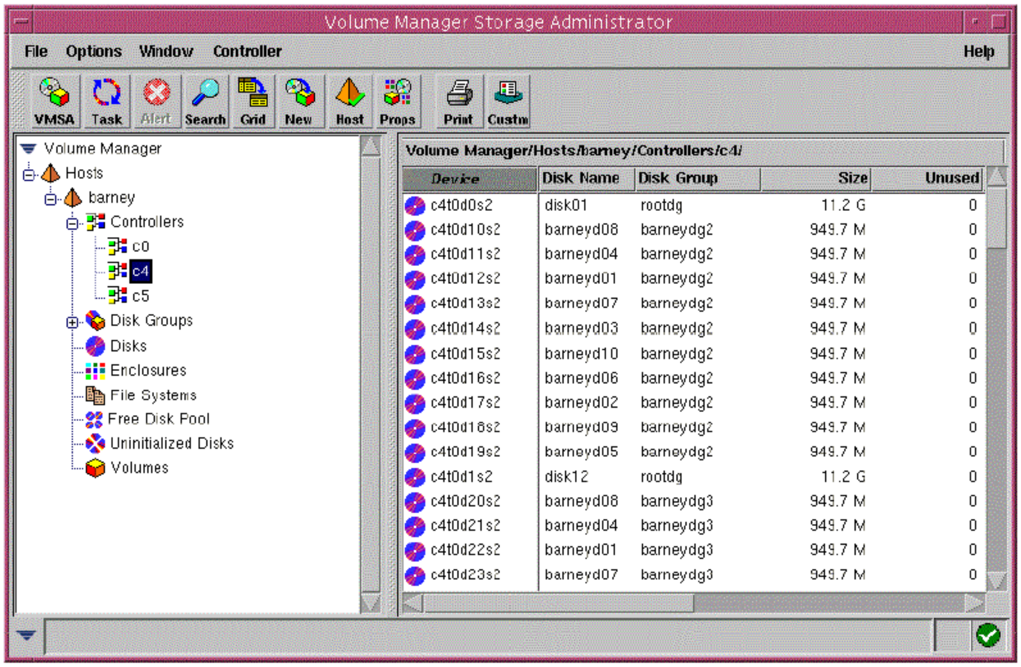


Figure 6-1 Volume Manager Storage Administrator

To open the Controllers icon in the left pane, simply double-click the line with the Controllers icon, or click the plus sign in the box to the left of the Controllers icon. On Sun Solaris systems, controller c0 is always the root or boot controller.

Clicking one of the other controllers under this view — in this case, either c4 or c5 — will produce a list of devices attached to that controller in the right pane. If any of these devices are under Veritas control, then the Disk Name and Disk Group columns will have that information. This will produce a view similar to Figure 6-2.

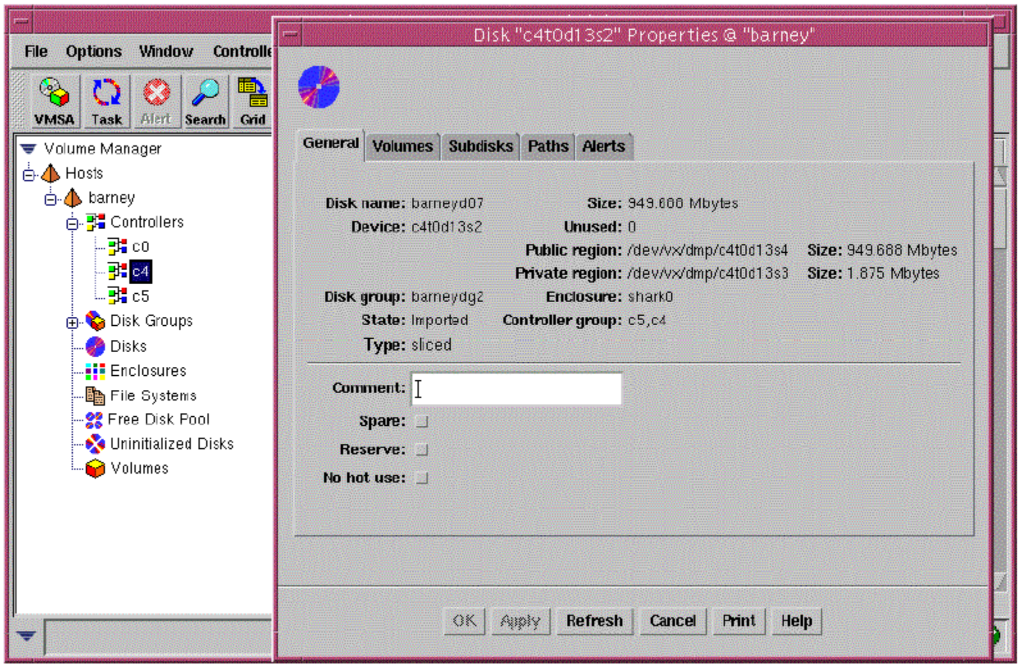


Volume Manager/Hosts/barney/Controllers/c4/

Device	Disk Name	Disk Group	Size	Unused
c4t0d0s2	disk01	rootdg	11.2 G	0
c4t0d10s2	barneyd08	barneydg2	949.7 M	0
c4t0d11s2	barneyd04	barneydg2	949.7 M	0
c4t0d12s2	barneyd01	barneydg2	949.7 M	0
c4t0d13s2	barneyd07	barneydg2	949.7 M	0
c4t0d14s2	barneyd03	barneydg2	949.7 M	0
c4t0d15s2	barneyd10	barneydg2	949.7 M	0
c4t0d16s2	barneyd06	barneydg2	949.7 M	0
c4t0d17s2	barneyd02	barneydg2	949.7 M	0
c4t0d18s2	barneyd09	barneydg2	949.7 M	0
c4t0d19s2	barneyd05	barneydg2	949.7 M	0
c4t0d1s2	disk12	rootdg	11.2 G	0
c4t0d20s2	barneyd08	barneydg3	949.7 M	0
c4t0d21s2	barneyd04	barneydg3	949.7 M	0
c4t0d22s2	barneyd01	barneydg3	949.7 M	0
c4t0d23s2	barneyd07	barneydg3	949.7 M	0

Figure 6-2 VMSA controller view

Right-clicking one of the devices in the right pane will produce a pop-up menu with several options including Properties. Select Properties to view a window similar to Figure 6-3.



Disk name: barneyd07 **Size:** 949.688 Mbytes
Device: c4t0d13s2 **Unused:** 0
Public region: /dev/vx/dmp/c4t0d13s4 **Size:** 949.688 Mbytes
Private region: /dev/vx/dmp/c4t0d13s3 **Size:** 1.875 Mbytes
Disk group: barneydg2 **Enclosure:** shark0
State: Imported **Controller group:** c5,c4
Type: sliced

Comment:
Spare: □
Reserve: □
No hot use: □

Figure 6-3 VMSA disk properties

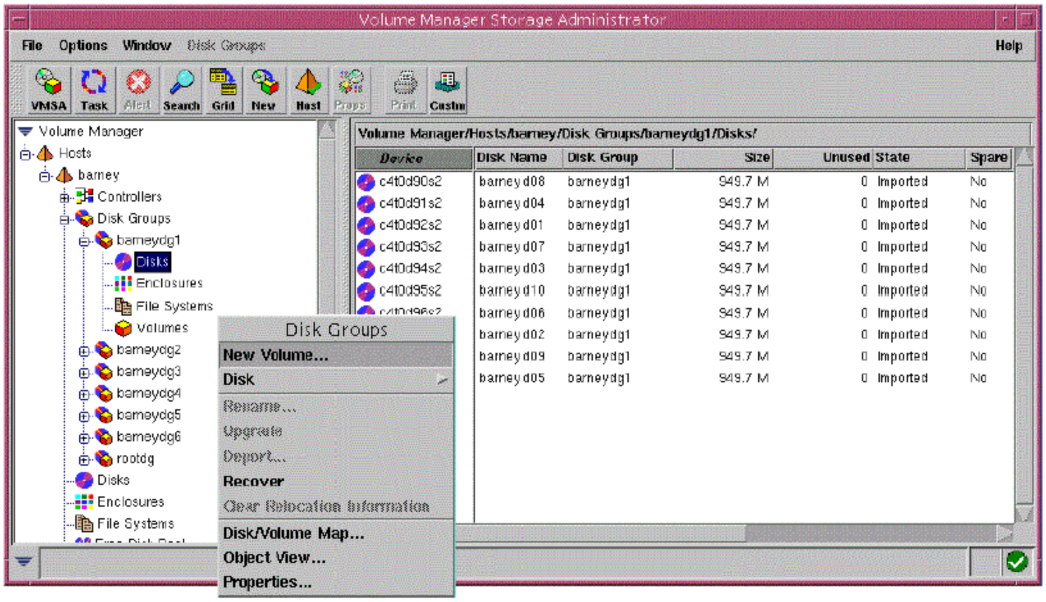
Before creating any volume, it is important to know if and/or how the device(s) being contemplated for use are currently being used. The Disk "XXXXX" Properties view provides a great deal of information pertaining to the selected device.

Selecting any of the tabs across the top row also provides information pertaining to the associated device, such as Volumes, Subdisks, Paths, and Alerts. For example, within the General view above, the Controller group field near the center of the display shows both c5, c4. This is indicative of Sun Veritas DMP on the host, as both controllers point to the same set of logical volumes on the ESS.

All Sun Veritas Volume Manager (VxVM) installations come with a rootdg disk group by default. This disk group is so important that the rootdg disk group cannot be removed. The VxVM application will not work properly should the rootdg become corrupt or become missing for any reason. For that reason alone, only the boot disk, the chosen mirror for the boot disk (on an internal SCSI adapter), and any on-board drives (again, on an internal SCSI adapter) that will be used for swap/paging space should be included in the rootdg disk group.

While it is possible to include ESS logical volumes in the rootdg disk group, we highly recommend that all logical volumes and all fibre-attached storage of any type be associated with other disk groups. This will eliminate any possibility of damage or corruption to the Sun Veritas Volume Manager database(s) on the rootdg disk group should the fibre-attached storage be unavailable, for any reason.

Right-click the rootdg icon or text to display the Disk Group pop-up menu similar to Figure 6-4.

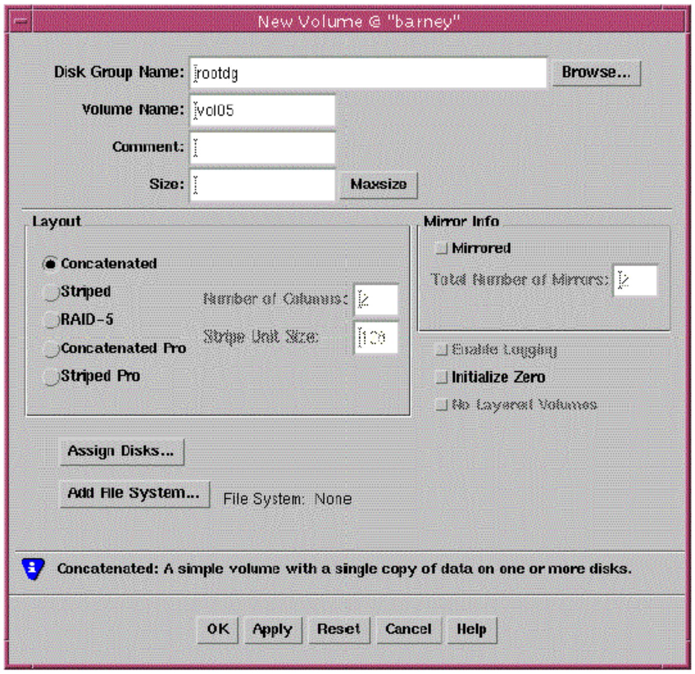


Volume Manager/Hosts/barney/Disk Groups/barneydg1/Disks/

Device	Disk Name	Disk Group	Size	Unused	State	Spare
c4t0d90s2	barneyd08	barneydg1	949.7 M	0	Imported	No
c4t0d91s2	barneyd04	barneydg1	949.7 M	0	Imported	No
c4t0d92s2	barneyd01	barneydg1	949.7 M	0	Imported	No
c4t0d93s2	barneyd07	barneydg1	949.7 M	0	Imported	No
c4t0d94s2	barneyd03	barneydg1	949.7 M	0	Imported	No
c4t0d95s2	barneyd10	barneydg1	949.7 M	0	Imported	No
c4t0d96s2	barneyd06	barneydg1	949.7 M	0	Imported	No
	barneyd02	barneydg1	949.7 M	0	Imported	No
	barneyd09	barneydg1	949.7 M	0	Imported	No
	barneyd05	barneydg1	949.7 M	0	Imported	No

Figure 6-4 VMSA right-click on rootdg to get Disk Group menu

Selecting New Volumes from the pop-up menu will open a view similar to Figure 6-5.



Disk Group Name: rootdg

Volume Name: vol05

Comment:

Size:

● Concatenated

○ Striped

○ RAID-5

○ Concatenated Pro

○ Striped Pro

Number of Columns: 2

Stripe Unit Size: 128

□ Mirrored

Total Number of Mirrors: 2

□ Enable Logging

□ Initialize Zero

□ No Layered Volumes

File System: None

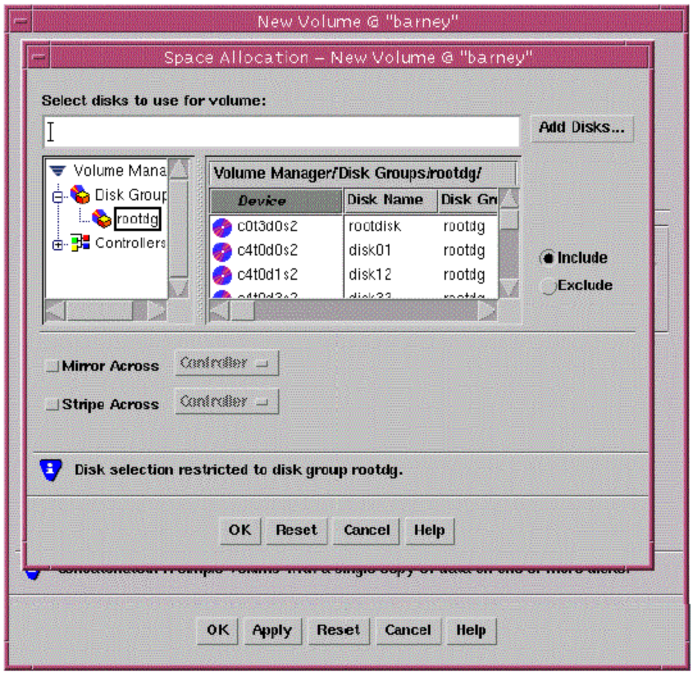
Concatenated: A simple volume with a single copy of data on one or more disks.

Figure 6-5 VMSA New Volume view

Even though the rootdg disk group was used to open this view, it is possible to change to another disk group by either typing in the new disk group name or using the Browse button to select from the disk groups configured on this host.

Also, the Volume Name can be renamed to anything the administrator desires. However, a naming convention that is illustrative of the use of the volume and/or the location of the devices that make up the volume is recommended. Using a naming convention that makes it visually obvious where the volume derives its storage (if possible) will only make the job as system administrator easier during changes, additions, or problem management in the future.

Select Assign Disks to get a listing of devices within the rootdg disk group. Expand the Disk Group icon in the left pane and then click the rootdg icon to display all the devices available in the rootdg disk group. What is visible should be a view similar to Figure 6-6.



Select disks to use for volume:

Volume Manager/Disk Groups/rootdg/

Device	Disk Name	Disk Gr
c0t3d0s2	rootdisk	rootdg
c4t0d0s2	disk01	rootdg
c4t0d1s2	disk12	rootdg

● Include
○ Exclude

□ Mirror Across Controller

□ Stripe Across Controller

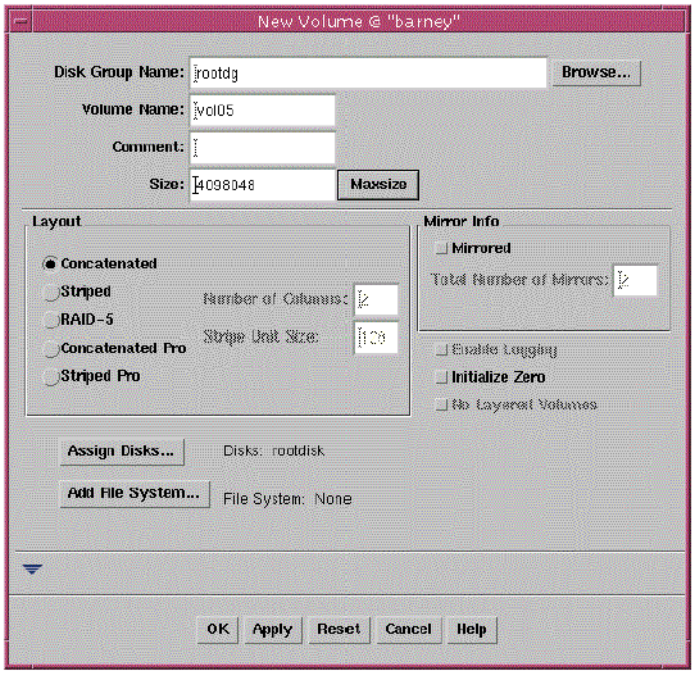
Disk selection restricted to disk group rootdg.

Figure 6-6 VMSA Assign Disks view

To find a device with available storage (assuming that the devices required for the new volume are not known yet), simply move the slider at the bottom of the right pane to the right. This will bring up the Available column, which will show how much space is available for use in the creation of a new volume on each of the devices.

In this example, a single disk is selected. Click OK to return to the New Volume view with the disk name displayed to the right of the Assign Disks button as in Figure 6-7.

Also in Figure 6-7, the Maxsize button can be clicked to utilize all the available free space on the disk for this new volume.



Disk Group Name: rootdg

Volume Name: vol05

Comment:

Size: 4098048

Layout

● Concatenated
○ Striped
○ RAID-5
○ Concatenated Pro
○ Striped Pro

Number of Columns: 2

Stripe Unit Size: 128

Mirror Info

□ Mirrored

Total Number of Mirrors: 2

□ Enable Logging
□ Initialize Zero
□ No Layered Volumes

Disks: rootdisk

File System: None

Figure 6-7 VMSA New Volume with disk information

Another option would be to enter some value, in 512-byte blocks, which is less than the space available on the selected disks.

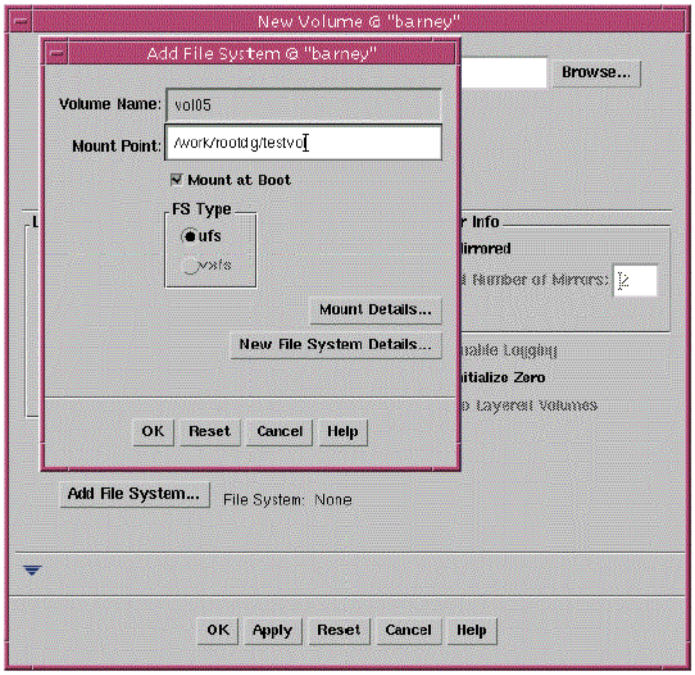
One method of finding the total space available on a group of disks would be to select the disks and return to this view. Then, clicking Maxsize would display the maximum space available using a Concatenated layout as illustrated above. Use some value less than the system-obtained maximum size for the new volume.

Also, while it is possible to set up striped, mirrored, or both striped and mirrored volumes, great care must be taken by the administrator to ensure that the layout of the volume will actually improve performance. It is entirely possible that a layout other than concatenated could be created that would actually degrade performance with regard to storage on the ESS.

The design of the disk read/write and cache algorithms within the ESS allows for faster data access, as long as most of those reads are sequential in nature. Even if the reads are non-sequential, the RAID-5 layout of the logical volumes on the ranks within the ESS allow more drives to be accessed than is possible using a JBOD configuration.

Suffice it to say that most volumes created using ESS logical volumes perform very well using the concatenated layout. Only during those special cases where a mirror is necessary for “snapshot” or offline backups should anything other than simple concatenation be used.

If a filesystem is desired on the volume, it can be created and mounted within this session, as well. Clicking the Add File System button will produce a view similar to Figure 6-8.



Volume Name: vol05

Mount Point: /work/rootdg/testvo

☑ Mount at Boot

● ufs

○ vxfs

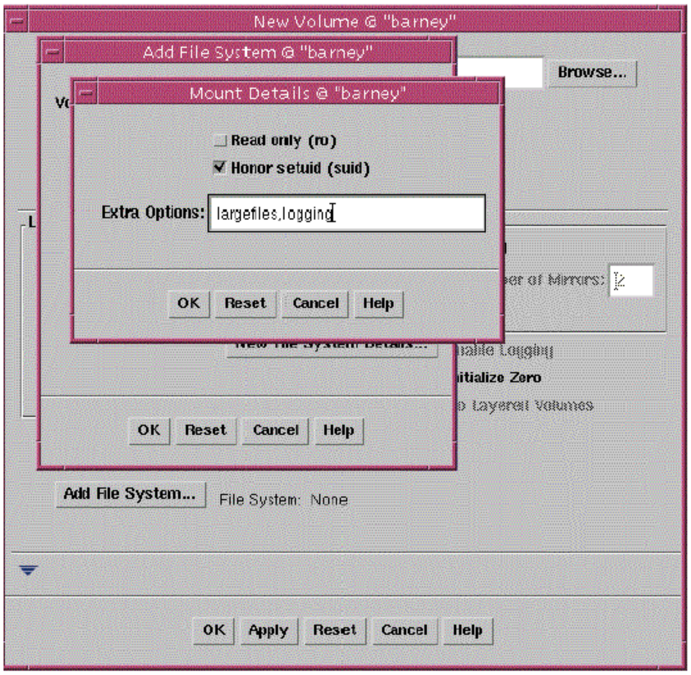
File System: None

Figure 6-8 VMSA Add File System view

It is not possible to change the volume name within this view. However, the mount point can be entered, as well as whether or not an entry should go in the /etc/vfstab file for automatic mount during system boot.

Should any other filesystem type be available, such as VxFS, it would be possible to select it using one of the radio buttons. However, in this example, the only choice is the UFS filesystem type.

Mount arguments can be entered by clicking the Mount Details button. A view similar to Figure 6-9 will be displayed.



□ Read only (ro)
☑ Honor setuid (suid)

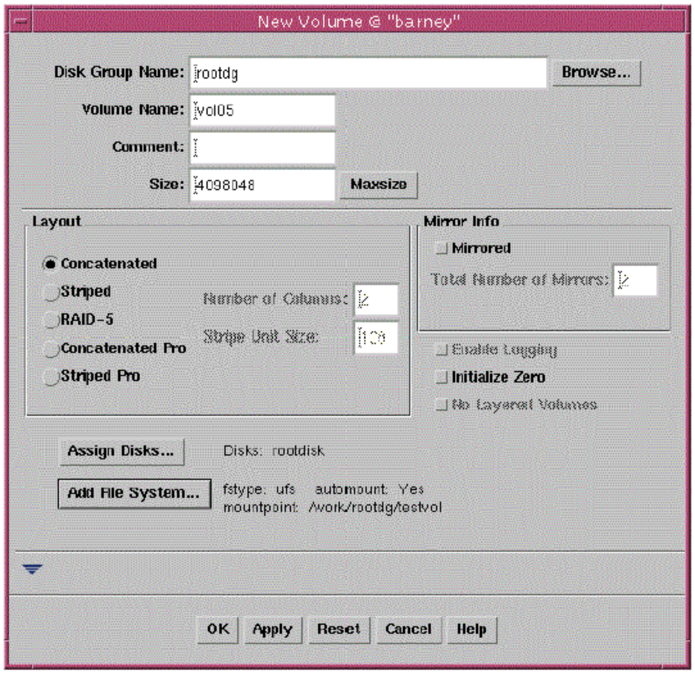
Extra Options: largefiles,logging

File System: None

Figure 6-9 VMSA Mount Details view

For most applications, the defaults are fine. However, for large volumes, we recommend that the large files option is used. Also, for all volumes, the logging option is highly recommended. See 6.5, “Installing and configuring Veritas Volume Manager” on page 134 and the operating system man pages for more information.

Click OK on each of the windows until the New Volume view is displayed with the filesystem information to the right of the Add File System button, as shown in Figure 6-10.



Disk Group Name: rootdg
Volume Name: vol05
Comment:
Size: 4098048

Layout
● Concatenated
○ Striped
○ RAID-5
○ Concatenated Pro
○ Striped Pro
Number of Columns: 2
Stripe Unit Size: 128

Mirror Info
□ Mirrored
Total Number of Mirrors: 2
□ Enable Logging
□ Initialize Zero
□ No Layered Volumes

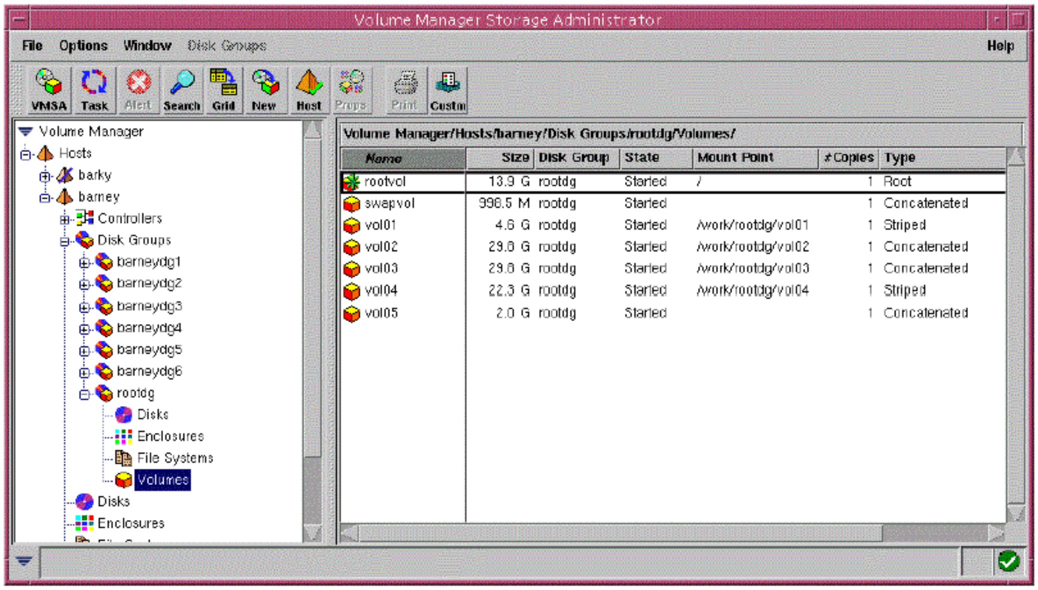
Disks: rootdisk

fstype: ufs automount: Yes
mountpoint: /work/rootdg/testvol

Figure 6-10 VMSA New Volume with disk and filesystem information

At this point, clicking OK or Apply will begin the volume and filesystem creation process. The OK button will cause the New Volume view to vanish, while the Apply button will retain the view and allow additional volumes to be created, as desired.

After clicking OK, the VMSA main view will be visible, as illustrated in Figure 6-11.

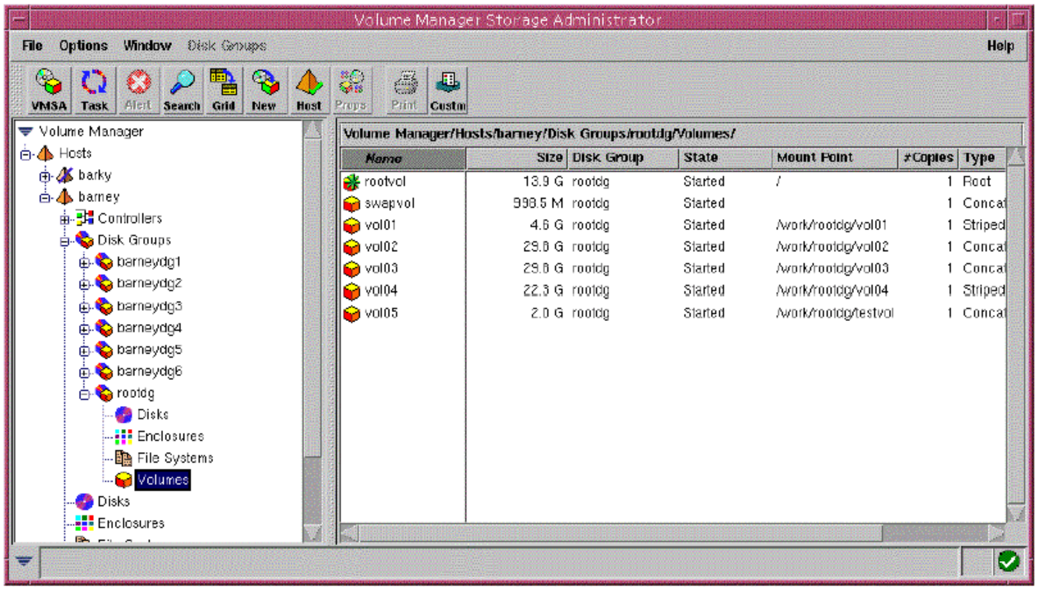


Volume Manager/Hosts/barney/Disk Groups/rootdg/Volumes/

Name	Size	Disk Group	State	Mount Point	#Copies	Type
rootvol	13.9 G	rootdg	Started	/	1	Root
swapvol	998.5 M	rootdg	Started		1	Concatenated
vol01	4.6 G	rootdg	Started	/work/rootdg/vol01	1	Striped
vol02	29.8 G	rootdg	Started	/work/rootdg/vol02	1	Concatenated
vol03	29.8 G	rootdg	Started	/work/rootdg/vol03	1	Concatenated
vol04	22.3 G	rootdg	Started	/work/rootdg/vol04	1	Striped
vol05	2.0 G	rootdg	Started		1	Concatenated

Figure 6-11 VMSA filesystem/volume creation in progress

Note that the new volume — vol05 — is still “under construction”, as the mount point is not yet visible. As soon as the volume has been created and mounted, the view will change to that in Figure 6-12.

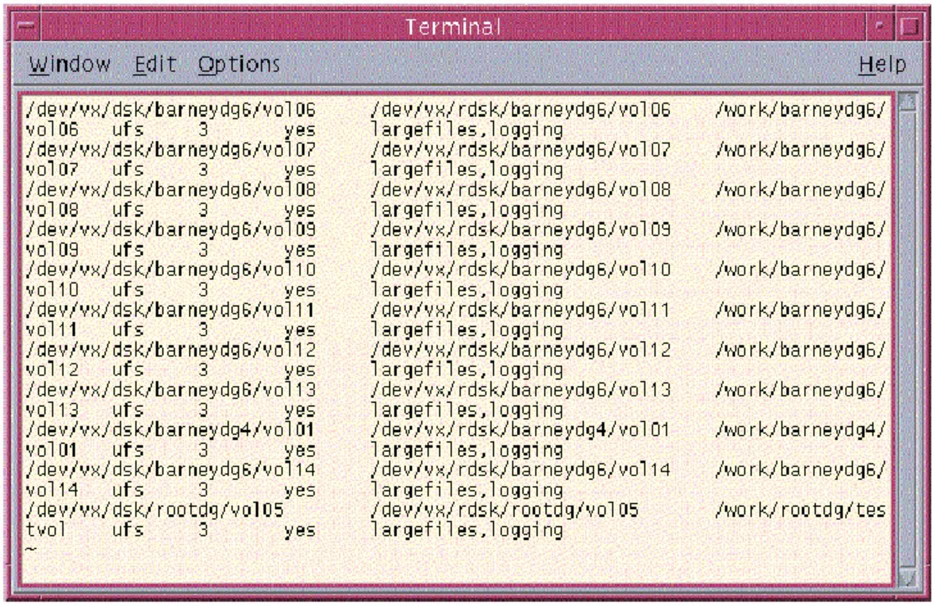


Volume Manager/Hosts/barney/Disk Groups/rootdg/Volumes/

Name	Size	Disk Group	State	Mount Point	#Copies	Type
rootvol	13.9 G	rootdg	Started	/	1	Root
swapvol	998.5 M	rootdg	Started		1	Concat
vol01	4.6 G	rootdg	Started	/work/rootdg/vol01	1	Striped
vol02	29.8 G	rootdg	Started	/work/rootdg/vol02	1	Concat
vol03	29.8 G	rootdg	Started	/work/rootdg/vol03	1	Concat
vol04	22.3 G	rootdg	Started	/work/rootdg/vol04	1	Striped
vol05	2.0 G	rootdg	Started	/work/rootdg/testvol	1	Concat

Figure 6-12 VMSA filesystem/volume creation complete

A quick look at the entry in the /etc/vfstab file verifies that the mount information is ready for the next system boot. The last line illustrates the entry, including the mount options for large files and logging (Figure 6-13).



```
/dev/vx/dsk/barneydg6/vol06     /dev/vx/rdsk/barneydg6/vol06    /work/barneydg6/
vol06   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol07     /dev/vx/rdsk/barneydg6/vol07    /work/barneydg6/
vol07   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol08     /dev/vx/rdsk/barneydg6/vol08    /work/barneydg6/
vol08   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol09     /dev/vx/rdsk/barneydg6/vol09    /work/barneydg6/
vol09   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol10     /dev/vx/rdsk/barneydg6/vol10    /work/barneydg6/
vol10   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol11     /dev/vx/rdsk/barneydg6/vol11    /work/barneydg6/
vol11   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol12     /dev/vx/rdsk/barneydg6/vol12    /work/barneydg6/
vol12   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol13     /dev/vx/rdsk/barneydg6/vol13    /work/barneydg6/
vol13   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg4/vol01     /dev/vx/rdsk/barneydg4/vol01    /work/barneydg4/
vol01   ufs     3       yes     largefiles,logging
/dev/vx/dsk/barneydg6/vol14     /dev/vx/rdsk/barneydg6/vol14    /work/barneydg6/
vol14   ufs     3       yes     largefiles,logging
/dev/vx/dsk/rootdg/vol05        /dev/vx/rdsk/rootdg/vol05       /work/rootdg/tes
tvol    ufs     3       yes     largefiles,logging
~
```

Figure 6-13 View of /etc/vfstab entries

6.5.2 Sun Veritas and ESS logical volumes

Within the VxSA GUI, almost everything has options that become available when the device, volume, diskgroup, and so on, are right-clicked. The pop-up menus that appear provide the available commands for that member under the current conditions.

If additional disk groups are required, simply right-clicking the Disk Groups icon in the left pane will produce a pop-up menu that has, as one of its options, New Disk Group. Follow the prompts under this view to create the disk group desired. Repeat this process for as many disk groups as are required.

If the VxVM software was installed after the installation of the ESS and the logical volumes were brought under VxVM control, then it will not be necessary to initialize the ESS logical volumes. However, it may be necessary to remove the logical volumes from the rootdg. Under the rootdg icon, click the Disks icon to view the devices associate with the rootdg disk group.

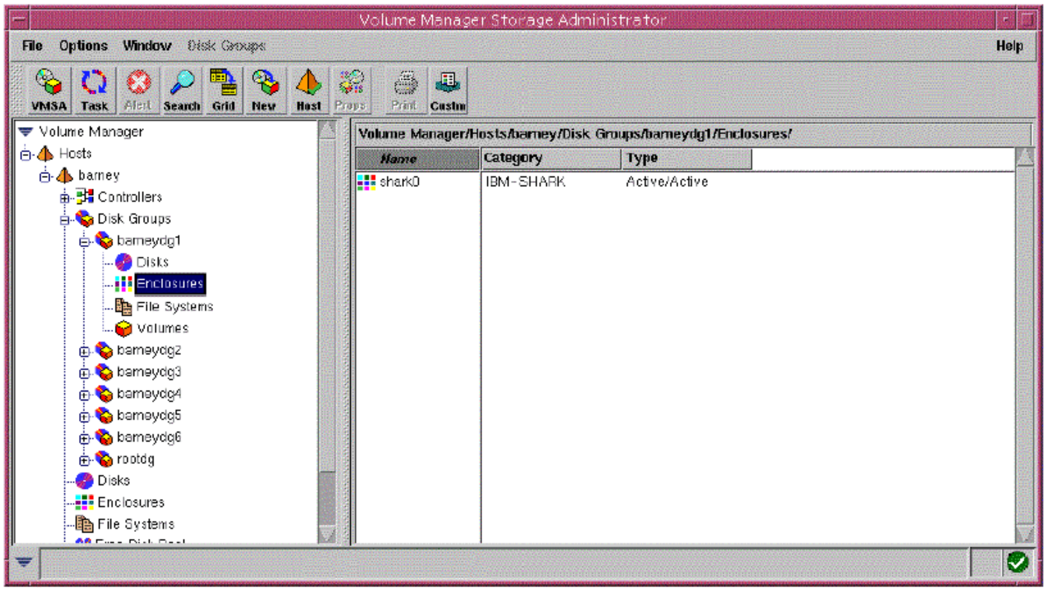
Using the Shift-click and/or the Ctrl-click method, select the devices that will be moved to another disk group or to the Free Pool. Then, right-click one of the selected devices and click Move to Free Disk Pool. The VxVM software will then move all the free devices within the rootdg to the Free Disk Pool area. This step is necessary, as VxVM will not allow a device within one disk group to be used in the creation of a volume in another disk group.

Once the devices are in the Free Disk Pool, they can be moved to any of the disk groups using the technique described above.

When the logical volumes are in the appropriate disk group, follow the steps outlined in 6.5.1, "Creating a filesystem under Veritas" on page 134 to create the volumes desired.

6.5.3 ESS identification under Veritas

When a system like the ESS is brought under Veritas control, it will show up under the Enclosures icon, as seen in Figure 6-14.



Volume Manager/Hosts/barney/Disk Groups/barneydg1/Enclosures/

Name	Category	Type
shark0	IBM-SHARK	Active/Active

Figure 6-14 VMSA Enclosure view

The ESS will be given the name shark0 to identify the array to the VxVM software. Multiple ESSs will be named consecutively.

6.6 Installing and configuring cluster software

The steps to install and configure Veritas Cluster or Sun Cluster with the IBM 2105 are available in the Sun Cluster manuals (Sun Cluster Software Installation Guide) at:

http://docs.sun.com

Note: If you are not familiar with the steps in the Sun Solaris documentation, we strongly recommend that the system administrator attend the Sun Solaris class about high availability features.

7

SDD installation and configuration on HP-UX servers

This chapter describes the concepts to use multipathing and clustering software on servers running the HP-UX operating system. We show how to install and configure the operating system, multipathing and clustering software. In addition, we describe:

- Hardware and software requirements
- Preparing the environment
- Installing and uninstalling the SDD

7.1 Concepts of multipathing and clustering on HP-UX platform

Our prime objective was to demonstrate the capability and effectiveness of the ESS in an HP-UX environment and to provide basic information to allow you to do this in your environment. We explored this by designing, implementing, and testing a high availability (HA) system using the disk storage capacity and speed of the ESS for storing and restoring the information. The ESS uses state of the art technology for data transmission.

For implementation, we used two HP servers for clustering and the ESS to test its compatibility and reliability with HP hardware and software. We configured two servers in an HA configuration using HP-UX version 11.00 and MC/Service Guard for clustering. The servers and software were installed and configured according to HP recommended practices. In addition to base HP-UX version 11.00, we used Logical Volume Manager (LVM).

7.2 Hardware and software requirements

The IBM Subsystem Device Driver has the following hardware and software requirements:

7.2.1 Hardware

- The IBM Enterprise Storage Server (ESS)
- HP 9000 Enterprise Servers

  The listed hardware or software has been thoroughly tested and certified for use. If you use unsupported hardware or software you can have unpredictable results. For the latest IBM supported hardware and software, please see the supported server Web sites:

  http://www.storage.ibm.com/hardsoft/products/ess/supserver_summary_open.htm
  http://www.storage.ibm.com/hardsoft/products/ess/pdf/1012-01.pdf

> **Important:** The TotalStorage Enterprise Storage System (ESS) is supported in many environments and customers are responsible for ensuring that the specific host system configuration used (that is, server model, operating system level and host adapter combination) is a valid and supported configuration by the server manufacturer.
>
> To confirm the list of supported servers, go to “ESS Supported Servers” at:
>
> http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

- SCSI and/or Fibre Channel adapters and cables
- Fibre Channel Switch, if using Fibre Channel adapters (not necessary)

7.2.2 Software

- HP-UX 10.20, 11.x with appropriate packages installed
- SCSI and Fibre Channel device driver installed
- MC/Service Guard

7.2.3 Non supported environments

The following environments are not supported by the Subsystem Device Driver:

- A host server with a single-path Fibre Channel connection to an ESS is not supported. There is no reason to install SDD when only one path is available.

Note: A host server with a single fibre adapter that connects through a switch to multiple ESS ports is considered a multipath Fibre Channel connection and therefore is a supported environment.

- A host server with SCSI channel connections and a single-path Fibre Channel connection to an ESS.
- A host server with both a SCSI channel and Fibre Channel connection to a shared LUN is not supported.

Note: The Subsystem Device Driver also supports one SCSI adapter on the host system. With single-path access, concurrent download of licensed internal code is supported with SCSI devices. However, the load balancing and failover features are not available.

For current information about the SCSI/Fibre Channel adapters which can attach to your HP host system go to:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

7.3 Preparing the environment

In this section we describe the procedures that you have to follow in order to successfully install and administer multipathing.

7.3.1 Pre-installation planning

Using SWAP space for HP-UX on the ESS is not supported. Installation suggestions include Journaled File System (JFS), also called Veritas File System (VxFS). HP-UX supports only HFS for /stand. The OnLine JFS software will not be able to perform online tasks on the /stand. We followed HP's instructions for the installation of Logical Volume Manager (LVM), OnLine JFS for online disk maintenance, and MC/Service Guard for clustering.

7.3.2 Installing the operating system

For the steps to install and configure the HP-UX operating system, see the Hewlett Packard documentation at the following Web sites:

- Installing and updating HP-UX 10.x
  http://www.docs.hp.com/hpux/os/10.x/index.html
- Installing and updating HP-UX 11.0
  http://www.docs.hp.com/hpux/os/11.0/index.html
- Installing and updating HP-UX 11.i
  http://www.docs.hp.com/hpux/os/11i/index.html
- Installing and updating HP-UX 11.i (Version 1.5)
  http://www.docs.hp.com/hpux/os/11iV1.5/index.html

Attention: Verify that the LVM (PV-Link) is installed in the machine, which is necessary to configure the SDD.

7.3.3 Installing and confirming HP-UX installation patches

Confirm that the proper version of HP-UX is installed, as in Example 7-1.

Example 7-1 Confirming the HP-UX version

```
# uname -rs
HP-UX B.11.00
```

Check that the proper software products to support the environment are installed, as in Example 7-2.

Example 7-2 List of packages for HP-UX version 11.00

```
# swlist | grep 1100
QPK1100 B.11.00.51.01 Quality Pack for HP-UX 11.00 (December 2000)
XSWGR1100 B.11.00.52.2 HP-UX General Release Patches, March 2001
XSWHWCR1100 B.11.00.52.3 HP-UX Hardware Enablement and Critical Patches, March 2001
```

Attention: The presented list of packages is valid at the date of this book’s publishing. We recommend that you always install the latest versions of packages, maintenance level fixes, and microcode updates, see:

`http://www.software.hp.com/SUPPORT_PLUS/`

7.3.4 Confirming correct installation of the host bus adapter

Shut down the HP hosts and install the HBAs according to HP’s directions. However, do not cable the HBAs at this time.

After booting the system, verify that the driver has discovered the HBAs, and the links reflecting the controllers are created in /dev. Note that Tachyon Lite adapters (A5158A, A6684A, A6685A) will be listed as /dev/tdX, while Tachyon adapters (A3404A, A3591B) will be /dev/fcmsX.

For Tachyon adapters, the adapter class is “lan”. For Tachyon Lite adapters, it is “fc”. Example 7-3 demonstrates a Tachyon configuration.

Example 7-3 Tachyon configuration

```
# ioscan -funC lan
Class I H/W Path Driver S/W State H/W Type Description
===================================================================
lan 0 8/8.5 fcT1_cntl CLAIMED INTERFACE HP Fibre Channel Mass Storage Cntl /dev/fcms0
lan 1 8/12.5 fcT1_cntl CLAIMED INTERFACE HP Fibre Channel Mass Storage Cntl /dev/fcms1
lan 2 10/12/6 lan2 CLAIMED INTERFACE Built-in LAN /dev/diag/lan2 /dev/ether2
```

The following HP-UX command will confirm installation of the Fibre Channel driver and show the WWPN of each card, where device is /dev/fcmsX or /dev/tdX and where X is the instance number of the adapter, as shown in the Example 7-4:

```
# fcmsutil <device>
```

Example 7-4 Confirming the Fibre Channel installation

```
#fcmsutil /dev/fcms0
Local N_Port_ID is = 0x000001
N_Port Node World Wide Name = 0x10000060B0F90172
N_Port Port World Wide Name = 0x10000060B0F90172
Topology = IN_LOOP
Speed = 1062500000 (bps)
HPA of card = 0xFFB48000
EIM of card = 0xFFFA000D
Driver state = READY
Number of EDB's in use = 0
Number of OIB's in use = 0
Number of Active Outbound Exchanges = 1
Number of Active Login Sessions = 3
```

Note: A significant amount of information is available about using the `fcmsutil` command. Most useful are the WWPN and topology of the adapters. An adapter without a fibre connection will be listed as offline. If an HBA continues to be displayed as offline and the cables are connected to the ESS or fabric, check your connections.

Additionally, Tachyon-Lite adapters are fabric aware, and will automatically determine the protocol to use (FC-AL or PTP) depending on the connection type. There is no configuration file to set for these adapters.

7.3.5 Setting up the ESS

The ESS comes with network-based software called the ESS Specialist to manage and monitor the ESS. ESS Specialist must be accessed from a PC located on a secure network with the ESS. Any browser supporting Java 1.1.6 or higher may be used. See these IBM Redbooks for a detailed discussion of configuring ESS with ESS Specialist.

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *IBM Enterprise Storage Server*, SG24-5465
- *Implementing ESS Copy Services on UNIX and Windows NT/2000*, SG24-5757

On the ESS Specialist Open System Storage window it is possible to add and remove host connections, configure host adapter ports, set up disk groups, add and remove volumes and modify volume assignments.

7.3.6 Attaching an ESS to a Hewlett Packard 9000

This section describes the host system requirements and provides procedures to attach an ESS to a Hewlett Packard 9000 host system with Fibre Channel adapters. This Web site shows all the steps to connect the servers to an ESS:

`http://www.storage.ibm.com/hardsoft/products/ess/pubs/f2ahs04.pdf`

This document describes the connectivity of Hewlett Packard D/K/L/N/V class hosts using the Fibre Channel switched (FC-SW) protocol on optical fiber media, HP Tachyon-Lite HBAs, and the IBM 2109 switch to the IBM Enterprise Storage Server (ESS). Direct connection between all supported HP 9000/s800 hosts and ESS using Fibre Channel arbitrated loop (FC-AL) protocol is also described.

The reader is expected to be able to administer an HP-UX 11.x system, as well as configure and administer an ESS on an IBM SAN. Detailed information on these topics will not be covered in this document. For a more detailed discussion of ESS configuration and SAN topics, please see the following IBM Redbooks:

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *Implementing Fibre Channel Attachment on the ESS*, SG23-6113
- *Implementing an Open IBM SAN*, SG24-6116

Additionally, please refer to the following HP documentation available from:

`http://www.docs.hp.com`

- *HP Fibre Channel Mass Storage Adapters Service and User Manual (HP-UX 10.x, HP-UX 11.0, HP-UX 11i)*
- *HP Fibre Channel Fabric Migration Guide (HP-UX 11.0, HP-UX 11i)*
- *HP A5158A Fibre Channel Adapter Release Notes (HP-UX 11.0)*
- *HP A6684A and A6685A HSC Fibre Channel Adapter Release Notes (HP-UX 10.x, HP-UX 11.0, HP-UX 11i)*

7.3.7 Confirming storage connectivity

The next stage in the process is to confirm that the host is connected to the storage. The procedure is slightly different for switched fabric and Fibre Channel Arbitrated Loop.

Switched fabric

In the our test environment (HP L2000 with A5158A HBAs), the HBA cards are on buses 0/2/0/0 and 0/7/0/0.

There are five defined LUNs on the ESS. There is one connection from the switch to each of two HAs in the ESS. The switch is zoned so each LUN is seen on only one HA from each HBA (two instances of each LUN).

As root at the HP-UX prompt, issue an `ioscan` command to reset and probe the host for storage devices, as in Example 7-5.

Example 7-5 Scan the system and list all the devices belonging to the disk device class

```
# ioscan -fC disk
Class I H/W Path Driver S/W State H/W Type Description
==================================================================
disk 0 0/0/2/0.0.0 sdisk CLAIMED DEVICE SEAGATE ST318203LC
disk 1 0/0/2/0.2.0 sdisk CLAIMED DEVICE SEAGATE ST318404LC
disk 2 0/0/2/1.2.0 sdisk CLAIMED DEVICE HP DVD-ROM 6x/32x
disk 814 0/2/0/0.18.15.0.32.0.0 sdisk CLAIMED DEVICE IBM 2105F20
disk 815 0/2/0/0.18.15.0.32.0.1 sdisk CLAIMED DEVICE IBM 2105F20
disk 816 0/2/0/0.18.15.0.32.0.2 sdisk CLAIMED DEVICE IBM 2105F20
```

Issue the `insf -e` command to assign logical controller numbers in the hardware tree to each of the newly discovered physical devices, as in Example 7-6.

Example 7-6 Reinstall the special files for pseudo-drivers and existing devices

```
# insf -e
insf: Installing special files for sdisk instance 3 address 0/2/0/0.18.15.0.32.0.0
insf: Installing special files for sdisk instance 3 address 0/2/0/0.18.15.0.32.0.1
insf: Installing special files for sdisk instance 3 address 0/2/0/0.18.15.0.32.0.2
```

Notes: This output is edited to only include information about the ESS LUNs. Actual output will include information for all devices on the system.

The hardware paths may not be sequential due to the algorithm which HP uses to assign device IDs. This algorithm is detailed in the *HP Fibre Channel Mass Storage Adapters Service and User Manual*.

Confirm that two instances of each LUN are seen, and the correct device nodes have been created, as in Example 7-7.

Example 7-7 List all devices

```
# ls /dev/rdsk
c0t0d0 c0t1d0 c2t0d0 c2t0d1 c2t0d2 c3t0d0 c3t0d1 c30t0d2
```

Fibre Channel Arbitrated Loop

In the test system, the HP HBAs are on buses 8/8.8 and 8/12.8. There are five defined LUNs on the ESS and each HBA is connected to an HA in the ESS. Every LUN is seen for each HBA (two instances of each LUN should be seen).

The procedure to prepare the LUNs for use in an FC-AL environment is the same as the PTP environment. `ioscan` will search the system for each LUN, and `insf -e` will install the special files.

7.3.8 Tuning recommendations

Based on the work that we did in order to satisfy the test requirements we have to come to the following conclusions on the best tunable parameters in the HP-UX software.

Note: It is necessary to be an experienced HP-UX administrator before attempting to change kernel parameters.

nbuf and bufpages: If you have a static Input Output (I/O) buffer defined, a static buffer in HPUX is usually defined by setting these two parameters to some appropriate number. These two parameters are defined as the size of an I/O buffer. But we feel that a dynamic I/O buffer will serve your needs better without occupying a large chunk (size of nbuf*bufpages) of your server's physical memory. On the flip side, a dynamic buffer will fluctuate between the value of dbc_min_pct and dbc_max_pct. A dynamic I/O buffer can be defined by setting both nbuf and bufpages to zero (0) and giving appropriate values to dbc_min_pct (lower limit) and dbc_max_pct (upper limit). These two parameters are measured in terms of percentage of the total physical memory of your server. We suggest that you should start with low numbers, for example, dbc_min_pct = 5 and dbc_max_pct = 15 or 20. Then monitor your system's resources usage, and adjust according to your needs.

maxfiles: With the storage capacity of the ESS at your disposal, you will have the capability to run large applications, which may cause more than 60 files to be open concurrently. By default, this parameter is set to 60. However, care needs to be exercised with older K class machines, as they will not have the processing power to handle this many files.

maxvgs: By default this parameter is set to nine (9). Which will limit you to ten (0-9) Volume Groups (VG). In order to keep storage organized you may need to create more than ten VGs. Your requirements will suggest the value of this parameter.

memswap_on: If your system is not used for real time computing. We suggest that you turn off the memory swap, and create an interleaved device swap space twice the size of server's physical memory. If the server is running a database application then follow the database vender's recommendations. Otherwise create device swap space approximately twice the size of server's physical memory, followed by a kernel rebuild and reboot of the server.

7.3.9 Installing and configuring cluster software

This section describes how to configure a Hewlett Packard host system for clustering.

The steps to configure MC/Service Guard with the IBM 2105 are the same as the steps in the Hewlett Packard high availability documentation located at:

http://www.docs.hp.com/hpux/ha/index.html

> **Note:** If you are not familiar with the steps in the Hewlett Packard documentation, IBM recommends that the system administrator attend the Hewlett Packard class about high availability features.

After configuring your host for normal operating system access, the 2105 acts as a normal disk device in the MC/Service Guard configuration. IBM recommends that you create volume groups that contain the volumes using the Hewlett Packard logical volume manager. This method of disk management is more reliable, easier, and more flexible to manage than whole-disk management techniques.

7.4 Installing and uninstalling the SDD

In this section we discuss how to install and uninstall SDD. Before we do this we explain how PV-links, the HP multipathing solution, works and compare it to SDD.

7.4.1 What is PV-Link?

Host failover using HP-UX native multipathing requires use of the HP Logical Volume Manger (LVM), and its built-in functionality called PV-Links.

PV-Links is Hewlett Packard's built in multipathing solution for storage attached to servers running HP-UX versions 10.20 and higher. It is built into the Logical Volume Manager (LVM), and is integrated into the operating system. This feature enables multiple paths to be connected between a server and its storage, using alternate paths to the same device for fault tolerance. Some static load balancing may also be established through careful implementation of PV-Links. It is a largely undocumented feature, but can be extremely useful when attaching HP systems to Storage Area Networks (SAN).

Alternate paths are normally used in disaster tolerance to the storage. When a primary path to a storage device is lost, the HP host looks at the VG for alternates. When the first alternate to the path is located, the host fails the I/O over to the new path, and continues normal operation. The host polls the primary path and, once it is restored to operation, automatically fails the I/O back to the primary.

Integrating PV-Links requires some manipulation of LVM's Volume Groups. When planning a multipath topology, careful attention should be paid to the applications and data to be transferred over the SAN. By analyzing the data to be transferred, it is often possible to establish effective, static load balancing through the use of PV-Links. When creating a VG, the first path established to a device is always used as the primary path during normal operation. Any subsequent definition to the same device is used as an alternate path. Therefore, in a balanced I/O configuration with two paths to the storage, assigning half of the storage volumes to one path as primary, and the other half to the second path, an average of fifty percent of the I/O will move down each path. Granted, this is an optimistic scenario, but it demonstrates the possibilities.

PV-Links may be established by the same means used to extend volume groups with single paths. Either SAM or the command line may be used. Simply extend the VG with the alternate paths to the storage, and LVM takes care of the details. After extending the VG with the alternate paths, a `vgdisplay -v` will show all primary and alternate paths. HP-UX versions 10.20 through 11.i support up to eight alternate paths to each primary. Each path may be a primary in a particular volume group, and an alternate in another. This can be very useful if multiple paths are used, and static load balancing is implemented. By using all possible paths as alternates to other volume groups, availability is maximized.

7.4.2 How to use PV-Links

Creating volume groups also allows the implementation of PV-Links, Hewlett Packard's built-in multipathing software for highly available disks such as the IBM 2105. To establish PV-Links, perform the following steps:

1. Create the volume group, using the path to the volumes that you want as the primary path to the data.
2. Extend the volume group with the path to the volumes intended as alternate paths.

   The logical volume manager reads the label on the disk and knows that it is an alternate path to one of the volumes in the group. The logical volume manager labels the volume.

   For example, if a host has access to a volume on a 2105 with the device nodes c2t0d0 and c3t0d0, you can use the c2 path as the primary and create the volume group using only the c2t0d0 path.
3. Extend the volume group to include the c3t0d0 path. When you issue a `vgdisplay -v <device>` command on the volume group, it lists c3t0d0 as an alternate link to the data.

7.4.3 PV-Link versus SDD

The most important difference between PV-Link and SDD is the load balance. PV-Link offers a static load balance, wherever the SDD is a dynamic load balance. Some customers prefer to use the SDD, because it is an IBM product and it is supported when using the ESS.

At the date of this book's publishing, the last version of SDD (version 1.2.0.5) on the IBM Web site was not supported for IBM when using in a cluster environment. Therefore, we obtained the pre-GA version 1.3.0.1 to provide the tests shown in this book.

This version appears satisfactory when used in a cluster environment, but now it is not yet formally supported by IBM. It is necessary to install the MC/Guard Service before installing SDD.

This clustering software, when installed, checks the physical name paths, but does not recognize the vpath as a default device, and then the installation will abort. It is necessary to follow these steps:

1. Install the MC/Service Guard for Clustering
2. Install the SDD
3. Configure all the Volume Groups using the vpaths

Important: Before starting the process to install the last SDD version, verify with an IBM representative or at the following Web site, to confirm if the product is supported to be used in a cluster environment:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm

7.4.4 Installing SDD

The following steps will show you how to install and configure the SDD and should be executed in one desktop window, configured to accept the SAM:

1. Log in as the root user.
2. Load your installation media into the appropriate device drive. Usually you will use the installation CD-ROM supplied with your IBM 2105 ESS server. You can download the latest version of SDD from the following Web site and install SDD from hard disk:

   http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/storsw
3. Type `swinstall -s <directory/driver_name>` and press **Enter** twice to go directly to the install window of SAM, as in Example 7-8.

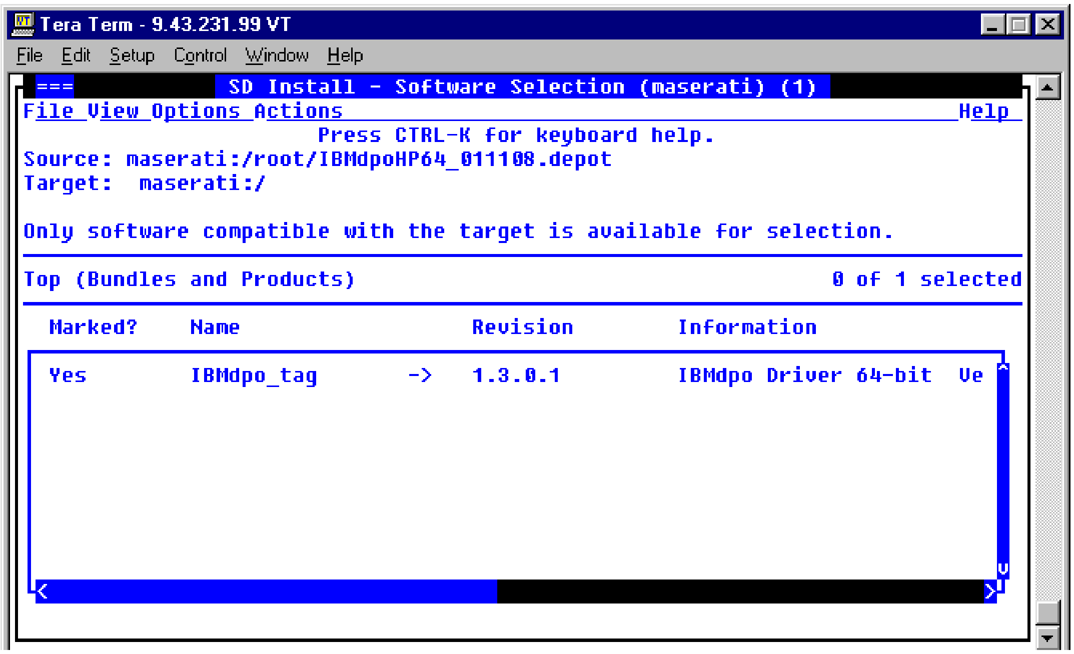
Example 7-8 Command to install the SDD package

```
swinstall -s /root/IBMdpoHP64_011108.depot
```

Note: The name of SDD file is composed for IBMdpoHP<bit>_<date>.depot, where:

- <bit> shows if it is a drive to 32bits or 64 bits
- <date> shows the date of this version (YYMMDD)

4. Select the IBMdpo_tag package using the space bar and press **M**, to mark the item. See Figure 7-1.


Tera Term - 9.43.231.99 VT
File Edit Setup Control Window Help

```
===               SD Install - Software Selection (maserati) (1)
File View Options Actions                                              Help
                       Press CTRL-K for keyboard help.
Source: maserati:/root/IBMdpoHP64_011108.depot
Target:  maserati:/

Only software compatible with the target is available for selection.

Top (Bundles and Products)                                    0 of 1 selected

  Marked?    Name                  Revision        Information

  Yes        IBMdpo_tag      ->    1.3.0.1         IBMdpo Driver 64-bit  Ve
```

Figure 7-1 Installing SDD window

5. Press the **TAB** to get up to the menu bar. The **FILE** word will be highlighted.

6. Type "a" to pull down the **Actions** menu or use the arrow keys and **Return** button to select the desired action and press **ENTER**.
7. Type "i" to pull down the **Install** menu or use the up and down arrows to move the cursor to the desired topic and press **ENTER**.
8. The system will analyze the prerequisites and install the SDD. If some error occur, check the **LOG FILE**.
9. To complete the installation process a system reboot is required. Please reboot your system.

Verifying the SDD Installation and preparing the vpaths

To verify that SDD has been successfully installed, issue the `swlist` command, as in Example 7-9.

Example 7-9 Verify if the SDD package are installed

```
# swlist | grep IBMdpo_tag
  IBMdpo_tag                    1.3.0.1        IBMdpo Driver 64-bit  Version: 1.3.0.1
Nov-08-2001 09:11
```

To finish the installation, it is necessary to create all the vpaths using the following command to the all vpaths, as in Example 7-10.

Example 7-10 Pvcreate command output

```
# pvcreate -f /dev/rdsk/vpath0
Physical volume "/dev/rdsk/vpath0" has been successfully created.
```

Creating Volume Groups using the SDD

The procedure to create a new Volume Group using the SDD is the same to create a Volume Group using the physical volume path.

Instead of using a default physical volume path, the Volume Group should be created using the vpath.

For more information to create a new VG, check the Hewlett Packard documentation Web sites at:

- Installing and updating HP-UX 10.x
  http://www.docs.hp.com/hpux/os/10.x/index.html
- Installing and updating HP-UX 11.0
  http://www.docs.hp.com/hpux/os/11.0/index.html
- Installing and updating HP-UX 11.i
  http://www.docs.hp.com/hpux/os/11i/index.html
- Installing and updating HP-UX 11.i (Version 1.5)
  http://www.docs.hp.com/hpux/os/11iV1.5/index.html

Note: The procedure to create a new Volume Group should be the same as listed in the HP documentation Web sites. The only difference is the command `vgcreate` which requires a device name which should be substituted for the vpath related to the physical volume name.

```
# vgcreate -v vg01 /de/vdsk/vpath
```

Configuring the Volume Groups using the SDD

The followings steps show how to change the physical volume (PV) names for the vpaths in the Volume Groups to provide the complete use of SDD.

1. Type `vgdisplay -v <volume_group>` so the Volume Groups will use the vpaths, as in Example 7-11.

Example 7-11 Verify if the vpaths are in use

```
# vgdisplay -v vg01
--- Volume groups ---
VG Name                     /dev/vg01
VG Write Access             read/write
VG Status                   available
Max LV                      255
Cur LV                      10
Open LV                     10
Max PV                      100
Cur PV                      41
Act PV                      41
Max PE per PV               238
VGDA                        82
PE Size (Mbytes)            4
Total PE                    7238
Alloc PE                    4000
Free PE                     3238
Total PVG                   0
Total Spare PVs             0
Total Spare PVs in use      0

   --- Logical volumes ---
   LV Name                     /dev/vg01/lvol1
   LV Status                   available/syncd
   LV Size (Mbytes)            1600
   Current LE                  400
   Allocated PE                400
   Used PV                     8


   --- Physical volumes ---

   PV Name                     /dev/dsk/c2t0d0
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d1
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On
   PV Name                     /dev/dsk/c2t0d2
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On
```

2. Choose the physical volume name which will be substituted for the vpath. For example, choose the /dev/dsk/c2t0d0.
3. Type `showvpath` to check the relationship between PV name and vpath, as in Example 7-12.

Example 7-12   Showvpath command output

```
root@maserati [/]
# cd /opt/IBMdpo/bin

root@maserati [/opt/IBMdpo/bin]
# ./showvpath
vpath0:
        /dev/rdsk/c2t0d0
        /dev/rdsk/c3t0d0
vpath1:
        /dev/rdsk/c2t0d1
        /dev/rdsk/c3t0d1
vpath2:
        /dev/rdsk/c2t0d2
        /dev/rdsk/c3t0d2
```

4. Identify the associated vpath for the selected physical volume. Insert the related vpath in the Volume Group. In the example, it will be the vpath0.
5. Extend the volume group with the path to the vpath intended as the alternate path, as in Example 7-13.

Example 7-13   Vgextend command output

```
# vgextend vg01 /dev/dsk/vpath0
Volume group "vg01" has been successfully extended.
Volume Group configuration for /dev/vg01 has been saved in /etc/lvmconf/vg01.conf
```

6. Check if the Volume Group has been successfully extended, as in Example 7-14.

Example 7-14   Check if the VG has been successfully extended

```
# vgdisplay -v vg01
--- Volume groups ---
VG Name                     /dev/vg01
VG Write Access             read/write
VG Status                   available
Max LV                      255
Cur LV                      10
Open LV                     10
Max PV                      100
Cur PV                      41
Act PV                      41
Max PE per PV               238
VGDA                        82
PE Size (Mbytes)            4
Total PE                    7238
Alloc PE                    4000
```

```
Free PE                     3238
Total PVG                   0
Total Spare PVs             0
Total Spare PVs in use      0

   --- Logical volumes ---
   LV Name                     /dev/vg01/lvol1
   LV Status                   available/syncd
   LV Size (Mbytes)            1600
   Current LE                  400
   Allocated PE                400
   Used PV                     8


   --- Physical volumes ---

   PV Name                     /dev/dsk/c2t0d0
   PV Name                     /dev/dsk/vpath0 Alternate Link
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d1
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d2
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On
```

7. Reduce the volume group with the path to the physical volume, as in Example 7-15.

Example 7-15 Vgreduce command output

```
# vgreduce vg01 /dev/dsk/c2t0d0
Volume group "vg01" has been successfully reduced.
Volume Group configuration for /dev/vg01 has been saved in /etc/lvmconf/vg01.conf
```

8. Check if the Volume Group has been successfully reduced, as in Example 7-16.

Example 7-16 Check if the VG has been successfully reduced

```
# vgdisplay -v vg01
--- Volume groups ---
VG Name                     /dev/vg01
VG Write Access             read/write
VG Status                   available
Max LV                      255
Cur LV                      10
Open LV                     10
Max PV                      100
Cur PV                      41
Act PV                      41
Max PE per PV               238
VGDA                        82
PE Size (Mbytes)            4
```

```
Total PE                    7238
Alloc PE                    4000
Free PE                     3238
Total PVG                   0
Total Spare PVs             0
Total Spare PVs in use      0

   --- Logical volumes ---
   LV Name                     /dev/vg01/lvol1
   LV Status                   available/syncd
   LV Size (Mbytes)            1600
   Current LE                  400
   Allocated PE                400
   Used PV                     8


   --- Physical volumes ---

   PV Name                     /dev/dsk/vpath0
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d1
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d2
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On
```

9. It is necessary to execute all the physical volumes in the Volume Group to provide several benefits of SDD to this VG.

7.4.5 Uninstalling SDD

The following steps will show you how to install and configure the SDD and should be executed in one desktop window, configured to accept the SAM.

We recommend that you remove the drive before installing MC/Service Guard for Clustering.

1. Log in as the root user.
2. Verify if the SDD are installed. Type `swlist` and look for the IBMdpo_tag, as in Example 7-17.

Example 7-17 Verify if the SDD package are installed

```
# swlist | grep IBMdpo_tag
  IBMdpo_tag                    1.3.0.1        IBMdpo Driver 64-bit  Version: 1.3.0.1
Nov-08-2001 09:11
```

3. Verify if the system is using the vpaths. Type `vgdisplay -v <volume_group>` to all the Volume Groups configured in the machine and check if there is some PV name using `/dev/rdsk/vpath`, as in Example 7-18.

Example 7-18 Verify if the vpaths are in use

```
# vgdisplay -v vg01
--- Volume groups ---
VG Name                     /dev/vg01
VG Write Access             read/write
VG Status                   available
Max LV                      255
Cur LV                      10
Open LV                     10
Max PV                      100
Cur PV                      41
Act PV                      41
Max PE per PV               238
VGDA                        82
PE Size (Mbytes)            4
Total PE                    7238
Alloc PE                    4000
Free PE                     3238
Total PVG                   0
Total Spare PVs             0
Total Spare PVs in use      0

   --- Logical volumes ---
   LV Name                     /dev/vg01/lvol1
   LV Status                   available/syncd
   LV Size (Mbytes)            1600
   Current LE                  400
   Allocated PE                400
   Used PV                     8


   --- Physical volumes ---

   PV Name                     /dev/dsk/vpath0
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d1
   PV Name                     /dev/dsk/vpath1 Alternate Link
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/vpath2
   PV Name                     /dev/dsk/c2t0d2 Alternate Link
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On
```

4. Check what kind of configuration you have in your environment. The example above shows three configuration types.
   - Physical volumes are just using the vpath
   - Physical volumes are using the PV name as a primary path
   - Physical volumes are using the vpath as a primary path
5. If you have the second and third configuration, skip to step 9.
6. Type `showvpath` to check the relationship between PV name and vpath, as in Example 7-19.

Example 7-19 Showvpath command output

```
root@maserati [/]
# cd /opt/IBMdpo/bin

root@maserati [/opt/IBMdpo/bin]
# ./showvpath
vpath0:
        /dev/rdsk/c2t0d0
        /dev/rdsk/c3t0d0
vpath1:
        /dev/rdsk/c2t0d1
        /dev/rdsk/c3t0d1
vpath2:
        /dev/rdsk/c2t0d2
        /dev/rdsk/c3t0d2
```

7. Extend the volume group with the path to the vpath intended as alternate path, as in Example 7-20.

Example 7-20 Vgextend command output

```
# vgextend vg01 /dev/dsk/c2t0d0
Volume group "vg01" has been successfully extended.
Volume Group configuration for /dev/vg01 has been saved in /etc/lvmconf/vg01.conf
```

8. Check if the Volume Group has been successfully extended, as in Example 7-21.

Example 7-21 Check if the VG has been successfully extended

```
# vgdisplay -v vg01
--- Volume groups ---
VG Name                     /dev/vg01
VG Write Access             read/write
VG Status                   available
Max LV                      255
Cur LV                      10
Open LV                     10
Max PV                      100
Cur PV                      41
Act PV                      41
Max PE per PV               238
VGDA                        82
PE Size (Mbytes)            4
Total PE                    7238
Alloc PE                    4000
Free PE                     3238
Total PVG                   0
Total Spare PVs             0
Total Spare PVs in use      0
```

```
   --- Logical volumes ---
   LV Name                     /dev/vg01/lvol1
   LV Status                   available/syncd
   LV Size (Mbytes)            1600
   Current LE                  400
   Allocated PE                400
   Used PV                     8


   --- Physical volumes ---

   PV Name                     /dev/dsk/vpath0
   PV Name                     /dev/dsk/c2t0d0 Alternate Link
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/c2t0d1
   PV Name                     /dev/dsk/vpath1 Alternate Link
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On

   PV Name                     /dev/dsk/vpath2
   PV Name                     /dev/dsk/c2t0d2 Alternate Link
   PV Status                   available
   Total PE                    118
   Free PE                     118
   Autoswitch                  On
```

9. Reduce the volume group with the path to the physical volume, as in Example 7-22.

Example 7-22 Vgreduce command output

```
# vgreduce vg01 /dev/dsk/vpath0
Volume group "vg01" has been successfully reduced.
Volume Group configuration for /dev/vg01 has been saved in /etc/lvmconf/vg01.conf
```

10. Check if the Volume Group has been successfully reduced, as in Example 7-23.

Example 7-23 Check if the VG has been successfully extended

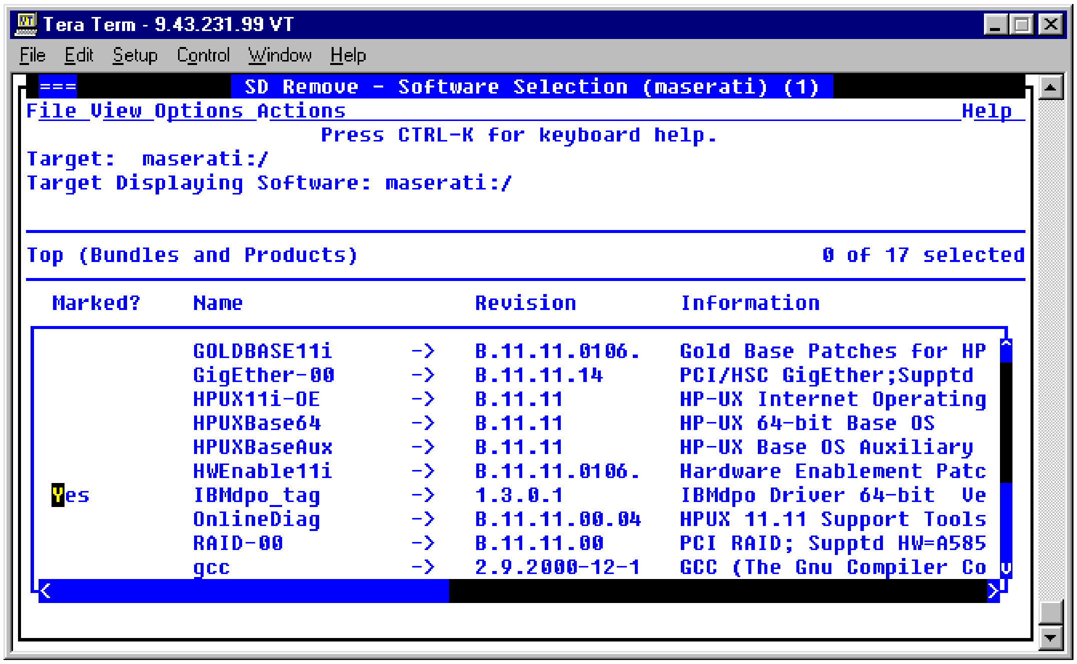
```
# vgdisplay -v vg01
--- Volume groups ---
VG Name                     /dev/vg01
VG Write Access             read/write
VG Status                   available
Max LV                      255
Cur LV                      10
Open LV                     10
Max PV                      100
Cur PV                      41
Act PV                      41
Max PE per PV               238
VGDA                        82
PE Size (Mbytes)            4
Total PE                    7238
Alloc PE                    4000
Free PE                     3238
```

```
Total PVG                0
Total Spare PVs          0
Total Spare PVs in use   0

   --- Logical volumes ---
   LV Name                  /dev/vg01/lvol1
   LV Status                available/syncd
   LV Size (Mbytes)         1600
   Current LE               400
   Allocated PE             400
   Used PV                  8


   --- Physical volumes ---

   PV Name                  /dev/dsk/c2t0d0
   PV Status                available
   Total PE                 118
   Free PE                  118
   Autoswitch               On

   PV Name                  /dev/dsk/c2t0d1
   PV Name                  /dev/dsk/vpath1 Alternate Link
   PV Status                available
   Total PE                 118
   Free PE                  118
   Autoswitch               On

   PV Name                  /dev/dsk/vpath2
   PV Name                  /dev/dsk/c2t0d2 Alternate Link
   PV Status                available
   Total PE                 118
   Free PE                  118
   Autoswitch               On
```

11.It is necessary to execute all the physical volumes in all Volume Group to remove all the dependencies of SDD.

12.Type `swremove` and press **Enter** twice to go directly to the uninstall screen of SAM. See Figure 7-2.

13.Select the IBMdpo_tag package using the space bar and press **M**, to mark the item.


Tera Term - 9.43.231.99 VT
File Edit Setup Control Window Help

```
 ===        SD Remove - Software Selection (maserati) (1)
File View Options Actions                                       Help
                       Press CTRL-K for keyboard help.
Target:  maserati:/
Target Displaying Software: maserati:/

Top (Bundles and Products)                             0 of 17 selected

  Marked?     Name                  Revision        Information

              GOLDBASE11i       ->  B.11.11.0106.   Gold Base Patches for HP
              GigEther-00       ->  B.11.11.14      PCI/HSC GigEther;Supptd
              HPUX11i-OE        ->  B.11.11         HP-UX Internet Operating
              HPUXBase64        ->  B.11.11         HP-UX 64-bit Base OS
              HPUXBaseAux       ->  B.11.11         HP-UX Base OS Auxiliary
              HWEnable11i       ->  B.11.11.0106.   Hardware Enablement Patc
  Yes         IBMdpo_tag        ->  1.3.0.1         IBMdpo Driver 64-bit  Ve
              OnlineDiag        ->  B.11.11.00.04   HPUX 11.11 Support Tools
              RAID-00           ->  B.11.11.00      PCI RAID; Supptd HW=A585
              gcc               ->  2.9.2000-12-1   GCC (The Gnu Compiler Co
```

Figure 7-2   Uninstalling SDD window

14. Press the **TAB** to get up to the menu bar. The **FILE** word will be highlighted.

15. Type "a" to pull down the **Actions** menu or use the arrow keys and **Return** button to select the desired action and press **ENTER**.

16. Type "r" to pull down the **Install** menu or use the up and down arrows to move the cursor to the desired topic and press **ENTER**.

17. The system will uninstall the SDD. If some error occurs, check the **LOG FILE**.

18. To complete the installation process, the system reboot is required. Please reboot your system.

Verifying the SDD uninstallation

To verify if that SDD has been successfully installed, issue the `swlist` command, as in Example 7-24.

Example 7-24   Verify if the SDD package are uninstalled

```
# swlist | grep IBMdpo_tag
#
```

8

Installing Fibre Channel and configuring multipathing on SGI IRIX 6.5

This chapter describes the connectivity of non-clustered SGI Origin hosts using Fibre Channel switched (FC-SW) protocol on optical fiber media, QLogic QLA2200F Host bus adapter (HBAs) and the IBM 2109 switch to the IBM Enterprise Storage Server (ESS).

Direct connection between SGI hosts and ESS using Fibre Channel Arbitrated Loop (FC-AL) protocol is also described.

The reader is expected to be able to administer an IRIX system as well as configure and administer an ESS on an IBM SAN as this information will not be covered here.

For a more detailed discussion of ESS configuration and SAN topics, please see these IBM Redbooks:

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *Implementing Fibre Channel Attachment on the ESS*, SG23-6113
- *Implementing an Open IBM SAN*, SG24-6116

ESS Specialist Copy Services for Open Systems is not described as it is not supported at this time with SGI hosts. IRIS Fail-safe (clustering) is not described as it is not supported at this time.

8.1 Supported configurations

Two configurations are supported: direct connection and switched fabric.

Direct connection

SGI Origin 200, Origin 2000 or Onyx servers with QLA2200F Fibre Channel HBAs are directly connected to the fibre HAs of ESS. Servers and ESS are configured as FC-AL.

Switched fabric

SGI Origin 200, Origin 2000 or Onyx servers with QLA2200F Fibre Channel HBAs are connected to an IBM 2109 switch (Model S08 or S16) or Brocade switches. Fibre HAs of ESS are connected to the switch. Servers, ESS and switches are configured as FC-SW. Cascading of switches is not supported.

IRIX FX limitation in switched fabric environments

Currently there is a limitation of the IRIX fx utility when used in switched fabric environments. The fx utility, used to partition and label disks, could only be used on LUNs below 10 (from lun0 up to lun9). This is due to an IRIX limitation of the internal argument list within fx (which currently only supports an argument list of 128 bytes). The utility will fail to start on all LUNs above lun9 in a fabric environment. The problem is known to SGI and will be fixed in IRIX release 6.5.13. (The SGI reference number to this issue is 822724). As a possible work around you might partition and label the disks using a direct fibre connection to the ESS and move to the fabric environment once you have finished partitioning and labeled all the disks.

This issue does not exist when connecting directly to the ESS.

Note: IRIS Failsafe (clustering) is not currently supported.

8.2 Installation of Fibre Channel on SGI IRIX 6.5

This chapter describes the installation and configuration of QLogic Fibre Channel adapter and its connection to ESS.

8.2.1 Confirming IRIX installation

Confirm that the proper version of IRIX is installed as shown in Example 8-1.

Example 8-1 Get IRIX version

```
# uname -Rs
IRIX64 6.5 6.5.10
#
```

8.2.2 Confirming QLA2200F installation

Shut down the Origin hosts and install the QLA2200F HBAs according to manufacturers directions. Do not cable the cards at this time.

Note: The IRIX Qlogic driver qlfc is always part of the IRIX operating system.

There are no additional driver packages. Although this Fibre Channel adapter is manufactured by Qlogic there is no formal support for IRIX by Qlogic itself. SGI is responsible for support and distribution of this adapter when used in SGI IRIX servers. After the Origin system reboots, verify that the links reflecting the QLogic controllers are created in /hw/scsi_ctlr. Note that pci/6 (PCI slot 6) is bus 3 and pci/7 (PCI slot 7) is bus 4 in Example 8-2.

Example 8-2   Get adapter information

```
# ls -l /hw/scsi_ctlr
total 0
lrw------- 1 root sys 61 Feb 25 13:23 0 ->
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/0/scsi_ctlr/0
lrw------- 1 root sys 61 Feb 25 13:23 1 ->
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/1/scsi_ctlr/0
lrw------- 1 root sys 61 Feb 25 13:23 2 ->
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/5/scsi_ctlr/0
lrw------- 1 root sys 61 Feb 25 13:23 3 ->
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/6/scsi_ctlr/0
lrw------- 1 root sys 61 Feb 25 13:23 4 ->
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/7/scsi_ctlr/0
#
```

The following IRIX command, where `device` is /hw/scsi_ctlr/N/bus and where `N` is the number of the PCI bus, will confirm installation of the qlfc driver and show the WWPN of each card, as in Example 8-3:

```
scsiha -w {device}
```

Example 8-3   Confirm installation of the qlfc driver

```
#scsiha -w /hw/scsi_ctlr/3/bus
/hw/scsi_ctlr/3/bus Portname: 210000e08b022d6e
#scsiha -w /hw/scsi_ctlr/4/bus
/hw/scsi_ctlr/4/bus Portname: 210000e08b022b6e
#
```

Once it is confirmed that the cards and driver are installed, the driver must be configured.

The /var/sysgen/master.d/qlfc file must be edited to let the driver know if the configuration uses FC-SW or FC-AL protocol. In that file, set qlfc_use_connection_mode to 0 for FC-AL or set it to 1 for FC-SW.

In Example 8-4 the connection mode is set to FC-SW.

Note: The /var/sysgen/master.d/qlfc output might vary, based on the IRIX version.

Example 8-4   /var/sysgen/master.d/qlfc file

```
# cd /var/sysgen/master.d
# vi qlfc
*#ident "master.d/qlfc: $Revision $"
*
* MEM
*
*FLAG PREFIX SOFT #DEV DEPENDENCIES
nsc qlfc_ - - scsi
+thread_class scsi
$$$
/*
```

```
* qlfc_use_connection_mode:
* 3 - point to point preferred, then loop
* 2 - loop preferred, then point to point
* 1 - point to point only
* 0 - loop mode
…etc…
*/
int qlfc_use_connection_mode = 1;
int qlfc_debug = 0;
int qlfc_watchdog_time = 5;
int qlfc_probe_wait_loop_up = 90;
int qlfc_trace_buffer_entries = 1024;
int qlfc_target_missing_timeout = 240;
int qlfc_controller_drain_timeout = 40;
```

Finally, if the configuration will use FC-AL, set each controller in the configuration into loop mode using the IRIX command, as in Example 8-5:

```
scsiha -l {bus_number | device}
```

Example 8-5 Configuration into FC-AL

```
# scsiha -l 3 4
#
```

8.2.3 Setting up ESS

The ESS comes with network-based software called the ESS Specialist to manage and monitor the ESS. ESS Specialist must be accessed from a PC located on a secure network with the ESS. Any browser supporting Java 1.1.8 JDK. See these IBM Redbooks for detailed discussion of configuring ESS with ESS Specialist.

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *IBM Enterprise Storage Server*, SG24-5465
- *Implementing ESS Copy Services on UNIX and Windows NT/2000*, SG24-5757

On the ESS Specialist Open System Storage window it is possible to add and remove host connections, configure host adapter ports, set up disk groups, add and remove volumes and modify volume assignments.

Adding or modifying SGI connections

To add, remove or modify the SGI connections select the **Modify Host Systems** button on the ESS Specialist Open System Storage window. It will be necessary to create as many unique Host Systems (connections) as there are QLA2200F HBAs installed in the SGI hosts attached to the ESS.

When adding a connection it is necessary to select a *host type*. At the time of writing this redbook there was no SGI host type. Until an SGI host type is defined, select the **Sun (Solaris with 32 LUN support)** option as the host type. All testing of SGI connectivity was done with this host type option. However, if there is an SGI host type option available, select that option.

When adding a connection it is necessary to specify the World-Wide Port Name of the host connection. Each installed QLA2200F should have a label with the World-Wide Port Name of the card on the visible (rear) portion of the card. It is also possible to obtain the World-Wide Port Name of each card with the IRIX command:

```
scsiha -w {bus_number | device}
```

In the test system, the QLA2200F cards are on buses 3 and 4.

The World-Wide Port Names are determined as in Example 8-6.

Example 8-6 Get World-Wide Port Names of the adapter

```
# scsiha -w 3 4
3 Portname: 210000e08b022d6e
4 Portname: 210000e08b022b6e
#
```

Configuring host adapter ports

To configure the ESS Host Adapter (HA) ports for SGI connections, select the **Configure Host Adapter Ports** button on the ESS Specialist Open System Storage window. One by one, select and configure the relevant Fibre Channel HAs. The setting of the Storage Server Attributes field will depend on whether multiple hosts utilize the HA. Normally this would not be set to *Access_Restricted* with single host attaches.

If the configuration topology is a direct connection between the QLA2200F and the HA, set the FC Port Attributes field to Fibre Channel Arbitrated Loop. If the configuration utilizes a switch, set the FC Port Attributes field to Fibre Channel Point to Point.

> **Note:** If the configuration topology is ever changed from direct connection to switched or vice-versa, it will be necessary to reset the FC Port Attributes field. **Note**: It will be necessary to set the field to **Undefined** followed by selecting **Perform Configuration Update** to set the HA into service mode. After the update completes, the FC Port Attributes field may then be set from Undefined to the desired value and the Perform Configuration Update selected to update the attribute to the new value.

Adding and assigning volumes

Set up disk groups, create volumes on the ranks within the ESS and assign them to the defined SGI connections as described in the IBM Redbooks listed previously in “Setting up ESS” on page 170.

8.2.4 Installing optical cable

Now install the fibre optic cables.

Switched fabric

- Connect the QLA2200F Fibre Channel HBAs to an IBM 2109 switch or Brocade switch.
- Connect the fibre HAs of the ESS to the switch.

Fibre Channel Arbitrated Loop

Connect the QLA2200F Fibre Channel HBAs directly to the fibre HAs of ESS.

8.2.5 Confirming switch connectivity

Now confirm the switch connectivity.

Switched fabric only

Log in to the IBM/Brocade switch as administrator. Execute a `switchshow` command (Example 8-7) and confirm that:

- Each Qlogic HBA has performed a fabric login to the switch.
- Each ESS HA has performed a fabric login to the switch.

Example 8-7 The switchshow command on 2109 switch

```
SWITCH1:admin> switchshow
switchName: SWITCH1
switchType: 3.2
switchState: Online
switchRole: Principal
switchDomain: 2
switchId: fffc02
switchWwn: 10:00:00:60:69:20:02:72
port 0: sw Online F-Port 21:00:00:e0:8b:02:2b:6e
port 1: sw Online F-Port 21:00:00:e0:8b:02:2d:6e
port 2: sw Online F-Port 10:00:00:00:c9:22:16:ab
port 3: sw Online F-Port 10:00:00:00:c9:20:eb:65
port 4: id No_Light
port 5: id No_Light
port 6: -- No_Module
port 7: -- No_Module
value = 8 = 0x8
SWITCH1:admin>
```

Zoning

In our test setup we have set up zoning based on the unique World Wide Names of the host bus adapters and ESS Fibre Channel ports (known as Soft Zoning). For each path to the ESS LUNs (primary and secondary) we have setup a separate zone (SGI_Path1 and SGI_Path2).

Example 8-8 shows of a WWN based zoning, using the cfgShow command to display the zoning information of the switch.

Example 8-8 Example of a WWN based zoning

```
osvl2109c:admin> cfgShow
Defined configuration:
cfg: osvl_lab
dc_1; w2k_1; MS6000Cluster; MSHPCluster; Tivoli;
compaqzone1;
compaqzone2; MS8500Cluster; AIX_ZONE; OSPL3Zone;
MSCompaqCluster; SGI_Path1; SGI_Path2; NW;
Netfyzone1; Netfyzone2; Netfyzone3; Netfyzone4
...
zone: SGI_Path1
SGI01; ospl3_b1a2
zone: SGI_Path2
SGI02; ospl3_b4a2
...
alias: SGI01 21:00:00:e0:8b:02:2b:6e
alias: SGI02 21:00:00:e0:8b:02:2d:6e
alias: ospl3_b1a2
10:00:00:00:c9:20:eb:65
alias: ospl3_b4a2
10:00:00:00:c9:22:16:ab
...
Effective configuration:
```

```
cfg: osvl_lab
...
zone: SGI_Path1
21:00:00:e0:8b:02:2b:6e
10:00:00:00:c9:20:eb:65
zone: SGI_Path2
21:00:00:e0:8b:02:2d:6e
10:00:00:00:c9:22:16:ab
...
```

8.2.6 Confirming storage connectivity

Now confirm the storage connectivity.

Switched fabric LUN

In the test system, the QLA2200F cards are on buses 3 and 4. There are four defined LUNs on ESS. There are two connections from the switch to two HAs in the ESS. The switch is not zoned so each LUN is seen on each HA from each HBA (four instances of each LUN).

As root at the IRIX prompt, issue the following commands for each card to reset and probe the buses for storage devices, as in Example 8-9:

```
scsiha -r {bus_number | device}
scsiha -p {bus_number | device}
```

Example 8-9 Reset and probe the buses for storage devices

```
# scsiha -r 3 4
# scsiha -p 3 4
```

Issue the `ioconfig` command to assign logical controller numbers in the hardware graph to each of the newly discovered physical devices, as in Example 8-10:

```
ioconfig –d –f /hw
```

Example 8-10 The ioconfig command

```
# ioconfig -d -f /hw
start dir name = /hw
Found /var/sysgen/ioconfig/vme device file
dpt:class=13,type=0,state=-1,case=1 : dat:suffix=-1,pattern=-1,
start_num=1,ioctl=0x20007363
Found /var/sysgen/ioconfig/ifcl device file
dpt:class=0,type=0,state=0,case=1 : dat:suffix=-1,
pattern=xplink,start_num=0,ioctl=0xffffffff
Found /var/sysgen/ioconfig/README device file
/hw/module/1/slot/MotherBoard/node/prom: class 21 type 1 controller 0
unit 0 state 0
/hw/module/1/slot/MotherBoard/node/xtalk/8: class 11 type 3 controller
1 unit 1 state 0
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2: class 11 type 4
controller 4265 unit 3 state 2
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2/ioc3/sys_critical_pare
nt: class 11 type 4 controller 4265 unit 3 state
2
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2/ef: class 8 type 0
controller 19 unit 0 state 1
process_ioc_device_vertex: class=8,type=19,state=1,suffix=,pattern=-1,
start_num=1,ioctl=0xffffffff
FOUND NETWORK AT /hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2/ef
```

```
ioconfig_find: KEY PATTERN = ef
Line1 : 2 /hw/module/1/slot/MotherBoard/node/xtalk/8/pci/5/scsi_ctlr/0
……
```

Confirm that the four instances of each LUN are seen, as in Example 8-11:

Example 8-11 Show disk inventory

```
hinv –c disk
Integral SCSI controller 0: Version QL1040B (rev. 2), single ended
Disk drive: unit 1 on SCSI controller 0
Integral SCSI controller 1: Version QL1040B (rev. 2), single ended
CDROM: unit 6 on SCSI controller 1
Integral SCSI controller 2: Version Fibre Channel AIC-1160, revision 1
Integral SCSI controller 3: Version Fibre Channel QL2200A
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 0 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 1 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 2 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 3 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 0 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 1 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 2 on
SCSI controller 3
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 3 on
SCSI controller 3
Integral SCSI controller 4: Version Fibre Channel QL2200A
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 0 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 1 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 2 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 3 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 0 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 1 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 2 on
SCSI controller 4
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 3 on
SCSI controller 4
#
```

Verify that a directory named with the same name as the World-Wide Node Name of the ESS is created in /hw/rdisk, as in Example 8-12.

Example 8-12 Verify rdisk directory with ESS WWNN

```
ls -l /hw/rdisk
total 0
drwxr-xr-x 2 root sys 0 Feb 25 13:24
5005076300c003b4
…
```

```
#
```

Verify that directories for each LUN are seen in the /hw/rdisk/<wwnn> directory, as in Example 8-13.

Example 8-13 Verify LUNs in the rdisk/ <wwnn> directory

```
ls -l /hw/rdisk/5005076300c003b4
total 0
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun0vh
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun0vol
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun1s0
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun1s1
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun1vh
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun1vol
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun2s0
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun2s1
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun2vh
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun2vol
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun3s0
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun3s1
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun3vh
drwxr-xr-x 2 root sys 0 Feb 25 13:24 lun3vol
#
```

Note: The lunxs0 and lunxs1 slices will only be seen if these slices have been configured with fx. Lunxvh and lunxvol will always be available.

Verify that character special files are created in each lun* directory, as in Example 8-14.

Example 8-14 Verify character special files are created in each lun directory*

```
ls -lR lun*
lun0s0:
total 0
crw------- 1 root sys 0,1384 Mar 6 11:33 c3p10000000c920eb65
crw------- 1 root sys 0, 633 Mar 6 11:33 c3p10000000c92216ab
crw------- 1 root sys 0,1454 Mar 6 11:33 c4p10000000c920eb65
crw------- 1 root sys 0, 813 Mar 6 11:33 c4p10000000c92216ab
lun0s1:
total 0
crw------- 1 root sys 0,1387 Mar 6 11:33 c3p10000000c920eb65
crw------- 1 root sys 0, 639 Mar 6 11:33 c3p10000000c92216ab
crw------- 1 root sys 0,1457 Mar 6 11:33 c4p10000000c920eb65
crw------- 1 root sys 0, 816 Mar 6 11:33 c4p10000000c92216ab
lun0vh:
total 0
crw------- 1 root sys 0, 955 Mar 6 09:20 c3p10000000c920eb65
crw------- 1 root sys 0, 171 Mar 2 09:19 c3p10000000c92216ab
crw------- 1 root sys 0,1173 Mar 6 09:20 c4p10000000c920eb65
crw------- 1 root sys 0, 395 Mar 6 09:18 c4p10000000c92216ab
…etc…
#
```

Fibre Channel Arbitrated Loop

In the test system, the QLA2200F cards are on buses 3 and 4. There are four defined LUNs on ESS. Each HBA is connected to an HA in the ESS. Every LUN is seen from each HBA (two instances of each LUN should be seen).

As root at the IRIX prompt, issue the following commands for each card to reset and probe the buses for storage devices, as in Example 8-15:

```
scsiha -r {bus_number | device}
scsiha -p {bus_number | device}
```

Example 8-15 Reset and probe the buses for storage devices (FC-AL)

```
# scsiha -r 3 4
# scsiha -p 3 4
```

Issue the `ioconfig` command to assign logical controller numbers in the hardware graph to each of the newly discovered physical devices, as in Example 8-16:

```
ioconfig –d –f /hw
```

Example 8-16 The ioconfig command (FC-AL)

```
# ioconfig -d -f /hw
start dir name = /hw
Found /var/sysgen/ioconfig/vme device file
dpt:class=13,type=0,state=-1,case=1 : dat:suffix=-1,pattern=-1,
start_num=1,ioctl=0x20007363
Found /var/sysgen/ioconfig/ifcl device file
dpt:class=0,type=0,state=0,case=1 : dat:suffix=-1,
pattern=xplink,start_num=0,ioctl=0xffffffff
Found /var/sysgen/ioconfig/README device file
/hw/module/1/slot/MotherBoard/node/prom: class 21 type 1 controller 0
unit 0 state 0
/hw/module/1/slot/MotherBoard/node/xtalk/8: class 11 type 3 controller
1 unit 1 state 0
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2: class 11 type 4
controller 4265 unit 3 state 2
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2/ioc3/sys_critical_pare
nt: class 11 type 4 controller 4265 unit 3 state
2
/hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2/ef: class 8 type 0
controller 19 unit 0 state 1
process_ioc_device_vertex: class=8,type=19,state=1,suffix=,pattern=-1,
start_num=1,ioctl=0xffffffff
FOUND NETWORK AT /hw/module/1/slot/MotherBoard/node/xtalk/8/pci/2/ef
ioconfig_find: KEY PATTERN = ef
Line1 : 2 /hw/module/1/slot/MotherBoard/node/xtalk/8/pci/5/scsi_ctlr/0
...
```

Confirm that the two instances of each LUN are seen, as in Example 8-17.

Example 8-17 Hardware inventory of disks (FC-AL)

```
# hinv –c disk
Integral SCSI controller 0: Version QL1040B (rev. 2), single ended
Disk drive: unit 1 on SCSI controller 0
Integral SCSI controller 1: Version QL1040B (rev. 2), single ended
Integral SCSI controller 2: Version Fibre Channel AIC-1160, revision 1
Integral SCSI controller 3: Version Fibre Channel QL2200A
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 0 on
SCSI contrler 3
Fabric Disk: node 5005076300c003b4 port 10000000c92216ab lun 1 on
SCSI contrler 3
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 0 on
SCSI contrler 3
Fabric Disk: node 5005076300c003b4 port 10000000c920eb65 lun 1 on
```

```
SCSI contrler 3
#
```

Verify that the appropriate block special files are created in /hw/disk for each LUN, as in Example 8-18.

Example 8-18 Verify that the appropriate block special files are created (FC-AL)

```
# ls -l /hw/disk
total 0
brw------- 1 root sys 0,511 Feb 25 12:15 dks0d1s0
brw------- 1 root sys 0,516 Feb 25 12:09 dks0d1s1
brw------- 1 root sys 0,533 Feb 25 12:15 dks1d6s7
brw------- 1 root sys 0,537 Feb 25 12:15 dks3d0l1s0
brw------- 1 root sys 0,540 Feb 25 12:15 dks3d0l1s1
brw------- 1 root sys 0,544 Feb 25 12:15 dks3d0l2s0
brw------- 1 root sys 0,547 Feb 25 12:15 dks3d0l2s1
brw------- 1 root sys 0,551 Feb 25 12:15 dks3d0l3s0
brw------- 1 root sys 0,554 Feb 25 12:15 dks3d0l3s1
brw------- 1 root sys 0,558 Feb 25 12:15 dks3d0l4s0
brw------- 1 root sys 0,561 Feb 25 12:15 dks3d0l4s1
brw------- 1 root sys 0,593 Feb 25 12:15 dks4d0l1s0
brw------- 1 root sys 0,596 Feb 25 12:15 dks4d0l1s1
brw------- 1 root sys 0,600 Feb 25 12:15 dks4d0l2s0
brw------- 1 root sys 0,603 Feb 25 12:15 dks4d0l2s1
brw------- 1 root sys 0,607 Feb 25 12:15 dks4d0l3s0
brw------- 1 root sys 0,610 Feb 25 12:15 dks4d0l3s1
brw------- 1 root sys 0,614 Feb 25 12:15 dks4d0l4s0
brw------- 1 root sys 0,617 Feb 25 12:15 dks4d0l4s1
#
```

8.3 Configuring host path failover

The next step is to configure the host path failover. We first discuss the general factors to consider and then look at different situations in turn.

8.3.1 General considerations

Please note that the IRIX multipathing *does not perform any dynamic load balancing* and will provide path failover only.

Static load balancing can be achieved if the primary paths are evenly distributed across the controllers when setting up the failover groups in the /etc/failover.conf file.

Once there was a failover to the alternate path you *manually need to switch back* to the primary path using the `/sbin/scsifo -d` command.

Host failover using IRIX multipathing requires use of the XLV volume manager.

Confirm that failover is enabled, as in Example 8-19.

Example 8-19 Check failover is enabled

```
#chkconfig | grep failover
failover on
```

If the failover flag is set to off you have to turn it on and start the failover script (or reboot the server), as in Example 8-20.

Example 8-20 Enable failover

```
#chkconfig | grep failover
failover off
#chkconfig failover on
#/etc/init.d/failover init
#Configuring Failover.
...
```

8.3.2 Switched fabric

Edit the /etc/failover.conf file to add the paths to ESS LUNs as primary and secondary paths. Use the paths of the character special files described in 8.2.6, "Confirming storage connectivity" on page 173, starting with the WWNN of ESS. See Example 8-21.

Example 8-21 Failover group definition on switched fabric

```
GroupA 5005076300c003b4/lun0/c3p10000000c92216ab
5005076300c003b4/lun0/c4p10000000c920eb65
GroupB 5005076300c003b4/lun1/c3p10000000c92216ab
5005076300c003b4/lun1/c4p10000000c920eb65
GroupC 5005076300c003b4/lun1/c3p10000000c92216ab
5005076300c003b4/lun2/c4p1 0000000c920eb65
GroupD 5005076300c003b4/lun2/c3p10000000c92216ab
5005076300c003b4/lun3/c4p1 0000000c920eb65
```

Confirm proper failover configuration with the IRIX `/sbin/scsifo -d` command, as in Example 8-22.

Example 8-22 Confirm proper failover configuration

```
#scsifo -d
Group 0:
[P] 5005076300c003b4/lun3/c3p10000000c92216ab (190)
[ ] 5005076300c003b4/lun3/c4p10000000c920eb65 (883)
Group 1:
[P] 5005076300c003b4/lun2/c3p10000000c92216ab (181)
[ ] 5005076300c003b4/lun2/c4p10000000c920eb65 (874)
Group 2:
[P] 5005076300c003b4/lun1/c3p10000000c92216ab (172)
[ ] 5005076300c003b4/lun1/c4p10000000c920eb65 (865)
Group 3:
[P] 5005076300c003b4/lun0/c3p10000000c92216ab (163)
[ ] 5005076300c003b4/lun0/c4p10000000c920eb65 (856)
```

Note: [P] indicates the primary path within a failover group.

8.3.3 Fibre Channel Arbitrated Loop

Edit the /etc/failover.conf file to add the paths to ESS LUNs as primary and secondary paths, as in Example 8-23.

Example 8-23 Failover group definition (FC-AL)

```
GroupA sc3d0l11 sc4d0l11
GroupB sc3d0l17 sc4d0l17
```

Confirm proper failover configuration with the IRIX `/sbin/scsifo -d` command, as in Example 8-24.

Example 8-24 Confirm proper failover configuration (FC-AL)

```
#scsifo -d
Group A:
[P] sc3d0l11
[ ] sc4d0l11
Group B:
[P] sc3d0l17
[ ] sc4d0l17
#
```

8.3.4 Manually switch IO between the primary and alternate path

You can switch the primary path using the `sbin/scsifo –s` command and specify the path from where you want to switch within a failover group (use `sbin/scsifo –d` to determine), as in Example 8-25. This might be needed if you want to switch back to the initial primary path after a path failure.

Example 8-25 SGI - manually switch IO between the primary and alternate path

```
#scsifo -d
…
Group 37:
[P] 5005076300c003b4/lun2/c3p10000000c92216ab (181)
[ ] 5005076300c003b4/lun2/c4p10000000c920eb65 (549)
…
#scsifo –s 5005076300c003b4/lun2/c3p10000000c92216ab
New primary path
/hw/module/1/slot/Motherboard/node/xtalk/8/pci/6/scsi_ctlr/0/node/500...
#scsifo -d
…
Group 37:
[ ] 5005076300c003b4/lun2/c3p10000000c92216ab (181)
[P] 5005076300c003b4/lun2/c4p10000000c920eb65 (549)
...
```

8.4 Working with ESS volumes

We now look at how to work with ESS volumes with the SGI host.

8.4.1 Configuring storage

We first examine how to setup the storage area network.

Switched fabric

Use standard IRIX storage configuration utilities to partition and format ESS LUNs and create and mount file systems, as in Example 8-26.

Example 8-26 Configure storage on switched fabric

```
#/usr/bin/fx -x -d
/dev/rdsk/5005076300c003b4/lun1vol/c3p10000000c92216ab
fx version 6.5, Jul 7, 2000
...opening /dev/rdsk/5005076300c003b4/lun1vol/c3p10000000c92216ab
...drive selftest...OK
```

```
Scsi drive type == IBM 2105F20 1206
----- please choose one (? for help, .. to quit this menu)-----
[exi]t [d]ebug/ [l]abel/ [a]uto
[b]adblock/ [exe]rcise/ [r]epartition/
fx> auto
----- create sgiinfo-----
...creating default sgiinfo
* * * * * W A R N I N G * * * * *
about to destroy data on disk
/dev/rdsk/5005076300c003b4/lun1vol/c3p10000000c92216ab! ok? writing
label info to /hw/rdis
k/5005076300c003b4/lun1vol/c3p10000000c92216ab
----- exercise-----
sequential pass 1: scanning [0, 1953152] (1953152 blocks)
0%..........10%..........20%..........30%..........40%..........50%....
......60%..........70%..........80%..........
90%..........100%
butterfly pass 1: scanning [0, 1953152] (1953152 blocks)
0%..........10%...writing label info to
/dev/rdsk/5005076300c003b4/lun1vol/c3p10000000c92216ab
----- done-----
----- please choose one (? for help, .. to quit this menu)-----
[exi]t [d]ebug/ [l]abel/ [a]uto
[b]adblock/ [exe]rcise/ [r]epartition/
fx>exi
# xlv_make
xlv_make> vol Sharks0
Sharks0
xlv_make> data
Sharks0.data
xlv_make> plex
Sharks0.data.0
xlv_make> ve -force
"/dev/dsk/5005076300c003b4/lun0s0/c3p10000000c92216ab"
Sharks0.data.0.0
xlv_make> end
Object specification completed
```

We the use standard IRIX utilities to configure the storage as shown in Example 8-27.

Example 8-27 Configure storage on switched fabric (2)

```
xlv_make> create
xlv_make> show
Completed Objects
(1) VOL Sharks0 (empty) (node=NULL)
VE Sharks0.data.0.0 [empty]
start=0, end=1753745, (cat)grp_size=1
/dev/dsk/5005076300c003b4/lun0s0/c3p10000000c92216ab (1753746
blks)
xlv_make> sh
# xlv_assemble
xlv_assemble: Checking for Disk Plexing Option ... done
VOL Sharks0 flags=0x1, [complete] (node=sgiorigin200)
DATA flags=0x0() open_flag=0x0() device=(192, 10)
PLEX 0 flags=0x0
VE 0 [empty]
start=0, end=1753745, (cat)grp_size=1
/dev/dsk/5005076300c003b4/lun0s0/c3p10000000c92216ab (1753746
blks)
```

```
xlv_assemble: Setting kernel configuration ... done
# exit
xlv_make> quit
# ls -l /dev/xlv
total 0
brw------- 1 root sys 192, 10 Mar 12 16:06 Sharks0
# mkfs -t xfs /dev/xlv/Sharks0
meta-data=/dev/xlv/Sharks0 isize=256 agcount=8, agsize=27403
blks
data = bsize=4096 blocks=219218, imaxpct=25
= sunit=0 swidth=0 blks,
unwritten=1
naming =version 1 bsize=4096
log =internal log bsize=4096 blocks=1168
realtime =none extsz=65536 blocks=0, rtextents=0
# mkdir /lv1_mount
# mount -t xfs /dev/xlv/Sharks0 /lv1_mount
# df -k
Filesystem Type kbytes use avail %use Mounted on
/dev/root xfs 1961580 1750112 211468 90 /
/dev/xlv/Sharks0 xfs 872200 144 872056 1 /lv1_mount
#
```

Fibre Channel Arbitrated Loop

Use standard IRIX storage configuration utilities to partition and format ESS LUNs and create and mount file systems, as in Example 8-28.

Example 8-28 Configure storage on Fibre Channel Arbitrated Loop

```
# fx -x -d /dev/rdsk/dks3d0l17s0
...drive selftest...OK
Scsi drive type == IBM 2105F20 1206
----- please choose one (? for help, .. to quit this menu)-----
[exi]t [d]ebug/ [l]abel/ [a]uto
[b]adblock/ [exe]rcise/ [r]epartition/
fx>exi
# xlv_make
xlv_make> vol Sharks0
Sharks0
xlv_make> data
Sharks0.data
xlv_make> plex
Sharks0.data.0
xlv_make> ve -force "/dev/rdsk/dks3d0l17s0"
Sharks0.data.0.0
xlv_make> end
Object specification completed
xlv_make> create
xlv_make> show
Completed Objects
(1) VOL Sharks0 (empty) (node=NULL)
VE Sharks0.data.0.0 [empty]
start=0, end=1753745, (cat)grp_size=1
/dev/rdsk/dks3d117s0 (1753746 blks)
xlv_make> sh
# xlv_assemble
xlv_assemble: Checking for Disk Plexing Option ... done
VOL Sharks0 flags=0x1, [complete] (node=sgiorigin200)
DATA flags=0x0() open_flag=0x0() device=(192, 10)
PLEX 0 flags=0x0
```

```
VE 0 [empty]
start=0, end=1753745, (cat)grp_size=1
/dev/rdsk/dks3d117s0 (1753746 blks)
xlv_assemble: Setting kernel configuration ... done
# exit
xlv_make> quit
```

We then have to configure the second Fibre Channel loop as shown in Example 8-29.

Example 8-29 Configure storage on Fibre Channel Arbitrated Loop (2)

```
# ls -l /dev/xlv
total 0
brw------- 1 root sys 192, 10 Mar 12 16:06 Sharks0
# mkfs -t xfs /dev/xlv/Sharks0
meta-data=/dev/xlv/Sharks0 isize=256 agcount=8, agsize=27403
blks
data = bsize=4096 blocks=219218, imaxpct=25
= sunit=0 swidth=0 blks,
unwritten=1
naming =version 1 bsize=4096
log =internal log bsize=4096 blocks=1168
realtime =none extsz=65536 blocks=0, rtextents=0
# mkdir /lv1_mount
# mount -t xfs /dev/xlv/Sharks0 /lv1_mount
# df -k
Filesystem Type kbytes use avail %use Mounted on
/dev/root xfs 1961580 1750112 211468 90 /
/dev/xlv/Sharks0 xfs 872200 144 872056 1 /lv1_mount
#
```

8.4.2 Important SGI disk devices naming convention

The /hw directory is used to build the hardware graph. The hardware graph represents the collection of all significant hardware connected to a system. The /hw entries are not meant to be specified on commands pertaining to disk devices.

Instead the traditional `/dev/dsk` and `/dev/rdsk` entries should be used! See Example 8-30.

Example 8-30 SGI disk devices naming convention

```
#/usr/bin/fx -x -d
/dev/rdsk/5005076300c003b4/lun1vol/c3p10000000c92216ab
```

8.4.3 Tuning recommendations

Use the `systune –l` and `swap –l` commands, and check the memory and vm statistics to verify that the best performance is provided by LUNs located on ESS. In particular, note the following:

- Monitor and adjust swap space as necessary to improve performance.
- ESS may not be used as swap space.
- We recommend that you leave Command Tag Queuing to its default value to start, but if errors occur on high throughput applications adjust it as necessary.

8.4.4 Unsupported utilities

All commands beginning with *ses*, such as *sesmgr*, are not supported as ESS does not support SES (SCSI Enclosure Services). *sesmgr* may appear to provide some useful information, however, much hardware information *sesmgr* would normally supply is not available from ESS.

8.4.5 Useful SGI information

Here are some general points that we found useful during the project.

Useful IRIX commands

Here is a collection of commands that might be helpful when using the IRIX operating system (some of them were used and explained in this redbook). Check the man pages or online information for more details.

Table 8-1 Useful IRIX commands

Command	**Purpose**
Monitoring the system log file	tail –f /var/adm/SYSLOG
Print the IRIX OS level	uname –Rs
Check and change software configuration flags	chkconfig
List installed Software	showprods, versions
Monitor system activity	top, gr_top osview –a , gr_osview –a gmemusage
Print underlaying device to mount point	devnm
Print disk information	prtvtoc
Partition and label disk devices	fx (-x)
Control IRIX path failover	scsifo
Boot persistent mount points	/etc/fstab
XLV Logical Volume Manager	xlv_make xlv_mgr
Print hardware inventory	hinv hinv -c disk (for disk devices)
List and install Software	inst swmgr
Probe and control SCSI and Fibre Channel busses	scsiha
Qlogic Fibre Channel adapter configuration file	var/sysgen/master.d/qlfc

SGI information on the Web

Here are some Internet links providing more information about SGI and the IRIX operating system.

Table 8-2 SGI information on the Web

URL	Description
General information about SGI hardware and software products	`http://www.sgi.com`
SGI online support page	`http://support.sgi.com`
SGI IRIX online manuals, man-pages, FAQ	`http://support.sgi.com/othersupp/index.html`

9

Installing and configuring IBM SDD on Windows 2000

In this chapter we describe how to install and set up the Subsystem Device Driver on a Windows 2000 host system attached to an IBM Enterprise Storage Server. We will not separately describe installation and configuration process for Windows NT because it is very similar to Windows 2000. For updated and additional information not included in this chapter, see the README file on the compact disc included with your ESS or visit the Subsystem Device Driver Web site at:

http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/storsw

9.1 Pre-installation

Before installing the IBM Subsystem Device Driver, you must first configure the ESS for single-port or multiple-port access for each LUN. The Subsystem Device Driver requires a minimum of two independent paths that share the same logical unit to use the load balancing and failover features.

For information about configuring your ESS, see the following publications:

- *IBM TotalStorage ESS Introduction and Planning Guide*, GC26-7294
  `http://ssddom02.storage.ibm.com/disk/ess/documentation.html`
- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
  `http://publib-b.boulder.ibm.com/Redbooks.nsf/RedbookAbstracts/sg245420.html`
- Implementing Fibre Channel Attachment on the ESS, SG24-6113
  `http://publib-b.boulder.ibm.com/Redbooks.nsf/RedbookAbstracts/sg246113.html`

As we mentioned earlier in this redbook, the Subsystem Device Driver is not a disk driver itself. Figure 9-1 shows where the IBM SDD fits in the protocol stack.

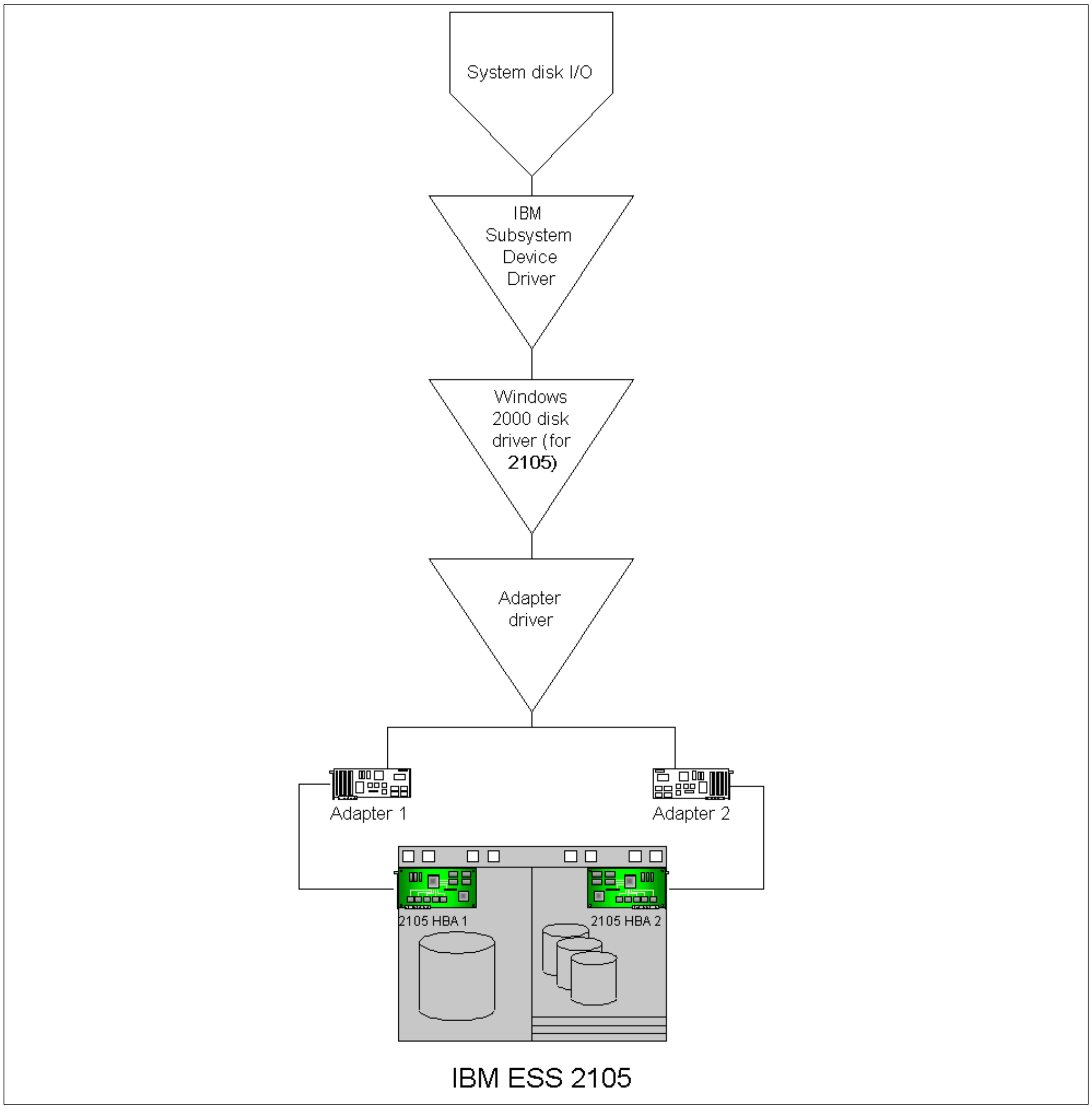


Figure 9-1 Where the IBM SDD fits in the protocol stack on Windows 2000

Note: IBM supports the use of the Enterprise Storage Server as a boot device in selected open system environments, provided that the customer configuration has been reviewed by IBM's engineering and test teams. The review is requested by using the RPQ process. The need for the review is driven by the many server software, host adapter and SAN fabric offerings in the marketplace, and complexity and interoperability issues that might arise using combinations of these offerings.

9.2 Hardware and software requirements

The IBM Subsystem Device Driver has following requirements:

- Hardware:
  - The IBM Enterprise Storage Server
  - Windows 2000 host system
  - SCSI and/or Fibre Channel adapters and cables

- Software:
  - Windows 2000 operating system with Windows 2000 service pack 2 installed
  - SCSI and/or Fibre Channel device drivers

9.2.1 SCSI requirements

To use the Subsystem Device Driver SCSI support, ensure your host system meets the following requirements:

- The maximum number of SCSI adapters that is supported is 32.
- A SCSI cable is required to connect each SCSI host adapter to an ESS port.
- The Subsystem Device Driver I/O load balancing and failover features require a minimum of two SCSI adapters.

> **Note:** The Subsystem Device Driver also supports one SCSI adapter on the host system. With single-path access, concurrent download of licensed internal code is supported with SCSI devices. However, the load balancing and failover features are not available.

For current information about the SCSI adapters that can attach to your Windows 2000 host system go to the Web site at:
`http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

9.2.2 Fibre Channel requirements

To use the Subsystem Device Driver Fibre Channel support, ensure your host system meets the following requirements:

- The maximum number of Fibre Channel adapters that are supported is 256.
- A fiber-optic cable is required to connect each Fibre Channel adapter to an ESS port.
- The Subsystem Device Driver I/O load balancing and failover features require a minimum of two Fibre Channel adapters.

For current information about the Fibre Channel adapters that can attach to your Windows 2000 host system go to the Web site at:
`http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

9.2.3 Non supported environments

The following environments are not supported by the Subsystem Device Driver:

- A host server with a single-path Fibre Channel connection to an ESS is not supported. There is no reason to install SDD when only one path is available.

> **Note:** A host server with a single fibre adapter that connects through a switch to multiple ESS ports is considered a multipath Fibre Channel connection and therefore is a supported environment.

- A host server with SCSI channel connections and a single-path Fibre Channel connection to an ESS is not supported.
- A host server with both a SCSI channel and Fibre Channel connection to a shared LUN is not supported.

9.3 Connecting and configuring SCSI adapters

Before we can install and use the Subsystem Device Driver in SCSI environment, we must configure SCSI adapters in host server. For SCSI adapters that attach boot devices, ensure that the BIOS for the adapter is *enabled*. For all other adapters that attach non-boot devices, ensure the BIOS for the adapter is *disabled*.

Note: When the adapter shares the SCSI bus with other adapters, the BIOS must be disabled.

SCSI specification requires that each device on an SCSI bus must have a unique SCSI ID. To avoid SCSI ID conflicts, before connecting host server SCSI adapters to the ESS HBAs we have to make sure that the SCSI ID on the host adapter is different from the SCSI ID on the ESS HBA. To do this we should check what ID is assigned to the adapter on the host server SCSI adapter and later reserve that ID on the ESS HBA — it will no longer be used for mapping.

SCSI devices attached to the ESS might be initiators (hosts) or target devices. The ESS supports a maximum of four SCSI target devices on any wide SCSI bus. IBM recommends that you use only one SCSI initiator per SCSI bus on an ESS. The number of SCSI devices that the ESS controller uses on the bus is determined by the number of targets specified in the logical configuration for that bus. The SCSI adapter card in the ESS operates in target-only mode. Figure 9-2 and Figure 9-3 show examples of possible ESS SCSI connections.

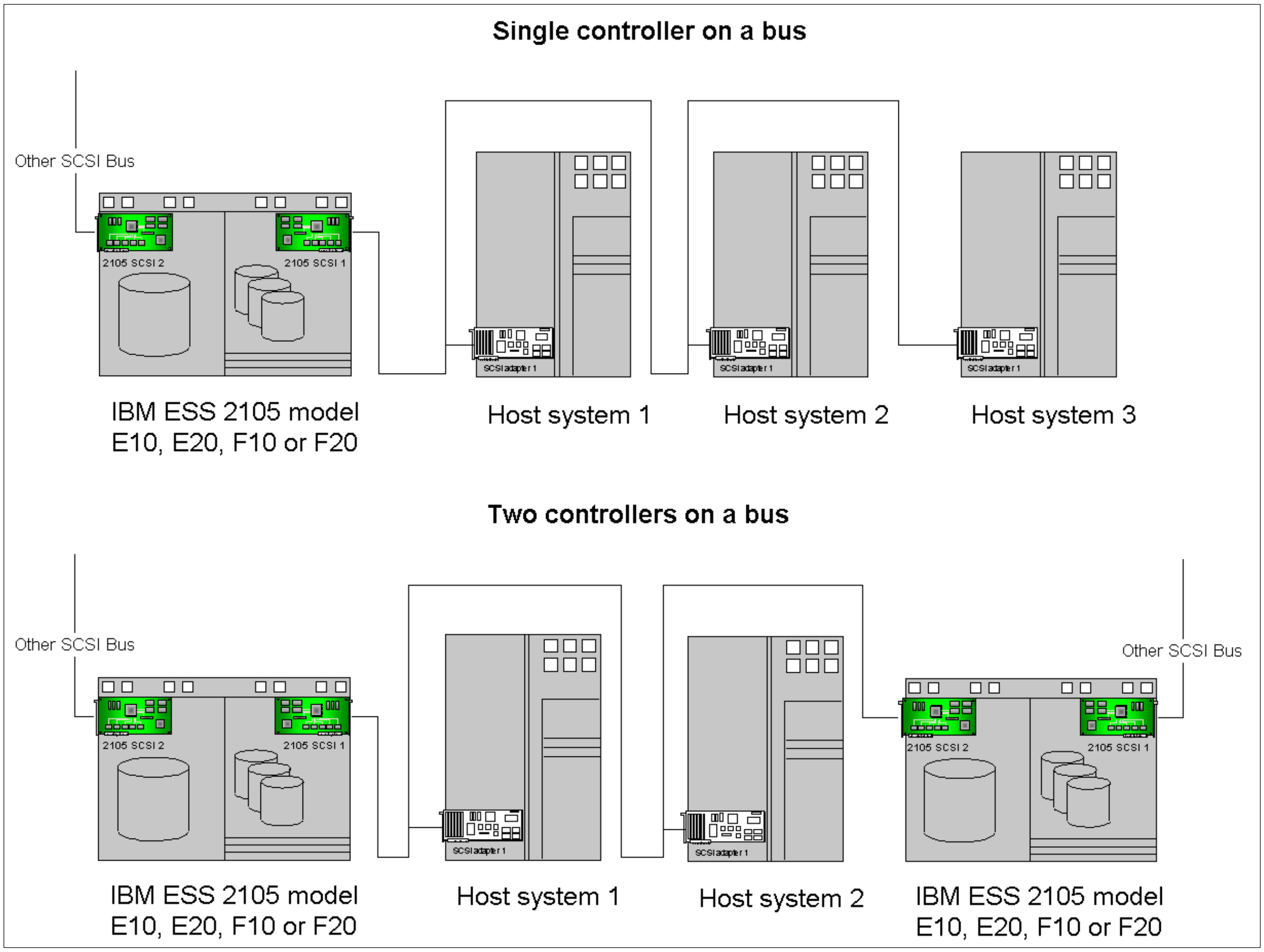


Figure 9-2 Examples of ESS SCSI host interconnections

> **Important:** In case of multiple hosts attached to the same SCSI bus, IBM strongly recommends that you use the same type of host. If you have different hosts on the same SCSI bus, you must use the same type of host adapter. For a list of adapters see this Web site: `http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

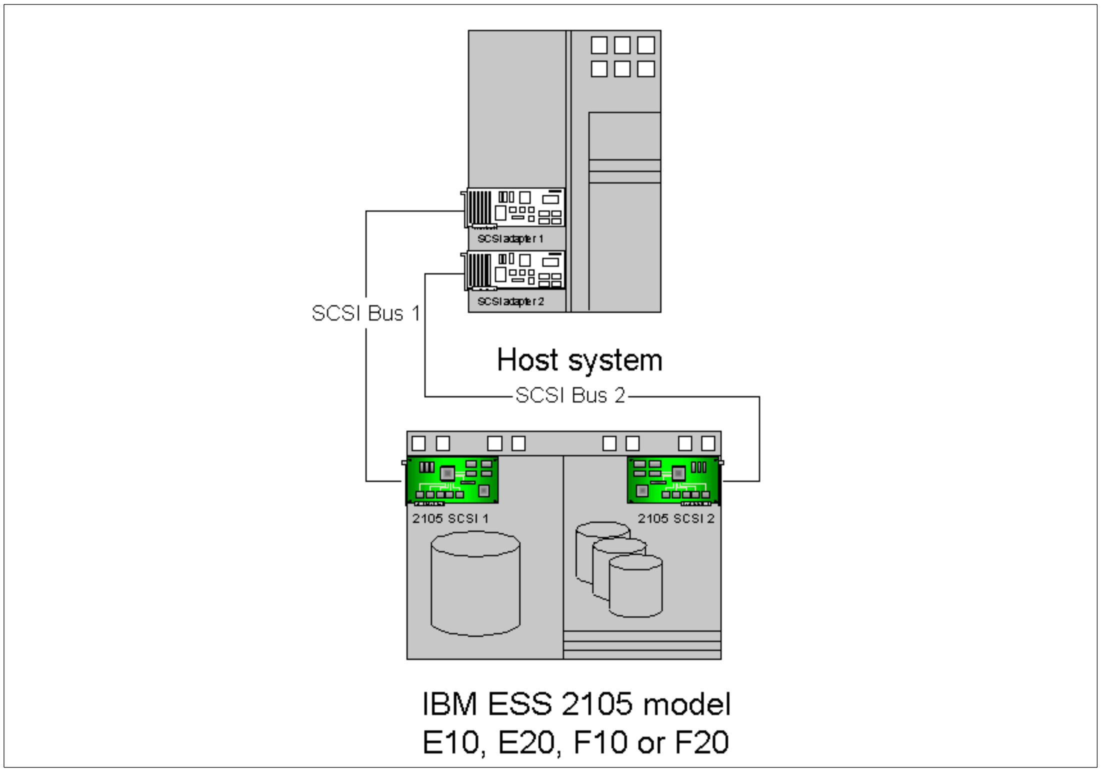


Figure 9-3 Example of multiple SCSI connections

Here we show some guidelines which have to be followed when connecting the IBM ESS to a system host equipped with SCSI adapters:

- Host time-outs might occur due to bus contention when there are too many initiators that try to drive excessive loads over a single bus. The four-initiator limit allows each host to run a significant amount of work without incurring time-outs on I/O operations.
- Your host system might have configuration requirements on the number and type of SCSI devices on the SCSI bus other than what you can do on the ESS.
- You can attach a host system to multiple ports through a separate SCSI bus cable and a separate SCSI adapter for each port.
- You cannot use the configuration in Figure 9-2 for the AS/400 and iSeries. See *Host Systems Attachment Guide 2105 Models E10, E20, F10 and F20,* SC26-7296, for information about how to configure an AS/400 and iSeries. You can find this documentation at:

  `http://ssddom02.storage.ibm.com/disk/ess/documentation.html`
- The SCSI adapter card in an ESS does not provide terminator power (TERMPWR) for the SCSI bus to which it is connected.
- Each host system you attach to a single SCSI bus must be a compatible host system.
- The SCSI adapter card in an ESS provides its own power for termination.
- The host adapter in the ESS has a built-in terminator. Therefore, it does not require external terminators.
- The SCSI adapter card in an ESS must always be at one end of the SCSI bus to which it is connected.
- Each device on a SCSI bus must have a unique ID. Before you attach any device to an SCSI bus, ensure that it has a unique ID for the bus to which you want to connect.

- When you attach a device to the end of your SCSI bus, you must terminate it. If you attach a device in the middle of a SCSI bus, you must not terminate it.
- Each SCSI bus requires at least one initiator. The SCSI specification requires initiators to provide TERMPWR to the SCSI bus.

Important: Before continuing with configuration of all adapters make sure that only one adapter is connected to storage or SAN device. All others adapters *must be connected after* the IBM Subsystem Device Driver or other multipathing software is installed. This will prevent against multiple access to the same disk device which, without any multipathing software, will be seen and treated by host operating system as two (or more) different devices.

Load balancing

In the Windows environment it is not possible to change the load balancing method as it is with AIX and Linux. The load balancing is done by calculating the load on each path and balancing across them. In a clustered environment the load balancing feature is disabled and only failover is supported.

9.4 Connecting and configuring Fibre Channel adapters

Before we can install and use the Subsystem Device Driver in Fibre Channel environment, we must configure Fibre Channel adapters in host server. For Fibre Channel adapters that attach boot devices, ensure that the BIOS for the adapter is enabled. For all other adapters that attach non-boot devices, ensure the BIOS for the adapter is disabled.

Fibre Channel transfers information between the sources and the users of the information. This information can include commands, controls, files, graphics, video and sound. Fibre Channel connections are established between Fibre Channel ports that reside in I/O devices, host systems, and the network that interconnect them. The network consists of elements like switches, hubs, bridges, and repeaters that are used to interconnect the Fibre Channel ports. The ESS architecture supports three basic topologies:

- Point-to-point
- Switched fabric
- Arbitrated loop

Note: If you have not configured the port, only the topologies for point-to-point and arbitrated loop are supported. If you have configured the port, and you want to change the topology, you must first unconfigure the port. After you configure the port, you can change the topology.

Before you start using Fibre Channel devices, check if within your host system correct Fibre Channel host bus adapters are installed or at least it has slots available for them. For a list of operating systems and the host bus adapters for Fibre Channel attachment, see the documentation available at:

`http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm`

9.4.1 Fibre Channel adapters, cables and node-to-node distances

You can order and have IBM install a maximum of 16 Fibre Channel adapters on ESS Models F10 and F20. Each adapter has a single host port.You can install both long-wave adapters and short-wave adapters in a single ESS. Table 9-1 lists the cables that are available for the long-wave and short-wave adapters. This table also lists the distances that the cables support. Feature codes of supported ESS Fibre Channel adapters are as follows:

- **3021** - Fibre Channel/FICON long-wave host adapter (optional 1-16) for open systems, AS/400, S/390 and zSeries hosts. This adapter includes a 31 m (101 ft), long-wave optics (9 micron) cable, P/N 08H2774 with an SC-type connector. It provides an interface that supports 100 Mbps full-duplex data transfer.

> **Note:** The long-wave Fibre Channel/FICON adapter is not available for the iSeries host system. You can, however, increase the distance to 10 km (32700 ft) with a link through cascaded hubs and a long-wave Fibre Channel adapter from the hub to the ESS.

- **3022** - Fibre Channel short-wave host adapter (optional 1-16) for open systems. Feature code 3022 comes with a complimentary 31 m (101 ft) 50-µmultimode fibre cable. You can also use a 62.5-micron fibre cable with the short-wave Fibre Channel card.
- **3023** - Fibre Channel/FICON short-wave host adapter (optional 1-16) for open-systems, AS/400, iSeries, S/390 and zSeries hosts. This adapter includes 31 m (101 ft), 50 micron cable, P/N 54G3384 with an SC-type connector. It provides an interface that supports 100 Mbps full-duplex data transfer.

Table 9-1 Fibre Channel cables for the ESS

Adapter	Cable type	Distance
FC 3023 (short wave Fibre Channel/FICON)	50 micron, multimode (500 MHz/km)	500 m (1635 ft)
	62.5 micron, multimode (200 MHz/km)	300 m (984 ft)
	62.5 micron, multimode (160 MHz/km)	250 m (817 ft)
FC 3022 (short wave Fibre Channel)	50 micron, multimode (500 MHz/km)	500 m (1635 ft)
	62.5 micron, multimode	175 m (575 ft)
FC 3021 (long wave Fibre Channel/FICON)	9 micron, singlemode	50 km (31 miles)
	62.5 micron, multimode (500 MHz/km)	550 m (1799 ft)
	62.5 micron, multimode (200 MHz/km)|	550 m (1799 ft)
	62.5 micron, multimode (160 MHz/km)	550 m (1799 ft)

Table notes:

1. FC 3021 can support point-to-point distances up to 20 km (12 miles) with an RPQ (Request for Price Quotation).
2. FC 3021 and FC 3023 can support distances up to 100 km (62 miles) with the appropriate SAN components.
3. FC 3021 and FC 3023 have a SC connector type. You can use fibre cable couplers to attach a cable with a SC connector type to a server or fabric component port with a LC connector

type.

4. A Mode Conditioning Patch (MCP) cable is required to use FC 3021 with existing 50 micron or 62.5 micron multimode fiber optic cables terminated with ESCON duplex connectors. The cable should be terminated at one end with a 9 micron singlemode SC connector type and at the opposite end with a 62.5 micron multimode ESCON duplex receptacle.
5. A jumper cable is required to use FC 3023 with existing 50 micron or 62.5 micron multimode fiber optic cables terminated with ESCON duplex connectors. The cable requires a male SC connector type on one end and ESCON duplex receptacle on the other end.

For Fibre Channel, the maximum distance between fabric switches, fabric hubs, link extenders and a host Fibre Channel port or an ESS Fibre Channel port is limited to 11 km (6 miles). The maximum distance might be greater than 11 km (6 miles) when a link extender provides appropriate target initiator or controller emulation functions such that the length of the connection perceived by the ESS does not exceed 11 km (6 miles). Link extenders with emulation functions should not be used on links over which synchronous PPRC operations are performed because of the additional path delay introduced by these units.

For details on how to install the different types of supported adapters please refer to Appendix A, “Installing adapters on a Windows 2000 host” on page 221.

9.4.2 LUN affinity, targets and LUNs

For Fibre Channel attachment, LUNs have an affinity to the host’s Fibre Channel adapter through the Worldwide Port Name (WWPN) for the host adapter. In a switched fabric configuration, a single Fibre Channel host could have physical access to multiple Fibre Channel ports on the ESS. In this case, you can configure the ESS to allow the host to use either:

- All physically accessible Fibre Channel ports on the ESS
- Only a subset of the physically accessible Fibre Channel ports on the ESS

In either case, the set of LUNs that are accessed by the Fibre Channel host are the same on each of the ESS ports that can be used by that host.

For Fibre Channel attachment, each Fibre Channel host adapter can architecturally attach up to 2^{64} LUNs. The ESS supports only a maximum of 4096 LUNs divided into a maximum of 16 logical subsystems each with up to 256 LUNs. If the software in the Fibre Channel host supports the SCSI command `Report LUNs`, then you can configure all 4096 LUNs on the ESS to be accessible by that host. Otherwise, you can configure no more than 256 of the LUNs in the ESS to be accessible by that host.

9.4.3 Installing and configuring the QLogic QLA2100F adapter card

The steps to install and configure adapter cards shown below are examples to give you an idea of how to configure QLogic QLA2100F adapter in your host system. Your configuration might be different.

Note: The arbitrated loop topology is the only topology available for the QLogic QLA2100F adapter card.

To install and configure the QLogic QLA2100F adapter card in the host system, perform the following steps:

1. Install the QLogic QLA2100F adapter card(s) in the host system.
2. Connect the cable to the ESS port.
3. Restart the host system.

4. Press Alt+Q to get to the FAST!Util menu.
5. From the Configuration Settings menu, select **Host Adapter Settings**.
6. From the Advanced Adapter Settings menu, press the Down Arrow to highlight **LUNs per Target**. Press Enter.
7. Use the Down Arrow to find and highlight **256**. Press Enter.
8. Press **ESC**.
9. To save the changes, highlight **Yes**. Press Enter.
10. . Restart the host system.
11. Load the QLogic QLA2100F driver and restart the system if instructed to do so.

9.4.4 Installing and configuring the QLogic QLA2200F adapter card

The steps to install and configure adapter cards shown below are examples to give you an idea how to configure QLogic QLA2200F adapter in your host system. Your configuration might be different.

To install and configure the QLogic QLA2200F adapter card in host-system perform the following steps:

1. Install the QLogic QLA2200F adapter card(s) in the host system.
2. Connect the cable to the ESS port.
3. Restart the host system.
4. Press `Alt+Q` to get to the FAST!Util menu.
5. From the Configuration Settings menu, select **Host Adapter Settings**. From the Host Adapter Settings menu, set the following parameters and values:
   a. Host adapter BIOS: **Disabled**

   > **Note:** Host adapter BIOS setting shown above is only an example. General rule of thumb is, that for SCSI adapters that attach boot devices BIOS for the adapter must be *enabled*. For all other adapters that attach non-boot devices, adapter's BIOS must be *disabled*.

   b. Frame size: **2048**
   c. Loop reset delay: **5** (minimum)
   d. Adapter hard loop ID: **Disabled**
6. From the Advanced Adapter Settings menu, press the Down Arrow to highlight **LUNs per target**, then press Enter. Set the parameters and values from the Advanced Adapter Settings menu as follows:
   a. Execution throttle: **100**
   b. Fast command posting: **Enabled**
   c. >4 GB addressing: **Disabled for 32 bit systems**
   d. LUNs per target: **0**
   e. Enable LIP reset: **No**
   f. Enable LIP full login: **No**

> **Note:** In a clustering environment, set Enable LIP full login to **Yes**.

g. Enable target reset: **Yes**

h. Login retry count: **20** (minimum)

i. Port down retry count: **20** (minimum)

j. Driver load RISC code: **Enabled**

k. Enable database updates: **No**

l. Disable database load: **No**

m. IOCB allocation: **256**

n. Extended error logging: **Disabled** (might be enabled for debugging)

> **Note:** The Enable LIP reset, Enable LIP full logon, and Enable target reset parameters control the behavior of the adapter when Windows 2000 tries to do a SCSI bus reset. You must perform a target reset to make cluster failover work. Use the SCSI bus device reset option to clear SCSI reservations. The SAN Data Gateway does not support LIP reset and full login is not necessary after the target reset.

7. Press ESC to return to the Configuration Settings menu.
8. From the Configuration Settings menu, scroll down to the Extended Firmware Settings menu. Press Enter.
9. From the Extended Firmware Settings menu, scroll down to Connection Options to open the Option and Type of Connection window.
10. Select the desired option:

    0: Loop only

    1: Point-to-point only

    2: Loop preferred, rather than point-to-point (If you cannot use arbitrated loop, then default to point-to-point)

    3: Point-to point, rather than loop (If you cannot use point-to-point, then default to arbitrated loop)

> **Note:** If you connect the ESS directly to the host system, the option you select must match the port connections on the ESS. However, if you connect through a switch, the options do not need to match the port connections because the ESS is point-to-point. The appropriate host bus adapter on the server must also support point-to-point connection on a direct connection. Currently, disparate vendors do not function properly in a direct point-to-point connection. This statement is not true if you connect through a switch because the ESS is point-to-point.

11. Press ESC.
12. Save the changes. Highlight **Yes**.
13. Restart the host system.
14. Load the QLogic QLA2100F driver and restart the system if instructed to do so.

9.4.5 Installing and configuring the Emulex LP8000 adapter card

The procedure for installing the Emulex LP8000 adapter is slightly different when compared to procedure for other supported Fibre Channel adapters. In this section we describe in detail how to install Emulex LP8000 Fibre Channel adapter in Windows 2000 host system. Some settings described later in this procedure might not be appropriate for your environment and can be changed if you are an experienced system administrator. However, for most environments, the settings in examples below are very common and should work properly.

The most important difference with the Emulex LP8000 adapter is that from the adapter BIOS level not all settings are available. A special software tool is provided with the Emulex device driver to configure the adapter. This software tool requires Windows 2000 operating system already up and running on the host server. Therefore, when installing Windows 2000 in local boot mode, the LP8000 adapter should be configured after the installation of the operating system is finished as well as device drivers for Emulex are loaded into the operating system.

When installing Windows 2000 in remote boot from an ESS server, the Emulex LP8000 adapter must be already present during installation. Because there is no operating system at this time and there is no possibility to run the software configuration tool, some additional steps have to be performed from the adapter BIOS level.

Downloading current version of Emulex LP8000 adapter device driver

In this section we describe how to download the current version of the Emulex LP8000 adapter device driver from the official EMULEX Web site:

1. Go to the EMULEX Web site at: `http://www.emulex.com`
2. From the Quick Links menu, click **Documentation, Drivers and Software**.
3. Click the host adapter type from the host adapter menu. In this case click **Emulex LP8000**.
4. Click **Drivers for Windows 2000**.
5. Click **Specialized Drivers**.
6. Click on driver with **“SCSI/IP Multi-Protocol...”** in name.
7. Click the **Download Now** button.
8. From the File Download window, click **Save this file to disk** and choose its destination folder. Ensure that the name of the file you want to download is displayed in the window. If the file name is not displayed in the window, go to step 1. Memorize the name of that file — it will vary according to current version of device driver.
9. Click **Save** to download and unzip the file to your hard drive. A window opens that indicates the progress of the download operation. When progress indicator window closes, the download is complete.
10. Unzip the file to desired location on your local hard disk drive or on a floppy diskette if you want to use the Emulex LP8000 device driver during initial installation of Windows 2000 operating system on remote ESS LUN.

Installing Emulex LP8000 device driver running on Windows 2000

Perform the following steps to install the Emulex Fibre Channel adapter device driver in your operating system environment.

Note: If you are installing the Fibre Channel adapter for the first time, you must specify the correct topology.You must also select the appropriate device mapping driver.

1. Install the Emulex LP8000 adapter card(s) in the host system.
2. From your desktop, click **Start** -> **Settings.**
3. Double-click **Control Panel**.
4. Double-click **Add/Remove Hardware** and click **Next** in **Add/Remove Hardware Wizard Welcome** panel.
5. Choose **Add/Troubleshoot a device** and click **Next**.
6. In Choose a Hardware Device panel choose **Add a new device** and click **Next**.
7. Question, `Do you want Windows to search for your new hardware?` will appear. Select **No, I want to select a hardware from a list** and click **Next**.
8. From the list of available hardware types select **SCSI and RAID controllers** and click **Next**.
9. Click **Have disk...** button and then enter the path to the folder containing unpacked Emulex LP800 device driver files. Click **OK** and **Next**.
10. In Select a Device Driver panel choose `Emulex LightPulse 8000 <Current Settings>` and click **Next**.
11. In Start Hardware Installation panel click **Next** to start installation.
12. After the installation is finished and you restart your host system, right-click on **My Computer** icon located on the desktop and choose **Manage**.
13. Double-click **Device Manager**.
14. Double-click **SCSI and RAID controllers**.
15. Verify that the `Emulex LP8000 <Current Settings>` host adapter is on the list. Double-click on it and choose the tab for **Drivers**.
16. Verify the proper Emulex driver is present.

Note: The driver will affect every Emulex LP8000 adapter in the system. If you have more than one adapter that requires a different device driver, then you must change the driver for that adapter. To do this you can use the **Update Driver...** button on that tab.

Configuring Emulex LP8000 device driver

After the Emulex LP8000 is installed in the host system and appropriate drivers are loaded, we must configure every adapter in our system and set-up it properly to connect to the IBM Enterprise Storage Server HBA and access ESS LUNs. Table 9-2 describes what the recommended settings are for that adapter to work properly with IBM ESS in most SAN environments.

Table 9-2 Recommended settings for Emulex LP8000 adapter

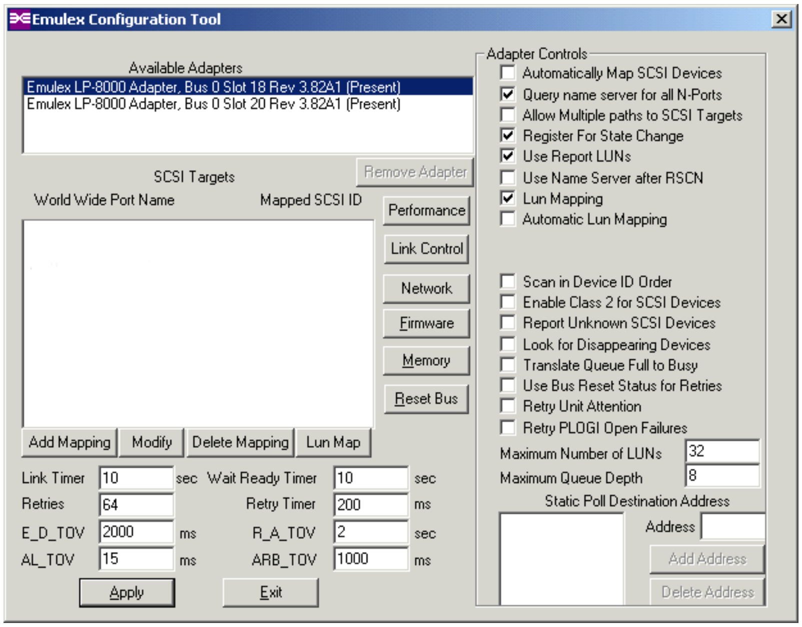
Parameters	Recommended settings
Automatically Map SCSI Devices	Checked (enabled)
Query Name Server for all N-Ports	Checked (enabled)
Allow Multiple Paths to SCSI Targets	Checked (enabled)
Register For State Change	Checked (enabled)

Parameters	Recommended settings
Use Report LUNs	Checked (enabled)
Use Name Server After RSCN	Checked (enabled) only if fabric attached using soft zoning
LUN Mapping	Checked (enabled)
Automatic Lun Mapping	Checked (enabled)
Scan in Device ID Order	Not checked (disabled)
Enable Class 2 for SCSI Devices	Not checked (disabled)
Report Unknown SCSI Devices	Not checked (disabled)
Look for Disappearing Devices	Not checked (disabled)
Translate Queue Full to Busy	Not checked (disabled)
Use Bus Reset Status for Retries	Not checked (disabled)
Retry Unit Attention	Not checked (disabled)
Retry PLOGI Open Failures	Not checked (disabled)
Maximum Number of LUNs	Equal to or greater than the number of the ESS LUNs available to the host bus adapter
Maximum Queue Depth	8
Link Timer	30 seconds
Retries	64
E_D_TOV	2000 milliseconds
AL_TOV	15 milliseconds
Wait Ready Timer	45 seconds
Retry Timer	2000 milliseconds
R_A_TOV	2 seconds
ARB_TOV	1000 milliseconds
Link Control	
Topology	▸ Point-to-point (for fabric switched) ▸ Arbitrated loop (for direct connection)
Link speed	Auto

Together with device driver files, Emulex provides a special software tool to configure adapters. You can run this tool by choosing **Start** -> **Programs** -> **Emulex Configuration Tool** or running `elxcfg.exe` (the full path is `C:\WINNT\system32\elxcfg.exe`). Figure 9-4 shows the initial window for Emulex Configuration Tool.

As you can see, there are no initial mappings between World Wide Port Names and SCSI ID. This result may vary depending on specific environment, topology and SAN devices used to connect system hosts to storage devices. In some cases mappings can be available even if no changes were made to adapter configurations. This is the proper result; however, we advise you to set all parameters for all Emulex LP8000 adapters accordingly to recommended settings, as presented in Table 9-2. To configure the adapter, choose it from the list of available adapters and make all the desired changes. After finishing applying the changes for all adapters, we advise you to reboot the system to automatically map all World Wide Port Names to SCSI ID and ESS LUNs to SCSI LUNs for all adapters. This is not necessary and experienced system administrators may want to manually configure all mappings; however, this should be done carefully.

Note: Link Control parameters are not shown in this example. To get to Link Control parameters, click the **Link Control** button.



Emulex Configuration Tool

Available Adapters

Emulex LP-8000 Adapter, Bus 0 Slot 18 Rev 3.82A1 (Present)
Emulex LP-8000 Adapter, Bus 0 Slot 20 Rev 3.82A1 (Present)

SCSI Targets

World Wide Port Name | Mapped SCSI ID

Adapter Controls

☐ Automatically Map SCSI Devices
☑ Query name server for all N-Ports
☐ Allow Multiple paths to SCSI Targets
☑ Register For State Change
☑ Use Report LUNs
☐ Use Name Server after RSCN
☑ Lun Mapping
☐ Automatic Lun Mapping

☐ Scan in Device ID Order
☐ Enable Class 2 for SCSI Devices
☐ Report Unknown SCSI Devices
☐ Look for Disappearing Devices
☐ Translate Queue Full to Busy
☐ Use Bus Reset Status for Retries
☐ Retry Unit Attention
☐ Retry PLOGI Open Failures

Maximum Number of LUNs: 32
Maximum Queue Depth: 8

Static Poll Destination Address

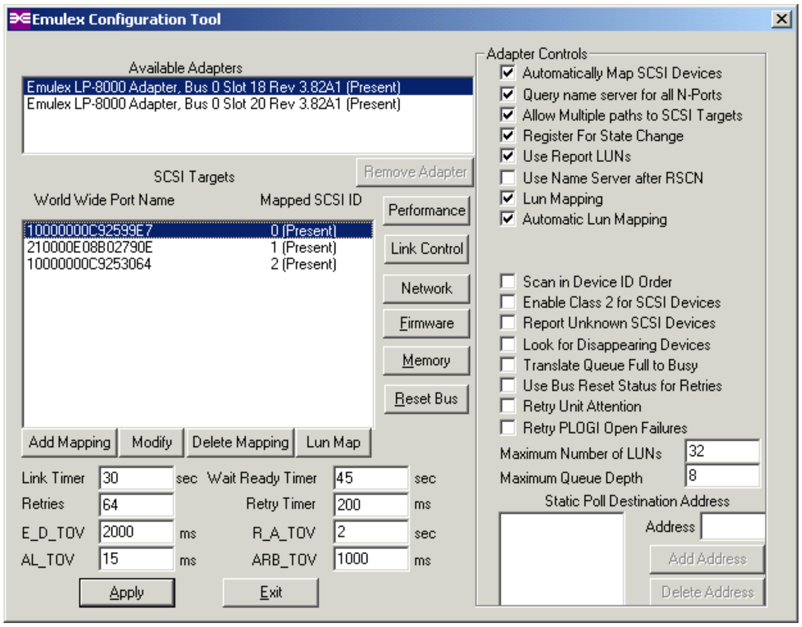
Address:

Link Timer: 10 sec
Retries: 64
E_D_TOV: 2000 ms
AL_TOV: 15 ms
Wait Ready Timer: 10 sec
Retry Timer: 200 ms
R_A_TOV: 2 sec
ARB_TOV: 1000 ms

Figure 9-4 Example of initial window for Emulex configuration tool

Tip: Before configuring next available adapter remember to apply changes to currently configured adapter.

Figure 9-5 shows the configuration of the adapters after applying all recommended changes and rebooting the system. As we can see in this example, all parameters for the adapter in Bus 0 Slot 18 are set accordingly to the recommended settings, as presented in Table 9-2. Also appropriate mappings of WWPN to SCSI ID are available. In case of any missing mappings, you can configure them manually. To do this, press the **Add Mapping** button. The window shown in Figure 9-6 will appear. Select the World Wide Port Name of the adapter you want to map to an SCSI ID and press **OK**. You will be prompted to choose an SCSI ID for that WWPN from list of available SCSI IDs. This is shown in Figure 9-7.



Available Adapters

Emulex LP-8000 Adapter, Bus 0 Slot 18 Rev 3.82A1 (Present)
Emulex LP-8000 Adapter, Bus 0 Slot 20 Rev 3.82A1 (Present)

SCSI Targets

World Wide Port Name	Mapped SCSI ID
10000000C92599E7	0 (Present)
210000E08B02790E	1 (Present)
10000000C9253064	2 (Present)

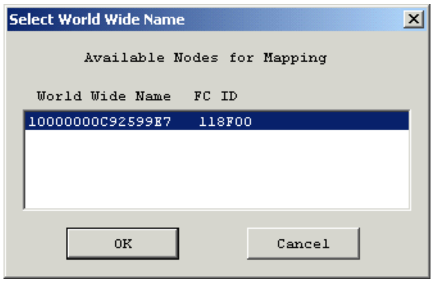
Link Timer 30 sec Wait Ready Timer 45 sec
Retries 64 Retry Timer 200 ms
E_D_TOV 2000 ms R_A_TOV 2 sec
AL_TOV 15 ms ARB_TOV 1000 ms

Adapter Controls

- ☑ Automatically Map SCSI Devices
- ☑ Query name server for all N-Ports
- ☑ Allow Multiple paths to SCSI Targets
- ☑ Register For State Change
- ☑ Use Report LUNs
- ☐ Use Name Server after RSCN
- ☑ Lun Mapping
- ☑ Automatic Lun Mapping
- ☐ Scan in Device ID Order
- ☐ Enable Class 2 for SCSI Devices
- ☐ Report Unknown SCSI Devices
- ☐ Look for Disappearing Devices
- ☐ Translate Queue Full to Busy
- ☐ Use Bus Reset Status for Retries
- ☐ Retry Unit Attention
- ☐ Retry PLOGI Open Failures

Maximum Number of LUNs 32
Maximum Queue Depth 8
Static Poll Destination Address
Address

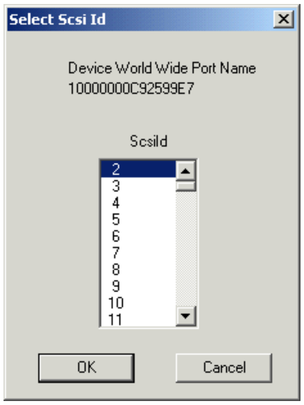
Figure 9-5 Recommended configuration of Emulex LP8000 adapter



Available Nodes for Mapping

World Wide Name	FC ID
10000000C92599E7	118F00

Figure 9-6 Adding WWPN to SCSI ID mapping - Step 1



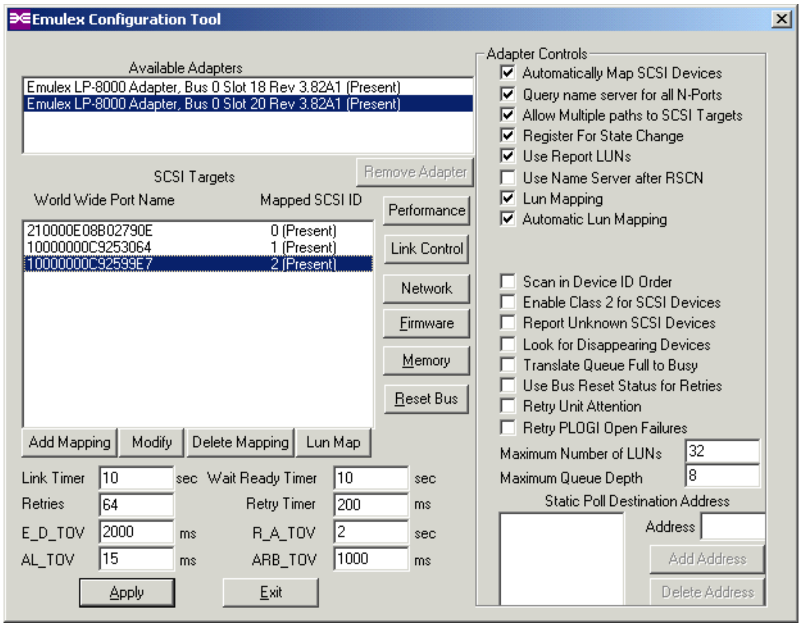
Device World Wide Port Name
10000000C92599E7

ScsiId

2
3
4
5
6
7
8
9
10
11

Figure 9-7 Adding WWPN to SCSI ID mapping - Step 2

You can also review or modify the Fibre Channel LUN to SCSI LUN mapping. To do this select the appropriate adapter from the list of available adapters and the desired SCSI ID from the list of WWPNs to SCSI ID mappings. This is shown in Figure 9-8. Pressing the **Lun Map** button will show a list of currently available LUN mappings, as shown in Figure 9-9. To add a Fibre Channel LUN to SCSI LUN mapping, click **Add** and follow the instructions, similar to adding SCSI ID mapping.



Available Adapters

Emulex LP-8000 Adapter, Bus 0 Slot 18 Rev 3.82A1 (Present)
Emulex LP-8000 Adapter, Bus 0 Slot 20 Rev 3.82A1 (Present)

SCSI Targets

World Wide Port Name	Mapped SCSI ID
210000E08B02790E	0 (Present)
10000000C9253064	1 (Present)
10000000C92599E7	2 (Present)

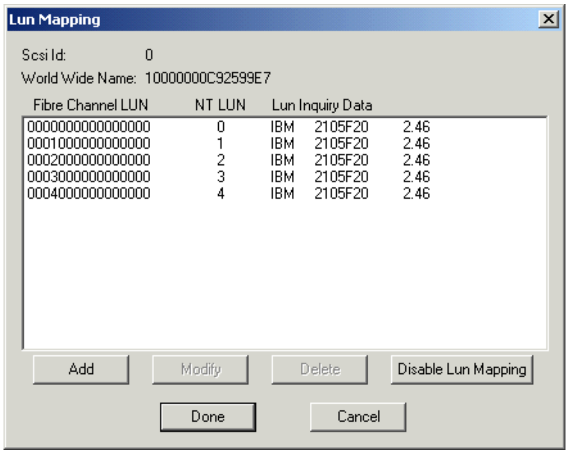
Link Timer 10 sec Wait Ready Timer 10 sec
Retries 64 Retry Timer 200 ms
E_D_TOV 2000 ms R_A_TOV 2 sec
AL_TOV 15 ms ARB_TOV 1000 ms

Adapter Controls

- ☑ Automatically Map SCSI Devices
- ☑ Query name server for all N-Ports
- ☑ Allow Multiple paths to SCSI Targets
- ☑ Register For State Change
- ☑ Use Report LUNs
- ☐ Use Name Server after RSCN
- ☑ Lun Mapping
- ☑ Automatic Lun Mapping
- ☐ Scan in Device ID Order
- ☐ Enable Class 2 for SCSI Devices
- ☐ Report Unknown SCSI Devices
- ☐ Look for Disappearing Devices
- ☐ Translate Queue Full to Busy
- ☐ Use Bus Reset Status for Retries
- ☐ Retry Unit Attention
- ☐ Retry PLOGI Open Failures

Maximum Number of LUNs 32
Maximum Queue Depth 8
Static Poll Destination Address
Address

Figure 9-8 Reviewing or adding Fibre Channel LUN to SCSI LUN mapping



Scsi Id: 0
World Wide Name: 10000000C92599E7

Fibre Channel LUN	NT LUN	Lun Inquiry Data		
0000000000000000	0	IBM	2105F20	2.46
0001000000000000	1	IBM	2105F20	2.46
0002000000000000	2	IBM	2105F20	2.46
0003000000000000	3	IBM	2105F20	2.46
0004000000000000	4	IBM	2105F20	2.46

Figure 9-9 Example of list of LUN mapping

9.5 A step for Installing Windows 2000 on remote ESS disks

Before we can proceed with a new installation of Windows 2000 system on remote disks located within an IBM Enterprise Storage Server, you have to make sure that proper device drivers for adapters installed in the host system are available on floppy disks. During the first phase of system installation you are prompted to press the F6 key to load an additional device driver, which may be required to access remote disks connected to SCSI or Fibre Channel adapters. A message will appear on the bottom line of the window for a few seconds:

```
Press F6 if you need to install a third-party SCSI or RAID driver
```

It is possible, that the installation image of the Windows 2000 operating system already contains appropriate adapter device drivers. If we do not install additional drivers and the installation of Windows 2000 quits with a message, that no disk are available for installation, you should run the system installation once again and load appropriate adapter device driver from a floppy diskette (pressing F6 when prompted).

9.6 Availability and recoverability for Windows2000

In this section we describe how to ensure optimum availability and recoverability when you attach an IBM ESS to a Windows 2000 host system. You must set the timeout value associated with the supported host bus adapters to 240 seconds. The setting is consistent with the configuration for IBM SSA adapters and disk subsystems when attached to Windows 2000 host system.

The host bus adapter uses the timeout parameter to bound its recovery actions and responses to the disk subsystem. The value exists in different places in the system configuration. You can retrieve and use it in different ways depending on the type of host bus adapter. The following instructions tell you how to modify the value safely in either the Windows 2000 registry or in the device adapter parameters.

Setting the TimeOutValue registry

The following instructions tell you how to set the timeout value registry:

1. From the Run menu or command prompt, type: `regedit.exe`
2. Navigate to the following registry key:
   `HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Disk`
3. Look for the value called `TimeOutValue`. If the value called `TimeOutValue` does not exist, follow the instructions in this step to add it to the Windows 2000 registry. If the value called `TimeOutValue` exists, go to step 6.
4. Click **Edit -> New -> DWORD Value**
5. For Name, type: `TimeOutValue`
6. If the value exists (or was added in step 3) and is less than 0x000000f0 (240 decimal), perform the following steps to increase it to 0xf0:
   a. Click **Edit -> Modify**
   b. Choose base: **Hexadecimal** (or **Decimal** if you want to enter a decimal value)
   c. For Value data type `f0` (or `240` if you choose a decimal base)

   > **Important:** Do not perform steps “b” and “c” in reversed order. Once a value for data is entered and you then change the base, the whole value will be recalculated. This may cause incorrect values to be stored in the Windows 2000 registry.

   d. Click **OK**.
   e. Exit the `regedit` program.
   f. Restart your Windows 2000 host system for the changes to take effect.

9.7 Installing IBM SDD on Windows 2000 host

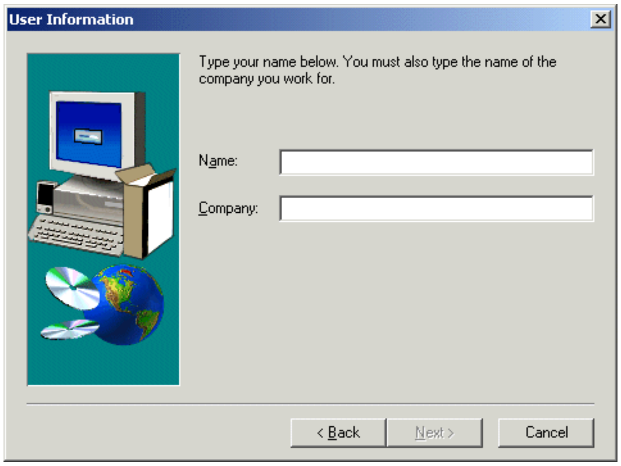
The process for the installation of IBM SDD in Windows 2000 environment is very simple. Before continuing with SDD installation, assume that all hardware and software requirements described earlier in this chapter are met.

To install IBM SDD you must have an SDD installation CD-ROM or a floppy diskette. You can also download the latest version of IBM SDD from the Internet:
http://ssddom01.storage.ibm.com/techsup/swtechsup.nsf/support/sddupdates

> **Important:** To install IBM SDD on your Windows 2000 operating system, you must log on as system administrator.

Perform the following steps to install the SDD device driver and application programs on your system:

1. Insert the SDD installation CD-ROM or floppy diskette into the selected drive and start the Windows 2000 Explorer program.
2. If you are installing IBM SDD from a CD disk, select the appropriate CD-ROM drive the `\win2k\IBMSdd` directory.
3. If you have previously downloaded the latest version of IBM SDD from the Internet, it is your responsibility to uncompress it to a desired directory.
4. Run the `setup.exe` program. The Install shield starts. A Welcome window will appear. Click **Next**.
5. The Software Licensing Agreement panel is displayed. If you accept the terms and conditions of product license, click **Yes**. The user information panel shown in Figure 9-10 is displayed. Type your name and your company name and click **Next**. You will not be able to proceed to the next window without entering the requested information.



User Information

Type your name below. You must also type the name of the company you work for.

Name:

Company:

Figure 9-10 SDD installation - user information

6. Choose Destination Location panel (shown in Figure 9-11) is displayed. Now you are able to select desired destination location; however, the default location is a good and reasonable choice:

   `C:\Program Files\IBM Corp.\Subsystem Device Driver`

   After finishing click **Next**.

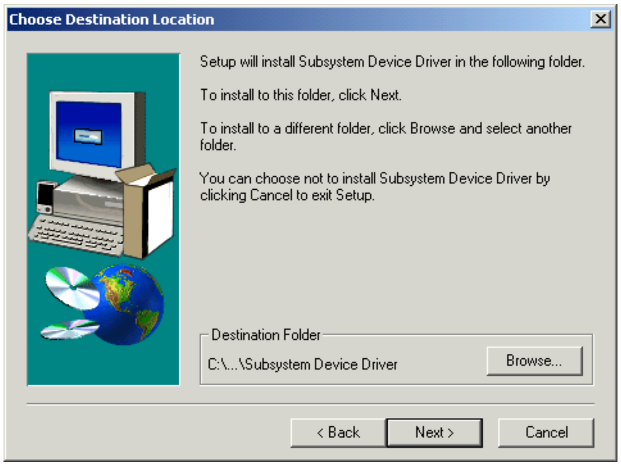
7. The Setup panel is displayed (as shown in Figure 9-12). Select the type of setup you prefer from the following setup choices:

   a. `Typical` - selects all options.

   b. `Compact` - selects the minimum required options only (the installation driver and README file).

   c. `Custom` - lets you to select manually options that you need.

   IBM recommends that you select `Typical`. The SDD installation does not consume large disk space (it uses only about 2 MB) so there is no reason not to select `Typical` installation. After finishing click **Next**.



Setup will install Subsystem Device Driver in the following folder.

To install to this folder, click Next.

To install to a different folder, click Browse and select another folder.

You can choose not to install Subsystem Device Driver by clicking Cancel to exit Setup.

Destination Folder: C:\...\Subsystem Device Driver

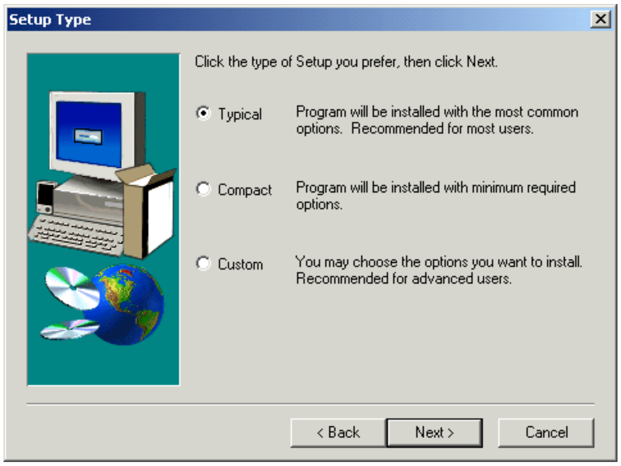
Figure 9-11 SDD installation - Choose Destination Location

8. The Setup Complete window is displayed. Click **Finish**. The SDD program prompts you to start your computer again. Click **Yes** to start your computer again.

9. When you log on again, you see a Subsystem Device Driver entry in your Program menu containing the following items:

   – Subsystem Device Driver Management
   – Subsystem Device Driver Manual
   – Readme

Note: You can verify that SDD has been successfully installed by issuing the `datapath query device` command. If the command executes, SDD is installed properly.



Click the type of Setup you prefer, then click Next.

● Typical — Program will be installed with the most common options. Recommended for most users.

○ Compact — Program will be installed with minimum required options.

○ Custom — You may choose the options you want to install. Recommended for advanced users.

Figure 9-12 SDD installation - setup type

When you finish installing IBM SDD and reboot the host system, you can plug-in additional SCSI or Fibre Channel cables to physically set-up the multipathing environment. When all cables are connected and all paths to disk devices are available, rescan your disks to configure all devices. To do this right-click on My Computer icon, then choose **Manage** -> **Device Manager**. Double-click on **Disk drives**. Choose **Action** -> **Scan for hardware changes**. After scanning is done, a window similar to that as shown in Figure 9-13 should appear.

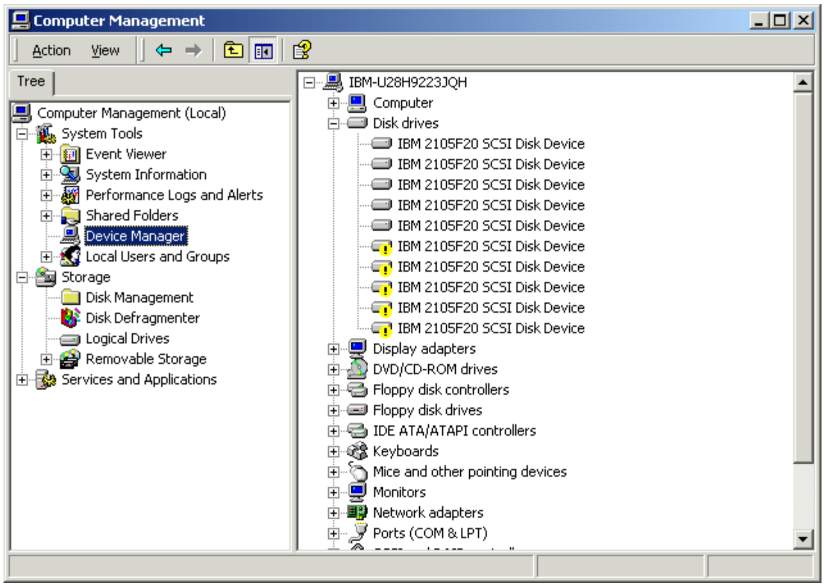

Computer Management
Action
View
Tree
Computer Management (Local)
System Tools
Event Viewer
System Information
Performance Logs and Alerts
Shared Folders
Device Manager
Local Users and Groups
Storage
Disk Management
Disk Defragmenter
Logical Drives
Removable Storage
Services and Applications
IBM-U28H9223JQH
Computer
Disk drives
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
IBM 2105F20 SCSI Disk Device
Display adapters
DVD/CD-ROM drives
Floppy disk controllers
Floppy disk drives
IDE ATA/ATAPI controllers
Keyboards
Mice and other pointing devices
Monitors
Network adapters
Ports (COM & LPT)

Figure 9-13 Disk devices status with SDD installed

Tip: If your host system is not plug-and-play compliant, you should restart your system again instead of rescanning the hardware, when all paths to disk devices are available.

All disks marked with “!” are multiplied paths to disk devices and will not be used by Windows 2000 Disk Manager. The presence of disks marked with “!” is fully correct with IBM SDD installed. Only SDD will use them to balance the load and failover the path in case of another path disaster. In the example shown in Figure 9-13, Windows 2000 Disk Manager will see and use only five logical disks, preventing the operating system from use other five disk as different devices.

9.8 Uninstalling or upgrading IBM SDD

To uninstall IBM SDD from your operating system perform the following steps:

1. Log on as the administrator user.
2. Close any applications and any system services which are accessing disk devices.

   Attention: This step must be carried out with caution. After the deinstallation of SDD, if there are multiple paths to a LUN, the LUN will be seen multiple times by the host. Each LUN that the host sees will be treated as a separate device. This can cause data corruption.

3. Click **Start** -> **Settings** -> **Control Panel**. The Control Panel window appear.
4. Select **Add/Remove Programs** in Control Panel. The Add/Remove Programs window opens.
5. In the Add/Remove Programs window, select the **Subsystem Device Driver** from the Currently installed programs selection list. Click **Change/Remove**.

Attention: After uninstalling the IBM SDD from the operating system, you *must immediately* install the new version of SDD or remove multiple paths to avoid any potential data loss.

To upgrade IBM SDD, you must first uninstall the previous version of SDD and then install the new version, as described in 9.7, “Installing IBM SDD on Windows 2000 host” on page 204.

9.9 Displaying the current version of the SDD

You can display the current version of SDD on a Windows 2000 host system by viewing the sddpath.sys file properties. Perform the following steps to view the properties of sddpath.sys file:

1. Click **Start** -> **Run** -> **Programs** -> **Accessories** -> **Windows Explorer** to open Windows Explorer.
2. In Windows Explorer, go to Winnt\system32\drivers directory.
3. Right-click on the sddpath.sys file and then click **Properties**. The sddpath.sys properties window opens.
4. In the sddpath.sys properties window, click the **Version** panel. The file version and copyright information about sddpath.sys displays.

9.10 Managing and troubleshooting IBM SDD on Windows 2000

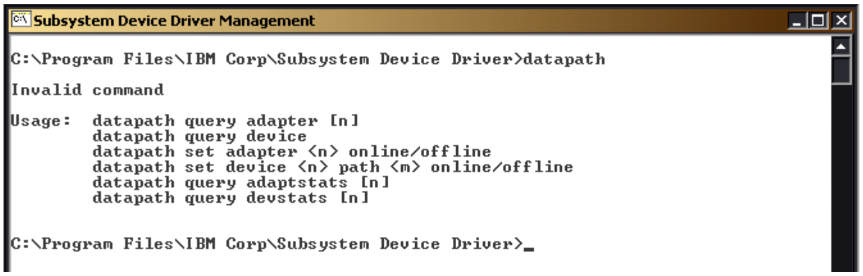
All software tools provided with IBM Subsystem Device Driver are located under installation directory. The default location is C:\Program Files\IBM Corp.\Subsystem Device Driver, unless changed during installation.

You can launch Subsystem Device Driver Management window by clicking on **Start** -> **Programs** -> **Subsystem Device Driver** -> **Subsystem Device Driver Management**. This will invoke a shell window, where you can run all tools which are available for managing and troubleshooting IBM SDD. These tools are:

- `datapath.exe` command
- `pathtest.exe` command

Figure 9-14 shows the usage of IBM SDD `datapath` command. The following options are available:

- `datapath query adapter [n]` — shows the status and basic statistic of all adapters or n-th adapter installed in system hosts.
- `datapath query adaptstats [n]` — shows more detailed statistics of all adapters or n-th adapter in the system.
- `datapath query device` [n] — shows the status and basic statistic of all disk devices or n-th disk device installed in system host.
- `datapath query devstats [n]` — shows more detailed statistics of all disk devices or n-th disk in the system.
- `datapath set adapter <n> online/offline` — forces to set manually status of n-th adapter in the system to online or offline.
- `datapath set device <n> path <m> online/offline` — forces to set manually status of m-th path to n-th disk device to online or offline.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath

Invalid command

Usage:  datapath query adapter [n]
        datapath query device
        datapath set adapter <n> online/offline
        datapath set device <n> path <m> online/offline
        datapath query adaptstats [n]
        datapath query devstats [n]

C:\Program Files\IBM Corp\Subsystem Device Driver>_
```

Figure 9-14 Usage of IBM SDD datapath command

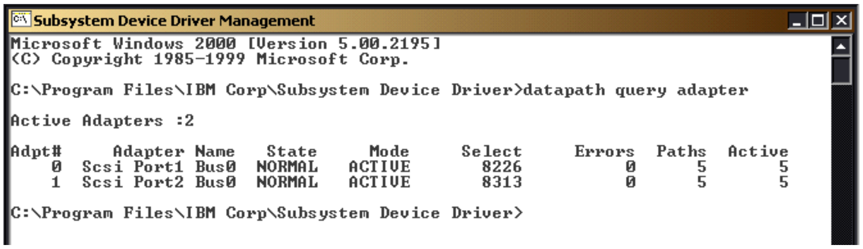
9.10.1 Use of datapath query adapter command

The `datapath query adapter` command displays basic information about the status of all adapters or a single adapter in a host system. Returned information also contains basic statistic information. The syntax for that command is shown in Example 9-1.

Example 9-1 Syntax for datapath query adapter command

```
datapath query adapter [adapter_number]

   Parameters:
      adapter_number - the number of adapter for which you want the information to be
      displayed. If you do not enter an adapter number, information about all adapters is
      displayed.
```

Figure 9-15 shows an example of `datapath query adapter` command output.



```
Microsoft Windows 2000 [Version 5.00.2195]
(C) Copyright 1985-1999 Microsoft Corp.

C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adapter

Active Adapters :2

Adpt#     Adapter Name   State     Mode      Select     Errors  Paths   Active
    0   Scsi Port1 Bus0  NORMAL   ACTIVE      8226          0      5        5
    1   Scsi Port2 Bus0  NORMAL   ACTIVE      8313          0      5        5

C:\Program Files\IBM Corp\Subsystem Device Driver>
```

Figure 9-15 Example of datapath query adapter command output

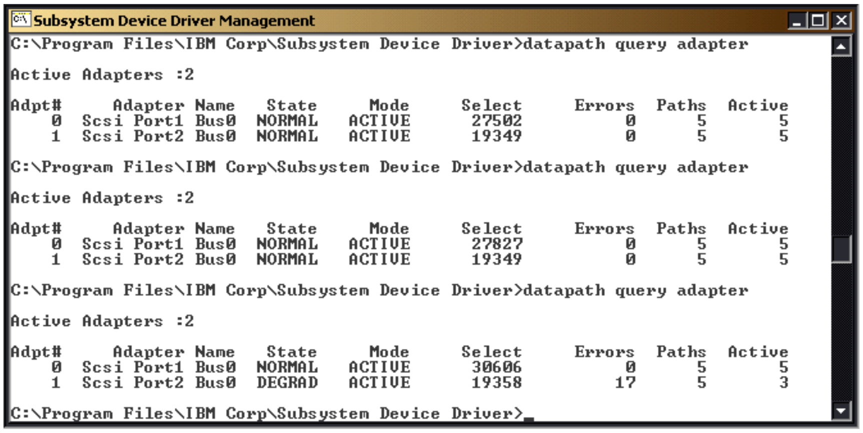
As you can see in this example, the host system is equipped with two adapters, both of them working properly. The meaning of the individual columns are as follows:

- `Adpt#` — the number of the adapter in the system.
- `Adapter Name` — the name of the adapter.
- `State` — the condition of the named adapter. It can be either:
  - `Normal` — adapter is in use.
  - `Degraded` — one or more paths are not functioning.
  - `Failed` — the adapter is no longer being used by Subsystem Device Driver.
- `Mode` — the mode of the named adapter, which is either Active or Offline.
- `Select` — the number of times this adapter was selected for input or output.
- `Errors` — the number of errors on all paths that are attached to this adapter.
- `Paths` — the number of paths that are attached to this adapter.

> **Note:** In the Windows NT host system, this is the number of physical and logical devices that are attached to this adapter.

- Active — the number of functional paths that are attached to this adapter. The number of functional paths is equal to the number of paths minus any that are identified as failed or offline.

Figure 9-16 shows an example of a degraded adapter.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adapter

Active Adapters :2

Adpt#     Adapter Name   State     Mode      Select     Errors  Paths   Active
    0   Scsi Port1 Bus0  NORMAL   ACTIVE     27502          0      5        5
    1   Scsi Port2 Bus0  NORMAL   ACTIVE     19349          0      5        5

C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adapter

Active Adapters :2

Adpt#     Adapter Name   State     Mode      Select     Errors  Paths   Active
    0   Scsi Port1 Bus0  NORMAL   ACTIVE     27827          0      5        5
    1   Scsi Port2 Bus0  NORMAL   ACTIVE     19349          0      5        5

C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adapter

Active Adapters :2

Adpt#     Adapter Name   State     Mode      Select     Errors  Paths   Active
    0   Scsi Port1 Bus0  NORMAL   ACTIVE     30606          0      5        5
    1   Scsi Port2 Bus0  DEGRAD   ACTIVE     19358         17      5        3

C:\Program Files\IBM Corp\Subsystem Device Driver>_
```

Figure 9-16 Example of degraded adapter

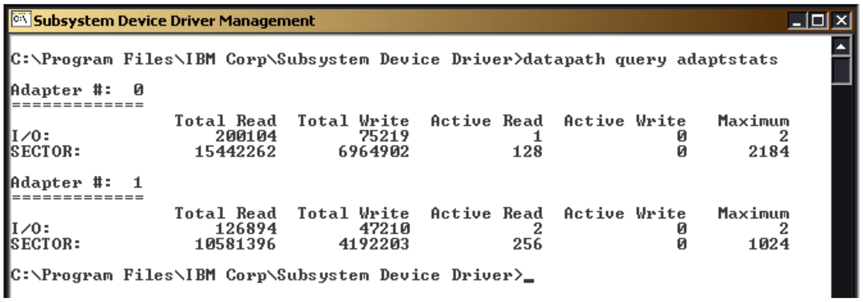
9.10.2 Use of datapath query adaptstats command

The `datapath query adaptstats` command displays performance information for all SCSI and Fibre Channel adapters that are attached to Subsystem Device Driver devices. The syntax for that command is shown in Example 9-2.

Example 9-2 Syntax for datapath query adaptstats command

```
datapath query adaptstats [adapter_number]

   Parameters:
      adapter_number - the number of adapter for which you want the information to be
      displayed. If you do not enter an adapter number, information about all adapters is
      displayed.
```

Figure 9-17 shows an example of `datapath query adaptstats` command output.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adaptstats

Adapter #:  0
=============
                Total Read  Total Write  Active Read  Active Write   Maximum
I/O:                200104        75219            1             0         2
SECTOR:           15442262      6964902          128             0      2184

Adapter #:  1
=============
                Total Read  Total Write  Active Read  Active Write   Maximum
I/O:                126894        47210            2             0         2
SECTOR:           10581396      4192203          256             0      1024

C:\Program Files\IBM Corp\Subsystem Device Driver>_
```

Figure 9-17 Example of datapath query adaptstats command output

The meaning of the individual columns are as follows:

- `Total Read`
  - `I/O` — total number of completed read requests
  - `SECTOR` — total number of sectors that have been read
- `Total Write`
  - `I/O` — total number of completed write requests
  - `SECTOR` — total number of sectors that have been written
- `Active Read`
  - `I/O` — total number of read requests in process
  - `SECTOR` — total number of sectors to read in process
- `Active Write`
  - `I/O` — total number of write requests in process
  - `SECTOR` — total number of sectors to write in process
- `Maximum`
  - `I/O` — the maximum number of queued I/O requests
  - `SECTOR` — the maximum number of queued sectors to read/write

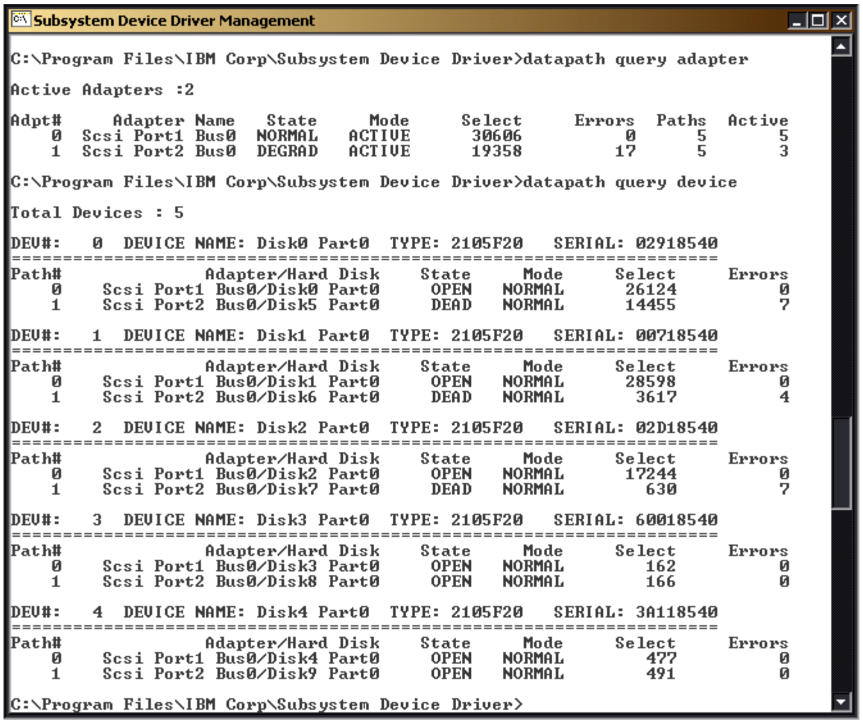
9.10.3 Use of datapath query device command

The `datapath query device` command displays basic information about the status of all disk devices or a single disk device that are under control of IBM Subsystem Device Driver. Returned information also contains basic statistic information. The syntax for that command is as shown in Example 9-3.

Example 9-3 Syntax for datapath query device command

```
datapath query device [device_number]

   Parameters:
      device_number - the number of device for which you want the information to be
      displayed. If you do not enter a device number, information about all devices is
      displayed.
```

Figure 9-18 shows an example of `datapath query device` command output.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query adapter

Active Adapters :2

Adpt#     Adapter Name   State     Mode      Select     Errors  Paths  Active
    0   Scsi Port1 Bus0  NORMAL   ACTIVE      30606          0      5       5
    1   Scsi Port2 Bus0  DEGRAD   ACTIVE      19358         17      5       3

C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query device

Total Devices : 5

DEV#:   0  DEVICE NAME: Disk0 Part0  TYPE: 2105F20   SERIAL: 02918540
===================================================================
Path#             Adapter/Hard Disk     State     Mode      Select     Errors
    0     Scsi Port1 Bus0/Disk0 Part0    OPEN   NORMAL       26124          0
    1     Scsi Port2 Bus0/Disk5 Part0    DEAD   NORMAL       14455          7

DEV#:   1  DEVICE NAME: Disk1 Part0  TYPE: 2105F20   SERIAL: 00718540
===================================================================
Path#             Adapter/Hard Disk     State     Mode      Select     Errors
    0     Scsi Port1 Bus0/Disk1 Part0    OPEN   NORMAL       28598          0
    1     Scsi Port2 Bus0/Disk6 Part0    DEAD   NORMAL        3617          4

DEV#:   2  DEVICE NAME: Disk2 Part0  TYPE: 2105F20   SERIAL: 02D18540
===================================================================
Path#             Adapter/Hard Disk     State     Mode      Select     Errors
    0     Scsi Port1 Bus0/Disk2 Part0    OPEN   NORMAL       17244          0
    1     Scsi Port2 Bus0/Disk7 Part0    DEAD   NORMAL         630          7

DEV#:   3  DEVICE NAME: Disk3 Part0  TYPE: 2105F20   SERIAL: 60018540
===================================================================
Path#             Adapter/Hard Disk     State     Mode      Select     Errors
    0     Scsi Port1 Bus0/Disk3 Part0    OPEN   NORMAL         162          0
    1     Scsi Port2 Bus0/Disk8 Part0    OPEN   NORMAL         166          0

DEV#:   4  DEVICE NAME: Disk4 Part0  TYPE: 2105F20   SERIAL: 3A118540
===================================================================
Path#             Adapter/Hard Disk     State     Mode      Select     Errors
    0     Scsi Port1 Bus0/Disk4 Part0    OPEN   NORMAL         477          0
    1     Scsi Port2 Bus0/Disk9 Part0    OPEN   NORMAL         491          0

C:\Program Files\IBM Corp\Subsystem Device Driver>
```

Figure 9-18 Example of datapath query device status

As you can see in this figure, some of the paths to devices are in *Dead* state (for devices number 0, 1 and 2), while others are still in *Open* state (devices 3 and 4). This may be caused by two different factors:

1. The only paths that are in Dead state are #1 paths for devices #0, #1 and #2. It is highly possible, that all #1 paths go through the same adapter and the adapter or link failed. Devices #3 and #4 are still in Open state, because there is no I/O traffic for them (see the chapter 3.1.1, “Path algorithms” on page 27 for details).
2. Devices #0, #1 and #2 are assigned to a single HBA within an IBM ESS and that HBA failed, while devices #3 and #4 are assigned to more than one HBA or to a different HBA.

The meaning of the individual columns are as follows:

- `Dev#` — the number of this device
- `Name` — the name of this device
- `Type` — the device product ID from inquiry data
- `Serial` — the logical unit number (LUN) for this device
- `Path` — the path number
- `Adapter` — the name of the adapter to which the path is attached

- `Hard Disk` — the name of the logical device to which the path is bound
- `State` — the condition of the named device:
  - `Open` — path is in use
  - `Close` — path is not being used
  - `Dead` — path is no longer being used. It was either removed by the IBM SDD due to errors or manually removed using the `datapath set device n path m offline` or `datapath set adapter n offline` command.
  - `Invalid` — path verification failed. The path was not opened.
- `Mode` — the mode of the named device. It is either Normal or Offline.
- `Select` — the number of times this path was selected for input or output
- `Errors` — the number of errors on a path that is attached to this device

> **Note:** Usually, the device number and the device index number are the same. However, if the devices are configured out of order, the two numbers are not always consistent. To find the corresponding index number for a specific device, you should always run the `datapath query device` command first.

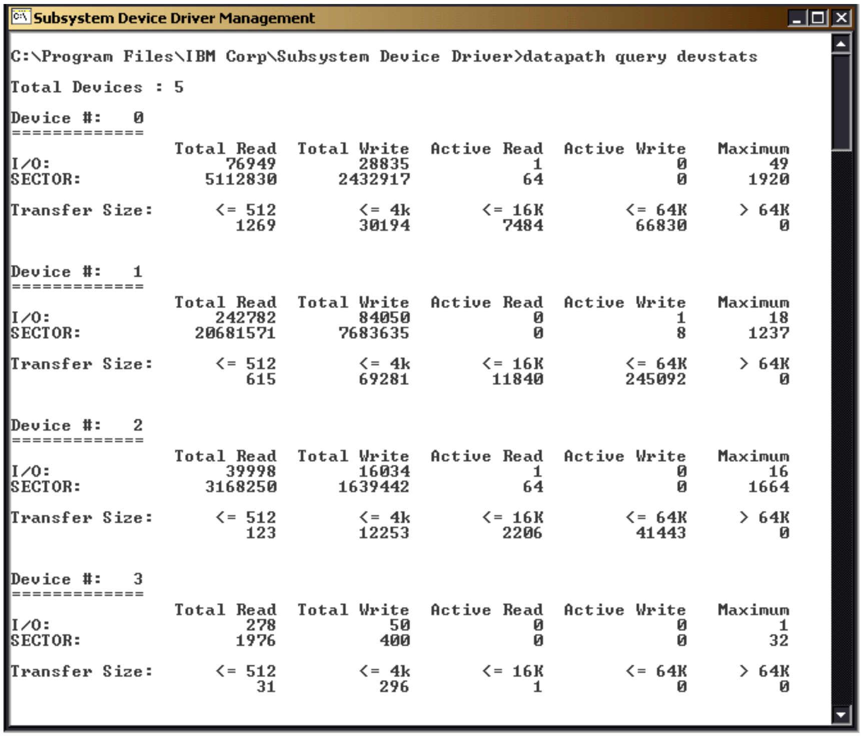
9.10.4 Use of datapath query devstats command

The `datapath query devstats` command displays the performance information for all disk devices or a single disk device that are under control of IBM Subsystem Device Driver. The syntax for that command is as shown in Example 9-4.

Example 9-4 Syntax for datapath query devstats command

```
datapath query devstats [device_number]

   Parameters:
      device_number - the number of device for which you want the information to be
      displayed. If you do not enter a device number, information about all devices is
      displayed.
```

Figure 9-19 shows an example of `datapath query devstats` command output.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query devstats

Total Devices : 5

Device #:   0
=============
                Total Read  Total Write  Active Read  Active Write    Maximum
I/O:                 76949        28835            1             0         49
SECTOR:            5112830      2432917           64             0       1920

Transfer Size:      <= 512        <= 4k       <= 16K        <= 64K      > 64K
                      1269        30194         7484         66830          0


Device #:   1
=============
                Total Read  Total Write  Active Read  Active Write    Maximum
I/O:                242782        84050            0             1         18
SECTOR:           20681571      7683635            0             8       1237

Transfer Size:      <= 512        <= 4k       <= 16K        <= 64K      > 64K
                       615        69281        11840        245092          0


Device #:   2
=============
                Total Read  Total Write  Active Read  Active Write    Maximum
I/O:                 39998        16034            1             0         16
SECTOR:            3168250      1639442           64             0       1664

Transfer Size:      <= 512        <= 4k       <= 16K        <= 64K      > 64K
                       123        12253         2206         41443          0


Device #:   3
=============
                Total Read  Total Write  Active Read  Active Write    Maximum
I/O:                   278           50            0             0          1
SECTOR:               1976          400            0             0         32

Transfer Size:      <= 512        <= 4k       <= 16K        <= 64K      > 64K
                        31          296            1             0          0
```

Figure 9-19   Example of datapath query devstats command output

The meaning of the individual columns are as follows:

- `Total Read`
  - `I/O` — total number of completed read requests
  - `SECTOR` — total number of sectors that have been read
- `Total Write`
  - `I/O` — total number of completed write requests
  - `SECTOR` — total number of sectors that have been written
- `Active Read`
  - `I/O` — total number of read requests in process
  - `SECTOR` — total number of sectors to read in process
- `Active Write`
  - `I/O` — total number of write requests in process
  - `SECTOR` — total number of sectors to write in process
- `Maximum`
  - `I/O` — the maximum number of queued I/O requests
  - `SECTOR` — the maximum number of queued sectors to read/write
- `Transfer size`

  <= 512: The number of I/O requests received, whose transfer size is 512 bytes or less

  <= 4K: The number of I/O requests received, whose transfer size is 4 KB or less, but greater then 512 bytes

  <= 16K: The number of I/O requests received, whose transfer size is 16 KB or less, but greater then 4 KB

<= 64K: The number of I/O requests received, whose transfer size is 64 KB or less, but greater then 16 KB

> 64K: The number of I/O requests received, whose transfer size is greater than 64 KB

9.10.5 Use of datapath set adapter command

The `datapath set adapter` command sets all device paths attached to the adapter either to `Online` or `Offline` state. The syntax for that command is shown in Example 9-5.

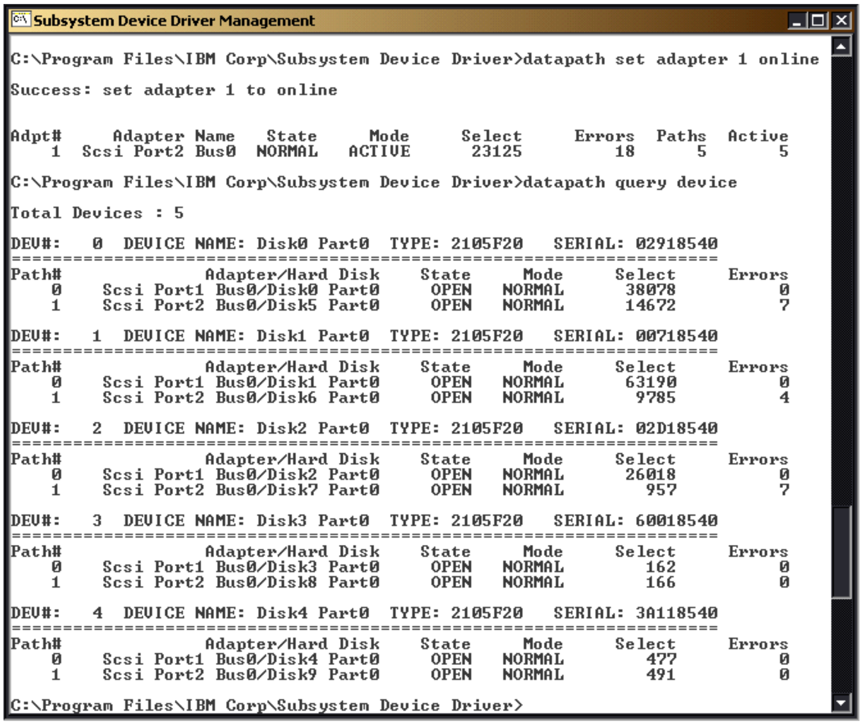
Example 9-5 Syntax for datapath set adapter command

```
datapath set adapter adapter_number online/offline

   Parameters:
      adapter_number - the number of the adapter for which you want to change the status,
      online - sets the adapter online,
      offline - sets the adapter offline
```

Restrictions: The following restrictions apply when issuing `datapath set adapter` command (see 3.1.1, “Path algorithms” on page 27 for details):

- This command will not remove the last path to a device.
- The `datapath set adapter offline` command fails if there is any device having the last path attached to this adapter.
- This command can be issued even when the devices are closed.
- If all paths are attached to a single Fibre Channel adapter that connects to multiple ESS ports through a switch, the `datapath set adapter 0 offline` command fails and all the paths are not set offline.

Figure 9-20 shows an example of usage `datapath set adapter` command. Assume, that before issuing the command, the status of adapter #1 and its corresponding paths to disk devices is shown in Figure 9-18. As you can see, after issuing the command all paths to disk devices are set to *Open* state and adapter state is changed from *Degraded* to `Normal`.



```
C:\Program Files\IBM Corp\Subsystem Device Driver>datapath set adapter 1 online

Success: set adapter 1 to online


Adpt#     Adapter Name   State     Mode     Select     Errors  Paths  Active
    1  Scsi Port2 Bus0  NORMAL   ACTIVE      23125         18      5       5

C:\Program Files\IBM Corp\Subsystem Device Driver>datapath query device

Total Devices : 5

DEV#:   0  DEVICE NAME: Disk0 Part0  TYPE: 2105F20   SERIAL: 02918540
=====================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk0 Part0     OPEN   NORMAL      38078          0
    1    Scsi Port2 Bus0/Disk5 Part0     OPEN   NORMAL      14672          7

DEV#:   1  DEVICE NAME: Disk1 Part0  TYPE: 2105F20   SERIAL: 00718540
=====================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk1 Part0     OPEN   NORMAL      63190          0
    1    Scsi Port2 Bus0/Disk6 Part0     OPEN   NORMAL       9785          4

DEV#:   2  DEVICE NAME: Disk2 Part0  TYPE: 2105F20   SERIAL: 02D18540
=====================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk2 Part0     OPEN   NORMAL      26018          0
    1    Scsi Port2 Bus0/Disk7 Part0     OPEN   NORMAL        957          7

DEV#:   3  DEVICE NAME: Disk3 Part0  TYPE: 2105F20   SERIAL: 60018540
=====================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk3 Part0     OPEN   NORMAL        162          0
    1    Scsi Port2 Bus0/Disk8 Part0     OPEN   NORMAL        166          0

DEV#:   4  DEVICE NAME: Disk4 Part0  TYPE: 2105F20   SERIAL: 3A118540
=====================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0    Scsi Port1 Bus0/Disk4 Part0     OPEN   NORMAL        477          0
    1    Scsi Port2 Bus0/Disk9 Part0     OPEN   NORMAL        491          0

C:\Program Files\IBM Corp\Subsystem Device Driver>
```

Figure 9-20 Example of datapath set adapter command

9.10.6 Use of datapath set device command

The `datapath set device` command sets the path to the device either to *Online* or *Offline* state. The syntax for that command is shown in Example 9-6.

Example 9-6 Syntax for datapath set device command

```
datapath set device device_number path path_number online/offline

   Parameters:
      device_number - the index number for a device for which you want to change the
      status,
      path_number - the number of the path to that device, for which you want to change
      the status
      online - sets the path online,
      offline - sets the path offline
```

Restrictions: The following restrictions apply when issuing `datapath set device` command (see 3.1.1, “Path algorithms” on page 27 for details):

- This command will not remove the last path to a device.
- This command can be issued even when the devices are closed.

In the case presented in Figure 9-18 on page 211, issuing the command `datapath set device 2 path 1 online` will change the status of path #1 for device #2 to Open.

9.10.7 Adding additional paths to Subsystem Device Driver devices

Path topology in Windows 2000 operating system is built automatically when the system starts up. To add a new path to a device perform the following steps:

1. Set up the new path to a device (reconfigure your ESS LUN to HBA assignments, install any additional hardware in your hosts system or the ESS, reconfigure your SAN switch settings, and so on).
2. Restart the Windows 2000 server.
3. Verify, that the path is added correctly. To do this run `datapath query device` command.

Attention: Ensure that the Subsystem Device Driver is installed *before* you add additional paths to a device. Otherwise, the same disk seen on separate paths will be treated as separate disk devices. This may cause data inconsistency or loss of access to data existing on the disk. Only one path to a device can be configured without any multipathing software running on the host server.

9.11 Using ESS with Veritas VxVM and DMP

As we mentioned in Chapter 3, “Multipathing software” on page 25, Veritas Volume Manager with its built-in Dynamic MultiPathing is an alternate software which can be used with IBM Enterprise Storage Server 2105. In this section we describe how to install VxVM and use it in multiple path environments to access ESS LUNs.

9.11.1 Installing Veritas VxVm

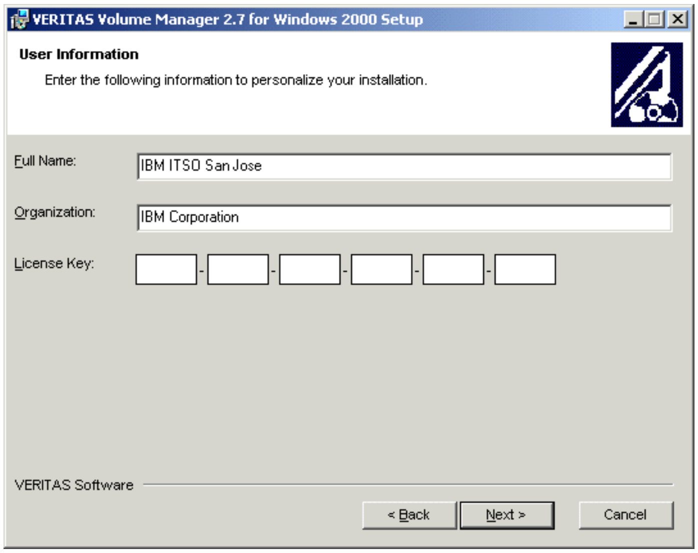
Before continuing with Veritas VxVM installation we assume that all adapters within the host system are configured properly and that required LUNs are configured within the ESS.

Attention: Ensure that *before* you install Veritas VxVM software with its built-in multipathing DMP software, *only one* path to each of the ESS LUNs is configured. Without multipathing software, the same disk seen on separate paths will be treated as separate disk devices. This may cause data inconsistency or loss of access to data existing on the disk. You *must* add additional paths to the LUNs *after* installation of Veritas VxVM is complete.

The easiest way to meet this requirement is to assign the LUNs within the ESS to only one host system HBA and later (after the installation is complete) to add assignments to the other HBA(s).

To install Veritas VxVM you must have the installation CD-ROM. Perform the following steps to install Veritas VxVM on your system:

1. Insert the VxVM installation CD-ROM into the selected drive.
2. The *Welcome to the VERITAS Volume Manager X.X for Windows 2000 Installation Wizard* window appears. The X.X is the version of Veritas VxVM you are installing. Click **Next**.
3. The Select Installation Type window appears. Select **Server** and click **Next**.
4. The User Information window appears, as shown in Figure 9-21. Please fill out all desired fields and click **Next**.



VERITAS Volume Manager 2.7 for Windows 2000 Setup

User Information

Enter the following information to personalize your installation.

Full Name: IBM ITSO San Jose

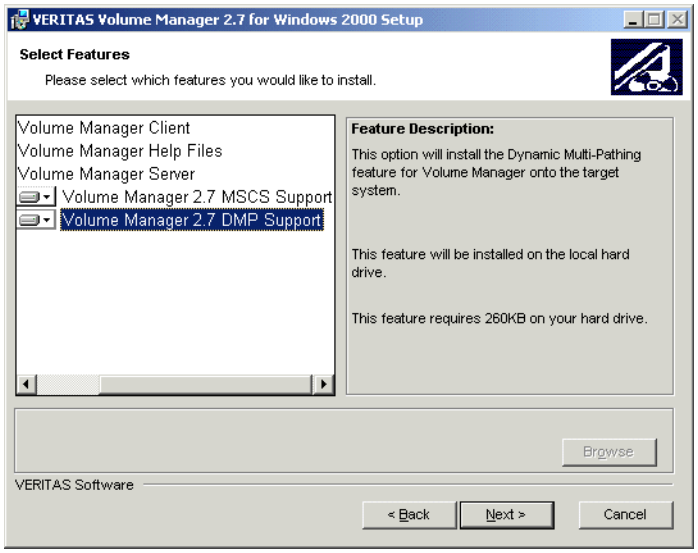
Organization: IBM Corporation

License Key: [ ]-[ ]-[ ]-[ ]-[ ]-[ ]

VERITAS Software

Figure 9-21 Veritas VxVM personalization and licensing window

5. The Select Features window appears as shown in Figure 9-22. Select the desired features and do not forget to select the **Volume Manager X.X DMP Support** feature. Otherwise, you will be unable to use the multipathing feature of the VxVM.



VERITAS Volume Manager 2.7 for Windows 2000 Setup

Select Features

Please select which features you would like to install.

- Volume Manager Client
- Volume Manager Help Files
- Volume Manager Server
- Volume Manager 2.7 MSCS Support
- Volume Manager 2.7 DMP Support

Feature Description:

This option will install the Dynamic Multi-Pathing feature for Volume Manager onto the target system.

This feature will be installed on the local hard drive.

This feature requires 260KB on your hard drive.

VERITAS Software

Figure 9-22 Veritas VxVM features to install

6. Depending on the features you have selected, additional windows may appear and prompt you to enter the cluster name to be remotely administered. This window appears *only* if you have selected VxVM feature Volume Manager X.X MSCS Support.

7. You are ready to begin VxVM installation. The Ready to Install the Application window appears. Click **Next** to begin the installation.

8. The window — VERITAS Volume Manager X.X for Windows 2000 has been successfully installed — appears. Click **Finish** to finish the installation.

> **Note:** Please remember that in order to complete the installation you must restart your host system. The pop-up window displays informing you that you can restart the system now or later.

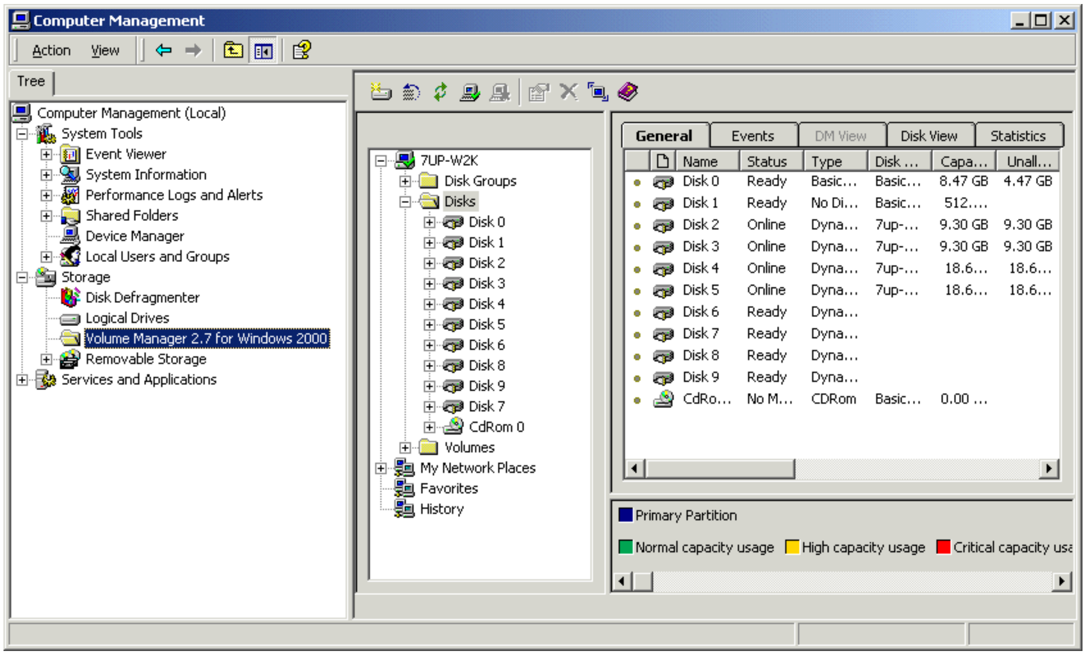
9. After the installation is complete and the system is restarted, the Found New Hardware Wizard window appears informing you that Windows has finished installing the software for VERITAS Volume Manager. You must once again restart the system.

9.11.2 Using Veritas Volume Manager and configuring DMP

With VxVM and DMP software installed you are now able to establish additional paths to the ESS LUN(s). You *must* reconfigure your ESS or perform any other required steps to properly set-up multiple paths to the LUN(s).

When Veritas Volume Manager is installed and operational, it replaces the standard Windows 2000 Disk Management tool. To get to the Veritas Volume Manager right-click on My Computer icon located on your desktop and choose **Manage** -> **Volume Manager X.X for Windows 2000**.

The Dynamic MultiPathing feature of VxVM is disabled by default even though DMP is installed. You have to manually enable multiple paths to the LUN(s). Before you do this, when multiple paths are physically established, the operating system will report the incorrect number of disk devices installed. This is because the same disks seen on separate paths are treated by the operating system as separate disk devices, which results in the number of reported disk devices multiplied by number of paths available to each device. This is shown in Figure 9-23, where disks 2 through 5 are in fact primary paths to the LUNs, while disks 6 through 9 are additional (secondary) paths to the same ESS LUNs.



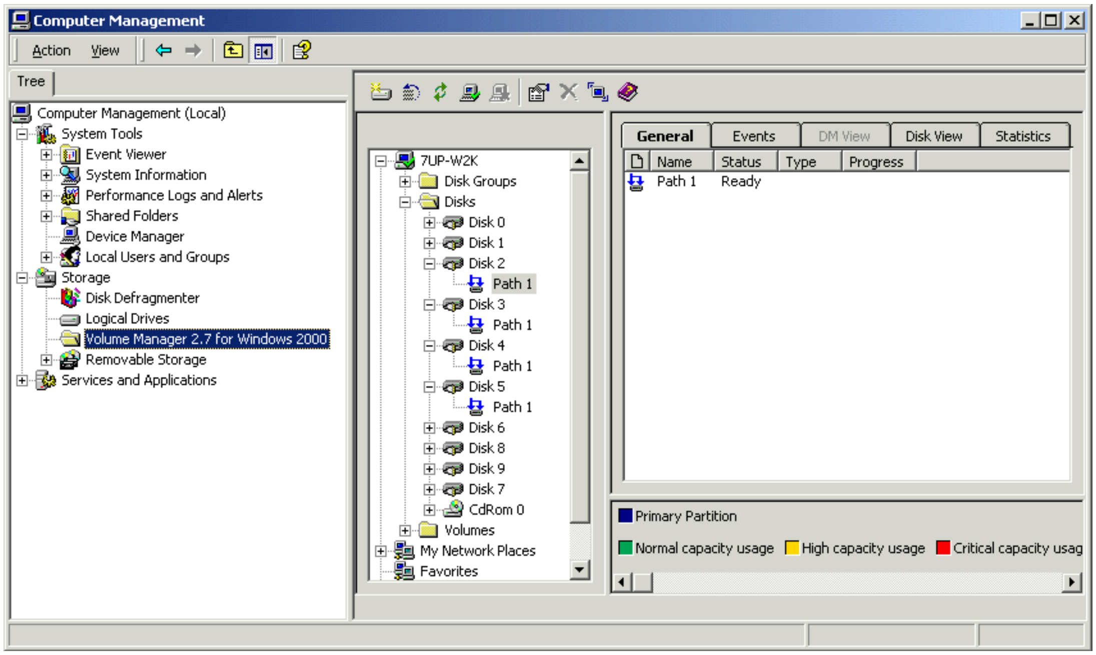
Name	Status	Type	Disk ...	Capa...	Unall...
Disk 0	Ready	Basic...	Basic...	8.47 GB	4.47 GB
Disk 1	Ready	No Di...	Basic...	512....	
Disk 2	Online	Dyna...	7up-...	9.30 GB	9.30 GB
Disk 3	Online	Dyna...	7up-...	9.30 GB	9.30 GB
Disk 4	Online	Dyna...	7up-...	18.6...	18.6...
Disk 5	Online	Dyna...	7up-...	18.6...	18.6...
Disk 6	Ready	Dyna...			
Disk 7	Ready	Dyna...			
Disk 8	Ready	Dyna...			
Disk 9	Ready	Dyna...			
CdRo...	No M...	CDRom	Basic...	0.00 ...	

Figure 9-23 Multiplied disk instances with DMP installed and not enabled

It is required now to enable DMP for all LUNs for which multiple paths are physically established. You do not need to enable DMP for additional paths, but only for primary paths to the LUNs. Once DMP for the specific LUN is enabled, it will automatically be enabled for all additional paths to that LUN.

To enable the DMP feature for a LUN, expand the list of all available disks as shown in Figure 9-23 and select the desired disk. Now expand the list of available paths to that disk as shown in Figure 9-24. There is only one path available, since the DMP feature is disabled by default. Right-click on that path and select **Properties**. The Dynamic Multipathing window appears. Switch to the Array tab and uncheck the **Exclude** box as shown in Figure 9-25. Click **OK**. This will automatically enable the DMP feature for all disks in that array.

> **Important:** You do not need to manually enable DMP for all disk devices with multiple paths established. When you enable DMP for one LUN in the array, multiple paths for all other LUNs in that array will be enabled automatically.
>
> If you select any single path to any LUN and disable it selectively, this will not disable multiple paths to other LUNs in that array.

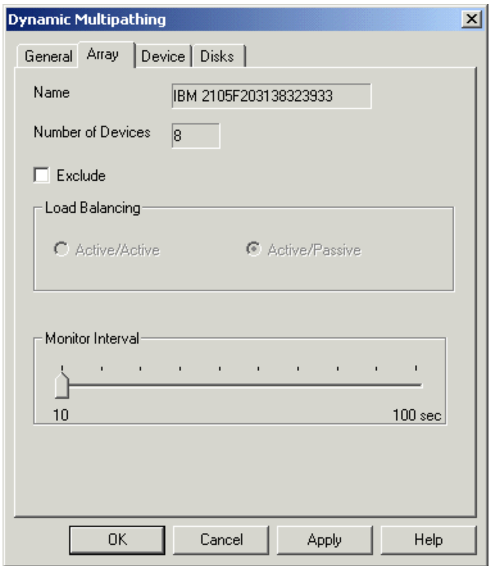


Name	Status	Type	Progress
Path 1	Ready		

Figure 9-24 Enabling DMP with VxVM - Step 1

Please note also, that when DMP is disabled, the load balancing feature of VxVM is automatically disabled. When you enable the DMP feature, it will automatically enable the load balancing. You can check this when you select the path properties for the second time — *after* enabling DMP. The Active/Active configuration should now be enabled. To learn more about Active/Active and Active/Passive configurations of DMP, refer to 3.3, “Veritas VxVM built-in Dynamic MultiPathing software (DMP)” on page 40.

Figure 9-26 shows a properly configured DMP environment. As you can see, with DMP enabled, Veritas Volume Manager (and also the operating system) reports the correct number of disk instances. In the case shown in Figure 9-26, disks 2-5 are DMP disks, while disk 0 and disk 1 are local disks.



Name: IBM 2105F203138323933

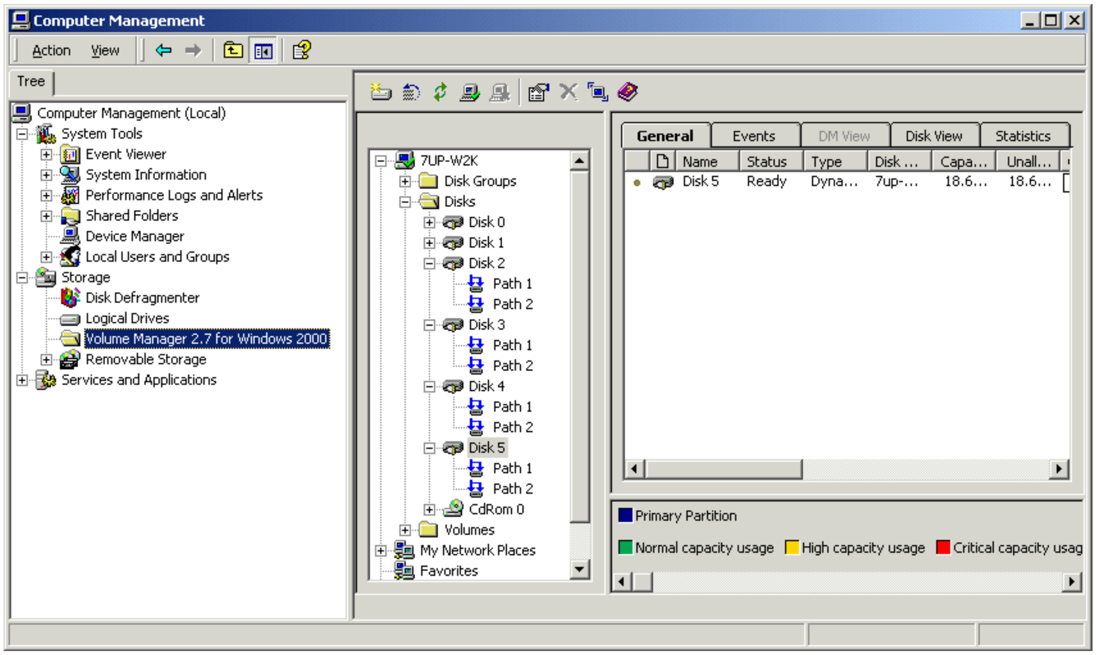
Number of Devices: 8

□ Exclude

Load Balancing: ○ Active/Active ● Active/Passive

Monitor Interval: 10 – 100 sec

Figure 9-25 Enabling DMP with VxVM - Step 2



Name	Status	Type	Disk ...	Capa...	Unall...
Disk 5	Ready	Dyna...	7up-...	18.6...	18.6...

🟦 Primary Partition

🟩 Normal capacity usage 🟨 High capacity usage 🟥 Critical capacity usag

Figure 9-26 Proper configuration of DMP

Installing adapters on a Windows 2000 host

In this appendix we describe how to install different adapters in hosts that are running Windows 2000.

Installing and configuring Adaptec AHA-2944UW adapter card

The steps to install and configure adapter cards shown here are examples that give you an idea of how to configure an Adaptec AHA-2944UW adapter in your host system. Your configuration might be different.

To install and configure the Adaptec AHA-2944UW adapter card in host-system perform the following steps:

1. Install the Adaptec AHA-2944UW in the server.
2. Connect the cable to the ESS port.
3. Start the server.
4. Press `Ctrl+A` to get to the SCSISelect menu and the list of adapter cards to configure.
5. From the SCSISelect menu, select **Configure/View Host Adapter Settings**.
    a. Set the parameters on the Configure/View Host Adapter Settings panel as follows:
        i. Host Adapter SCSI ID: **7**
        ii. SCSI Parity Checking: **Enabled**
        iii. Host Adapter SCSI Termination: **Automatic**
6. Select **SCSI Device Configuration**.
    a. Set the parameters on the SCSI Device Configuration panel as follows:
        i. Sync Transfer Rate (megabytes per second): **40.0**
        ii. Initiate Wide Negotiation: **Yes**
        iii. Enable Disconnection: **Yes**
        iv. Send Start Unit Command: **No**
        v. Enable Write Back Cache: **No**
        vi. BIOS Multiple LUN Support: **Yes**
        vii. Include in BIOS Scan: **Yes**
7. Select **Advanced Configuration Options**.
    a. Set the parameters on the Advanced Configuration Options panel as follows:
        i. Reset SCSI BIOS at IC Int: **Enabled**
        ii. Display Ctrl+A Message During BIOS: **Enabled**
        iii. Extend BIOS translation for DOS drives >1GB: **Enabled**
        iv. Verbose or Silent Mode: **Verbose**
        v. Host Adapter BIOS: **Disabled**

        > **Note:** The host adapter BIOS setting shown above is only an example. The general rule of thumb is that for SCSI adapters that attach boot devices, BIOS for the adapter must be *enabled*. For all other adapters that attach non-boot devices, the adapter's BIOS must be *disabled*.

        vi. Support Removable Disks under Basic Input/Output System (BIOS) as fixed disks: **Disabled**
        vii. BIOS support for bootable CD-ROM: **Disabled**
        viii.BIOS support for INT 13 extensions: **Enabled**

8. Save the changes and select **SCSISelect** again to verify that you saved the changes.
9. Restart the server.
10. In local boot mode, load the Adaptec device driver and restart the system if instructed to do so. When installing Windows 2000 system on a remote ESS LUN, follow the Windows 2000 installation procedure to load appropriate Adaptec device driver during initial installation.

Installing and configuring Symbios 8751D adapter card

The steps to install and configure adapter cards shown here are examples that give you an idea how to configure Symbios 8751D adapter in your host system. Your configuration might be different.

To install and configure the Symbios 8751D adapter card in host-system perform the following steps:

1. Install the Symbios 8751D in the server.
2. Connect the cable to the ESS port.
3. Start the server.
4. Press `Ctrl+C` to get to the Symbios Configuration Utility menu.
5. From the Symbios Configuration Utility menu, select **LSI Logic Host Bus Adapters**.
   a. Set the parameters on the LSI Logic Host Bus Adapters panel as follows:
      i. Press F2 at the first panel.
      ii. Select the Boot Adapter list option to display the boot adapter list. See Example A-1 for an example of the boot adapter list.

> **Note:** The boot adapter list shows only user-definable parameters.

Example: A-1 Example of boot adapter list for the Symbios 8751D adapter

```
Boot Order [0]
NextBoot[Off]
```

6. Perform the following steps to change the BIOS settings:
   a. Highlight Next Boot and then click **On** to change the setting to **On**.
   b. Restart the host.
   c. Select the **Symbios Configuration Utility** again and make the changes.
   d. After you make the changes, highlight and then click **Off** to change the setting back to **Off**.
   e. Restart the host.
7. Set the parameters on the Global Properties panel as follows:
   a. Pause When Boot Alert Displayed: **[No]**
   b. Boot Information Display Mode: **[Verbose]**
   c. Negotiate With Devices: **[Supported]**
   d. Video Mode: **[Color]**

   e. Restore Defaults - use this option only if you want to restore global properties default settings

8. Set the parameters on the Adapters Properties panel as follows:
   a. SCSI Parity: **[Yes]**
   b. Host SCSI ID: **[7]**
   c. SCSI Bus Scan Order: **[Low to High (0..Max)]**
   d. Removable Media Support: **[None]**
   e. CHS Mapping: **[SCSI Plug and Play Mapping]**
   f. Spinup Delay (Secs): **[2]**
   g. Secondary Cluster Server: **[No]**
   h. Termination Control: **[Auto]**
   i. Restore Defaults - use this option only if you want to restore adapter properties default settings
9. Set the parameters on the Device Properties panel as follows:
   a. MT or Sec: **[20]**
   b. Data Width: **[16]**
   c. Scan ID: **[Yes]**
   d. Scan LUNs>0: **[Yes]**
   e. Disconnect: **[On]**
   f. SCSI Timeout: **240**
   g. Queue Tags: **[On]**
   h. Boot Choice: **[No]**
   i. Format: **[Format]**
   j. Verify: **[Verify]**
   k. Restore Defaults - use this option only if you want to restore device properties default settings
10. Save the changes and select **Symbios Configuration Utility** again to verify that you saved the changes.
11. Restart the server.
12. In local boot mode, load the Symbios device driver and restart the system if instructed to do so. When installing Windows 2000 system on remote ESS LUN, follow the Windows 2000 installation procedure to load appropriate Symbios device driver during initial installation.

Installing and configuring the QLogic adapter card

The steps to install and configure adapter cards shown here are examples that give you an idea how to configure QLogic QLA1041 adapter in your host system. Your configuration might be different.

To install and configure the QLogic QLA1041 adapter card in host-system perform the following steps:

1. Install the QLogic QLA1041 adapter card in the server.

2. Connect the cable to the ESS port.
3. Start the server.
4. Press `Alt+Q` to get to the FAST!Util menu.
   a. From the Configuration Settings menu, select **Host Adapter Settings**. Set the following parameters:
      i. Host Adapter: **Enabled**
      ii. Host Adapter BIOS: **Disabled**

      > **Note:** The host adapter BIOS setting shown above is only an example. The general rule of thumb is that for SCSI adapters that attach boot devices, BIOS for the adapter must be *enabled*. For all other adapters that attach non-boot devices, the adapter's BIOS must be *disabled*.

      iii. Host Adapter SCSI ID: **7**
      iv. PCI Bus direct memory access (DMA) Burst: **Enabled**
      v. Compact disc Boot: **Disabled**
      vi. SCSI Bus Reset: **Enabled**
      vii. SCSI Bus Reset Delay: **5**
      viii.Concurrent Command or Data: **Enabled**
      ix. Drivers Load RISC Code: **Enabled**
      x. Adapter Configuration: **Auto**
   b. Set the parameters in the SCSI Device Settings menu as follows:
      i. Disconnects OK: **Yes**
      ii. Check Parity: **Yes**
      iii. Enable LUNS: **Yes**
      iv. Enable Devices: **Yes**
      v. Negotiate Wide: **Yes**
      vi. Negotiate Sync: **Yes**
      vii. Tagged Queueing: **Yes**
      viii.Sync Offset: **8**
      ix. Sync Period: **12**
      x. Exec Throttle: **16**
   c. Save the changes and select **FAST!Util** again to verify that you saved the changes.
5. Restart the server.

In local boot mode load the QLogic device driver and restart the system if instructed to do so. When installing Windows 2000 system on remote ESS LUN, follow the Windows 2000 installation procedure to load appropriate QLogic device driver during initial installation.

Related publications

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

IBM Redbooks

For information on ordering these publications, see "How to get IBM Redbooks" on page 229.

- *IBM TotalStorage Enterprise Storage Server: Implementing the ESS in Your Environment*, SG24-5420
- *Implementing Fibre Channel Attachment on the ESS*, SG24-6113
- *Implementing ESS Copy Services on UNIX and Windows NT/2000*, SG24-5757
- *Implementing an Open IBM SAN*, SG24-6116
- *Exploiting HACMP 4.4: Enhancing the Capabilities of Cluster Multi-Processing*, SG24-5979
- *RS/6000 SP Cluster: The Path to Universal Clustering*, SG24-5374
- *IBM e(logo)server xSeries Clustering Planning Guide*, SG24-5845
- *Universal Clustering Problem Determination Guide*, SG24-6602
- *Linux HPC Cluster Installation*, SG24-6041
- *Installing and Managing Microsoft Exchange 2000 Clusters*, SG24-6265
- *RSCT Group Services: Programming Cluster Applications*, SG24-5523
- *HACMP Enhanced Scalability Handbook*, SG24-5328
- *HACMP/ES Customization Examples*, SG24-4498
- *HACMP Enhanced Scalability: User-Defined Events*, SG24-5327
- *IBM Enterprise Storage Server*, SG24-5465

Other resources

These publications are also relevant as further information sources:

- *IBM TotalStorage ESS Introduction and Planning Guide*, GC26-7294
- *Host Systems Attachment Guide 2105 Models E10, E20, F10, and F20*, SC26-7296, at:

  `http://www.storage.ibm.com/hardsoft/products/ess/pubs/f2ahs04.pdf`
- *PSSP for AIX: Administration Guide*, SA22-7348
- *PSSP for AIX: Diagnosis Guide*, GA22-7350
- *HP Fibre Channel Mass Storage Adapters Service and User Manual* (HP-UX 10.x, HP-UX 11.0, HP-UX 11i)
- *HP Fibre Channel Fabric Migration Guide* (HP-UX 11.0, HP-UX 11i)
- *HP A5158A Fibre Channel Adapter Release Notes* (HP-UX 11.0)

- *HP A6684A and A6685A HSC Fibre Channel Adapter Release Notes* (HP-UX 10.x, HP-UX 11.0, HP-UX 11i)

Referenced Web sites

These Web sites are also relevant as further information sources:

- Tivoli Storage Network Manager Web site

  http://www.tivoli.com/products/index/storage_net_mgr

- IBM ESS supported servers page

  http://www.storage.ibm.com/hardsoft/products/ess/supserver

- JNI home page

  http://www.jni.com

- IBM ESS pdf page

  http://www.storage.ibm.com/hardsoft/products/ess/pdf/1012-01.pdf

- Subsystem Device Driver Installation and User's Guide

  http://ftp.software.ibm.com/storage/subsystem/tools/f2asdd00.htm

- Veritas support site

  http://seer.support.veritas.com/docs/180452.htm

- Index to HP support documentation

  http://www.docs.hp.com/hpux/ha/index.html

- Support documentation for HP-UV 11.0

  http://www.docs.hp.com/hpux/os/11.0/index.html

- Support documentation for HP-UX 10.x

  http://www.docs.hp.com/hpux/os/10.x/index.html

- Home page for HP-UC Support Plus

  http://www.software.hp.com/SUPPORT_PLUS

- IBM TotalStorage Support page

  http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/storsw

- SGI Home page

  http://www.sgi.com

- Publications on Linux

  http://www.kernel.org/pub/linux/kernel/v2.4

- 99.9% Availability Guarantee Program

  http://www.pc.ibm.com/ww/netfinity/999guarantee.html

- The Beowulf Project

  http://www.beowulf.org

- Linux Virtual Server Project

  http://www.linuxvirtualserver.org

- High-Availability Linux Project

  http://linux-ha.org

- Redhat High Availability Server Project

  http://ha.redhat.com
- Sistina Products and Support — Global file system

  http://www.sistina.com/products_gfs.htm
- InterMezzo Home Page

  http://www.inter-mezzo.org
- System and options books

  http://www-1.ibm.com/servers/eserver/pseries/library/hardware_docs
- pSeries & RS/6000 License Agreement for Machine Code

  http://www.rs6000.ibm.com/support/micro/flicense.html
- Adapter Microcode

  http://www.rs6000.ibm.com/support/micro/download.html#adapter
- ESS Technical Support

  http://ssddom02.storage.ibm.com/disk/ess/documentation.html
- xSeries Support

  http://techsupport.services.ibm.com/server/support
- AIX Fix Distribution Service

  http://techsupport.services.ibm.com/rs6k/fixdb.html
- Sun Product Documentation

  http://docs.sun.com
- Emulux Corporation

  http://www.emulex.com
- Support Resources for IRIX

  http://support.sgi.com/othersupp/index.html
- Linux patch — kernel 2.4.9

  http://www.kernel.org/pub/linux/kernel/people/andrea/kernels/v2.4/2.4.9aa3/00_bh-async-3
- Linux patch — kernel 2.4.10

  http://www.kernel.org/pub/linux/kernel/people/andrea/kernels/v2.4/2.4.10aa1/00_vm-tweaks-1
- Linux.com

  http://www.linux.com/howto/SCSI-2.4-HOWTO/kconfig.html
- openMosix Project

  http://openmosix.sourceforge.net

How to get IBM Redbooks

Search for additional Redbooks or Redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

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

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Special notices

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

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

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

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

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

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

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

The following terms are trademarks of other companies:

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

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

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

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

PC Direct is a trademark of Ziff Communications Company in the United States and/or other

Index

Numerics

- 99.999 4, 46, 49
  - definition 9

A

- access profiles 18
- access-any 18
- access-restricted 18
- Adaptec 222
- adapter
  - profiles 22
- adapter failure 5
- adapters
  - Tachyon 150
- adaptstats 210
- adding volumes 171
- additional paths
  - adding 215
- addpaths 97
- ADSM 4
- AIX xiii, 7, 75
  - device driver installation 80
  - diagnostic software 89
  - Fibre Channel problems 87
  - fixes 78
  - hardware and software requirements 77
  - host 101
  - install device driver software 80
  - install hardware adapter 80
  - install required APARs 80
  - installing Fibre Channel 79
  - installing IBM SDD 92
  - installing SDD 91
  - load balancing 96
  - non supported environments 79
  - pre-installation checks 76
  - problem determination 86
  - protocol stack 76
  - SDD installation package filesets 91
  - upgradingFibre Channel firmware 88
  - verify hardware installation 82
  - verify software installation 82
  - verifying the SDD installation 93
- AIX 4.2.1 98
- AIX 4.3.2 97
- AIX SCSI
  - requirements 78
- alternate paths 154
- APARs 78
- architecture 2
- ARP 56
- assigning volumes 171
- automap 128
- autoswitch 40
- availability 3, 5, 8, 12, 13, 36, 46, 49, 50, 52
  - definition 3
  - increasing 13
  - single system 13
- availability and recoverability 203

B

- back-end database 66
- Beowulf 54
- bottleneck 5
- bufpages 153

C

- cable failure 5
- cache 2
- cfgmgr 107
- CLI 3
- cluster
  - hardware 51
    - shared disk 51
    - shared nothing 51
  - local 14
  - software 50
  - stretched 14
- clustering 1, 8, 11, 46
  - benefits 46
  - concepts 14
  - why it is used 9
- command tag queuing 182
- configuration manager 107
- configuring storage 179
- connectivity graph 70
- connectivity message 70
- continuous availability 12
- continuous operations 12
- controllers 135
- Copy Services 2, 3
- CPU 5

D

- data availability 26
- Data Path Device Configuration window 101
- Data Path Optimizer 76
- datapath
  - query adapter 31
  - query adaptstats 32
  - query device 32
  - query devstats 33
  - set adapter 33
  - set device 34
- datapath command 31, 210
  - query adapter 208
  - query device 210

query devstats 212
set adapter 214
set device 215
datapath query adapter 208
datapath query adaptstats 208
datapath query device 208
datapath query devstats 208
datapath set adapter 208
datapath set device 208
direct connection 168
direct routing 59
disaster recovery 3
disk mirroring 7
DMP 40, 130
configuring 218
reconfiguration 43
downtime 12
DPO 76

E
EM 72
EM Configuration Database 73
EMCDB 73
Emulex 198
Emulex adapter 88, 89
Emulex LP7000E adapter 80
Emulex LP8000 adapter card 197
Emulex LP8000 adapter device 197
ERP 49
ESCON 2
ESS xiii, 1, 2, 126, 151
AIX attach 117
HP attachment 151
SGI attach 170
SUN attach 130
Sun attachment 130
supported servers 148
Windows attach 189
ESS Copy Services 3
ESS identification under Veritas 144
ESS Specialist 3, 130, 151
Event Management 72
Resource Monitors 72

F
Fail Over Service 55, 56
failover 55, 97, 129
failover protection
lost 106
FC 15, 65
FC 6227 features 82
FC 6228 features 84
fca_npor 129
FC-AL 151, 167, 178, 181
connectivity 153
fcmsutil 151
fcp-on 128
FC-SW 167
Fibre Channel 2, 15, 17, 18, 26, 87
access modes 18
connections 20
LUNs 18
requirements 188
targets 18
Fibre Channel Arbitrated Loop 152, 175
Fibre Channel Switch 126, 148
FICON 2
fixes for AIX 78
FlashCopy 2, 3, 4
FOS 55

G
GFS 65
Global File System 65
Group Services 71
GS 71
GS subsystem 71
client subsystems 71
group membership list 71
Group Services API (GSAPI) 71
group state value 71
list of the providers 71
nameserver 72

H
HACMP 91, 116, 117, 119
concurrent mode 117
non-concurrent mode 117
support concurrent mode 117
support concurrent mode AIX 5.1.0 118
support non-concurrent mode 118
support non-concurrent mode AIX 5.1.0 119
HACMP support matrix 122
HACMP/6000 116, 120, 121, 122
support 117, 119
HACMP/ES 69
haemd daemon 72
hagsd daemon 71
hardware 126
hardware mirroring 7
HBA 127, 150
heartbeat message 69
heartbeat ring 69
high availability 1, 4, 8, 9, 12, 14, 31, 46, 47, 48, 49, 52, 53, 54, 55, 67, 68, 69, 126
definition 4
downtime/year 49
High Availability Cluster Multi-Processing 116
high performance computing 54
host adapter ports
configuring 171
host definition 22
hot-swappable 3
HP
switched fabric 152
tuning recommendations 153
HP 9000 151
HP 9000 Enterprise Server 148

HP_UX
  tuning recommendations 153
HPC 54
HP-UX xiii, 7, 34, 147, 154, 155
  cluster software 154
  creating volume groups 157
  installation of patches 150
  installing SDD 156
  non supported environments 148
  pre-installation planning 149
  setting up the ESS 151
  verifying SDD Installation 157
HPUX 148
HP-UX 10.20 148
HP-UX confirming storage connectivity 152

I
IBM cluster strategy 53
install directory 131
installing optical cable 171
Intermezzo 65, 66
introduction 1
ip_disable 129
IRIX 168
IRIX fx utility 168

J
Java 3
JNI 129

K
kernel 129

L
large files 134
latency 4
LC connector 85
LIC
  concurrent download 124
Linux clustering 53
Linux Virtual Server 54
load balancing 6, 8, 26, 31, 34, 35, 36, 38, 41, 43, 46, 47, 54, 55, 58, 67, 192
  definition 9
  dynamic 6
  failover only 6
  static 6
LUN 15, 16, 19, 153, 174, 194
  definition 22
  masking 15, 22
LUN affinity 194
lun-queue-depth 128
LV Timeout value 35
LVM 4, 7, 148
LVS 54

M
MAC 56
manageability 48
maxfiles 153
maxphys 128
MC/Service Guard 148, 154, 155
memswap_on 154
metadevice 41
metanodes 41
migrating non-SDD 112
migrating to ESS SDD multipath 112
mirror 37
Mirror Disk/UX 7
mirroring 7
  hardware 7
  software 7
MOSIX 55
mount 98
MPI 54
multipathing 1, 5, 6, 11
multipathing versus mirroring 8
multiple data paths 3
multiple-path mode 27

N
nameserver 72
NAT 61
NBD 65
nbuf 153
Network Address Translation 61
Network File System 64
network-on 128
NFS 64
node-to-node 193
non supported environments 188
NS 72

O
Object Database Management 100
ODM 100, 107
OLTP 53
outage 12

P
parallel processing 8, 46
  definition 9
PCI 84
Peer-to-Peer Remote Copy 3
performance 5, 47
Persistent Reserve Command set 120
pmand client 72
point in time
  PIT 4
PPRC 2, 3, 4, 15
pseudo parent (dpo) 120
public_loop 129
PV Timeout value 35
PV-Link versus SDD 155

PV-Links 34, 36, 154, 155
PVM 54

Q
QLogic adapter 224
QLogic QLA2100F adapter card 194
QLogic QLA2200F adapter card 195

R
Raid controller 7
RAID1 7
RAID5 2, 3
RAM 4
Redbooks Web site 229
  Contact us xv
redundancy 3
redundant systems 2
Reliable Messaging 70
Reliable Messaging Library 71
remote ESS disks 202
Resource Monitors 72
round robin 6, 97
RS/6000 Cluster Technology 68
RSCT 68
rsync 63

S
SAM 36, 155, 156
SAN 19
SAN Data Gateway 4, 87
scalability 2, 46, 47, 51, 54, 58, 65, 69
  definition 2
scientific computing 54
SCSI 2, 16, 26, 126
  attachment 16
  connection 19
  host system limitations 16
  limitations 16
  requirements 188
  targets 16
SCSI adapter 191
SCSI bus 191
SCSI channel connection 127
scsi_probe_delay 129
sd_io_time 128
sd_max_throttle 127
SDD 26, 31
  error log 115
  installation for AIX 82
  major files installed 94
  uninstall 207
  uninstallation 133, 166
  upgrading 123
  verify inistallation 157
SDD configuration
  verify 96
SDD devices
  unconfigure 100
SDD error log
  VPATH_DEVICE_OFFLINE 116
  VPATH_DEVICE_ONLINE 116
  VPATH_DEVICE_OPEN 116
  VPATH_FAIL_RELPRESERVE 116
  VPATH_OUT_SERVICE 116
  VPATH_PATH_OPEN 116
  VPATH_RESV_CFLICT 116
  VPATH_XBUF_NOMEM 115
SDD utility programs 114
  addpath 114
  dpovgfix 115
  extendvg4vp 115
  hd2vp 114
  lsvpcfg 115
  mkvg4vp 115
  vp2hd 114
Seascape 2
server cluster 9
  definition 53
server consolidation 50
  logical 50
  re-centralization 50
server farm 53
servers
  active 52
  hybrid 53
  passive 52
SGI
  configuring host path failover 177
  disk naming convention 182
  general considerations 177
  information 183
  information on the Web 184
  IRIX commands 183
  manual switch 179
  manual switching 177
  switched fabric 177, 178
  tuning 182
  unsupported utilities 183
  working with ESS volumes 179
SGI connections
  adding 170
  modifying 170
single-path Fibre Channel 127
Single-path mode 27
SMIT 100
SMP 46
software 126, 148
Solaris 7
Solstice Disk Suite 7
SPOF 14
storage connectivity 173
storage subsystem 8
Subsystem Device Driver 77
  removing 100
Sun 125
  assign disks 138
  clustering concepts 126
  logging 134

multipathing and clustering 125
non supported environments 126
post-installation 133
rootdg disk group 137
SDD installation 131
Sun Enterprise Servers 126, 130
Sun Solaris xiii, 126
Sun Veritas and ESS logical volumes 144
Sun Veritas Clustering Software 126
Sun Veritas DMP 137
Sun Veritas Dynamic Multi Pathing 126
Sun Veritas Fast Filesystem 126
Sun Veritas Volume Manager 126
supercomputers 2
supported configurations 168
supported services 63
swap 182
switched fabric 152, 168
confirming connectivity 171
Symbios 223
synchronous 4
System Administration and Management 36
system availability 49
System Management Interface Tool 100
systune 182

T

T(0) 4
targets 16, 194
TimeOutValue 203
Tivoli 23
topology 128
Topology Services 69
Crown Prince 70
Generic 70
Group Leader 70
Mayor 70
topology variable 128
traffic manager 58, 59, 60, 61, 67
troubleshooting 208
TS 69
TSM 4
tuning recommendations 153
tunneling 62

U

upgrading SDD 122
UPS 67
using the SDD 125

V

Veritas Cluster Manager 130
Veritas Volume Manager 218
Veritas VxVM 216
VG 154
Virtual Memory System 17
VMS 17
VMSA 135
volume groups 155
configuring 158
creating 157
Volume Manager Storage Administrator 135
VxSA 144
VxVM 7, 40
Veritas Volume Manager 7

W

Windows
configuring Fibre Channel adapters 192
datapath command 208
FC-AL requirements 188
hardware and software requirements 187
pre-installation 186
SCSI requirements 188
Windows 2000 185
Windows 2000 Disk Management 7
WWPN 17, 18, 19, 150

X

XRC 2

Z

zone-rscn 129
zoning 22, 172

IBM

Redbooks

Fault Tolerant Storage: Multipathing and Clustering Solutions

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

Fault Tolerant Storage

Multipathing and Clustering Solutions for Open Systems for the IBM ESS

All you need to know about multipathing software and more!

Integrate multipathing with clustering

Improve your storage availability

Clustered environments are growing in popularity in today's computing environments. Many customers require that their businesses are running 24/7/365. This IBM Redbook tells you how to integrate multiple paths to the ESS disks in a clustered environment while improving availability and throughput of your disk channels.

The IBM Enterprise Storage Server is a highly available, scalable and reliable, SAN-ready storage server, and its success is growing in the storage marketplace. Inside this book you will find how multipathing can improve total system reliability, and which multipathing software is supported for use with the IBM ESS. This combination gives you fault tolerant storage.

This redbook answers some important questions:

- Are your connections to the ESS reliable enough?
- Did you eliminate all single points of failure in your environment?
- Are some of your connectivity channels to the ESS overloaded, while others are idle?
- Do you need to improve your data paths, but you don't know how to do it?

You will find how disks are seen in a multiple path environment and how they are treated by the operating system. You can learn how to load-balance your channels and establish multiple paths to a single disk, while still maintaining data consistency on this disk. You'll discover all of this using the ESS storage server, on many operating systems, including IBM AIX, Microsoft Windows 2000, HP-UX, Sun Solaris and others.

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

SG24-6295-00 ISBN 073842336X