IBM Parallel System Support Programs for AIX
Managing Shared Disks

Document Number SC23-3849-02

Note:

Before using this information and the product it supports, be sure to read the general information in “Notices” on page ix.

Third Edition (October 1996)

This edition applies to

- Version 2 Release 2 of the IBM Parallel Systems Support Programs for AIX (PSSP) Licensed Program (5765–529)
- Version 1 Release 2 of the IBM Recoverable Virtual Shared Disk Licensed Program (5765–444)

and to all subsequent releases and modifications until otherwise indicated in new editions.

Order publications through your IBM representative or the IBM branch office serving your locality. Publications are not stocked at the address below.

IBM welcomes your comments. A form for your comments appears at the back of this publication. If the form has been removed, address your comments to the following address:

IBM Corporation, Department 55JA, Mail Station P384
522 South Road
Poughkeepsie, NY 12601-5400
United States of America

FAX (United States and Canada): 1+914+432-9405

FAX (Other Countries): Your International Access Code +1+914+432-9405

IBMLink (United States customers only): KGNVMC(MHVRCFS) IBM Mail Exchange: USIB6TC9 at IBMMAIL Internet: mhvrcfs@vnet.ibm.com

If you would like a reply, be sure to include your name, address, telephone number, or FAX number.

Make sure to include the following in your comment or note:

- Title and order number of this book
- Page number or topic related to your comment

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

© **Copyright International Business Machines Corporation 1994, 1996. All rights reserved.**
Note to U.S. Government Users — Documentation related to restricted rights — Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

Contents

Notices ix
Trademarks ix

About This Book xi
Who Should Use This Book xi
What You Need to Know First xi
How to Use This Book xi
Overview of Contents xi
Typography and Terminology xiii
Obtaining Additional Copies of This Book xiii

Chapter 1. An Overview of IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk Licensed Program 1
IBM Virtual Shared Disk Overview 1
IBM Virtual Shared Disk Restrictions 1
Data Striping Device Restrictions 2
Data Striping Device Overview 2
HSD and LVM Striping 2
The HSD Software 3
When to Use the HSD 3
IBM Recoverable Virtual Shared Disk Licensed Program Overview 3
The IBM Recoverable Virtual Shared Disk Software 4
IBM Virtual Shared Disk Recovery with Twin-tailed Disks 4
Twin-tailed Volume Groups 5
What's New in This Release? 5

Chapter 2. Installing IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk Licensed Program 7
Planning to Install IBM Virtual Shared Disks 7
Installation Requirements 7
Migrating to the IBM Virtual Shared Disk Upgraded Software 7
Before You Install the New Release 8
Installing the IBM Virtual Shared Disk Software 8
Step 1: Planning Your IBM Virtual Shared Disk Configuration 8
Step 2:. Installing at the Control Workstation 8
Step 3: Installing at IBM Virtual Shared Disk Clients and IBM Virtual Shared Disk Servers 9
Other Considerations before You Use IBM Virtual Shared Disks 9
The Communication Protocol Number 9
Installing the Data Striping Device 9
Installation Requirements 10
Installation Using SMIT 10
Installing the IBM Recoverable Virtual Shared Disk Licensed Program 10
Hardware and Software Requirements 11
Detailed Space Requirements 11
An Overview of Planning and Installing IBM Recoverable Virtual Shared Disk and Creating Twin-Tailed Volume Groups 12
Installing the IBM Recoverable Virtual Shared Disk Software 12
Installation Using SMIT 13
Migrating IBM Virtual Shared Disk Data 13

© Copyright IBM Corp. 1994, 1996

Chapter 3. Introducing the IBM Virtual Shared Disk Perspective 15
The IBM Virtual Shared Disk Perspective Primary Window 15
The Menu Bar 15
The Tool Bar 15
The Work Areas 16
The Information Area 17

Chapter 4. Creating and Managing Shared Disks 19
Preparing to Create IBM Virtual Shared Disks and Data Striping Devices . . . 19
Authorization 20
Deleting Old Rollback Files 20
Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR 20
Creating IBM Virtual Shared Disks with the createvsd Command 22
Examples of Creating IBM Virtual Shared Disks with createvsd 26
Creating Data Striping Devices with the createhsd Command 27
Examples of Creating Data Striping Devices with createhsd 30
Defining Data Striping Devices in a Node-Pair Configuration 32
Configuring and Unconfiguring IBM Virtual Shared Disks and Data Striping Devices 33
Verifying an IBM Virtual Shared Disk 33
Displaying Information about IBM Virtual Shared Disk and Data Striping Device Nodes 34
Displaying Information about IBM Virtual Shared Disk Clients 35
Displaying Information about Data Striping Device Definitions and Clients . . . 35
Displaying IBM Virtual Shared Disk Node Statistics 36
Displaying Data Striping Device Node Statistics 37
Resetting IBM Virtual Shared Disk and Data Striping Device Statistics 37
Running IBM Virtual Shared Disk Diagnostics 37
Changing the States of IBM Virtual Shared Disks 38
Starting an IBM Virtual Shared Disk 38
Preparing an IBM Virtual Shared Disk 38
Resuming an IBM Virtual Shared Disk 38
Suspending an IBM Virtual Shared Disk 38
Stopping an IBM Virtual Shared Disk 39
Removing an IBM Virtual Shared Disk or Data Striping Device 40
Adding a Server Node to an Existing IBM Virtual Shared Disk Configuration . 40
Removing a Node from an Existing IBM Virtual Shared Disk Configuration . . 41

Chapter 5. Application Programming Considerations 43
Application Programming Considerations for IBM Virtual Shared Disks 43
Data Integrity 43
IBM Virtual Shared Disk Transmission Protocol 43
Getting IBM Virtual Shared Disk Information with C Interfaces 45
Application Programming Considerations for Data Striping Devices 45
Getting HSD Information with C Interfaces 45
Making Your Application Recoverable 46

Chapter 6. Performance and Tuning Considerations for IBM Virtual Shared Disks and Data Striping Devices 49
Tuning IBM Virtual Shared Disk Performance 49
IBM Virtual Shared Disk Tunable Parameters 50
The High Performance Switch/Scalable POWERparallel Switch 50
Logical Volume Manager (LVM) Tuning Considerations 51

Buffer Allocation . . . 52
The Cache Buffer . . . 52
Buddy Buffers . . . 53
Maximum I/O Request Size . . . 53
Request Blocks . . . 54
pbufs . . . 54
mbufs and the Switch Pool . . . 55
Tuning Data Striping Device Performance . . . 56
IBM Virtual Shared Disk Tuning Effects and Data Striping Devices . . . 57

Chapter 7. How the IBM Recoverable Virtual Shared Disk Software Works 59
The Group Services Component of Availability Sevices . . . 59
HA Topology Services and Recovery . . . 60
The Membership List . . . 60
Recovery Scripts . . . 60
Quorum . . . 60
Active Groups . . . 61
The IBM Recoverable Virtual Shared Disk Recovery Subsystems . . . 61
The rvsd Subsystem . . . 61
The hc Subsystem . . . 62
Startup Time . . . 63
Command Considerations . . . 64
IBM Virtual Shared Disk Data Striping Device Considerations . . . 66

Chapter 8. IBM Recoverable Virtual Shared Disk Recovery Scenarios . . . 67
Recognizing Recovery . . . 67
Planning For Recovery . . . 67
IBM Virtual Shared Disk Client or Server Node Failure . . . 68
Switch Failure Scenario . . . 69
Disk Controller Failure Scenario . . . 69
HA Topology Services or Recovery Service Daemon Failure . . . 70
IBM Virtual Shared Disk Failure . . . 70
Recovery . . . 71

Chapter 9. Using HACMP with IBM Recoverable Virtual Shared Disk . . . 73
Configuring IBM Recoverable Virtual Shared Disk with HACMP/6000 . . . 73
Restrictions . . . 73

Chapter 10. Diagnosis and Messages Information . . . 75
Diagnosis Information . . . 75
Messages . . . 75

Appendix A. IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands . . . 77
cfghsd . . . 77
cfghsdvsd . . . 77
cfgvsd . . . 77
cksumvsd . . . 78
createhsd . . . 78
createvsd . . . 78
ctlhsd . . . 78
ctlvsd . . . 79
defhsd . . . 79
defvsd . . . 79

ha.vsd . . . 79
hc.vsd . . . 80
hsdatalst . . . 80
hsdvts . . . 80
lshsd . . . 80
lsvsd . . . 80
preparevsd . . . 81
removehsd . . . 81
removevsd . . . 81
resumevsd . . . 81
spvsd . . . 82
startvsd . . . 82
statvsd . . . 82
stopvsd . . . 82
suspendvsd . . . 83
ucfghsd . . . 83
ucfghsdvsd . . . 83
ucfgvsd . . . 83
undefhsd . . . 83
undefvsd . . . 84
updatehsd . . . 84
updatevsdnode . . . 84
updatevsdtab . . . 84
verparvsd . . . 85
vsdatalst . . . 85
vsddiag . . . 85
vsdelnode . . . 85
vsdelvg . . . 86
vsdnode . . . 86
vsdsklst . . . 86
vsdvg . . . 86
vsdvts . . . 87

Appendix B. Single-Node Command and SMIT Interfaces for IBM Virtual Shared Disks, Data Striping Devices and the IBM Recoverable Virtual Sha red Disk Licensed Product . . . 89
Command and SMIT Interfaces for IBM Virtual Shared Disks . . . 89
Entering Node Information in the SDR for One Node . . . 89
Entering Global Volume Group Information in the SDR for One Node . . . 91
Entering IBM Virtual Shared Disk Information in the SDR for One Node . . . 92
An Example of Defining an IBM Virtual Shared Disk on One Node at a Time 93
Displaying IBM Virtual Shared Disk Information Stored in the SDR . . . 95
Listing Node Information . . . 96
Listing Global Volume Group Information . . . 96
Listing Defined IBM Virtual Shared Disks . . . 96
Deleting IBM Virtual Shared Disk Information from the SDR . . . 96
Managing IBM Virtual Shared Disks . . . 97
Configuring an IBM Virtual Shared Disk . . . 99
Starting an IBM Virtual Shared Disk . . . 99
Preparing an IBM Virtual Shared Disk . . . 100
Resuming an IBM Virtual Shared Disk . . . 100
Suspending an IBM Virtual Shared Disk . . . 100
Stopping an IBM Virtual Shared Disk . . . 101
Unconfiguring an IBM Virtual Shared Disk . . . 101

Displaying Managed IBM Virtual Shared Disk Characteristics 102
Displaying Managed IBM Virtual Shared Disk Statistics 102
Displaying IBM Virtual Shared Disk Device Driver Statistics 102
Setting and Displaying IBM Virtual Shared Disk Device Driver Operational Parameters 103
Testing to Verify an IBM Virtual Shared Disk (the Verification Test Suite) . 104
Command and SMIT Interfaces for Data Striping Devices 104
Defining HSD Device Driver Information in the SDR 104
An Example of Creating an HSD 105
Displaying Data Striping Device Information in the SDR 106
Undefining HSD Information in the SDR 107
Managing Data Striping Devices 107
Verifying Data Striping Device Installation 109
Management and Usage Notes 110
Command and SMIT Interfaces for the IBM Recoverable Virtual Shared Disk Licensed Program 110
Defining Twin-tailed LVM Components in the IBM Recoverable Virtual Shared Disk Environment 110
Coordinating the IBM Recoverable Virtual Shared Disk Environment . . . 111
Overview of Shared Volume Groups Administrative Tasks 111
Prerequisite Tasks 111
Maintaining Twin-Tailed Volume Groups 112
Creating a Twin-tailed Volume Group 112
Adding Physical Volumes To a Twin-tailed Volume Group 115
Adding or Removing Logical Volumes to or from a Twin-tailed Volume Group 119
Reducing a Twin-tailed Volume Group 121
Removing a Twin-tailed Volume Group 126

Appendix C. Setting Up 7134 Disks as Twin-tailed Volume Groups . . . 129

Appendix D. Glossary 131

Index 133

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 product, program, or service that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program, or service. Evaluation and verification of operation in conjunction with other products, except those expressly designated by IBM, are the user's responsibility.

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:

IBM Director of Licensing
IBM Corporation
500 Columbus Avenue
Thornwood, NY 10594
USA

Licensees of these programs 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
Mail Station P300
522 South Road
Poughkeepsie, NY 12601-5400
USA
Attention: Information Request

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

Trademarks

The following terms are trademarks of the IBM Corporation in the United States or other countries or both:

AIX

IBM

RS/6000

RS/6000 Scalable POWERparallel Systems

Windows is a trademark of Microsoft Corporation.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

© Copyright IBM Corp. 1994, 1996

Other company, product, and service names, which may be denoted by a double asterisk (**), may be trademarks or service marks of others.

About This Book

This book describes how to install, manage, and use the disk management facilities of the SP system—the IBM Virtual Shared Disk and Data Striping Device components of PSSP and the IBM Recoverable Virtual Shared Disk licensed program. It also provides an overview of these components and explains how they can add reliability to your data storage.

This book applies to PSSP Version 2 Release 2 and to IBM Recoverable Virtual Shared Disk Version 1 Release 2.

Who Should Use This Book

Use this book if you are considering using IBM Virtual Shared Disks, Data Striping Devices, or the IBM Recoverable Virtual Shared Disk software. You can use the information in this book to install, manage, and use these shared disk management components and to write application programs that use their facilities.

What You Need to Know First

You may need to be familiar with and have the documentation for the following:

- IBM Parallel System Support Programs for AIX
- The AIX operating system, including the Logical Volume Manager subsystem, especially: *IBM AIX Version 4.1 System Management Guide: Operating System and Devices*, SC23-2525.
- Installation procedures for this release in *IBM Parallel System Support Programs for AIX: Installation Guide*, GC23-3897
- The IBM Virtual Shared Disk commands in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*, GC23-3900
- The installation and configuration information that came with your twin-tailed disk hardware
- Planning information in *IBM Parallel System Support Programs for AIX: System Planning Guide*, GC23-3902
- Availability services information in *IBM Parallel System Support Programs for AIX: Group Services Programming Guide and Reference*, SC28–1675

How to Use This Book

Overview of Contents

This book is divided into the following chapters:

- Chapter 1, " An Overview of IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk Licensed Program" on page 1 provides an overview of the IBM Virtual Shared Disk and Data Striping Device components of the SP system and of the IBM Recoverable Virtual Shared Disk licensed program.

© Copyright IBM Corp. 1994, 1996

- Chapter 2, " Installing IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk Licensed Program" on page 7 lists the installation requirements and describes the procedure for installing, configuring, and starting up IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk software.
- Chapter 3, " Introducing the IBM Virtual Shared Disk Perspective" on page 15 describes the graphical user interface for managing IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk software.
- Chapter 4, " Creating and Managing Shared Disks" on page 19 tells how to create and manage IBM Virtual Shared Disks, Data Striping Devices, and IBM Recoverable Virtual Shared Disks.
- Chapter 5, "Application Programming Considerations" on page 43 contains programming considerations for IBM Virtual Shared Disks, Data Striping Devices, and recoverability.
- Chapter 6, "Performance and Tuning Considerations for IBM Virtual Shared Disks and Data Striping Devices" on page 49 describes performance and tuning considerations for IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk licensed program.
- Chapter 7, "How the IBM Recoverable Virtual Shared Disk Software Works" on page 59 describes the operation of the IBM Recoverable Virtual Shared Disk licensed program.
- Chapter 8, "IBM Recoverable Virtual Shared Disk Recovery Scenarios" on page 67 describes IBM Recoverable Virtual Shared Disk recovery procedures.
- Chapter 9, "Using HACMP with IBM Recoverable Virtual Shared Disk" on page 73 describes how recovery is handled when you have HACMP/6000 installed on your system.
- Chapter 10, "Diagnosis and Messages Information" on page 75 contains IBM Recoverable Virtual Shared Disk error messages.
- Appendix A, "IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands" on page 77 contains reference pages for the IBM Virtual Shared Disk, Data Striping Device, and IBM Recoverable Virtual Shared Disk commands.
- Appendix B, "Single-Node Command and SMIT Interfaces for IBM Virtual Shared Disks, Data Striping Devices and the IBM Recoverable Virtual Sha red Disk Licensed Product" on page 89 contains older procedures for defining and managing IBM Virtual Shared Disks, Data Striping Devices, and RVSDs, to help you maintain scripts.
- Appendix C, "Setting Up 7134 Disks as Twin-tailed Volume Groups" on page 129 contains procedures for setting up IBM 7134 disks as twin-tailed volume groups.
- Appendix D, "Glossary" on page 131 is a glossary of terms used in this book.

Typography and Terminology

This book uses the following typographical conventions:

Convention	Usage
Bold	**Bold** words or characters represent system elements that you must use literally, such as commands, subcommands, flags, path names, directories, file names, values, and selected menu options.
Italic	• *Italic* words or characters represent variable values that you must supply. • *Italics* are used for book titles • *Italics* are used for general emphasis
`Monospace`	All of the following are displayed in `monospace` type: • Displayed information • Message text • Example text • Specified text typed by the user • Field names as displayed on the screen • Prompts from the system • References to example text

The name of the product, IBM Recoverable Virtual Shared Disk, is occasionally referred to in the text as simply RVSD. In all cases, this expression is equivalent to IBM Recoverable Virtual Shared Disk.

The term *SP* or *SP system* refers to the current system partition, if you have partitioned your system. Otherwise, the term refers to the physical SP system.

Obtaining Additional Copies of This Book

This book is available in PostScript format. You can obtain PostScript files for all SP product documentation by:

- Downloading them from the **lscftp.pok.ibm.com** anonymous FTP server. The files for xyz_prod_name can be found in the directory called **/pub/pps/xyz_ftp_prod_dir**. This book, for example, can be found in **/pub/pps/rvsd_ftp_prod_dir**.
- Printing the PostScript documentation files that are shipped on the product media.

You can also view or download this book from the IBM RS/6000 web site at **http://www.rs6000.ibm.com**.

To order additional copies of this book from the IBM distribution center, request order number SC23-3849-02.

Chapter 1. An Overview of IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk Licensed Program

This chapter provides a basic overview of the IBM Virtual Shared Disk and Data Striping Device components of the SP system and of the IBM Recoverable Virtual Shared Disk Licensed Product. Later chapters give more detailed information about these software features and how to manage them.

IBM Virtual Shared Disk Overview

An IBM Virtual Shared Disk lets application programs executing at different nodes of a system partition access a raw logical volume as if it were local at each of the nodes. See Figure 1 for an illustration of a simplified IBM Virtual Shared Disk implementation. The logical volume is actually local at *one* of the nodes, called the *server* node. The IBM Virtual Shared Disk software routes I/O requests from the other nodes, called *client* nodes, to the server node and returns results to the client nodes.

The I/O routing is done by the IBM Virtual Shared Disk device driver layer that sits on top of the Logical Volume Manager (LVM). The device driver is loaded as a kernel extension on each node. Thus, logical volumes can be made globally accessible.

The application program interface to an IBM Virtual Shared Disk is the character (raw) special device.

BETH3

Figure 1. An IBM Virtual Shared Disk IP Network Implementation

IBM Virtual Shared Disk Restrictions

- In systems at IBM Parallel Systems Support Programs levels earlier than Version 2.2, all IBM Virtual Shared Disk nodes within a system partition must run at the same software levels; no intermixing is allowed. In systems at IBM Parallel Systems Support Programs Version 2.2, one system level earlier than Version 2.2 can be supported in the same system partition with Version 2.2. Interoperability among IBM Virtual Shared Disk nodes at different levels is not supported.
- You cannot create an AIX Journaled File System (JFS) that you want other nodes to be able to mount and access on an IBM Virtual Shared Disk.
- The maximum number of IBM Virtual Shared Disks that can be defined in a system partition is 32767 (32K).

© Copyright IBM Corp. 1994, 1996

Data Striping Device Restrictions

- Striping is supported across IBM Virtual Shared Disks only.
- The function **lseek(...,...,END)** is not supported because the file may be composed of several IBM Virtual Shared Disks.
- All the underlying IBM Virtual Shared Disks in one HSD must be the same size.
- The size of each IBM Virtual Shared Disk in one HSD must be a multiple of the stripe size in the HSD.
- The size of an HSD (*number_of_vsds* × *size_of_vsd*) cannot be greater than the maximum file size in the filesystem.
- After you have written data to an HSD, do not directly access any of the underlying IBM Virtual Shared Disks.
- The address of your I/O buffer must be aligned on a page boundary. If it is not, the **hsd_read** and **hsd_write** routines will **not** parallelize I/O requests to underlying IBM Virtual Shared Disks. Performance will be less than optimal. Set the environment variable MALLOCTYPE to **3.1** to make all new allocated buffers at the page boundary.

Data Striping Device Overview

The Data Striping Device is a device driver that lets application programs stripe data across the physical disks in multiple IBM Virtual Shared Disks, thus reducing I/O bottlenecks and hot spots.

Instead of writing all the data from one I/O request on one IBM Virtual Shared Disk at a specific location, data striping writes or reads blocks of the data on each of several separate IBM Virtual Shared Disks.

IBM Virtual Shared Disk support provides shared access to disks physically attached at any node.

The principal benefit of striping is that it provides automatic distribution or partitioning of data across physical disks and nodes while requiring minimal administrator intervention. When the data is well partitioned among the disk drives, the application can get better performance from the underlying disk drives.

Your application can conduct concurrent I/O processing on the different stripes at the same time. This all is done by the striping software and is transparent to the application.

The data striping device is sometimes called the “hashed shared disk” (HSD).

HSD and LVM Striping

AIX 4.1 introduces striping capabilities for the Logical Volume Manager. A logical volume can be striped across two or more physical disks. HSD can stripe data across two or more IBM Virtual Shared Disks (an IBM Virtual Shared Disk corresponds to a single logical volume).

The advantage of HSD striping over LVM striping is that HSD can stripe data across multiple nodes, while LVM striping is limited to local physical disks.

We recommend that you use LVM for local striping and HSD for global striping. For example, if you have two physical disks on 10 nodes and want to stripe the data across all 20 physical disks, you should create a local logical volume that spans both physical disks on each node. After creating IBM Virtual Shared Disks that point to these 10 logical volumes, create one HSD that spans all 10 IBM Virtual Shared Disks. You now have one HSD that spans 10 IBM Virtual Shared Disks (one on each node) and each IBM Virtual Shared Disk points to a logical volume that is striped across two physical disks.

Refer to *AIX Versions 3.2 and 4.1 Performance Tuning Guide* SC23-2365, for more information on how to tune striped logical volumes.

The HSD Software

The HSD driver is a pseudo device driver layer above the IBM Virtual Shared Disk layer. It is used only as a raw device driver interface with the IBM Virtual Shared Disk devices.

Each HSD is defined with a list of IBM Virtual Shared Disks and a stripe size. The stripe size indicates the maximum size of a block of data to be stored on a single IBM Virtual Shared Disk in one I/O operation. The HSD uses a hash function based on the following parameters to determine which IBM Virtual Shared Disk to place data on.

- The offset in the HSD file
- The number of IBM Virtual Shared Disks defined in the HSD.
- The stripe size

When to Use the HSD

Whether or not to use the HSD depends on your configuration of IBM Virtual Shared Disks and the I/O characteristics of your application programs.

If the I/O load to a specific IBM Virtual Shared Disk is too heavy (hot spots), then you can use HSD to distribute the load to other IBM Virtual Shared Disks and nodes.

Use the IBM Virtual Shared Disk software initially to determine if you have hot spots.

IBM Recoverable Virtual Shared Disk Licensed Program Overview

In order to explain the benefits of the IBM Recoverable Virtual Shared Disk licensed program, let's look at a system *without* the IBM Recoverable Virtual Shared Disk software. The IBM Virtual Shared Disk function lets all nodes in the SP system or system partition access a given disk, even though that specific disk is physically attached to only one node. When a read or write operation is issued to an IBM Virtual Shared Disk, the device driver determines the server node and routes the request to that node, whether the node issuing the I/O request is local or remote to the server node. If that single serving node should fail, access to the disk is lost until the serving node is rebooted by the administrator.

The IBM Recoverable Virtual Shared Disk Software

The IBM Recoverable Virtual Shared Disk software lets you use twin-tailed disks and configure nodes as primary and secondary IBM Virtual Shared Disk server nodes. It provides transparent switchover to a secondary server if the primary server node for a set of IBM Virtual Shared Disks fails.

A *twin-tailed disk* is a disk or group of disks that are attached to two nodes of an SP. For IBM Recoverable Virtual Shared Disk purposes, only one of these nodes serves the disks at any time. The secondary or backup node serves the disks only if the primary node fails or is powered off.

The IBM Recoverable Virtual Shared Disk software automatically manages your IBM Virtual Shared Disks by detecting node failures and enabling a secondary node. The software configures your IBM Virtual Shared Disk definitions and moves IBM Virtual Shared Disks among states as necessary. If at any time the primary IBM Virtual Shared Disk server node fails, an administrator can reboot that node; the primary IBM Virtual Shared Disk server node rejoins the system, and the I/O load is switched from the secondary node back to the primary node for which it was originally configured.

This can mean less down time because the IBM Recoverable Virtual Shared Disk support provides automated recovery through a secondary node with twin-tailed access to the failing (primary) node's disks.

Users with applications exploiting the IBM Recoverable Virtual Shared Disk software can recover more easily from node failures and have continuous access to the data on the twin-tailed disks.

Note: In order for the primary node to resume operations after a failure, you must issue the **Estart** or **Eunfence** command (depending on the switch type) after rebooting the node. Refer to the switch chapter in the *IBM Parallel System Support Programs for AIX: Administration Guide* for information on starting the each type of switch, node isolation, and so forth.

IBM Virtual Shared Disk Recovery with Twin-tailed Disks

The following three figures show a simple system with one pair of IBM Recoverable Virtual Shared Disks. Figure 2 shows the basic configuration. In Figure 3, the secondary node is active following a failure of the primary node.

twtl1

Figure 2. Simplified View of a Twin-tailed Disk

twtl2

Figure 3. The Secondary Node Serves after a Primary Node Failure

In Figure 4, the primary node has been recovered and has automatically taken over the disk from the secondary node.

twtl1

Figure 4. The Primary Node Is the Server Again after Recovery

Recovery is transparent to an application running on a client node – there is no disruption of service, only a slight delay while takeover occurs. When the primary server node is rebooted, the IBM Recoverable Virtual Shared Disk software switches the I/O load back to the primary server node. The IBM Recoverable Virtual Shared Disk software provides interfaces that enable your application to be recoverable.

Notes:

1. Do not put the root volume group (**rootvg**) or any other volume group that contains bootable logical volumes on a twin-tailed disk.
2. We recommend allocating only IBM Virtual Shared Disks to twin-tailed volume groups. Applications accessing data on non-IBM Virtual Shared Disk logical volumes (a file system, for example) will lose access to the logical volume when the IBM Virtual Shared Disk server node changes.
3. If a SCSI bus is used for twin-tailed DASD and a tape drive is using the same bus, a long tape job can be reset if the DASD (or any other device on the shared SCSI bus) is powered down for service.

Twin-tailed Volume Groups

A twin-tailed volume group is a volume group that contains disks that are accessible to two nodes. Both nodes must be in the same system or system partition.

What's New in This Release?

- New commands that allow you to create IBM Virtual Shared Disk configurations and Data Striping Devices, (including identifying secondary direct clients if you have the IBM Recoverable Virtual Shared Disk licensed product installed), update the characteristics of those IBM Virtual Shared Disks and Data Striping Devices, and view their attributes across multiple nodes in a system partition, are supplied.
- A graphical user interface, Perspectives, has been developed to help you manage and monitor system resources, including IBM Virtual Shared Disks and Data Striping Devices, more easily.
- The monvsd tool is no longer available. It has been replaced by IBM Virtual Shared Disk monitoring functions in Perspectives. This book describes how to use Perspectives to monitor and manage IBM Virtual Shared Disks (see Chapter 4, " Creating and Managing Shared Disks" on page 19).
- Improvements to IP performance for the SP2 allow the system to perform better without the need for buddy buffers. Buddy buffers are generally needed only during the first few requests that IBM Virtual Shared Disk processes.
- The **ctlvsd** command has two new options, **-t** and **-T**. The -t option lists the current routing table and mbuf headers cached by the IBM Virtual Shared Disk device driver. The -T option clears or releases all cached routes.
- Three new IBM Virtual Shared Disk statistics are defined; indirect I/O, 64–byte unaligned reads, and communications buffer pool shortages.
- A new diagnostic tool, **vsddiag**, lets you gather useful information before calling IBM service with an IBM Virtual Shared Disk problem.

- IBM Recoverable Virtual Shared Disk no longer uses heartbeat directly. Instead, Group Services determines which nodes in the IBM Virtual Shared Disk cluster are active and manages the recovery process. *IBM Parallel System Support Programs for AIX: Group Services Programming Guide and Reference* contains more information about Group Services.
- Two new commands, **ha.vsd** and **hc.vsd**, are used in the diagnosis and recovery of IBM Virtual Shared Disks by the IBM Recoverable Virtual Shared Disk licensed program. These commands are described in Chapter 7, "How the IBM Recoverable Virtual Shared Disk Software Works" on page 59 and in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*. The command **hb** with the **reset** option has been replaced in the IBM Recoverable Virtual Shared Disk recovery scenarios by the command **ha.vsd reset**.

  Note: If your system is running HACMP, IBM Recoverable Virtual Shared Disk will not use Group Services or the **ha.vsd** and **hc.vsd** commands. See Chapter 9, "Using HACMP with IBM Recoverable Virtual Shared Disk" on page 73 for more information.

Chapter 2. Installing IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk Licensed Program

This section describes how to install IBM Virtual Shared Disks, Data Striping Devices, and IBM Recoverable Virtual Shared Disk. For IBM Virtual Shared Disk, you'll install **ssp.csd**, which contains **ssp.csd.vsd** and **ssp.csd.cmi**. For Data Striping Devices, you'll install **ssp.csd.hsd**, which is also contained in the **ssp.csd** package. For IBM Recoverable Virtual Shared Disk, you'll install the **rcsd** install package, which containes **rcsd.vsd, rcsd.rvsd, rcsd.hc,** and **rcsd.hahc.**

Planning to Install IBM Virtual Shared Disks

Before you install the IBM Virtual Shared Disk software, determine which applications are to have access to the IBM Virtual Shared Disks you plan to create. Consider the size of the applications you will run. Consider also how you want to spread data across the disks. Refer to the *IBM Parallel System Support Programs for AIX: System Planning Guide* for planning considerations. See “IBM Virtual Shared Disk Restrictions” on page 1 and “Data Integrity” on page 43 for considerations in selecting applications to exploit your IBM Virtual Shared Disks.

Installation Requirements

Before you can set up your IBM Virtual Shared Disk, the SP system software, (IBM Parallel Systems Support Programs (PSSP), Program Number 5765-529) must have been installed and operational, **ssp.basic** and **ssp.sysctl** in particular. Sysctl must also be configured. See *IBM Parallel System Support Programs for AIX: Installation Guide* for installation information and *IBM Parallel System Support Programs for AIX: Administration Guide* for information about sysctl.

Read the documentation that accompanies the installation media to make sure you have the appropriate version and level of all software. That documentation will list any required APARs (service updates for the product) that must be installed along with the IBM Virtual Shared Disk software.

Migrating to the IBM Virtual Shared Disk Upgraded Software

Read this section if you currently are using IBM Virtual Shared Disks and plan to upgrade from AIX Release 3.2.5 to AIX Release 4.1 with IBM Parallel System Support Programs for AIX, Version 2 Release 2.

The AIX Logical Volume Manager limits the number of physical partitions per disk to 1016. Older releases of AIX did not always enforce this limit, but AIX 4.1 does. If a disk larger than 4GB is added to a volume group that uses the default physical partition size of 4MB, the disk addition fails.

If you are using large disks, define a larger physical partition size. If you already have disks that are over the 1016 limit and intend to move to AIX 4.1 or later AIX releases, you can redefine your volume groups with a larger partition size.

© Copyright IBM Corp. 1994, 1996

Before You Install the New Release

Before you install the new release, check to make sure the following steps have been taken:

Are the Existing IBM Virtual Shared Disks Unconfigured?

Unconfigure all existing IBM Virtual Shared Disks before installing the new IBM Virtual Shared Disk software. See “Configuring and Unconfiguring IBM Virtual Shared Disks and Data Striping Devices” on page 33 and the **ucfghsdvsd** and **ucfgvsd** command reference pages for more information. Then install the new IBM Virtual Shared Disk software, which will update the SDR. Then reconfigure and restart the IBM Virtual Shared Disks. The current IBM Virtual Shared Disk software release must be installed on the SP nodes that you will use as either IBM Virtual Shared Disk server nodes or IBM Virtual Shared Disk client nodes.

Is the ssp.csd Package Installed?

The **ssp.csd** package must already be installed on all SP nodes to be used as either IBM Virtual Shared Disk server or client nodes.

Are You Using Buddy Buffers?

If you had large numbers of buddy buffers defined, you should increase your switch pool buffer by the amount of space represented by that number of buddy buffers and redefine four buddy buffers. See the recommendations in “Buddy Buffers” on page 53 and “mbufs and the Switch Pool” on page 55 for more information.

Installing the IBM Virtual Shared Disk Software

Follow these steps to install, define, and activate the IBM Virtual Shared Disk software.

Step 1: Planning Your IBM Virtual Shared Disk Configuration

You must install this software at *all* client and server nodes as well as at the control workstation if you use the CWS to manage the IBM Virtual Shared Disks. See *IBM Parallel System Support Programs for AIX: Installation Guide* for details.

Plan and decide which logical volumes should be made globally accessible as IBM Virtual Shared Disk server nodes.

Plan the volume groups containing these logical volumes and plan which server nodes these volume groups should be on.

Step 2:. Installing at the Control Workstation

Install the IBM Virtual Shared Disk software at the control workstation for configuration management. The control workstation is neither a server nor a client.

Use the **ssp.csd installp** package to install both options:

ssp.csd.vsd The IBM Virtual Shared Disk's device driver and commands.

ssp.csd.cmi The IBM Virtual Shared Disk's configuration management support, including the SMIT panels.

You must install both options.

Issue the command:

```
installp  -aXd ssp.csd.vsd ssp.csd.cmi
```

or

```
installp  -aXd ssp.csd
```

to install the IBM Virtual Shared Disk software.

If you plan to install the Data Striping Device (HSD) later, you must have installed both **ssp.csd.vsd** and **ssp.csd.cmi**.

Step 3: Installing at IBM Virtual Shared Disk Clients and IBM Virtual Shared Disk Servers

Install the IBM Virtual Shared Disk software at the client and server nodes you identified in the first step. The only modules you need at the client and server nodes are in **ssp.csd**.

Other Considerations before You Use IBM Virtual Shared Disks

Before you create your IBM Virtual Shared Disks, check the IP protocol number you are using.

The Communication Protocol Number

The IBM Virtual Shared Disk implements its own IP protocol, similar to TCP and UDP. You must use protocol number **63**. Don't change the default, which is the line

```
#define IPPROTO_LOCAL     63
```

in the /**usr/include/netinet/in.h**. header file.

Note: **All nodes must use protocol number 63** or the IBM Virtual Shared Disk will not work properly.

Installing the Data Striping Device

Before you define the Data Striping Devices:

1. Read “Tuning Data Striping Device Performance” on page 56 in this chapter to understand tuning considerations.
2. Determine what the hot spots are on the IBM Virtual Shared Disks.
3. Determine what nodes and disks will be used to eliminate the hot spots.
4. Determine the stripe size.
5. Install the HSD software on all nodes with IBM Virtual Shared Disks.

The HSD is the **ssp.csd.hsd** option in the **ssp.csd** install image. You can install **ssp.csd.hsd** either when you install the IBM Virtual Shared Disk (**ssp.csd.vsd**) or at a later time.

Installation Requirements

The HSD requires 0.5MB disk space and will be installed in /**usr**/**lpp**/**csd**/**bin**, /**etc**/**drivers**, and the message catalog directories.

Before you install the Data Striping Device, review *IBM Parallel System Support Programs for AIX: Installation Guide*, to understand the general requirements for optional install images.

Installation Using SMIT

You can use either SMIT or the **installp** command to install the **ssp.csd.hsd** option. See the **installp** reference page to use that installation method.

Follow these steps to install the **ssp.csd.hsd** option using SMIT.

TYPE SMIT

▲ The System Management menu appears.

SELECT Software Installation and Maintenance

▲ The Software Installation Maintenance menu appears.

SELECT Install / Update Software

▲ The Install / Update Software menu appears.

SELECT Install / Update Selectable Software (Custom Install)

▲ The Install / Update Selectable Software (Custom Install) menu appears.

SELECT Install Software Products at Latest Available Level

▲ The Install Software Products at Latest Available Level panel appears.

PRESS **List** for the input device / directory for the software and press enter.

SELECT **ssp.csd.hsd** and press enter.

PRESS Enter to start the installation.

▲ **OK** will appear when the install is completed successfully.

If your installation failed, refer to /**smit.log** file to determine the error.

PRESS **CANCEL** to exit SMIT.

ENTER /**usr**/**lpp**/**csd**/**bin**/**hsdatalst** to verify the installation. The command should show that you have no HSDs defined yet.

The HSD commands, such as **defhsd** and **cfghsd**, are in /**usr**/**lpp**/**csd**/**bin**. The HSD is in /**etc**/**drivers**.

Installing the IBM Recoverable Virtual Shared Disk Licensed Program

This section describes how to install and configure the IBM Recoverable Virtual Shared Disk software.

Hardware and Software Requirements

The IBM Recoverable Virtual Shared Disk software operates on IBM Parallel System Support Programs for AIX and requires twin-tailed disks.

The following are the hardware requirements for the IBM Recoverable Virtual Shared Disk:

- Any twin-tailed disk supported by the Logical Volume Manager running on AIX 4.1 is supported

The following are the software requirements for the IBM Recoverable Virtual Shared Disk:

- The AIX 4.1 or 4.2 (when available) operating system
- The IBM Parallel System Support Programs for AIX (PSSP), Version 2 Release 2 (program number 5765-529), which includes the IBM Virtual Shared Disk support package
- The IBM Recoverable Virtual Shared Disk Licensed Program (program number 5765-444).

Detailed Space Requirements

The individual components of the IBM Recoverable Virtual Shared Disk software require the following amounts of space in the specified directories:

Table 1. Space Used by Installation Process

Option Name	root file system	var file system	usr file system
rcsd.vsd	0	0	70KB
rcsd.rvsd	0	0	150KB
rcsd.hahc	0	0	150KB
rcsd.docs	0	0	2MB

Table 2. Additional Space Used during Execution on the Control Workstation

Option Name	root file system	var file system	usr file system
rcsd.vsd	0	0	0
rcsd.rvsd	0	0	0
rcsd.hahc	0	0	0
rcsd.docs	0	0	0

Table 3. Additional Space Used during Execution on the Nodes

Option Name	root file system	var file system	usr file system
rcsd.vsd	0	0	0
rcsd.rvsd	100 KB	0	0
rcsd.hahc	0	0	0
rcsd.docs	0	0	0

The additional space used during execution is for log entries.

An Overview of Planning and Installing IBM Recoverable Virtual Shared Disk and Creating Twin-Tailed Volume Groups

Perform the following steps after all your hardware is operational and the base AIX operating system and the PSSP are installed and operational. The steps involved in defining twin-tailed volume groups on the system are:

1. Plan your twin-tailed volume groups and decide which twin-tailed disks will be in each group.
2. Ensure that the supported disks are properly installed and connected to both nodes.
3. Ensure that the lastest AIX base maintenance level is installed on all the nodes.
4. Ensure that the IBM Virtual Shared Disk software is installed. This is an option of the PSSP described in "Planning to Install IBM Virtual Shared Disks" on page 7.
5. Define your IBM Virtual Shared Disk node information in the System Data Repository (SDR).
6. Create IBM Virtual Shared Disks, identifying the primary and secondary nodes for each. This step automatically creates twin-tailed volume groups and logical volumes. See "Creating IBM Virtual Shared Disks with the createvsd Command" on page 22 for detailed instructions and examples.
7. Disable any user-supplied scripts that issue IBM Virtual Shared Disk commands that change the state of IBM Virtual Shared Disks.
8. Install the IBM Recoverable Virtual Shared Disk software on all IBM Virtual Shared Disk nodes.
9. Do one of the following:
   - Reboot all the IBM Recoverable Virtual Shared Disk nodes.

   or

   - Make sure all the IBM Virtual Shared Disk volumes are stopped; then use the **ha_vsd reset** command on all IBM Virtual Shared Disk nodes. See the **ha_vsd reset** reference page in *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for the complete syntax of the command.
10. If you currently have IBM Virtual Shared Disks defined, see "Migrating IBM Virtual Shared Disk Data" on page 13 for migration instructions.

Installing the IBM Recoverable Virtual Shared Disk Software

This section describes the IBM Recoverable Virtual Shared Disk software install image and how to install the software on a single node. You should install the image on all nodes used as either IBM Virtual Shared Disk server nodes or IBM Virtual Shared Disk client nodes. Refer to the "Performing Software Maintenance" chapter of the *IBM Parallel System Support Programs for AIX: Installation Guide* for details on how to install software on all SP nodes.

You cannot have both recoverable and nonrecoverable IBM Virtual Shared Disk nodes in a single SP system or system partition.

Installation Using SMIT

Follow these steps to install the **rcsd** image using SMIT:

TYPE **smit install_latest**

▲ The install Software Products at Latest Available Level window appears.

ENTER The input device for the IBM Recoverable Virtual Shared Disk software.

PRESS **Do** to display the default install parameters.

PRESS **List** to show options.

SELECT **ALL** to install all the options.

Note: You can select the options individually. You may not want to install the documentation image (**rcsd.docs**) at all the nodes.

PRESS **OK** to complete option selection.

PRESS **Do** to begin installation.

Installation Notes

All the options in the install package, **rcsd.vsd**, **rcsd.rvsd**, **rcsd.hahc**, and **rcsd.hc**, must be installed.

The IBM Parallel System Support Programs for AIX (PSSP) is a prerequisite for all the options in the **rcsd installp** package. If you are running PSSP Version 2 Release 1 on AIX 4.1, **ssp.csd.vsd**, **ssp.csd.cmi**, and **ssp.basic** must all be installed. If you are running PSSP Version 1 Release 2 on AIX 3.2.5, **csd.vsd**, **csd.cmi**, and **ssp.basic** must all be installed.

Note: **ssp.csd.cmi** or **csd.cmi** are only installed on the control workstation.

The **rcsd** package must be installed on all SP nodes that are to be used as either IBM Virtual Shared Disk server nodes or IBM Virtual Shared Disk client nodes. Such nodes must be reset as described in step 11 of “An Overview of Planning and Installing IBM Recoverable Virtual Shared Disk and Creating Twin-Tailed Volume Groups” on page 12.

Migrating IBM Virtual Shared Disk Data

Follow the steps in this section to migrate IBM Virtual Shared Disk data to a twin-tailed logical volume.

1. Bring down any application that uses the old IBM Virtual Shared Disks for the duration of this migration.
2. Suspend, stop, and unconfigure all IBM Virtual Shared Disks involved.
3. Use the **createvsd** command to create the new volume groups, logical volumes, and IBM Virtual Shared Disks.
4. Perform test reads and writes to make sure you've successfully created your new IBM Virtual Shared Disks.

5. Copy your data from your existing (old) IBM Virtual Shared Disks to the newly-defined IBM Virtual Shared Disks by copying the data directly from one logical volume to the other, using the new volume group name from the **createvsd** command. For example:

   ```
   cplv [-v new_volume_group]-y new_logical_volume \
   old_logical_volume
   ```

6. After making sure that the data is copied to the new logical volume, use the **removevsd** command to remove the old IBM Virtual Shared Disk.

Note: If you are using the IBM Virtual Shared Disk in an IBM Virtual Shared Disk striping device (HSD), you must use the **createhsd** command to create the new HSD as well as the underlying IBM Virtual Shared Disks.

Chapter 3. Introducing the IBM Virtual Shared Disk Perspective

This chapter introduces you to the IBM Virtual Shared Disk Perspective, a graphical user interface that lets you perform IBM Virtual Shared Disk tasks. If you have the IBM Recoverable Virtual Shared Disk licensed program installed, you can view and change IBM Recoverable Virtual Shared Disk options with the Perspective, too.

To begin working with the IBM Virtual Shared Disk Perspective, type `spvsd`. The full path for this command is **/usr/lpp/ssp/bin/spvsd**.

The IBM Virtual Shared Disk Perspective Primary Window

Like other windows, Perspectives primary windows have panes. From the top of the primary window, the different areas and panes are:

- The menu bar
- The tool bar
- Three work areas:
  - The Syspars pane
  - The IBM VSD Nodes and Clients pane
  - The IBM VSD/HSD Definitions pane
- The information area

The Menu Bar

This area contains four menu choices: **Window**, **Actions**, **View**, and **Options**.

- From **Window**, you can close the IBM Virtual Shared Disk Perspective primary window and exit from the Perspective
- From **Actions**, you can select the actions that apply to any of the three work areas (IBM VSD Nodes, IBM VSD/HSD Definitions, and Syspars)
- From **View**, you can change how the objects in the IBM Virtual Shared Disk Perspective are displayed; for example, one option lets you select all the objects in a work area
- From **Options**, you can change how the IBM Virtual Shared Disk Perspective itself is displayed; for example, you can set fonts and colors

For more information on the View and Options menu choices, see the Perspectives chapter in *IBM Parallel System Support Programs for AIX: Administration Guide*.

The Tool Bar

The tool bar is displayed right below the menu bar in the primary window. The tool bar consists of icons that represent the most frequently used actions from the Actions choice of the menu bar. For the IBM Virtual Shared Disk Perspective, these are:

© Copyright IBM Corp. 1994, 1996

- The **notebook** icon, which lets you view and modify attributes of each type of object shown in the work areas: syspars (system partitions), IBM VSD nodes, IBM VSD and HSD clients, and IBM VSD and HSD definitions.
- The **select clients** icon, which lets you select the client or clients that are associated with a particular IBM VSD or HSD definition.
- The **expand tree** icon, which lets you fully expand the tree of objects that are related to the selected object. For example, you can click on this icon to show all the clients of a selected IBM VSD node or nodes.
- The **collapse tree** icon, which lets you fully collapse the tree of objects that are related to the selected object (hide them from the work area).
- The **run diagnostics** icon, which lets you run IBM Virtual Shared Disk diagnostics for an IBM VSD node.
- The **select definition** icon, which lets you select the IBM VSD or HSD definition that is associated with a particular IBM VSD or HSD client or clients.
- The **monitor** icon, which lets you monitor conditions associated with IBM VSD nodes, IBM VSD and HSD definitions, and syspars.

The Work Areas

The work areas display the objects that you can work with. There are three work areas in the IBM Virtual Shared Disk Perspective :

- The **syspars** work area, where you can select the current system partition. Each partition will have a different set of associated nodes, clients, and definitions.
- The **IBM VSD nodes** work area, which contains IBM VSD node icons, non-IBM-VSD-node icons, IBM VSD client icons, and HSD client icons. Non-IBM-VSD-node icons represent nodes that do not have IBM Virtual Shared Disks defined on them. The IBM VSD and HSD client icons are hidden until you either
  - click on the arrow to the left of node icons that have clients configured,
  - or select Expand Tree from the View menu item,
  - or click on the Expand Tree icon in the tool bar.

  If no clients are configured on a node, no client icons will be displayed or displayable.
- The **IBM VSD/HSD definitions** work area, where you can select IBM VSD definitions and HSD definitions. HSD definitions are made up of IBM VSD definitions and can be expanded to show the underlying IBM VSD definitions.

If all three work areas are not displayed in your primary window when you start the IBM Virtual Shared Disk Perspective, you can use the Add Pane option at the top of the primary window to add any missing work areas.

The Information Area

The information area gives you a line of help information about each icon in the tool bar. As you move your cursor over an icon, a line of text that describes it appears in the information area.

Chapter 4. Creating and Managing Shared Disks

This chapter tells you how to:

- Enter information about IBM Virtual Shared Disks in the System Data Repository (SDR), using the IBM Virtual Shared Disk Perspective
- Create IBM Virtual Shared Disks, using the **createvsd** command
- Assign backup nodes to IBM Virtual Shared Disks, using the **createvsd** command, if you have the IBM Recoverable Virtual Shared Disk licensed program installed
- Create Data Striping Devices (HSDs), using the **createhsd** command
- Configure and unconfigure IBM Virtual Shared Disks, using the IBM Virtual Shared Disk Perspective
- Verify IBM Virtual Shared Disks, using the IBM Virtual Shared Disk Perspective
- Display information about IBM Virtual Shared Disks and Data Striping Devices, using the IBM Virtual Shared Disk Perspective
- Display and reset IBM Virtual Shared Disk and Data Striping Device node and client statistics, using the IBM Virtual Shared Disk Perspective
- Run IBM Virtual Shared Disk diagnostics, using the IBM Virtual Shared Disk Perspective
- Change the states of IBM Virtual Shared Disks, when you do **not** have the IBM Recoverable Virtual Shared Disk licensed program installed, using the IBM Virtual Shared Disk Perspective
- Remove IBM Virtual Shared Disks and Data Striping Devices, using the **removevsd** and **removehsd** commands.
- Add a server node to an existing IBM Virtual Shared Disk configuration
- Remove a server node from an existing IBM Virtual Shared Disk configuration.

Preparing to Create IBM Virtual Shared Disks and Data Striping Devices

Before you use the **createvsd** and **createhsd** commands, you must have the correct authorization and you must set up the nodes in your IBM Virtual Shared Disk and Data Striping Device configuration by putting some control information into the System Data Repository (SDR) on each node.

Note: If you have already defined IBM Virtual Shared Disks and Data Striping Devices using the commands and SMIT functions that were available before Version 2 Release 2 of the SP System and you want to continue to use those commands and SMIT functions, see Appendix B, “Single-Node Command and SMIT Interfaces for IBM Virtual Shared Disks, Data Striping Devices and the IBM Recoverable Virtual Sha red Disk Licensed Product” on page 89.

© Copyright IBM Corp. 1994, 1996

Authorization

In order to issue commands, such as **createvsd**, that operate on multiple nodes, you must have Kerberos authority and sysctl authorization. Your system administrator should run **/usr/lpp/kerberos/bin/add_principal** to put your ID into the authorized list. The administrator must also add your principal ID to the **/etc/sysctl.acl** and **/etc/sysctl.vsd.acl** files. Then the administrator must run **sysctl svcrestart** to complete the database update for the new entries. The administrator will have to perform these authorization steps on the control workstation and on all nodes that will be IBM Virtual Shared Disk clients or servers.

When you next logon, you may need to run **kinit** to get the correct Kerberos ticket. You can verify your ticket with the **klist** command and run **sysctl whoami** to check your sysctl authorization. To check whether you can run the IBM Virtual Shared Disk and Data Striping Device commands that operate on multiple nodes, use **sysctl sysctl_vsdcheck**. Issue **vsdsklst** as a check before you use any other commands that operate on multiple nodes. If your authorization is correct, it will return information about the IBM Virtual Shared Disks that are defined on the SP system or system partition.

Deleting Old Rollback Files

If you issue an IBM Virtual Shared Disk command that operates on multiple nodes, such as **createvsd** or **createhsd**, and the command fails, a rollback file will be created as part of the recovery processing. A command that is issued against multiple nodes fails if any of the nodes cannot execute it. Before you re-execute such a command, be sure there are no rollback files from failed invocations of these commands on your system. If there are any, they can be found in **/usr/lpp/csd/vsdfiles/vsd_rollback**. You can delete them with the **rm** command. Old rollback files can prevent proper recovery if new invocations of these commands fail.

Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR

The IBM Virtual Shared Disk section of the Perspectives graphical user interface, which you can get to by issuing **/usr/lpp/spp/bin/spvsd** or by double-clicking on the IBM Virtual Shared Disk icon on the Perspectives initial panel, has a notebook page that displays the setup information in the SDR required for each node before you create IBM Virtual Shared Disks or Data Striping Devices. To reach the notebook page for a node that has not been initialized for IBM Virtual Shared Disks,

- Click on the node icon in the IBM Virtual Shared Disk Nodes work area to select it
- Click on the Actions menu item at the top of the IBM Virtual Shared Disk primary window
- Click on IBM Virtual Shared Disk Nodes
- Select the Initialize Node for IBM Virtual Shared Disk action.

The following information must be entered into the SDR before you can create IBM Virtual Shared Disks or Data Striping Devices. For more information about these options, see the description of the **vsdnode** command in *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.

- Node number

Each node is identified by a *node_number* that represents its position in the SP rack.

- IBM Virtual Shared Disk Adapter Name

  The name of the adapter to be used for IBM Virtual Shared Disk communications with this node.

  The IBM Virtual Shared Disk component uses this name and the related node number to get the IBM Virtual Shared Disk IP address associated with each node from the SDR **Adapter** object. Either the High Performance Switch or the Scalable Parallel (SP) Switch (both are defined as **css0**) is recommended as the IBM Virtual Shared Disk adapter for best performance.

  The set of nodes and addresses in the IBM Virtual Shared Disk configuration (the clients and servers) must define a fully-connected network, that is, all nodes can send messages directly to each other. For the best performance, however, all the nodes should be directly connected to each other over the network.

- Initial cache buffer count

  The IBM Virtual Shared Disk device driver implements an optional write-through cache of pinned kernel memory with a block size of 4KB bytes. When the first cached IBM Virtual Shared Disk is configured on a node, the cache is created, and it contains the specified number of blocks. The minimum value is 1. The recommended value is 64, which results in a 256KB cache.

- Maximum cache buffer count

  The number of buffers in the cache can be increased (using **ctlvsd**) up to this value. You cannot decrease the number of cache blocks. You can start over by unconfiguring all the IBM Virtual Shared Disks, changing the value of the this number in the SDR, and then, when you configure the IBM Virtual Shared Disks, the initial number of IBM Virtual Shared Disk buffer cache blocks will be created. The recommended value is 256, resulting in a 1MB cache.

- Number of outstanding logical volume read/write requests

  Specifies the maximum number of outstanding requests the IBM Virtual Shared Disk device driver will allow for each underlying logical volume. Each such request uses a pbuf. A pbuf is a little bigger than a buf structure, which is approximately 104 bytes. The IBM Virtual Shared Disk device driver allocates this number of pbufs in pinned kernel memory whenever an IBM Virtual Shared Disk that has this node as its primary is configured. The minimum value is 1. The recommended value for a primary or secondary node is 48. If the node is not going to be used as a primary or secondary, this value is irrelevant.

- Minimum buddy buffer size

  The smallest buddy buffer a server will use to satisfy a remote request to an IBM Virtual Shared Disk. This value must be a power of 2 and greater than or equal to 4096. The recommended value is 4096 (4KB). For a 512-byte request, 4KB is clearly overkill but a buddy buffer is used only for the short period of time a remote request is being processed at the direct-client node.

  For more information on buddy buffers, see “Buddy Buffers” on page 53.

- Maximum buddy buffer size

  The largest buddy buffer a server will use to satisfy a remote request. This value must be a power of 2 and greater than or equal to the

min_buddy_buffer_size. The maximum value, and the recommended value, is 65536 (64KB). This value must be the same on all nodes within a system partition.

- Number of buddy buffers

  The buddy buffer is pinned kernel memory allocated when the IBM Virtual Shared Disk device driver is loaded the first time an IBM Virtual Shared Disk is configured (and freed when the last IBM Virtual Shared Disk is unconfigured). If you do not use the switch as your IP adapter, the size of the buddy buffer affects the number of remote requests the IBM Virtual Shared Disk server node can handle at one time. Remote requests can queue waiting for a buddy buffer. The **statvsd** command and the IBM Virtual Shared Disk Statistics notebook page of the IBM Virtual Shared Disk Perspective report this queuing as buddy buffer shortages. See "Buddy Buffers" on page 53 for more information about buddy buffers and how to determine how many to define. The recommended value when the switch is used as the IP adapter is no more than 4, which results in a 256KB buddy buffer size in combination with the recommended maximum buddy buffer size of 64KB. Tuning may be required on the switch.

- Request block count

  Specifies the maximum number of outstanding IBM Virtual Shared Disk requests originating on the node. The IBM Virtual Shared Disk device driver allocates this number of request blocks in pinned kernel memory the first time an IBM Virtual Shared Disk is configured. If the number is too small, local requests will queue up waiting for a request block to become available. The recommended value is 256. The size of the block is approximately 76 bytes.

- Maximum IP message size

  The maximum IP message size for IBM Virtual Shared Disks, in bytes. The default value is 24KB (24,576). If you are using the switch as your IBM Virtual Shared Disk adapter, use a value of 60KB (61,440).

- Level of parallelism

  The IBM Virtual Shared Disk component divides I/O requests into smaller units. This allows one unit to be transmitted, while another is being placed in the buffer. The default value is 9 and it should always be used.

You can display and change the IBM Virtual Shared Disk node attributes in the SDR from this notebook page at any time. You must unconfigure and reconfigure the IBM Virtual Shared Disk definitions on the node before the changes take effect, however.

You can also use the **vsdnode** command or SMIT to enter node information in the SDR. To see this and other SMIT interfaces for managing IBM Virtual Shared Disks, see "Command and SMIT Interfaces for IBM Virtual Shared Disks" on page 89.

Creating IBM Virtual Shared Disks with the createvsd Command

You can create several IBM Virtual Shared Disks with a single command, including secondary disks if you have the IBM Recoverable Virtual Shared Disk licensed program installed. You must first have used the IBM Virtual Shared Disk Perspective or the **vsdnode** command to set up information in the SDR about each node involved in this IBM Virtual Shared Disk configuration.

The syntax of the **createvsd** command is:

createvsd –**n** {*node_list* | **ALL**}
–**s** *size_in_MB* –**g** *volume_group_name*
[{–**c** *number_of_vsds_per_node* | –**L**}]
[–**o cache** | **nocache**][–**m** *mirror_count*]
[–**p** *lvm_stripe_size*] [–**v** *vsd_name_prefix*]
[–**l** *lv_name_prefix*] [–**T** *lp_size*]

Note: Some examples shown in this list do not contain enough flags to be executable. They are shown in an incomplete form to illustrate specific flags.

–**n** Specifies the nodes on which you are creating IBM Virtual Shared Disks. The backup node cannot be the same as the primary node.

node_list Given in the format **P**/**S:***hdisk_list1*+*hdisk_list2*/, where **P** is the primary node, **S**, if specified, is the backup (secondary) node, *hdisk_list1* is the list of local physical disks in the logical volume on the primary, and *hdisk_list1*+*hdisk_list2* is the list of local physical disks in the volume group on the primary, if you want to have more disks in the volume group than are needed for the logical volume. The sequence in which nodes are listed determines the names given to the IBM Virtual Shared Disks. For example:

```
createvsd -n 1,6,4 -v PRE
```

(with the *vsd_prefix* PRE) creates IBM Virtual Shared Disks PRE1n1 on node 1, PRE2n6 on node 6, and PRE3n4 on node 4.

To create a volume group that spans hdisk2, hdisk3, and hdisk4 on node 1, with a backup on node 3, enter:

```
createvsd -n 1/3:hdisk2,hdisk3,hdisk4/ -v DATA
```

This command creates:

- IBM Virtual Shared Disk DATA1n1b3 with logical volume lvDATA1n1b3 on a volume group with the global volume group name DATA1n1b3 on node 1, exported to node 3. The Logical Volume Manager (LVM) volume group name is DATA. The logical volumes span hdisk2, hdisk3, and hdisk4.

To create volume groups just like that one on nodes 1, 2, and 3 of a system with backup on nodes 4, 5, and 6 of the same system, enter:

```
createvsd -n 1/4:hdisk1,hdisk2,hdisk3/,2/5:hdisk5,hdisk6, \
hdisk7/,3/6:hdisk2,hdisk4,hdisk6/ -v DATA
```

This command is shown on two lines here, but you must enter it without any spaces between the items in *node_list*.

The command creates:

- IBM Virtual Shared Disk DATA1n1b4 with logical volume lvDATA1n1b4 on a volume group with the local volume group name DATA on node 1, exported to node 4. The global volume group name is DATAn1b4.
- IBM Virtual Shared Disk DATA2n2b5 with logical volume lvDATA2n2b5 on a volume group with the local volume group

name DATA on node 2, exported to node 5. The global volume group name is DATAn2b5.

- IBM Virtual Shared Disk DATA3n3b6 with logical volume lvDATA3n3b6 on a volume group with the local volume group name DATA on node 3, exported to node 6. The global volume group name is DATAn3b6.

To create an IBM Virtual Shared Disk where the logical volume spans only two of the physical disks in the volume group, enter:

```
createvsd -n 1/3:hdisk1,hdisk2+hdisk3/ -v DATA
```

This command creates the IBM Virtual Shared Disk DATA1n1b3 with logical volume lvDATA1n1b3 spanning hdisk1 and hdisk2 in the volume group DATA, which includes hdisk1, hdisk2, and hdisk3. It exports the volume group DATA to node 3.

If a volume group is already created and the combined physical hdisk lists contain disks that are not needed for the logical volume, those hdisks are added to the volume group. If the volume group has not already been created, **createvsd** creates a volume group that spans *hdisk_list1+hdisk_list2*.

Backup nodes cannot use the same physical disk as the primary does to serve IBM Virtual Shared Disks.

ALL Specifies that you are creating IBM Virtual Shared Disks on all nodes in the system or system partition. No backup nodes are assigned if you use this operand. The IBM Virtual Shared Disks will be created on all the physical disks attached to the nodes in *node_list* (you cannot specify which physical disks to use.)

–s Specifies the size in megabytes of each IBM Virtual Shared Disk

–g Specifies the Logical Volume Manager (LVM) volume group name. This name is concatenated with the node number to produce the global volume group name. For example:

```
createvsd -n 6 -g VSDVG
```

creates a volume group with the local volume group name VSDVG and the global volume group name VSDVG1n6 on node 6. The node number is added to the prefix to avoid name conflicts when a backup node takes over a volume group. If a backup node exists, the global volume group name will be concatenated with the backup node number as well as the primary. For example:

```
createvsd -n 6/3/ -g VSDVG
```

creates a volume group with the local volume group name VSDVG and the global volume group name VSDVGn6b3. The primary node is node 6 and the backup node for this volume group is node 3.

–c Specifies the number of IBM Virtual Shared Disks to be created on each node. If *number_of_vsds_per_node* is not specified, one IBM Virtual Shared Disk is created for each node specified on **createvsd**. If more than one IBM Virtual Shared Disk is to be created for each node, the names will be allocated alternately. For example:

```
createvsd -n 1,6 -c 2 -v DATA
```

creates IBM Virtual Shared Disks DATA1n1 on node 1, DATA2n6 on node 6, DATA3n1 on node 1, and DATA4n6 on node 6.

–**L** The **-L** option allows you to create one IBM Virtual Shared Disk on each node without using sequential numbers, for locally-accessed IBM Virtual Shared Disks.

–**A** Specifies that IBM Virtual Shared Disk names will be allocated to each node in turn, for example:

```
createvsd -n 1,6 -c 2 -h DATA
```

creates DATA1n1 and DATA2n1 on node 1, and DATA3n6 and DATA4n6 on node 6.

–**o** Specifies either the **cache** or the **nocache** option.

–**m** Specifies the LVM mirroring count. The mirroring count sets the number of physical partitions allocated to each logical partition. The range is from 1 to 3 and the default value is 1.

–**p** Specifies the LVM stripe size. If this flag is not specified, the logical volumes are not striped. In order to use striping, the node on which the IBM Virtual Shared Disks are defined must have more than one physical disk.

–**v** Specifies a prefix to be given to the names of the created IBM Virtual Shared Disks. This prefix will be concatenated with the IBM Virtual Shared Disk number, node number, and backup node number, if a backup disk is specified. For example, if the prefix PRE is given to an IBM Virtual Shared Disk created on node 1 and there are already two IBM Virtual Shared Disks with this prefix across the partition, the new IBM Virtual Shared Disk name will be PRE3n1. The name given to the underlying logical volume will be lvPRE3n1, unless the –**l** flag is used. The **createvsd** command continues to sequence IBM Virtual Shared Disk names from the last PRE-prefixed IBM Virtual Shared Disk

If **-v** is not specified, the prefix `vsd` is used.

Note: The last character of the *vsd_name_prefix* cannot be a digit. Otherwise, the 11th IBM Virtual Shared Disk with the prefix PRE would have the same name as the first IBM Virtual Shared Disk with the prefix PRE1. Nor can the *vsd_name_prefix* contain the character '**.**' because '**.**' can be any character in regular expressions.

–**l** Overrides the prefix *lvx* that is given by default to a logical volume by the **createvsd** command, where *x* is the IBM Virtual Shared Disk name prefix specified by *vsd_name_prefix* or the default (`vsd`). For example:

```
createvsd -n 1 -v DATA
```

creates one IBM Virtual Shared Disk on node 1 named DATA1n1 with an underlying logical volume lvDATA1n1. If the command

```
createvsd -n 1 -v DATA -l new
```

is used, the IBM Virtual Shared Disk on node 1 is still named DATA1n1, but the underlying logical volume is named lvnew1n1.

It is usually more helpful **not** to specify **-l**, so that your lists of IBM Virtual Shared Disk names and logical volume names are easy to associate with each other and you avoid naming conflicts.

–**T** Specifies the size of the physical partition in the Logical Volume Manager logical volume group and also the logical partition size (they will be the same) in megabytes. You must select a power of 2 in the range 2—256. The default is 4MB.

The Logical Volume Manager limits the number of physical partitions to 1016 per disk. If a disk is greater than 4 gigabytes in size, the physical partition size must be greater than 4MB to keep the number of partitions under the limit.

Information about IBM Virtual Shared Disk definitions is stored in the System Data Repository (SDR) **VSD_Global_Volume_Group** object and the **VSD_Table** object. For the complete command description, see *IBM Parallel System Support Programs for AIX: Commands and Technical Reference*.

Examples of Creating IBM Virtual Shared Disks with createvsd

The following examples range from simple to fairly complex.

- To create identical IBM Virtual Shared Disk definitions on each of three disks in a system partition, type:

  ```
  createvsd –n 1,3,5 -s 4 -g SYS2VG -v SYS2VSD
  ```

  This creates the following IBM Virtual Shared Disk definitions:

  – SYS2VSD1n1 on node 1. The local volume group name on node 1 is SYS2VG. The global volume group name is SYS2VGn1. The logical volume is lvSYS2VSD1n1.

  – SYS2VSD2n3 on node 3. The local volume group name on node 3 is SYS2VG. The global volume group name is SYS2VGn3. The logical volume is lvSYS2VSD2n3.

  – SYS2VSD3n5 on node 5. The local volume group name on node 5 is SYS2VG. The global volume group name is SYS2VGn5. The logical volume is lvSYS2VSD3n5.

  No secondary disks are defined. The space allocated to the IBM Virtual Shared Disk is spread across all the physical disks (hdisks) connected to each node (1,3, and 5).

- To assign each disk in the previous example a secondary direct client node, (with the IBM Recoverable Virtual Shared Disk licensed product installed) type:

  ```
  createvsd –n 1/2/,3/4/,5/6/ -s 4 -g SYS2VG -v SYS2VSD
  ```

  This adds the backup node number to the naming convention and creates the following IBM Virtual Shared Disk definitions:

  – SYS2VSD1n1b2 on node 1 with a twin-tailed connection to a second direct client on node 2. The local volume group name on node 1 is SYS2VG. The global volume group name is SYS2VGn1b2. The logical volume is lvSYS2VSD1n1b2.

  – SYS2VSD2n3b4 on node 3 with a twin-tailed connection to a second direct client on node 4. The local volume group name on node 3 is SYS2VG. The

global volume group name is SYS2VGn3b4. The logical volume is lvSYS2VSD2n3b4.

  - SYS2VSD3n5b6 on node 5 with a twin-tailed connection to a second direct client on node 6. The local volume group name on node 5 is SYS2VG. The global volume group name is SYS2VGn5b6. The logical volume is lvSYS2VSD3n5b6.

- To create three IBM Virtual Shared Disk definitions on primary and secondary nodes (with the IBM Recoverable Virtual Shared Disk licensed product installed), where the logical volume created on nodes 3 and 4 spans two disks and the volume group spans three disks, type:

  ```
  createvsd -n 3/4:hdisk1,hdisk2+hdisk3/,5/6/,7/8/ -s 4 \
            -g datavg -v USER
  ```

  This command creates the following IBM Virtual Shared Disk definitions:

  - USER1n3b4, with logical volume lvUSER1n3b4 defined on a volume group with the global volume group name datavgn3b4 on node 3, imported to node 4. The volume group datavgn3b4 spans hdisk1, hdisk2, and hdisk3. The logical volume lvUSER2n3b4 spans hdisk1 and hdisk2.
  - USER2n5b6, with logical volume lvUSER2n5b6 defined on a volume group with the global volume group name datavgn5b6 on node 5, imported to node 6.
  - USER3n7b8, with logical volume lvUSER3n7b8 defined on a volume group with the global volume group name datavgn7b8 on node 7, also imported to node 8.

    The volume groups datavgn5b6 and datavgn7b8 are created with one 4MB partition from a single physical disk. The volume group datavgn3b4 is created with one 4MB partition from the three physical disks hdisk1, hdisk2, and hdisk3.

Creating Data Striping Devices with the createhsd Command

You can create a whole configuration of Data Striping Devices with a single command, including secondary disks for the underlying IBM Virtual Shared Disks if you have the IBM Recoverable Virtual Shared Disk licensed program installed. The syntax of the **createhsd** command is:

createhsd –**n** {*node_list* | ALL}
–**s** *size_in_MB* –**g** *volume_group_name*
–**t** *stripe_size_in_KB* [{–**c** *number_of_vsds_per_node* | –**A**}]
[–**o cache** | **nocache**] [–**m** *mirror_count*]
[–**d** *hsd_name_prefix*]
[–**l** *lv_name_prefix*] [–**S**]
[–**T** *lp_size*]

Note: Some examples shown in this list do not contain enough flags to be executable. They are shown in an incomplete form to illustrate specific flags.

–**n** Specifies the numbers of the nodes on which you are creating a data striping device (HSD). The backup node for the underlying IBM Virtual Shared Disks cannot be the same as the primary node.

node_list Given in the format **P**/**S:***hdisk_list1*+*hdisk_list2*/, where **P** is the primary node, **S**, if specified, is the backup (secondary) node, *hdisk_list1* is the list of local physical disks in the logical volume on the primary, and *hdisk_list1*+*hdisk_list2* is the list of local physical disks in the volume group on the primary, if you want to have more disks in the volume group than are needed for the logical volume.

The sequence in which nodes are listed determines the names given to the IBM Virtual Shared Disks; for example:

```
createhsd -n 1,6,4 -d DATA
```

(with the *hsd_prefix* DATA) creates IBM Virtual Shared Disks DATA1n1 on node 1, DATA2n6 on node 6, and DATA3n4 on node 4, which make up a single HSD named DATA. To create volume groups that span specified disks on nodes 1, 2, and 3 of a system with backup on nodes 4, 5, and 6 of the same system, and that make up a single HSD, enter:

```
createhsd -n 1/4:hdisk1,hdisk2,hdisk3/,2/5:hdisk5,hdisk6, \
          hdisk7/,3/6:hdisk2,hdisk4,hdisk6/ \
          -d DATA -s 12 -g OLD -t 4096
```

This command is shown on two lines here, but you must enter it without any spaces between the items in *node_list*.

The command creates:

- IBM Virtual Shared Disk DATA1n1b4 with logical volume lvDATA1n1b4 on a volume group with the local volume group name OLD on node 1, exported to node 4. The global volume group name is OLDn1b4.
- IBM Virtual Shared Disk DATA2n2b5 with logical volume lvDATA2n2b5 on a volume group with the local volume group name OLD on node 2, exported to node 5. The global volume group name is OLDn2b5.
- IBM Virtual Shared Disk DATA3n3b6 with logical volume lvDATA3n3b6 on a volume group with the local volume group name OLD on node 3, exported to node 6. The global volume group name is OLDn3b6.
- An HSD named DATA that comprises the three IBM Virtual Shared Disks listed previously. The HSD is 12MB in total usable size and has a 512KB stripe size.

If a volume group is already created and the combined physical hdisk lists contain disks that are not needed for the logical volume, those hdisks are added to the volume group. If the volume group has not already been created, **createhsd** creates a volume group that spans *hdisk_list1*+*hdisk_list2*.

Backup nodes cannot use the same physical disk as the primary does to serve IBM Virtual Shared Disks.

ALL Specifies that you are creating HSDs on all nodes in the system or system partition. If you use **ALL**, you can't assign backup nodes for the disks.

-s Specifies the total usable size of the HSD in MB. Unless **-S** is specified, **createhsd** adds at least a stripe size to each IBM Virtual Shared Disk's size for each HSD.

-g Specifies the Logical Volume Manager (LVM) volume group name, or local volume group name. This name is concatenated with the node number to form the global volume group name (VSD_GVG). For example:

```
createhsd -n 6 -g VSDVG
```

creates a volume group with the IBM Virtual Shared Disk global volume group name VSDVGn6 and the local AIX volume group name VSDVG. The node number is added to the prefix for the global volume group name to avoid name conflicts with the name used for this volume group on other nodes. If a backup node exists, the global volume group name will be concatenated with the backup node number as well as the primary. For example:

```
createhsd -n 6/3/ -g VSDVG
```

creates VSDVGn6b3, where the primary node is node 6 and the backup node for this volume group is node 3.

-c Specifies the number of IBM Virtual Shared Disks to be created on each node. If *number_of_vsds_per_node* is not specified, one IBM Virtual Shared Disk is created for each node specified on **createvsd**. If more than one IBM Virtual Shared Disk is to be created for each node, the names will be allocated cyclically. For example:

```
createhsd -n 1,6 -c 2 -d DATA
```

creates IBM Virtual Shared Disks DATA1n1 on node 1, DATA2n6 on node 6, DATA3n1 on node 1, and DATA4n6 on node 6 and uses them to make up the HSD DATA.

-A Specifies that IBM Virtual Shared Disk names will be allocated to each node in turn. For example:

```
createhsd -n 1,6 -c 2 -g DATA
```

creates DATA1n1 and DATA2n1 on node 1, and DATA3n6 and DATA4n6 on node 6.

-o Specifies the use of either the **cache** or the **nocache** option for the underlying IBM Virtual Shared Disks.

-m Specifies the LVM mirroring count. The mirroring count sets the number of physical partitions allocated to each logical partition. The range is from 1 to 3. If **-m** is not specified, the count is set to 1.

-t Specifies the stripe size in kilobytes that an HSD will use. The stripe size must be a multiple of 4KB and less than or equal to 1GB.

-d Specifies the name assigned to the created HSD. It is used as the IBM Virtual Shared Disk prefix name (the **-v** in **createvsd**). If an HSD name is not specified, a default name, *x*HsD is used, where x denotes a sequence number.

The command:

```
createhsd -n 1,2 -h DATA
```

creates two IBM Virtual Shared Disks, DATA1n1 and DATA2n2. These IBM Virtual Shared Disks make up one HSD named DATA.

–l Overrides the prefix *lvx* that is given by default to a logical volume by the **createvsd** command, where *x* is the IBM Virtual Shared Disk name prefix specified by *vsd_name_prefix* or the default (`vsd`). For example:

```
createhsd -n 1 -v DATA
```

creates one IBM Virtual Shared Disk on node 1 named DATA1n1 with an underlying logical volume lvDATA1n1. If the command

```
createhsd -n 1 -v DATA -l new
```

is used, the IBM Virtual Shared Disk on node 1 is still named DATA1n1, but the underlying logical volume is named lvnew1n1.

It is usually more helpful **not** to specify **–l**, so that your lists of IBM Virtual Shared Disk names and logical volume names are easy to associate with each other and you avoid naming conflicts.

–S Specifies that the HSD overrides the default skip option and does not skip the first stripe to protect the first LVM Control Block (LVCB).

–T Specifies the size of the physical partition in the Logical Volume Manager logical volume group and also the logical partition size (they will be the same) in megabytes. You must select a power of 2 in the range 2—256. The default is 4MB.

The Logical Volume Manager limits the number of physical partitions to 1016 per disk. If a disk is greater than 4 gigabytes in size, the physical partition size must be greater than 4MB to keep the number of partitions under the limit.

For the detailed command description, see *IBM Parallel System Support Programs for AIX: Commands and Technical Reference*.

Examples of Creating Data Striping Devices with createhsd

The following examples range from simple to fairly complex.

- To create an Data Striping Device that stripes data across three identical IBM Virtual Shared Disk definitions on each of three disks in a system partition, type:

  ```
  createhsd –n 1,3,5 -s 12 -g SYS2VG -t 4 -d SYS2HSD
  ```

  This creates the Data Striping Device definition SYS2HSD and its underlying IBM Virtual Shared Disk definitions:

  – SYS2HSD1n1 on node 1. The local volume group name on node 1 is SYS2VG. The global volume group name is SYS2VGn1. The logical volume is lvSYS2HSD1n1.

  – SYS2HSD2n3 on node 3. The local volume group name on node 3 is SYS2VG. The global volume group name is SYS2VGn3. The logical volume is lvSYS2HSD2n3.

  – SYS2HSD3n5 on node 5. The local volume group name on node 5 is SYS2VG. The global volume group name is SYS2VGn5. The logical volume is lvSYS2HSD3n5.

The usable Data Striping Device size is 12MB. The stripe size is 4KB. The first stripe on each disk is skipped to allow space for the LVCB.

No secondary disks are defined. The space allocated to the Data Striping Device is spread across all the physical disks (hdisks) connected to each node (1,3, and 5).

- With the IBM Recoverable Virtual Shared Disk licensed product installed, to create a Data Striping Device
  definition whose underlying IBM Virtual Shared Disks are twin-tailed and have secondary direct client nodes assigned, type:

  ```
  createhsd –n 1/2/,3/4/,5/6/ -s 48 -g SYS2VG -t 8 -d SYS2HSD
  ```

  This adds the backup node number to the naming convention and creates Data Striping Device SYS2HSD with the following IBM Virtual Shared Disk definitions:

  - SYS2HSD1n1b2 on node 1 with a twin-tailed connection to a second direct client on node 2. The local volume group name on node 1 is SYS2VG. The global volume group name is SYS2VGn1b2. The logical volume is lvSYS2HSD1n1b2.
  - SYS2HSD2n3b4 on node 3 with a twin-tailed connection to a second direct client on node 4. The local volume group name on node 3 is SYS2VG. The global volume group name is SYS2VGn3b4. The logical volume is lvSYS2HSD2n3b4.
  - SYS2HSD3n5b6 on node 5 with a twin-tailed connection to a second direct client on node 6. The local volume group name on node 5 is SYS2VG. The global volume group name is SYS2VGn5b6. The logical volume is lvSYS2HSD3n5b6.

    The usable Data Striping Device size is 48MB. The stripe size is 8KB. The first stripe on each disk is skipped to allow space for the LVCB.

- With the IBM Recoverable Virtual Shared Disk licensed product installed, to create an Data Striping Device with three underlying IBM Virtual Shared Disk definitions on primary and secondary direct clients, where the logical volumes and volume groups span two physical disks, type:

  ```
  createhsd -n 3/4:hdisk1,hdisk2/,5/6:hdisk1,hdisk2/, \
            7/8:hdisk1,hdisk2/ \
            -s 48 -t 12 -S -g datavg -d USER
  ```

  This command creates the Data Striping Device USER, with the following underlying IBM Virtual Shared Disk definitions:

  - USER1n3b4, with logical volume lvUSER1n3b4 defined on a volume group with the global volume group name datavgn3b4 on node 3, imported to node 4. The volume group datavgn3b4 and the logical volume lvUSER2n3b4 span hdisk1 and hdisk2.
  - USER2n5b6, with logical volume lvUSER2n5b6 defined on a volume group with the global volume group name datavgn5b6 on node 5, imported to node 6. The volume group datavgn5b6 and the logical volume lvUSER2n5b6 span hdisk1 and hdisk2.
  - USER3n7b8, with logical volume lvUSER3n7b8 defined on a volume group with the global volume group name datavgn7b8 on node 7, imported to node 8. The volume group datavgn7b8 and the logical volume lvUSER2n7b8 span hdisk1 and hdisk2.

The usable size of the Data Striping Deviceis 48MB. The stripe size is 12KB. The first stripe on each disk is **not** skipped to allow space for the LVCB.

Defining Data Striping Devices in a Node-Pair Configuration

If your SP system or system partition is configured in a node-pair arrangement like that shown in Figure 5, where sets of physical disks are twin-tailed on nodes 1 and 2, nodes 3 and 4, nodes 5 and 6, and so forth, you will need to use a special process to create a single Data Striping Device for the system partition. In this configuration, odd-numbered nodes are the backups for IBM Virtual Shared Disks defined on even-numbered nodes and even-numbered nodes are the backups for IBM Virtual Shared Disks defined on odd-numbered nodes.

NPAIR

Figure 5. A Node-Pair Configuration

To set up the Data Striping Device for this configuration, you use **createvsd** to define one or more IBM Virtual Shared Disks on the odd-numbered nodes, and then to define one or more IBM Virtual Shared Disks on the even-numbered nodes. IBM Virtual Shared Disks on odd-numbered nodes must be defined separately from those on even-numbered nodes in this configuration if they are to be part of a single Data Striping Device. You then use **defhsd** to define the Data Striping Device.

For example, to define an Data Striping Device for the configuration in Figure 5, you could use the following sequence of commands:

```
createvsd 1/2/,3/4/,5/6/,7/8/ –s 8192 –g s21ovg –v ovsd
createvsd 2/1/,4/3/,6/5/,8/7/ –s 8192 –g s21evg –v evsd
```

```
defhsd protect_lvcb s21hsd 8192 ovsdn1b2 ovsdn3b4 ovsdn5b6 ovsdn7b8
evsdn2b1 evsdn4b3 evsdn6b5 evsdn8b7
```

Configuring and Unconfiguring IBM Virtual Shared Disks and Data Striping Devices

After you've created your IBM Virtual Shared Disks or Data Striping Devices, you must configure them so that all the clients can read from and write to them. To configure or unconfigure all the IBM Virtual Shared Disks or Data Striping Devices on a node:

- Select the node in the IBM Virtual Shared Disk Nodes work area
- Click on Actions in the menu bar
- Click on IBM Virtual Shared Disk Nodes
- From the Actions list, click on Configure all IBM Virtual Shared Disks, Unconfigure all IBM Virtual Shared Disks, Configure all Data Striping Devices, or Unconfigure all Data Striping Devices

To unconfigure and reconfigure the IBM Virtual Shared Disks or Data Striping Devices in a single step, click on Reconfigure all IBM Virtual Shared Disks or Reconfigure all Data Striping Devices.

You can also unconfigure or reconfigure an IBM Virtual Shared Disk or Data Striping Device client. To do so, select the client icon in the IBM Virtual Shared Disk Nodes work area and then follow the same steps as in the previous paragraph. The highlighted options in the Actions list will be Unconfigure and Reconfigure.

You can also use the **cfgvsd**, **cfghsdvsd**,**ucfgvsd**, or **ucfghsdvsd** commands or SMIT.

Verifying an IBM Virtual Shared Disk

CAUTION:
The verification procedure writes data to the IBM Virtual Shared Disks. Do not select this action after you have put real data on a IBM Virtual Shared Disk.

To test that you have successfully installed the IBM Virtual Shared Disk software and that you can successfully define, activate, read from, and write to an IBM Virtual Shared Disk, follow these steps after creating your IBM Virtual Shared Disk definition with **createvsd**:

1. Bring up the IBM Virtual Shared Disk Perspective by typing **spvsd** or by selecting the IBM Virtual Shared Disk icon from the IBM Virtual Shared Disk Perspectives main panel.
2. Look for the icon that represents the primary node for your new IBM Virtual Shared Disk definition in the IBM Virtual Shared Disk Nodes work area. If you created your IBM Virtual Shared Disk definitions successfully, the icon will look like a disk in color. It will not have an international negative sign over it. Click on the icon to select it.

3. Configure the IBM Virtual Shared Disk definition on your system by clicking on Actions in the menu bar and then double-clicking on Configure All IBM Virtual Shared Disks.
4. Change the state of all clients on this IBM Virtual Shared Disk to Active by selecting Actions again and choosing Change IBM Virtual Shared Disk States. Use the Change States dialog box to select Start. You'll see a warning message telling you not to use this command if you have the IBM Recoverable Virtual Shared Disk licensed program installed. This is because IBM Virtual Shared Disks are usually activated through Group Services when the IBM Recoverable Virtual Shared Disk product is in the system. You may use this Start action right after you create your IBM Virtual Shared Disks, however, in order to verify the configuration, even if you have the IBM Recoverable Virtual Shared Disk product installed.
5. The client icons change from stopped to active, confirming that this IBM Virtual Shared Disk is now in the Active state.
6. Display configuration information about the IBM Virtual Shared Disk by clicking on the notebook icon in the tool bar. Select the notebook page labelled IBM Virtual Shared Disk Node Attributes. The display should show that your IBM Virtual Shared Disk is in the active state.
7. Confirm that your IBM Virtual Shared Disks are working properly by exiting from the notebook page and double-clicking on Run Verification Tests under Actions. You'll see a couple of warning messages here, too, since the verification tests write data to the IBM Virtual Shared Disks and would overwrite any data you had on them. The operation should complete quickly, depending on the size of your IBM Virtual Shared Disk configuration, and you should get the message:

   ```
   VSD verification test successful!
   ```

You will need to use this procedure from each IBM Virtual Shared Disk client node.

You can also use the **vsdvts** command to verify all of your IBM Virtual Shared Disk clients.

Displaying Information about IBM Virtual Shared Disk and Data Striping Device Nodes

To display information about IBM Virtual Shared Disk and Data Striping Device nodes, select the node in the IBM Virtual Shared Disk Nodes work area. Bring your cursor up to the tool bar and click on the notebook icon. The IBM Virtual Shared Disk Node Attributes notebook page reflects the options set in the SDR for this node when it was defined (see “Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR” on page 20).

The HSD Node Attributes notebook page shows you the level of parallelism for this Data Striping Device.

Displaying Information about IBM Virtual Shared Disk Clients

To view client attributes for IBM Virtual Shared Disks, select the client whose attributes you want to view and click on the notebook icon in the tool bar. The Client Attributes page shows the following information:

- Name of the IBM Virtual Shared Disk
- Node number on which this client is configured
- Primary node for this IBM Virtual Shared Disk
- Secondary node for this IBM Virtual Shared Disk, if one is defined
- Minor number of this client
- Physical disks spanned by this IBM Virtual Shared Disk
- Local volume group name
- Global volume group name
- Logical volume name
- Logical volume minor number
- IBM Virtual Shared Disk Client reads
- IBM Virtual Shared Disk Client writes
- State of the IBM Virtual Shared Disk
- Owner of this IBM Virtual Shared Disk
- Group to which the owner of this IBM Virtual Shared Disk belongs

The Client Values page of this notebook shows you the values of the following information for this client:

- The adapter on which this IBM Virtual Shared Disk is defined
- Initial cache buffer count for this client
- Maximum cache buffer count
- Outstanding logical volume read and write requests
- Minimum buddy buffer size
- Maximum buddy buffer size
- Number of buddy buffers
- Request block count
- Maximum IP message size
- Level of parallelism
- Current setting of the cache option

See “Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR” on page 20 for descriptions of these parameters.

The System Values page of this notebook shows you the values of the information shown above that were set for the node on which this IBM Virtual Shared Disk was defined.

You can use the **vsdsklst** and **vsdatalst** commands or SMIT to get the same information.

Displaying Information about Data Striping Device Definitions and Clients

To view definition attributes for Data Striping Devices, select the definition whose attributes you want to view and click on the notebook icon in the tool bar. The IBM HSD Definition Attributes page shows the following information:

- Name of the HSD
- Minor number of this client
- Stripe size
- Status of the Protect LVCB option (protect or not-protect)

From the same notebook page, you can click on the HSD Clients tab to see the clients on which this definition has been configured.

To view client attributes for Data Striping Devices, select the HSD client icon whose attributes you want to view and click on the notebook icon in the tool bar. The system displays the IBM Client Attributes notebook page, which shows you the following information:

- Name of the HSD
- Node number on which this HSD client is configured
- Minor number of this client
- Stripe size
- HSD level of parallelism
- IBM Virtual Shared Disk client reads
- IBM Virtual Shared Disk client writes
- Owner of this HSD
- Group of this HSD

The Client Values—HSD Client Attributes, and System Values—HSD Client Attributes tabs show you the settings of the *protect_lvcb* option.

Displaying IBM Virtual Shared Disk Node Statistics

You can use the IBM Virtual Shared Disk Perspective to display the IBM Virtual Shared Disk node statistics and other information about the IBM Virtual Shared Disks defined on a node. The information is displayed on the node attributes notebook page. To open this page, click on the appropriate IBM Virtual Shared Disk node icon in the IBM Virtual Shared Disk Nodes work area to select it. Then click on the notebook icon in the tool bar. When the notebook page is displayed, click on the tab at the bottom of the page that says IBM Virtual Shared Disk Node Statistics. (You may have to page to the right, using the small arrow button at the bottom of the screen, to see the tab for IBM Virtual Shared Disk Node Statistics.) The following statistics are displayed:

- Status of the IP checksum counters
- SP system product level
- Level of parallelism
- Maximum IP message size
- Requests queued waiting for various resources, such as pbufs
- Average buddy buffer wait queue size
- Rejected operations and the reasons for the rejection
- 64–byte unaligned reads
- Communications buffer pool shortages
- Retries

Information about the status of each IBM Virtual Shared Disk node, including:

- Node name
- Volume group name

- Partition size
- Physical disks and the space in use on them
- IBM Virtual Shared Disks defined on this node

This information is also available from the **statvsd** and **vsdsklst** commands.

To display usage information about the IBM Virtual Shared Disk device driver using SMIT, the **statvsd** command, and the **vsdsklst** command, see Appendix B, "Single-Node Command and SMIT Interfaces for IBM Virtual Shared Disks, Data Striping Devices and the IBM Recoverable Virtual Sha red Disk Licensed Product" on page 89.

Displaying Data Striping Device Node Statistics

You can use the IBM Virtual Shared Disk Perspective to display Data Striping Device node statistics and other information about the HSDs defined on a node. The information is displayed on the node attributes notebook page. To open this page, click on the appropriate node icon in the IBM Virtual Shared Disk Nodes work area to select it. Then click on the notebook icon in the tool bar. When the notebook page is displayed, click on the tab at the bottom of the page that says Data Striping Device Node Statistics. (You may have to page to the right, using the small arrow button at the bottom of the screen, to see the tab for Data Striping Device Node Statistics.)

Resetting IBM Virtual Shared Disk and Data Striping Device Statistics

IBM Virtual Shared Disk and Data Striping Device statistics are cumulative. If you want to use them to see the results of specific parameter changes, you must reset them, make the change, and then read them after the appropriate time period.

You can reset IBM Virtual Shared Disk and Data Striping Device statistics with the IBM Virtual Shared Disk Perspective. In the IBM Virtual Shared Disk Nodes work area of the IBM Virtual Shared Disk Perspective:

- Click on the appropriate node icon to select it
- Click on Actions from the menu bar at the top of the page
- Select IBM Virtual Shared Disk Nodes
- Select Reset All IBM Virtual Shared Disk Statistics or Reset All Data Striping Device Statistics.

Running IBM Virtual Shared Disk Diagnostics

If you have problems with your IBM Virtual Shared Disk or Data Striping Device configuration, your IBM service representative may ask you to look at the IBM Virtual Shared Disk diagnostics. In the IBM Virtual Shared Disk Nodes Work Area of the IBM Virtual Shared Disk Perspective:

- Click on the appropriate IBM Virtual Shared Disk node icon to select it
- Click on Actions from the menu bar at the top of the page
- Select IBM Virtual Shared Disk Nodes
- Select Run IBM Virtual Shared Disk Diagnostics.

The diagnostics compare the settings of some IBM Virtual Shared Disk parameters across all the nodes in a IBM Virtual Shared Disk configuration and look for discrepancies that could affect the operation of your IBM Virtual Shared Disks.

You can also use the **vsddiag** command for this purpose.

Changing the States of IBM Virtual Shared Disks

If you haven't installed the IBM Recoverable Virtual Shared Disk licensed product, changes to your configuration or system, application, or network problems may require you to move your IBM Virtual Shared Disks from one state (stopped, suspended, active) to another. You can do this with the IBM Virtual Shared Disk Perspective. In the IBM Virtual Shared Disk Nodes work area:

- Click on the appropriate IBM Virtual Shared Disk node icon to select it
- Click on Actions from the menu bar at the top of the page
- Select IBM Virtual Shared Disk Nodes
- Select Change IBM Virtual Shared Disk States.

For information about using commands or SMIT to change the states of IBM Virtual Shared Disks, see "Command and SMIT Interfaces for IBM Virtual Shared Disks" on page 89.

Configuring puts a defined IBM Virtual Shared Disk in the stopped state, but does not make it available.

Starting an IBM Virtual Shared Disk

Starting an IBM Virtual Shared Disk puts a stopped one in the active (and available) state. (This is equivalent to preparing and resuming an IBM Virtual Shared Disk.) Note that for an IBM Virtual Shared Disk to be usable, it must be in the active state on both the server and client nodes.

Preparing an IBM Virtual Shared Disk

Preparing an IBM Virtual Shared Disk puts a stopped one in the suspended state. In the suspended state, open and close requests are honored. Read and write requests are held until the IBM Virtual Shared Disk is brought to the active state.

Resuming an IBM Virtual Shared Disk

Resuming an IBM Virtual Shared Disk puts a suspended one in the active state. The IBM Virtual Shared Disk remains available and read and write requests that were held are resumed.

Suspending an IBM Virtual Shared Disk

Suspending an IBM Virtual Shared Disk puts an active one in the suspended state. The IBM Virtual Shared Disk remains available. Read and write requests that were active are suspended and held. All read and write requests subsequent to those that were active are also held.

Stopping an IBM Virtual Shared Disk

Stopping an IBM Virtual Shared Disk puts a suspended one in the stopped state. The IBM Virtual Shared Disk becomes unavailable. All applications that have outstanding requests to a stopped IBM Virtual Shared Disk terminate in error.

Unconfiguring a stopped IBM Virtual Shared Disk makes it inaccessible. It does not, however, undefine or change the definition information for the IBM Virtual Shared Disk in the SDR.

If you use the IBM Recoverable Virtual Shared Disk, do not issue any commands that change the state of the IBM Virtual Shared Disks (**cfgvsd**, **ucfghsdvsd**, **startvsd**, **preparevsd**, **resumevsd**, **suspendvsd**, **stopvsd**, **ucfghsdvsd**, or **ucfgvsd**.) The IBM Recoverable Virtual Shared Disk handles state changes for you.

Figure 6 on page 40 shows the commands that move IBM Virtual Shared Disks from one state to another.

| BETH2

| *Figure 6. IBM Virtual Shared Disk States and Associated Commands*

Removing an IBM Virtual Shared Disk or Data Striping Device

To remove an IBM Virtual Shared Disk, use the **removevsd** command. To remove a Data Striping Device, use the **removehsd** command. The **removehsd** command also removes the underlying IBM Virtual Shared Disks of the HSD that is to be removed.

Adding a Server Node to an Existing IBM Virtual Shared Disk Configuration

To add a new node to an existing IBM Virtual Shared Disk configuration, follow these steps:

1. Enter the necessary information for the new node into the SDR, using the notebook page of the IBM Virtual Shared Disk Perspective. See “Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR” on page 20.
2. Shut down any applications that use IBM Virtual Shared Disks.
3. Exit from the Perspectives interface and type in the **createvsd** command for the new configuration, or use the SMIT fastpath, **smit vsd_data**. See “Creating IBM Virtual Shared Disks with the createvsd Command” on page 22
4. Reenter the IBM Virtual Shared Disk Perspective and reconfigure all the IBM Virtual Shared Disks on the system or system partition, using the Reconfigure All IBM Virtual Shared Disk Clients action from the IBM Virtual Shared Disk Nodes work area. When you do this, the new configuration and the new node information in the SDR take effect.
5. Start all the IBM Virtual Shared Disks on the system or system partition, using the “Start All IBM Virtual Shared Disk Clients” action from the IBM Virtual Shared Disk Nodes work area of the IBM Virtual Shared Disk Perspectives initial panel.
6. You can now restart any applications that use IBM Virtual Shared Disks.

See Appendix A, “IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands” on page 77 for the syntax of the **vsdnode**, **createvsd**, **ucfgvsd**, and **cfgvsd** commands. The commands can be used instead of the IBM Virtual Shared Disk Perspective in the description above.

Removing a Node from an Existing IBM Virtual Shared Disk Configuration

To remove a node from an existing IBM Virtual Shared Disk configuration, follow these steps:

1. Shut down any applications that use IBM Virtual Shared Disks.
2. Using the **removevsd** command, unconfigure and remove all IBM Virtual Shared Disks defined on this node. The **removevsd** command doesn't affect the data on the IBM Virtual Shared Disks, it just removes the IBM Virtual Shared Disk definition.
3. If the data on the physical disks must be moved, use the **smit migratepv** fastpath to move it.
   a. As the root user on the primary node, enter:

      smit migratepv

      SMIT returns a screen similar to the following:

```
          Move Contents of a Physical Volume
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*SOURCE physical volume name                         [ ]  +
```

   b. Enter the names of the physical volumes being moved in the SOURCE physical volume name field and press Enter.

SMIT returns a screen similar to the following. The physical volumes being migrated are entered in the SOURCE physical volume names field.

```
          Move Contents of a Physical Volume
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*SOURCE physical volume name                         [ ]
*DESTINATION physical volumes                        [ ]  +
 Move only data belonging to this                    [ ]  +
  LOGICAL VOLUME?
```

c. Enter field values as follows:

DESTINATION physical volumes
Enter the names of the physical volumes to which you want to move the data.

Move only data belonging to this LOGICAL VOLUME?
Set this field to **no**.

d. Press Enter to migrate the physical volumes.

e. After the command completes, press F10 to leave SMIT and return to the command line.

4. Type in the **createvsd** command for the new configuration, or use the SMIT fastpath, **smit vsd_data**. See Appendix A, "IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands" on page 77 for the syntax of **createvsd**.
5. Reenter the IBM Virtual Shared Disk Perspective and reconfigure all the IBM Virtual Shared Disks on the system or system partition, using the Reconfigure All IBM Virtual Shared Disk Clients action from the IBM Virtual Shared Disk Nodes work area of the IBM Virtual Shared Disk Perspectives primary window. When you do this, the new configuration takes effect.
6. Start all the IBM Virtual Shared Disks on the system or system partition, using the "Start All IBM Virtual Shared Disk Clients" action from the IBM Virtual Shared Disk Nodes work area of the IBM Virtual Shared Disk Perspectives primary window.
7. You can now restart any applications that use IBM Virtual Shared Disks.

See Appendix A, "IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands" on page 77 for the syntax of the **removevsd** and **cfgvsd** commands.

Chapter 5. Application Programming Considerations

This chapter tells you how to create or modify your application programs to take advantage of functions in the IBM Virtual Shared Disk and Data Striping Device components and in the IBM Recoverable Virtual Shared Disk licensed product.

Application Programming Considerations for IBM Virtual Shared Disks

Application programmers who work with IBM Virtual Shared Disks need to be aware of data integrity considerations, the IBM Virtual Shared Disk transmission protocol, and how to get IBM Virtual Shared Disk information with C language interfaces.

Data Integrity

Use the character (raw) device to access an IBM Virtual Shared Disk, because the block device user can be presented with stale data from the cache. For the block IBM Virtual Shared Disk device, this buffer cache exists on all nodes in the system partition, so even if all users use the block interface, data integrity is compromised.

An IBM Virtual Shared Disk is a device. On a single node, users must provide and use their own synchronization mechanisms for their reads and writes to insure data integrity. Since the IBM Virtual Shared Disk allows multiple nodes to have simultaneous access to devices, users must provide their own distributed synchronization mechanisms. Neither the IBM Virtual Shared Disk nor the SP system software provides a distributed synchronization mechanism.

The IBM Virtual Shared Disk device driver lets you calculate a checksum in software for all messages sent to remote nodes. The option is provided for use on networks that don't provide their own checksum processing. You can use the **cksumvsd** command to turn checksum processing on and off in the IBM Virtual Shared Disk device driver. Because the switch provides checksum processing in hardware, don't use **cksumvsd** to turn checksum processing on if you use **css0** as your IBM Virtual Shared Disk communication adapter.

Note: Checksum processing should be either on or off on all IBM Virtual Shared Disk nodes in a system partition.

IBM Virtual Shared Disk Transmission Protocol

The IBM Virtual Shared Disk implements its own UDP-like IP protocol. The requesting node implements an exponential back-off retransmission strategy. If the remote node does not service the request after about 15 minutes of retransmission, the IBM Virtual Shared Disk fails the user's I/O request. Suspending and resuming an IBM Virtual Shared Disk will give all requests a fresh start with the full 15 minutes of retransmission time.

Sequence Numbers

As part of the initial implementation of the IBM Virtual Shared Disk's communication protocol, sequence numbers are used and maintained. Each node keeps an expected and outgoing sequence number for all other nodes. Each time a node sends a request message to another node, it increments its outgoing sequence number for that node. If a node receives a message, and its sequence number is

© Copyright IBM Corp. 1994, 1996

not within the valid range, based on the expected number for that node, the message is discarded. The valid range is defined as the current value of the expected number for the node, plus a fairly large window size. When a node receives a valid request message from another node, the receiving node resets its expected sequence number for that node to the sequence number in the message.

Checking and Resetting Sequence Numbers: To check sequence numbers from the IBM Virtual Shared Disk Perspective, select the appropriate node in the IBM Virtual Shared Disk Nodes work area. Then click on the notebook icon in the tool bar. When the IBM Virtual Shared Disk Node Attributes notebook page is displayed, look for the IBM Virtual Shared Disk Node Statistics tab at the bottom of the page. You may have to click on the small arrow at the bottom of the page to see more tabs. To reset the statistics, select the appropriate node, click on Actions, then on IBM Virtual Shared Disk Nodes, then on Reset IBM Virtual Shared Disk Statistics.

Sequence numbers become significant if a node in the IBM Virtual Shared Disk cluster is rebooted after it has been engaged in IBM Virtual Shared Disk I/O. When the node comes back up, all its sequence numbers will be reset to 0, but all other nodes in the system partition will expect numbers greater than the last sequence number they received from the node before it rebooted, that is, nonzero. Therefore, none of the other existing nodes that received messages from the node before it crashed will now “listen” to the rebooted node because the sequence number is below the expected value. The new node will continue to retransmit, increasing the outgoing sequence number each time, until it eventually reaches the expected range on all the other nodes. Some requests, however, may exhaust their retry limits and fail. The rebooted node will probably accept requests from the existing nodes as their sequence number will be within the range of the expected count (0) and the (large) window size.

When one node's sequence numbers are out of synchronization with the rest of the IBM Virtual Shared Disk nodes, it will look like all IBM Virtual Shared Disk I/O to remote servers from the out-of-synch node is hung. After about 15 minutes of retransmissions, the I/O will start to fail.

You can also use the **-R** and **-r** options of the **ctlvsd** to reset the sequence number of all or some of the nodes to 0 (both the expected and the outgoing). Use **statvsd** on client and server nodes to check the sequence numbers. Use **ctlvsd** to reset sequence numbers as required. (See the **statvsd** and **ctlvsd** command reference pages for more details.) Before issuing **ctlvsd**, put the IBM Virtual Shared Disk in suspended state.

Casting Out

You can ignore or allow requests for data from IBM Virtual Shared Disks that are coming from other nodes in the system by “casting out” those nodes. The Actions menu items

- Ignore IBM Virtual Shared Disk Requests

  and
- Allow IBM Virtual Shared Disk Requests

cast nodes out and in. You can verify their status with the IBM Virtual Shared Disk Statistics notebook pages. To cast out nodes with commands, you can use **ctlvsd -k** to cast out *nodes_to_cast_out*; **ctlvsd -r** to cast nodes back in and reset their sequence numbers; **ctlvsd -K** to cast out all nodes (except the node on which you

issued the command—local requests are still processed); and **ctlvsd -R** to cast all nodes back in. Initially, when the IBM Virtual Shared Disks are configured, all nodes in the IBM Virtual Shared Disk network are cast in.

Checksum Processing on Unreliable Networks

The switch is a low-noise communication medium. The switch adapter performs checksum processing in hardware to provide a reliable network. Some networks do not provide checksum processing (for example, the Ethernet). The IBM Virtual Shared Disk software provides a checksum processing option for use on unreliable networks. You can use the IBM Virtual Shared Disk Perspective or the **cksumvsd** command to turn checksum processing on and off. For more information, see the reference page for the **cksumvsd** command.

Getting IBM Virtual Shared Disk Information with C Interfaces

Note: The header file **/usr/include/vsd_ioctl.h** is installed with the **ssp.csd.vsd** option. Structures defined in **vsd_ioctl.h**may change from release to release. This may require that you recompile your applications that utilize C interfaces to the IBM Virtual Shared Disk to run on new releases.

The IBM Virtual Shared Disk base code has a sample directory: **/usr/lpp/csd/samples**.

This directory contains the following files:

vsdinfo.c When compiled, generates a binary that shows IBM Virtual Shared Disk information for the **vsd_name** given from the command line. IBM Virtual Shared Disk information is retrieved from an open IBM Virtual Shared Disk's file descriptor.

vsdsinfo.c When compiled, generates a binary that lists information for all configured IBM Virtual Shared Disks.

Application Programming Considerations for Data Striping Devices

Application programmers need to be aware of data integrity considerations in programming applications for Data Striping Devices. Data Striping Device information is obtainable through C language interfaces.

Getting HSD Information with C Interfaces

The HSD base code has a sample directory: **/usr/lpp/csd/samples**.

This directory contains the following files:

hsdinfo.c When compiled, this file generates a binary that shows Data Striping Device information for the **hsd_name** given from the command line. Data striping device information is retrieved from an open Data Striping Device's file descriptor.

hsdsinfo.c When compiled, this file generates a binary that lists information for all configured Data Striping Devices.

Note: The header file **/usr/include/hsd_ioctl.h** is installed with the **ssp.csd.hsd** image. Structures defined in **hsd_ioctl.h** may change from release to release. This may require that you recompile your applications that utilize C interfaces to the HSD to run on new releases.

Making Your Application Recoverable

You can code your program to use the **hc** application programming interface of the IBM Recoverable Virtual Shared Disk software to aid in application recovery. A distributed lock manager (not provided with the IBM Recoverable Virtual Shared Disk software) is the type of program that might use **hc**'s services.

Your program should include the following **hc.h** header file:

```
#include <hc.h>
```

Your program connects to a socket where **hc** is listening. If a node fails or the client application on a node fails, **hc** sends a reconfiguration message according to the protocol defined in the **hc.h** header file. Your program should check for messages from that socket and react accordingly.

The hc subsystem can support mutliple applications by running multiple instances of hc. Each hc is invoked by a different **hc.vsd** script.

To create a new instance of the hc subsystem, you must make a copy of the **hc.vsd** script (from **usr/lpp/csd/bin**) and rename it with a meaningful name, *hc.myappname*. Then edit the newly-created *hc.myappname* and change the *instance* variable to *myappname*. You should also export the path to the well-known socket for communication with your application, inserting the following line into the script:

```
export CLIENT_PATH=PATH
```

where PATH is an absolute pathname. The hc subsystem then creates the socket **PATH/myappname** to communicate with your application.

As a convenience, the **hc** program will run an optional script, **/usr/lpp/csd/bin/hc.activate**, once when it initializes. The **hc.activate** script is passed two parameters; the local node number and the path name of the UNIX domain stream socket that it serves. You can start your recoverable application from this script.

If the script doesn't complete, **hc** won't communicate with the application. An application that checks connectivity through **hc** running on the other clients can't perform that check until the **hc.activate** script completes. If you need **hc** to continue processing, fork a background process from the **hc.activate** script. Then exit the **hc.activate** script with an exit code of 1, which indicates that communication with the application does not have to be complete for **hc** to continue processing. **hc.activate** is executed only once.

The **hc** program also runs the **/usr/lpp/csd/bin/hc.deactivate** script upon detecting failure of its client. The same parameters are passed as are passed to **/usr/lpp/csd/bin/hc.activate**, letting you restart the client if you desire. The client may fail either by exiting or by failure to respond to a **ping** message within 10 minutes.

Note: Ten minutes is the default for the **PING_DELAY**. That is, if a client application fails to respond to a ping sent from the **hc** program within 10 minutes, **hc** will consider the application to have failed and will invoke **hc.deactivate.**

After **hc.deactivate** runs, it sends a **node down** message to the other nodes.

Some application programs running under some system loads may require a longer **PING_DELAY** timeframe. To change the **PING_DELAY**, edit the **ha_vsd** script. For example, if you wish to increase the **PING_DELAY** to 15 minutes, change the **SCRIPT_PATH=** /**hc** line in the **ha_vsd** script to:

```
PING_DELAY=900 SCRIPT_PATH= /hc
```

If **hc** fails, the application will receive a zero-length message over the socket and should shut itself down.

Applications are responsible for cleaning up system resources when they complete or fail. **hc.deactivate** should be used for this purpose.

Chapter 6. Performance and Tuning Considerations for IBM Virtual Shared Disks and Data Striping Devices

System administrators and application programmers should be aware of the performance characteristics of IBM Virtual Shared Disks and Data Striping Devices and of how they can affect those characteristics.

Tuning IBM Virtual Shared Disk Performance

The IBM Virtual Shared Disk device driver layer passes all its requests to the underlying Logical Volume Manager and disk layers. This can cause your I/O subsystem to become overloaded. If you are not satisfied with IBM Virtual Shared Disk performance, you should first check that the I/O subsystem is not the bottleneck. If an overloaded I/O subsystem is degrading your system's performance, tuning the IBM Virtual Shared Disk will not help. In the case of I/O subsystem overload, consider spreading the I/O load over more disks or nodes.

We recommend taking the following steps to tune the IBM Virtual Shared Disk for best performance:

1. Use the defaults when defining IBM Virtual Shared Disks (refer to “Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR” on page 20)
2. Turn IBM Virtual Shared Disk caching off (if you are not using your system for Online Transaction Processing)
3. Do a performance run to collect statistics on the IBM Virtual Shared Disk, your I/O subsystem, and the CPU on all nodes. Use **iostat** to check your disk utilization.
   - If the system is I/O bound (meaning your disks are more than 50% utilized), add disks
   - If the system is CPU bound, add nodes or spread the workload. HSD could be used to spread the workload.
   - If nodes are doing excessive swapping due to insufficient memory, reduce cache size.
   - If requests are queuing because of a shortage of buddy buffers or pbufs, that's a symptom of disk bottlenecks. Spread the data or add disks to the node.
   - If your application issues requests that are larger than 64KB, set your maximum buddy buffer size to 256KB.
   - If you see too many retries, check for disk bottlenecks. If that's not the problem, increase the switch pool size (see “mbufs and the Switch Pool” on page 55).
4. Reset the statistics counter, using the IBM Virtual Shared Disk Perspective or the **ctlvsd command.** (See the *IBM Parallel System Support Programs for AIX: Command and Technical Reference*.)
5. Do another performance run.

You should generally turn IBM Virtual Shared Disk caching off. Memory is better allocated to the operating system itself, for paging, and to the cache belonging to

© Copyright IBM Corp. 1994, 1996

the application using the IBM Virtual Shared Disk. To turn IBM Virtual Shared Disk caching off,

1. Select one or more nodes
2. Go to the IBM Virtual Shared Disk Node Attributes notebook page
3. Change the *cache/nocache* option to nocache
4. Close the notebook page and select Reconfigure all IBM Virtual Shared Disks on a Node from the Actions list

This changes the *cache/nocache* option for all the IBM Virtual Shared Disks on the selected node or nodes.

IBM Virtual Shared Disk Tunable Parameters

The main IBM Virtual Shared Disk tunable parameters are:

- IP communications adapter (usually the switch)
- Logical Volume Manager (striping and other LVM characteristics)
- The IBM Virtual Shared Disk cache buffer
- The buddy buffer
- Maximum I/O request size
- Request blocks
- pbufs
- mbufs

In the following sections we define these parameters and discuss relevant tuning considerations. You should also consider the tunable characteristics of the application that uses IBM Virtual Shared Disks, especially its use of buffers.

The High Performance Switch/Scalable POWERparallel Switch

If you configure the IBM Virtual Shared Disk software to use the High Performance Switch or the SP Switch (**css0**), the maximum IP message size (utilized by the IBM Virtual Shared Disk driver) is 64KB (65536) You should ensure the maximum buddy buffer size is also 64KB. If your application issues requests that are larger than 64KB, you should set the maximum buddy buffer size to 256KB,since the value you assign to *maximum buddy buffer size* in the SDR also limits the maximum size of the packet IP will send across the nodes. For example, if you have:

- A request from a client to write 256KB of data to a file or database
- A maximum buddy buffer size of 64KB
- A maximum IP message (packet) size of 60KB

the following transmission sequence occurs:

1. The IBM Virtual Shared Disk component divides the 256KB of data into four 64KB blocks to accommodate the buddy buffer size
2. Each 64KB block of data becomes one 60 KB packet and one 4KB packet for transmission to the server via IP
3. At the server, the eight packets are reassembled into four 64KB blocks of data

4. The server then has to perform four 64KB write operations and return four acknowledgements to the client.

A better scenario for the same write operation would be:

- A request from a client to write 256KB of data to a file or database
- A maximum buddy buffer size of 128KB
- A maximum IP message (packet) size of 60KB

with the following transmission sequence:

1. The IBM Virtual Shared Disk component divides the 256KB of data into two 128KB blocks to accommodate the buddy buffer size
2. Each 128KB block of data becomes two 60 KB packets and one 8KB packet for transmission to the server via IP
3. At the server, the six packets are reassembled into two 128 KB blocks of data
4. The server then has to perform two 128 KB write operations and return two acknowledgements to the client.

The second scenario is preferable to the first because even though there are more switch packets in the second, the I/O operations at the server are minimized. A perfect scenario would be one where the client request and the maximum IP message size were equal and the maximum buddy buffer size was at least as large as the request and IP message sizes. For example:

- A request from a client to write 60KB of data to a file or database
- A maximum buddy buffer size of 128KB
- A maximum IP message (packet) size of 60KB

When you use the switch, buddy buffers will be used only when a shortage in the switch buffer pool occurs. If you see buddy buffer shortages, instead of increasing your buddy buffers, you need to increase your switch pool size. See “mbufs and the Switch Pool” on page 55.

Logical Volume Manager (LVM) Tuning Considerations

There is always an associated logical volume for every IBM Virtual Shared Disk device defined and configured in a system. Every IBM Virtual Shared Disk I/O request eventually becomes an I/O request to the associated logical volume (unless you get a cache hit at the server). This mapping of IBM Virtual Shared Disk device I/O requests to the associated logical volume I/O requests is handled transparently by the IBM Virtual Shared Disk software layer. All the performance tuning considerations that apply to a logical volume also apply to IBM Virtual Shared Disk devices. Refer to *AIX System Management Guide* (SC23-2525) and *AIX Performance Monitoring and Tuning Guide* (SC23-2365) for information on the performance tuning of logical volumes.

Buffer Allocation

Make all new allocated buffers at the page boundary. If your I/O buffer is not aligned on a page boundary, the IBM Virtual Shared Disk device drivers will not parallelize I/O requests to underlying IBM Virtual Shared Disks. Performance will be degraded. This is particularly important if you use C language programs for read operations.

The Cache Buffer

Each IBM Virtual Shared Disk device driver, that is, each node, has a single cache buffer, shared by all cacheable IBM Virtual Shared Disks configured on and served by the node. The cache buffer is used to store the most recently accessed data from the cached IBM Virtual Shared Disk devices (associated logical volumes) on the server node. The objective is to minimize physical disk I/O activity. If the requested data is found in the cache, it is read from the cache, rather than the corresponding logical volume.

Data in the cache is stored in 4KB blocks. The content of the cache is a replica of the corresponding data blocks on the physical disks. The cache contains only valid data. Write-through cache semantics apply; that is, the write operation is not complete until the data is on the disk.

When you define IBM Virtual Shared Disks with the **defvsd** command, you can specify **cache** or **nocache**. We recommend that you specify **nocache** (or make the cache buffer small) in most instances (especially in the cases of read-only or other than 4KB applications) for the following reasons:

- Requests that are not exactly 4KB and not aligned on a 4KB boundary will bypass the cache buffer, but will incur the overhead of searching the cache blocks for overlapping pages.
- Every 4KB I/O operation incurs the overhead of copying into or out of the cache buffer, as well as the overhead of moving program data from the processor cache due to the copy.
- There is overhead for maintaining an index on the blocks cached.
- Even when there is heavy write traffic that makes the use of a cache buffer look attractive, the processing and message overhead associated with transferring ownership of the data and transferring the data itself is so high that the elimination of the second I/O may have only a small impact on overall system performance.

If you are running an application that involves heavy writing followed immediately by reading, it may be advantageous to turn the cache buffer on for some IBM Virtual Shared Disks on a particular node. Choose the appropriate size for the cache based on the expected throughput and the expected time lag between writes and reads. For example, if the expected throughput is 100 4 KB-aligned I/O operations per second and reads lag writes by 0.5 seconds, calculate the cache buffer size by multiplying 100 X 0.5 and, as a safety factor, double it for a total of 100 cache blocks.

The IBM Virtual Shared Disk Node Statistics tab of the IBM Virtual Shared Disk Perspective gives detailed statistics on IBM Virtual Shared Disk cache hit and I/O activities. You can also use the **lsvsd -s** command. Both methods will tell you which IBM Virtual Shared Disks are heavily used.

Buddy Buffers

The IBM Virtual Shared Disk server node uses the buddy buffer to temporarily store data for I/O operations originating at a client node. In contrast to the data in the cache buffer, the data in a buddy buffer is purged immediately after the I/O operation completes.

Note: Buddy buffers are used only when a shortage in the switch buffer pool occurs or on certain networks (for example, the Ethernet).

The values associated with the buddy buffer are:

- Minimum buddy buffer size allocated to a single request
- Maximum buddy buffer size allocated to a single request
- Total size of the buddy buffer

Buddy buffer space is allocated in powers of two. If an I/O request size is not a power of two, the smallest power of two that is larger than the request is allocated. For example, for a request size of 24KB, 32KB are allocated on the server.

If you are using the switch as your adapter for IBM Virtual Shared Disks, we recommend setting 4KB (4096) and 64KB (65536), respectively, for minimum and maximum buddy buffer size allocated to a single request.

To define the total size of the buddy buffer, consider the remote I/O throughput for the server and specify the number of maximum-sized buddy buffers in the buffer. For example, if you expect the server to serve 10MB per second on behalf of remote clients and a request spends an average of 60 milliseconds on the server, multiply 10MB per second X 0.06 second and, for safety, double or triple the result for a total buddy buffer size of 1.8MB (30 64KB maximum buddy buffers).

If the IBM Virtual Shared Disk statistics consistently show requests queued waiting for buddy buffers, do not add more buddy buffers. Instead, increase the size of the switch send pool (see “mbufs and the Switch Pool” on page 55) or spread the data over disks attached to other nodes, to prevent a bottleneck.

Maximum I/O Request Size

The following factors limit the block size that the IBM Virtual Shared Disk software will use to process each I/O request:

- The largest block size the IBM Virtual Shared Disk software will use is *max_buddy_buffer_size*.
- If the IBM Virtual Shared Disk uses the switch as its adapter, the *max_IP_msg_size* that could be sent is 63.5KB (65024 bytes). 60KB (61440) is the default value for the IBM Virtual Shared Disk device driver when **css0** is defined as the IBM Virtual Shared Disk adapter in the SDR. The **ctlvsd -M** command can override the default. The *max_IP_msg_size* should be set to a value that is a multiple of 512 bytes and is less than or equal to 63.5KB (when the switch is used) and less than or equal to 24KB (when the switch is not used). The **statvsd** command displays the current value.

  Note: Setting the *max_IP_msg_size* to more than 24K when using communication adapters with small MTU (maximum transmission unit) could overflow the adapter driver's internal buffers, causing the IP layer

to drop requests. This will force the IBM Virtual Shared Disk device driver to retry, sometimes without success.

Every IBM Virtual Shared Disk I/O request is subject to all three limits. For example, with a *max_buddy_buffer_size* of 64K and a *max_ip_msg_size* of 61440, if an application requests a single 64K read to an IBM Virtual Shared Disk served by the local node, the request is passed down to the local LVM as one 61440 request and one 4096 request.

The atomicity of an I/O operation is gated by the size of the IBM Virtual Shared Disk request, rather than the size of the user request (if the IBM Virtual Shared Disk request is smaller than the user request the user requests would be split down to the size of IBM Virtual Shared Disk requests).

Request Blocks

The number of request blocks is the total number of physical I/O operations that have been issued by all processes on a node, but have not completed. The number includes requests to both local and remote devices. Because large requests may be broken up into smaller subrequests, the request block number may be several times greater than the actual number of pending read/write requests.

For example, if I/O requests are:

- issued at a rate of 1000 I/O operations per second
- broken on the average into three pieces
- responded to in 50 milliseconds

then the algorithm for calculating the number of request blocks is 1000 X 3 X 0.05 for a total of 150 request blocks. You may want to somewhat increase the number as a safety precaution to account for the possibility of workload surges and so forth. However, in general, we do not recommend specifying more than a few hundred request blocks.

Specifying an inordinately large number of request blocks could have a negative performance impact. A large number of request blocks could flood the network, causing servers to run out of mbufs and initiating unnecessary retransmissions.

Although the **statvsd** command reports the number of times there is no request block available, queueing for request blocks does not necessarily imply a performance bottleneck. If you increased the number or request blocks infinitely, queueing would occur elsewhere in the operating system.

pbufs

pbufs are used for actual physical I/O requests when the requests cannot be satisfied from the IBM Virtual Shared Disk cache (a cache miss). A pbuf shortage affects overall performance by causing context switches and additional rotational misses and device seeks.

pbufs are specified by read/write request count parameter of the **vsdnode** command as a way of controlling the number of pending device requests for a specific IBM Virtual Shared Disk on its server node. pbufs are specified on a per device basis.

A pbuf shortage reported by the Node Statistics notebook page of the IBM Virtual Shared Disk Perspective or by **statvsd** does not imply that increasing the number of pbufs will improve performance because requests will be queued at other parts of the operating system.

To check on the interactions among request blocks, pbufs, and cache blocks, using the IBM Virtual Shared Disk Perspective:

1. With a node selected, click on the Actions menu item.
2. Select IBM Virtual Shared Disk Nodes
3. From the list of available actions, choose Reset IBM Virtual Shared Disk Statistics
4. Go to the IBM Virtual Shared Disk Node Attributes notebook page
5. Select the IBM Virtual Shared Disk Node Statistics page
6. View the statistics and decide which parameters to tune

The **statvsd** command also gives detailed statistics on request shortages, pbuf shortages and cache block shortages. You can run this command together with **lsvsd -s** before and after an application execution to determine how to tune these parameters with respect to a particular application workload.

mbufs and the Switch Pool

mbufs are used for data transfer between the client and the server nodes by IBM Virtual Shared Disk's own protocol (the layer resembling UDP protocol) and IP protocol (which IBM Virtual Shared Disk uses). If you are using the switch (**css0**) as your communications adapter, IBM Virtual Shared Disk uses mbuf clusters to do I/O directly from the switch's send and receive pools.

If you notice that the indirect I/O statistic (from the IBM Virtual Shared Disk Statistics notebook page) is incremented consistently, check **errpt()**. If you see the line:

```
IFIOCTL_MGET(): send pool shortage
```

you should consider increasing the size of the send and receive pools.

To check the current sizes of the send and receive pools, type:

```
/usr/sbin/lsattr -l css0 -E
```

The default size for each pool is 524288 bytes.

To change the sizes of the pools to 4MB, type:

```
/usr/lpp/ssp/css/chgcss -l css0 -a spoolsize=4194304
/usr/lpp/ssp/css/chgcss -l css0 -a rpoolsize=4194304
```

This command increases the send and receive pool size to 4MB. You must reboot the node for the new sizes to take effect.

To see what your current system mbuf setting is, type:

```
/etc/no -a | grep thewall
```

We recommend allowing 16MB (16777216) for mbufs and clusters. You can set this value by issuing:

```
/etc/no -o thewall=16777216
```

System performance considerations regarding mbufs and mbuf clusters also apply to IBM Virtual Shared Disk environments. See *AIX Performance Monitoring and Tuning Guide* (SC23-2365) for more information.

Tuning Data Striping Device Performance

This section presents information for tuning the HSD.

Note: For more information on tuning the HSD, access the RS/6000 home page at **http://rs6000.austin.ibm.com**. Once you get to the home page, look under **Software** for information about IBM Virtual Shared Disks and HSDs.

We discuss HSD tuning in relationship to these performance considerations:

- Database ("hot table") access
- Disk throughput
- CPU overhead
- I/O completion speeds

The parameters that can be tuned to affect performance are:

- Parallelism
- Request size
- Stripe size

If HSD is configured to parallelize (**ctlhsd -p**) the split requests across multiple IBM Virtual Shared Disks (referred to as "striping"), disk throughput may increase even if the underlying disks are not fully utilized. You should understand, though, that parallelism is achieved at the expense of additional CPU cycles, especially when setting small stripe sizes. Note that parallelism benefits can only take place if the requests issued to the HSD device are on a 4KB user buffer memory boundary.

If you are using an HSD application that is not submitting 4KB memory aligned requests to HSD, you should use HSD only to resolve hot table bottleneck problems. In this case, HSD's parallelism should be set to **1**. In addition, the stripe size must be greater than the largest request size issued the application.

If you are using a 4KB memory aligned application, you can use HSD to both eliminate hot table bottlenecks and increase disk throughput (by using parallelism). Note that HSD (in PSSP Version 2 Release 1 and later) can handle a parallelism value of up to 64. To take advantage of the parallelism potential, HSD's stripe size should be smaller than the application's request size.

HSD and IBM Virtual Shared Disk request latency is primarily gated by disk I/O. Disk I/O can be separated into three distinct operations:

- Seek delay
- Rotational delay
- Transfer delay

Transfer delay always depends on the request's size. Rotational and seek delays depend on timing (rotational) and previous location of the disk heads (seek and rotational disk cache hit ratio). If you issue one disk request, the seek and rotational

delays will be the disk's average delays (assuming random requests). If you issue many HSD requests, the probability that one of those requests incur larger-than-average delays (seek or rotational) increases significantly. When you issue an HSD request to a set of IBM Virtual Shared Disks (disks), the total latency will be gated by the slowest disk (seek, rotational, and transfer delays). The only savings are the reduced transfer delays. Thus, if increased parallelism does not reduce the transfer delay enough to compensate for the probability of large rotational or seek delays, you may degrade performance.

If your system is CPU bound, using HSD (with parallelism equal to **1** and a large stripe size) can help only with eliminating hot table bottlenecks. Eliminating hot table bottlenecks may improve overall transaction throughput. You should consider spreading the workload by adding more nodes/disk arms.

If your system is I/O bound, HSD with striping may help. However, you should consider that:

- Processing more information from disks requires more CPU cycles and that HSD striping uses some of these CPU cycles. It is important to stay below the point of diminishing returns when you set the striping size and HSD parallelism level.
- The smaller the stripe size, the more disk arms an HSD should encompass.
- Disks should be on separate adapters because adapters (such as SCSI) limit the number of concurrent requests they can send to disks that are sharing the same bus.

HSD parallelism should be increased if your application issues large requests and uses a small HSD stripe size in conjunction with many disk arms.

The only way to fully optimize an HSD to your environment is to try out different configurations. The main obstacle in tuning an HSD is that it is not simple to add or remove IBM Virtual Shared Disks and change the stripe size after the HSD is loaded with data. The only way to do this is to back up the HSD (using **dd**), recreate it, and load it with the new parameters. Therefore, careful consideration should be given to HSD usage planning. If utilized properly, HSD can, in some circumstances, double IBM Virtual Shared Disk throughput performance.

IBM Virtual Shared Disk Tuning Effects and Data Striping Devices

All I/O requests are passed to the underlying IBM Virtual Shared Disks. Thus, the parameters selected for IBM Virtual Shared Disk configuration affect the HSD performance. Read this section in conjunction with “Tuning Data Striping Device Performance” on page 56.

Buddy Buffers

If the stripe size is greater than *max_IP_packet_size* or *max_buddy_buffer_size*, then one stripe will have to be broken up into more than one I/O request.

Caching and Stripe Size

Use a stripe size of 4KB if you want to exploit the IBM Virtual Shared Disk cache.

See also the tuning suggestions for IBM Virtual Shared Disks earlier in this chapter.

Chapter 7. How the IBM Recoverable Virtual Shared Disk Software Works

This chapter explains how the IBM Recoverable Virtual Shared Disk software works.

The Group Services Component of Availability Sevices

The Group Services subsystem is a system-wide, fault-tolerant, and highly available service that provides a general purpose facility for coordinating and monitoring changes to the state of an application that is running on a set of nodes.

A Group Services domain consists of the set of nodes that makes up a system partition. If there is only one system partition in the SP system, the Group Services domain consists of all of the nodes in the SP system. From the point of view of the Group Services subsystem, the system partition in which it is running is "the system."

Each group that is maintained by the Group Services subsystem is uniquely named. Any authorized process in a Group Services domain may create a new group. Any authorized process in the domain may ask to become a member of a group. This request is called a join request or joining the group. If the join request is successful, the process becomes a *provider* for the group.

Any authorized process in the domain can ask to monitor the group. This request is called a *subscribe request* or subscribing to the group. If the subscribe request is successful, the process becomes a *subscriber* for the group.

A group may have members on multiple nodes in the domain, and each node may have multiple members. For each group, the Group Services subsystem maintains consistent group state data. A group's state consists of two pieces of information:

- The membership list. The membership list of a group is the list of providers in the group. Each provider is identified by a provider identifier. The Group Services subsystem maintains the list in the following order: the oldest provider (that is, the first provider to join the group) is at the head of the list and the youngest is at the end. All of the group's providers and subscribers see the same ordering of the list.
- The group state value

The state value of a group is defined by the application. that is using the GS API and is controlled by the providers in a way that is meaningful to the application.

Providers cause the membership list to be modified by joining or leaving the group. In addition to voluntarily leaving a group, a provider may leave involuntarily, due to the failure of the provider process itself or the failure of the node on which it is running. An involuntary leave is called a *failure leave* and is initiated by the Group Services subsystem.

© Copyright IBM Corp. 1994, 1996

HA Topology Services and Recovery

The *HA topology services daemon*, **hatsd**, is a daemon that runs on each node in the SP system or system partition, including the control workstation. It exchanges *heartbeats* with its peers on other nodes. The **hatsd** daemon maintains a list of all the nodes in the system or system partition that are operational and can be reached through the network. This list is called the *HA topology services membership list*. The group services daemon imports this list and uses it to notify other daemons of membership changes in the system or system partition. If for any reason a node fails to successfully transmit a few heartbeats, the HA topology services daemons on the other nodes declare it inactive and notify the group services daemons, which in turn notify the IBM Recoverable Virtual Shared Disk daemons. The appropriate recovery scripts are then activated.

Like **hatsd**, the *rvsd daemon*, **rvsdd** runs on every node in the system partition. It uses Group Services utilities to manage IBM Virtual Shared Disk recovery.

The Membership List

The group services daemon, **hagsd**, is part of the High Availability subsystem of PSSP. The group services daemon allows distributed applications to coordinate their activities. A group services daemon on one node can subscribe to a specific group to receive notification of changes in that group's membership list. A subscription to a group is called a *join* and a termination of that subscription is called a *leave*. A subscription can be terminated deliberately by **hagsd** or the termination can result from a hardware or software failure on a node. A group can also maintain state information that has meaning for its members and can broadcast messages to members. The group services daemon uses **hatsd** to determine which nodes are active.

For more detailed information about Group Services and the group services daemon, see *IBM Parallel System Support Programs for AIX: Group Services Programming Guide and Reference*

Recovery Scripts

Recovery scripts perform recovery for the IBM Virtual Shared Disks in an IBM Virtual Shared Disk cluster or system partition. They set and change the states of the IBM Virtual Shared Disks and synchronize states across the cluster.

Quorum

When there is a *quorum*, the majority of the IBM Virtual Shared Disk nodes at the same software level in the system or system partition (including the control workstation) are active and responding to the network. Quorum is not applied across different software levels. The IBM Recoverable Virtual Shared Disk software uses the quorum to prevent both the primary and secondary nodes for a given IBM Virtual Shared Disk from simultaneously serving requests for that disk. The IBM Recoverable Virtual Shared Disk recovery scripts are run only when there is a quorum. If the quorum is lost, the recovery scripts put all IBM Recoverable Virtual Shared Disk disks into the stopped state and I/O requests fail.

The system administrator can use the **ha.vsd** command to set the quorum manually if that is necessary to diagnose and recover from problems. This is called *quorum override*.

Active Groups

An active group is one that has a quorum of the IBM Virtual Shared Disk nodes (including the control workstation) in a system or system partition.

The IBM Recoverable Virtual Shared Disk Recovery Subsystems

The IBM Recoverable Virtual Shared Disk recovery subsystems, rvsd and hc, respond to changes in the status of the system by running recovery scripts and notifying client applications. The subsystems operate as daemons named **rvsdd** and **hcd**. The following sections describe the functions of the rvsd and hc subsystems.

The rvsd Subsystem

The rvsd subsystem controls recovery for the IBM Recoverable Virtual Shared Disk licensed program. It invokes the recovery scripts whenever there is a change in the group membership. The **ha.vsd** command controls the rvsd subsystem. When a node goes down, rvsd notifies all surviving processes in the remaining IBM Virtual Shared Disk nodes, so they can begin recovery. If a node fails, recovery involves switching the ownership of a twin-tailed disk to the secondary node. When a node comes back up, recovery involves switching disk ownership back to the primary (or server) node.

The *primary node* is a node that is physically connected to a set of IBM Virtual Shared Disks and will always manage them if it is active. The *secondary node* is a node that is physically connected to a set of IBM Virtual Shared Disks and will manage them only if the primary node becomes inactive. Primary and secondary nodes are defined with the **createvsd** and **createhsd** commands.

The rvsd subsystem uses the notion of *quorum*, the majority of the IBM Virtual Shared Disk nodes, to cope with communication failures. If the nodes in a system partition are divided by a network failure, so that the nodes in one group cannot communicate with the nodes in the other group, rvsd uses the quorum to decide which group continues operating and which group is deactivated. In previous releases of the SP system, quorum was defined as a majority of all the nodes in a system or system partition. With PSSP 2.2, quorum is based on nodes that have been defined as IBM Virtual Shared Disk nodes only.

Table 4 on page 62 shows how the daemons in the nodes in a system partition react as inactive nodes come back up and as active nodes fail. A *client* is a node that has access to IBM Virtual Shared Disks but is not physically connected to them. Some nodes in a system partition may not have any access to IBM Virtual Shared Disks defined on that system and are neither servers nor clients.

In the following example, the changes that affect three of the nodes in a system that has more than three nodes are described.

Table 4. Recovery Actions When Nodes Fail

Nodes and Clients	Node Is Active	Recovery Scenario (Node Is Inactive)
Node1, primary	• Daemons running on the other nodes in the system partition accept node1 into the group. • All IBM Virtual Shared Disks on node1 become active. • All clients designate node1 as the manager for the node1 IBM Virtual Shared Disks and designate those IBM Virtual Shared Disks as active.	• All IBM Virtual Shared Disks defined on node1 are put into stopped state on all clients. • Daemons running on the other nodes in the system partition remove node1 from the group. • Node2 takes over the management of node1's IBM Virtual Shared Disks. • All clients designate node2 as the server for node1's IBM Virtual Shared Disks and put those IBM Virtual Shared Disks into active state.
Node2, secondary to node1	• Daemons running on the other nodes in the system partition accept node2 into the group. • If node1 is active, there is no change to the status of the node1 IBM Virtual Shared Disks. • If node1 is inactive, node2 takes over the management of the node1 IBM Virtual Shared Disks. The node1 IBM Virtual Shared Disks are designated as active and managed by node2 on all clients.	• If node1 is active, there is no change to the status of its IBM Virtual Shared Disks. • If node1 is inactive, the node1 IBM Virtual Shared Disks are put into stopped state on all clients. They remain in stopped state until node1 or node2 comes back up. • Daemons running on the other nodes in the system partition remove node2 from the group.
Node3, client	• Daemons running on the other nodes in the system partition accept node3 into the group. • All IBM Virtual Shared Disks on active nodes for which node3 is a client are put into active state from node3's point of view.	• All IBM Virtual Shared Disks defined on node3 are put into stopped state from node3's point of view. • Daemons running on the other nodes in the system partition remove node3 from the group.

The hc Subsystem

The hc subsystem is also called the Connection Manager. It supports the development of recoverable applications. Chapter 5, “Application Programming Considerations” on page 43 provides more information on how to write recoverable applications. The hc subsystem maintains a membership list of the nodes that are currently running hc processes and an incarnation number that is incremented every time the membership list changes. The hc subsystem shadows the rvsd subsystem; recording the same changes in state and management of IBM Virtual Shared Disks that rvsd records. The difference is that hc only records these changes after rvsd processes them, to assure that IBM Recoverable Virtual Shared Disk recovery activities begin and complete before the recovery of hc client applications takes place. This serialization helps ensure data integrity. It also explains why hc cannot run on a node where rvsd is not running.

Another important characteristic of the hc subsystem is that it will wait for its client application to connect before joining the hc group. It will also leave the group if the client application loses its connection. This means that the hc group's membership list corresponds to the list of nodes in which the client application is currently running.

Startup Time

When you start up your SP, the IBM Recoverable Virtual Shared Disk software does the following:

- Configures **all** defined IBM Virtual Shared Disks at system startup using **cfgvsd** –**a**. All the IBM Virtual Shared Disks on the nodes are in stopped state.
- When quorum is reached, the IBM Recoverable Virtual Shared Disk software activates all IBM Virtual Shared Disks with servers in the active group.
  - All IBM Virtual Shared Disks on the active nodes that also have a server in the active group will be in the active state.
  - If an IBM Virtual Shared Disk's primary server node is in the active group, the primary will be the server for that IBM Virtual Shared Disk.
  - If an IBM Virtual Shared Disk's primary server is not in the active group but the secondary server is, the secondary will be the server for that vsd.
  - All IBM Virtual Shared Disks without a server in the active group will be in the stopped state.
- If quorum is lost, all the IBM Virtual Shared Disks are put into the stopped state. When quorum is active again, the IBM Virtual Shared Disks are put into appropriate states based on the list above.

The IBM Recoverable Virtual Shared Disk software actively manages all configured IBM Virtual Shared Disks. The recovery scripts invoke the state-changing commands listed below, as well as the **varyonvg**, and **varyoffvg** commands asynchronously and automatically in the background as IBM Virtual Shared Disk server nodes become active or fail.

If you have any script that issues any IBM Virtual Shared Disk command, you should disable it so it does not interfere with IBM Recoverable Virtual Shared Disk recovery. The IBM Recoverable Virtual Shared Disk software issues these commands for you.

The commands are:

cfgvsd
cfghsdvsd
preparevsd
resumevsd
startvsd
stopvsd
suspendvsd
ucfgvsd
ucfghsdvsd

Note: In order for HA to be activated, there must be a quorum; that is, more than half of the IBM Virtual Shared Disk nodes defined in the system partition, including the control workstation, must be up and active.

Note: If your system is already running, you can use the **ha_vsd reset** command. See the **ha_vsd reset** reference page in *IBM Parallel System Support Programs for AIX: Command and Technical Reference* or in Appendix A, "IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands" on page 77, for the command syntax.

Command Considerations

Because the IBM Recoverable Virtual Shared Disk software is actively managing your IBM Virtual Shared Disks, there are some limitations on your use of the IBM Virtual Shared Disk commands. This section describes those limitations. Commands not listed should not interfere with IBM Recoverable Virtual Shared Disk operation.

In all cases, the same limitations or considerations apply whether a command is issued from the command line or through a corresponding SMIT panel.

Command	Consideration
cfghsd and cfghsdvsd	When used with the IBM Recoverable Virtual Shared Disk software, you must edit the **ha_vsd** script and uncomment the lines that refer to **cfghsdvsd** or **cfghsd**. See the Data Striping Device documentation in *IBM Parallel System Support Programs for AIX: Installation Guide*
cfgvsd	The IBM Recoverable Virtual Shared Disk software manages all configured IBM Virtual Shared Disks on a node. When the IBM Recoverable Virtual Shared Disk software is started, it issues **cfgvsd -a**. Only use this command to add a new IBM Virtual Shared Disk nondisruptively: 1. Define a new IBM Virtual Shared Disk. The primary node must be up. 2. Issue **cfgvsd** *vsd_name* on all nodes. 3. Issue **startvsd** *vsd_name* on all nodes. The only time you use the –**b** option is if the primary node fails after step 1. In that case, use the –**b** option in step 3.
preparevsd	Do not issue this command.
resumevsd	Do not issue this command.
startvsd	Only use this command when nondisruptively adding a new IBM Virtual Shared Disk: 1. Define the new IBM Virtual Shared Disk. The primary node must be up. 2. Issue **cfgvsd** *vsd_name* on all nodes. 3. Issue **startvsd** *vsd_name* on all nodes.

	The only time you use the –**b** option is if the primary node fails after step 1. In that case, use the –**b** option in step 3.
stopvsd	This command should be used only as part of the command sequence to undefine an IBM Virtual Shared Disk nondisruptively while the IBM Recoverable Virtual Shared Disk software is running: 1. Suspend the IBM Virtual Shared Disk on all nodes. 2. Stop the IBM Virtual Shared Disk on all nodes. 3. Unconfigure the IBM Virtual Shared Disk. 4. Undefine the IBM Virtual Shared Disk. This sequence removes an IBM Virtual Shared Disk definition, therefore, this IBM Virtual Shared Disk is no longer managed by the IBM Recoverable Virtual Shared Disk software.
suspendvsd	This command should only be used as part of the command sequence to undefine an IBM Virtual Shared Disk nondisruptively while the IBM Recoverable Virtual Shared Disk software is running: 1. Suspend the IBM Virtual Shared Disk on all nodes. 2. Stop the IBM Virtual Shared Disk on all nodes. 3. Unconfigure the IBM Virtual Shared Disk. 4. Undefine the IBM Virtual Shared Disk. This sequence removes an IBM Virtual Shared Disk definition, therefore, this IBM Virtual Shared Disk is no longer managed by the IBM Recoverable Virtual Shared Disk software.
ucfgvsd and ucfghsdvsd	These commands put IBM Virtual Shared Disks in the unconfigured state, causing the IBM Recoverable Virtual Shared Disk software to stop managing that IBM Virtual Shared Disk on that node until recovery services are restarted by **ha_vsd reset** or until the node is rebooted. At that time, the IBM Recoverable Virtual Shared Disk software resumes management of all defined IBM Virtual Shared Disks.
vsdelnode	This command is used to remove an IBM Virtual Shared Disk node and resynchronize the IBM Virtual Shared Disk load tables. The affected node is unable to access any IBM Virtual Shared Disks. Using this command with the IBM Recoverable Virtual Shared Disk can cause the IBM Recoverable Virtual Shared Disk software to hang across the whole system if it is not followed by an **ha_vsd reset** on all IBM Virtual Shared Disk nodes. After you issue this command, issue **ha_vsd reset** at both the current and former IBM Virtual Shared Disk nodes.

vsdnode	This command is used to add a new IBM Virtual Shared Disk node. Using this command with the IBM Recoverable Virtual Shared Disk can cause the IBM Recoverable Virtual Shared Disk software to hang across the whole system if it is not followed by an **ha_vsd reset** on all IBM Virtual Shared Disk nodes. After you issue this command, issue **ha_vsd reset** at both the current and former IBM Virtual Shared Disk nodes.
vsdvg	Refer to “Creating a Twin-tailed Volume Group” on page 112 for sequences showing **vsdvg** use. If you use **createvsd** or **createhsd** to set up IBM Virtual Shared Disks or Data Striping Devices with secondary direct clients, you don't need to use **vsdvg** directly.

IBM Virtual Shared Disk Data Striping Device Considerations

When an IBM Virtual Shared Disk node fails and its secondary node takes over, the takeover is transparent to the HSDs. The HSDs work with the IBM Virtual Shared Disks and do not know where the IBM Virtual Shared Disk servers are.

If the HSD driver is working with the IBM Recoverable Virtual Shared Disk software, the HSDs can also be made recoverable. To do this, edit the **/usr/lpp/csd/bin/ha_vsd** file and uncomment the lines that refer to **cdghsdvsd** or **cfghsd**. Now, when a node is rebooted, all the defined HSDs will be configured after all the defined IBM Virtual Shared Disks are configured.

Chapter 8. IBM Recoverable Virtual Shared Disk Recovery Scenarios

The IBM Recoverable Virtual Shared Disk software consists of several interdependent components. In order to plan for successful recovery from unexpected system events in a multinode application, you must consider the relationship and dependencies of these components. The following recovery scenarios describe some of the events that could occur and some of the steps you and your system could take to recover.

The recovery scenarios describe what an administrator or operator may see, and tells how the IBM Recoverable Virtual Shared Disk software can recover.

Recognizing Recovery

The suspended state of an IBM Virtual Shared Disk is a transient state while recovery is in progress. You know that the IBM Recoverable Virtual Shared Disk software is performing recovery processing when you see IBM Virtual Shared Disks in the suspended state after they have been active when you did not put them there. The IBM Virtual Shared Disk Perspective main panel displays the states of your IBM Virtual Shared Disk clients. To see them, select a node in the IBM Virtual Shared Disk Nodes work area and then click on the expansion arrow. All the clients on that node will be displayed.

If you recognize that recovery is not taking place normally, verify that the recovery daemon is active on all nodes. Use **ha.vsd query** and **hc.vsd query** to see if the rvsd and hc subsystems are active.

You can issue **ha_vsd reset** to restart the IBM Recoverable Virtual Shared Disk software. You may be able to bring the IBM Virtual Shared Disks to the proper state, or you may need to reboot all the IBM Virtual Shared Disk nodes to fully recover.

Planning For Recovery

You should have disabled or removed all user-provided scripts that issue change-of-state commands to IBM Virtual Shared Disks.

Monitoring the IBM Recoverable Virtual Shared Disk activity with the IBM Virtual Shared Disk Perspective can help you to be aware of potential problems earlier.

The format for each of the following recovery scenarios is:

1. Symptom
2. Detection
3. Affected components
4. Recovery steps
5. Restart

© Copyright IBM Corp. 1994, 1996

IBM Virtual Shared Disk Client or Server Node Failure

- **Symptoms**

  A node has either hung or has failed.

- **Detection**

  Node failure was detected by the **hatsd** daemon. An operator may have seen the change in state displayed by the IBM Virtual Shared Disk Perspective. Failure to communicate with the node over the Ethernet can cause this symptom.

- **Affected Components**

  The affected components can include everything accessing the IBM Virtual Shared Disks, including the IBM Virtual Shared Disks themseoves, the application and its components, the **hatsd** daemon and its related functions, and the IBM Recoverable Virtual Shared Disk software itself running on the failed nodes.

- **Recovery Steps**

  The recovery steps are:

  1. The **hatsd** daemons running on surviving nodes notice the node failure. Recovery then begins on the surviving IBM Virtual Shared Disk nodes.
  2. The recovery services of surviving nodes suspend IBM Virtual Shared Disks served by the failing node. The IBM Virtual Shared Disk recovery process configures available backup servers, and the suspended IBM Virtual Shared Disks are made active.
  3. Optional subscribers to **hc** on surviving nodes are informed of the membership change.

- **Restarting the Failed Node**

  Some of the following actions must be done manually; others are done automatically by the IBM Recoverable Virtual Shared Disk software. You should have procedures in place to instruct operators when and how to do the manual operations.

  1. An operator reboots the failed node.
  2. An operator issues the **Estart** command (if you are using the High Performance Switch) to enable the rebooted node to access the switch. This is a manual step, asynchronous to all the automatic steps that follow, although it must be done to enable IBM Virtual Shared Disk access from this node.
  3. The IBM Recoverable Virtual Shared Disk software components are automatically brought up on reboot.
  4. The rebooted node rejoins the active group. If it was a primary node, it takes over again as a primary node.
  5. The **hc.activate** script, if present, is invoked.
  6. Optional subscribers to **hc** on surviving nodes are informed of the membership change.
  7. If the IBM Virtual Shared Disks use the SP switch as their IP adapter, an operator issues **Eunfence** to give clients access to the rebooted node.

Switch Failure Scenario

- **Symptoms**

  Remote IBM Virtual Shared Disk I/O requests hang and then fail after about 15 minutes. The IBM Virtual Shared Disk clients see a time out error.

- **Detection**

  The IBM Recoverable Virtual Shared Disk software neither detects nor handles switch failure. An operator using the IBM Virtual Shared Disk Perspective may recognize that **switch_responds** for the node is off.

- **Affected Components**

  Applications using IBM Virtual Shared Disks will hang and then fail after about 15 minutes.

- **Recovery Steps and Restart**

  An operator issues the **Estart** or **Eunfence** command to restart the switch.

  If **Estart** or **Eunfence** fails, use standard diagnostic methods for handling switch problems (see the switch error logs in *IBM Parallel System Support Programs for AIX: Diagnosis and Messages*).

Disk Controller Failure Scenario

- **Failure Symptoms**

  Requests to the disk corresponding to the failed controller return errors.

- **Detection**

  The IBM Recoverable Virtual Shared Disk software neither detects nor handles disk controller failure.

  Issue the **errpt** command to check for hardware failure.

- **Affected Components**

  The IBM Virtual Shared Disk application cannot access the data on the affected disk.

- **Recovery Steps**

  1. The operator must shut down the node with the failed disk controller.
  2. The node failure recovery procedure takes place. The IBM Recoverable Virtual Shared Disk software uses the secondary node (and its disk controller) to access the affected disk.
  3. Follow your local procedures for hardware repair.

- **Restart**

  Do one of the following:

  1. The operator must replace the disk controller on the failed node. Then follow the steps in on page 68 starting with step 2 on page 68.

  or

2. If you cannot fix the component but you want to bring up the node, use SMIT or the IBM Virtual Shared Disk Perspective to make sure the node is not the IBM Virtual Shared Disk server.

HA Topology Services or Recovery Service Daemon Failure

1. If any one of the heartbeat or recovery service daemons fails and exits, the complete set is automatically restarted.
   a. At the node where the failure occurred, all the IBM Virtual Shared Disks stop and restart, causing I/O errors to the application.
   b. At other nodes, the IBM Virtual Shared Disks served by the problem node are switched to their secondary servers briefly, then returned to the primary server.
2. In the unlikely event that the recovery service daemons hang, no recovery is performed. You must manually issue **ha_vsd reset** on the problem node. If this is insufficient, reboot the problem node.
   a. At the node where failure exists, the IBM Virtual Shared Disks remain in the active state, but subsequent node failures or reboots can cause I/O requests to remote IBM Virtual Shared Disks to hang and fail after 15 minutes. Issue the **ps** command at that node to check for the **rvsd** and **hc** daemons.
   b. At the other nodes, some IBM Virtual Shared Disks remain indefinitely in the suspended state. All the IBM Recoverable Virtual Shared Disk daemons are present and their presence can be verified with **ps**.

IBM Virtual Shared Disk Failure

- **Failure Symptoms**

  The symptoms of an IBM Virtual Shared Disk software failure are I/O errors. Any hardware, LVM, or network error may give the same symptoms.

- **Detection**

  There is no automatic method to determine that IBM Virtual Shared Disk failures are occurring.

- **Affected Components**

  The affected components are the IBM Virtual Shared Disk client applications. When failure occurs:

  – The IBM Virtual Shared Disk could return errors to the application. The operator must resolve the IBM Virtual Shared Disk problem.
  – The IBM Virtual Shared Disk failure could cause a node to crash. This situation should be treated as a node failure. Follow the steps in “IBM Virtual Shared Disk Client or Server Node Failure” on page 68.

Recovery

Reboot the node if necessary, then follow the procedures for restarting a failed node. If that action fails, follow your local procedures for obtaining service.

Chapter 9. Using HACMP with IBM Recoverable Virtual Shared Disk

If you are using HACMP with IBM Recoverable Virtual Shared Disk, the hc subsystem does not run under the control of Availability Services; it is controlled by HACMP. Under HACMP, hc is independent of Availability Services and Group Services. The **hc.vsd** command does not work when HACMP is present in the system partition.

You don't need to configure the **hc.activate** and **hc.deactivate** scripts when IBM Recoverable Virtual Shared Disk is managed by HACMP/6000. Instead, you define your application to HACMP/6000 as an application server.

Configuring IBM Recoverable Virtual Shared Disk with HACMP/6000

Follow the instructions in *HACMP/6000 Administration Guide* to configure your IBM Recoverable Virtual Shared Disk cluster and adapters.

HACMP will not manage the IBM Recoverable Virtual Shared Disks unless the IP addresses used for IBM Virtual Shared Disk communication are defined to it. This step sets up the interfaces to your application. You should use the switch as your IP adapter for IBM Virtual Shared Disks. The IP address used for IBM Virtual Shared Disk definitions and traffic must be defined to HACMP as adapters (service addresses). The adapters need not be part of a resource group. Remember to configure any switch addresses with the string "HPS" in the hacmp network name. Use the switch NIM and define the network as *service* and *private*. Make sure you also specify the HACMP node name where that address resides.

IBM Recoverable Virtual Shared Disk manages the underlying volume groups upon failover. The information that controls failover is contained in the IBM Virtual Shared Disk fields in the SDR. No further configuration is required in HACMP.

Note: Do not use **hc.activate** and **hc.deactivate** when HACMP/6000 is managing IBM Recoverable Virtual Shared Disk. If you were previously using **hc.activate** and **hc.deactivate** scripts, back up these scripts and ensure that they are either deleted or empty (cannot be invoked). This prevents HACMP and IBM Recoverable Virtual Shared Disk from both trying to start and stop your applications.

HACMP needs an extra configuration step to allow it to manage the IBM Recoverable Virtual Shared Disks. In order to start the **hcd** daemon automatically with HACMP present, you must issue the following command:

```
mkitab "hacmp_hc:2:respawn:/usr/lpp/csd/bin/hacmp_hc_start\
>/dev/console 2>&1"
```

Restrictions

Please note the following points concerning the use of IBM Recoverable Virtual Shared Disk with HACMP/6000:

© Copyright IBM Corp. 1994, 1996

- A node with HACMP/6000 installed on it will never activate the daemons that control IBM Recoverable Virtual Shared Disk without HACMP.
- IBM Recoverable Virtual Shared Disk controlled by HACMP/6000 and IBM Recoverable Virtual Shared Disk controlled by Group Services cannot coexist within the same system partition.

Note: If a set of SP nodes run HACMP and the rest of the nodes run IBM Recoverable Virtual Shared Disks with the group services daemons, the following rules must be observed:

- No IBM Recoverable Virtual Shared Disks can be managed from the HACMP/6000 nodes. These nodes should be removed from the IBM Virtual Shared Disk configurations if they exist (in other words, there should be no IBM Virtual Shared Disk information in the SDR for these nodes).
- HACMP/6000 can be configured and used normally, but must not manage IBM Virtual Shared Disks in this system.

The following table describes dependencies and the supported environment of HACMP/6000-controlled IBM Recoverable Virtual Shared Disks.

Table 5. Control Functions with and without HACMP/6000

	HACMP/6000 and IBM Recoverable Virtual Shared Disk	**IBM Recoverable Virtual Shared Disk Alone**
IBM Recoverable Virtual Shared Disk is controlled by HACMP?	Yes	No
IBM Recoverable Virtual Shared Disk is controlled by **hatsd**?	Not supported	Yes
IP address takeover?	Yes	No
Other volume group management or recovery?	Yes	No
Error notification on adapter errors?	Yes	Yes

The IBM Recoverable Virtual Shared Disk licensed program managed through Group Services can accommodate up to a 128-node IBM Virtual Shared Disk configuration. Since the limit of the HACMP-controlled IBM Recoverable Virtual Shared Disk product is eight nodes, this is the limit on the size of an HACMP/6000-controlled IBM Recoverable Virtual Shared Disk configuration. Multiple HACMP/6000-controlled IBM Recoverable Virtual Shared Disk configurations in a system partition are permitted, provided that the IBM Virtual Shared Disks defined are independent (not shared) among the configurations.

Note: All IBM Recoverable Virtual Shared Disks are configured. HACMP only activates the IBM Recoverable Virtual Shared Disks served within its cluster.

Chapter 10. Diagnosis and Messages Information

This chapter provides information on debug files and messages that can aid you in diagnosing and resolving IBM Recoverable Virtual Shared Disk software problems.

Diagnosis Information

There are two IBM Recoverable Virtual Shared Disk script debug files that provide information that can help you in diagnosing problems with the IBM Recoverable Virtual Shared Disk software:

/var/adm/SPlogs/csd/vsd.log
: Provides a history of actions taken by the recovery scripts (driven by rvsd)

/var/tmp/vsd.debuglog
: Provides verbose descriptions of actions taken by the recovery scripts (driven by rvsd)

Messages

2506-001 Variable text.

Explanation: These messages come from the **rvsdd** daemon, which provides services to the IBM Recoverable Virtual Shared Disk software on the SP The text of the message will reveal either a usage, configuration, resource, or internal problem.

User Response: The **rvsdd** daemon is started automatically by the internal SP heartbeat service on all SP rack nodes. This daemon is only required on nodes defined with a non-null VSD_adapter in the SDR, and will exit otherwise.

If the **hc** daemon is running on a node, it will monitor the condition of the **rvsdd** daemon. If **hc** detects **rvsdd** exiting, if **rvsdd** detects **hc** exiting, or if **init** detects **heartbeatd** exiting, automatic recovery will restart **rvsdd**. Therefore, no action is normally required.

If you believe **rvsdd** is broken on a node, you can issue the **/usr/lpp/ssp/bin/ha_vsd reset** command to cause **heartbeatd** and thus **rvsdd** and **hc** recovery to be attempted manually.

If the problem persists, follow your local procedures for obtaining service.

2506-002 Variable text.

Explanation: These messages come from the **hc** daemon, which provides services to a recoverable service on the SP. The text of the message will reveal either a usage, configuration, resource, or internal problem.

User Response: The **hc** daemon is started automatically by the internal SP heartbeat service on those SP rack nodes that have a non-null VSD_adapter value defined in the SDR.

If the **hc** daemon is running on a node, it will monitor the condition of the **rvsdd** daemon. If **hc** detects **rvsdd** exiting, if **rvsdd** detects **hc** exiting, or if **init** detects **heartbeatd** exiting, automatic recovery will restart **hc**. Therefore, no action is normally required.

If you believe **hc** is broken on a node, you can issue the **/usr/lpp/ssp/bin/ha_vsd reset** command to cause **heartbeatd** and thus **hc** and **rvsdd** recovery to be attempted manually.

If the problem persists, follow your local procedures for obtaining service.

© Copyright IBM Corp. 1994, 1996

Appendix A. IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands

This appendix lists the purpose and syntax for the IBM Virtual Shared Disk, Data Striping Device, and IBM Recoverable Virtual Shared Disk commands.

To access the SP online manual pages, set the MANPATH environment variable as follows:

for ksh

```
export MANPATH=$MANPATH:/usr/lpp/ssp/man
```

for csh

```
setenv MANPATH $MANPATH\:/usr/lpp/ssp/man
```

cfghsd

Purpose

cfghsd – Configures a data striping device (HSD) for an IBM Virtual Shared Disk.

Syntax

cfghsd –**a** *hsd_name* ...

cfghsdvsd

Purpose

cfghsdvsd – Configures a data striping device (HSD) and the IBM Virtual Shared Disks that comprise it on one node.

Syntax

cfghsdvsd –**a** | {*hsd_name*...}

cfgvsd

Purpose

cfgvsd – Configures an IBM Virtual Shared Disk.

Syntax

cfgvsd –**a** | *vsd_name* ...

© Copyright IBM Corp. 1994, 1996

cksumvsd

Purpose

cksumvsd – Views and manipulates the IBM Virtual Shared Disk checksum parameters.

Syntax

cksumvsd [–**s**] [–**R**] [–**i** | –**l**]

createhsd

Purpose

createhsd – Creates one data striping device (HSD) that encompasses two or more IBM Virtual Shared Disks.

Syntax

createhsd –**n** {*node_list* | **ALL**} –**s** *size_in_MB* –**g** *vg_name*
–**t** *stripe_size_in_KB* [–**c** *vsds_per_node* [–**A**]] [–**m** *mirror_cnt*]
[–**d** *hsd_name*] [–**l** *lv_name_prefix*] [–**S**] [–**o cache** | **nocache**]
[–**T** *lp_size_in_MB*]

createvsd

Purpose

createvsd – Creates a set of IBM Virtual Shared Disks, with their associated logical volumes, and puts information about them into the System Data Repository (SDR).

Syntax

createvsd –**n** {*node_list* | **ALL**}
–**s** *size_in_MB* –**g** *volume_group_name*
[{–**c** *number_of_vsds_per_node* | –**L**}]
[–**o cache** | **nocache**] [–**m** *mirror_count*]
[–**p** *lvm_stripe_size*] [–**v** *vsd_name_prefix*]
[–**l** *lv_name_prefix*] [–**T** *lp_size_in_MB*]

ctlhsd

Purpose

ctlhsd – Sets the data striping device (HSD) for the IBM Virtual Shared Disks operational parameters or reset statistics.

Syntax

ctlhsd [–**p** *parallel_level* | –**v** *hsd_name* ... | –**C** | –**V**]

---

ctlvsd

Purpose

ctlvsd – Sets IBM Virtual Shared Disk operational parameters.

Syntax

ctlvsd [–**t**] | [–**T**] | [–**c** *NewCacheSize* | –**r** *node_number*... | –**R** | –**k** *node_number*... –**K** | –**p** *1—9* | –**M** *max_IP_msg_size* | –**v** *vsd_name* ... | –**C** | –**V**]

---

defhsd

Purpose

defhsd – Defines a data striping device (HSD).

Syntax

defhsd **protect_LVCB** | **not_protect_LVCB** *hsd_name stripe_size vsd_name...*

---

defvsd

Purpose

defvsd – Defines an IBM Virtual Shared Disk.

Syntax

defvsd *logical_volume_name global_group_name vsd_name* [**nocache** | **cache**]

---

ha.vsd

Purpose

ha.vsd – Queries and controls the rvsd subsystem of IBM Recoverable Virtual Shared Disk.

Syntax

ha.vsd {**debug**[*n*] | **query** | **quorum** *n* | **qsrc** | **refresh** | **reset** | **start** | **stop** }

hc.vsd

Purpose

hc.vsd – Queries and controls the hc subsystem of IBM Recoverable Virtual Shared Disk.

Syntax

hc.vsd {**debug**[*n*] | **mksrc** | **query** | **qsrc** | **reset** | **start** | **stop** }

hsdatalst

Purpose

hsdatalst – Displays data striping device (HSD) data for the IBM Virtual Shared Disks from the System Data Repository (SDR).

Syntax

hsdatalst [–**G**]

hsdvts

Purpose

hsdvts – Verifies that a data striping device (HSD) for the IBM Virtual Shared Disks has been correctly configured and works.

Syntax

hsdvts *hsd_name*

lshsd

Purpose

lshsd – Displays configured data striping device (HSD) for the IBM Virtual Shared Disks and their characteristics.

Syntax

lshsd [–**l** | –**s**] [*hsd_name* ...]

lsvsd

Purpose

lsvsd – Displays configured IBM Virtual Shared Disks and their characteristics.

Syntax

lsvsd [–**l** | –**s**] [–**i**] [*vsd_name*...]

---

preparevsd

Purpose

preparevsd – Makes an IBM Virtual Shared Disk available.

Syntax

preparevsd –**a** | *vsd_name*...

---

removehsd

Purpose

removehsd – Removes one or more data striping devices, the IBM Virtual Shared Disks associated with them, and the System Data Repository (SDR) information for IBM Virtual Shared Disks on the associated nodes. If the only volume groups defined on the associated nodes are the ones used by the data striping devices that are being removed, those volume groups are also removed.

Syntax

removehsd {–**v** {*hsd_names* | –**a**}} [–**f**]

---

removevsd

Purpose

removevsd – Removes a set of IBM Virtual Shared Disks that are not part of any data striping device (HSD).

Syntax

removevsd –**v** *vsd_names* | –**a** [–**f**]

---

resumevsd

Purpose

resumevsd – Makes an available IBM Virtual Shared Disk active.

Syntax

resumevsd –**a** | *vsd_name*...

---

spvsd

Purpose

spvsd – Directly launches the IBM Virtual Shared Disk window of the SP Perspectives graphical user interface (GUI).

Syntax

spvsd [–**userProfile** *name*] [–**systemProfile** *name*] [–**noProfile**]
[{–**backgroundColor** | **bg**} *colorName*]
[{–**foregroundColor** | –**fg**} *colorName*] [–**fontFamily** *name*]
[–**fontSize** *size*] [–**fontBold**] [–**fontItalic**]

---

startvsd

Purpose

startvsd – Makes an IBM Virtual Shared Disk available and activates it.

Syntax

startvsd [–**p** | –**b**] –**a** | *vsd_name* ...

---

statvsd

Purpose

statvsd – Displays IBM Virtual Shared Disk device driver statistics of a node.

Syntax

statvsd

---

stopvsd

Purpose

stopvsd – Makes an IBM Virtual Shared Disk unavailable.

Syntax

stopvsd –**a** | *vsd_name* ...

suspendvsd

Purpose

suspendvsd – Deactivates an available IBM Virtual Shared Disk.

Syntax

suspendvsd –**a** | *vsd_name*...

ucfghsd

Purpose

ucfghsd – Makes a data striping device (HSD) for the IBM Virtual Shared Disks unavailable.

Syntax

ucfghsd –**a** | *hsd_name*...

ucfghsdvsd

Purpose

ucfghsdvsd – Makes a data striping device (HSD) and the IBM Virtual Shared Disks that comprise it unavailable on one node.

Syntax

ucfghsdvsd –**a** | {*hsd_name*...}

ucfgvsd

Purpose

ucfgvsd – Makes an IBM Virtual Shared Disk unavailable.

Syntax

ucfgvsd –**a** | *vsd_name* ...

undefhsd

Purpose

undefhsd – Undefines a data striping device (HSD).

Syntax

undefhsd *hsd_name*...

---

undefvsd

Purpose

undefvsd – Undefines an IBM Virtual Shared Disk.

Syntax

undefvsd *vsd_name* ...

---

updatehsd

Purpose

updatehsd – Lets you change the option in the System Data Repository (SDR) that prevents overwriting the Logical Volume Control Block (LVCB) for specified data striping devices (HSDs).

Syntax

updatehsd {–**d** *hsd_names* | –**a}** [–**f**]
–**o** {**protect_lvcb** | **not_protect_lvcb**}

---

updatevsdnode

Purpose

updatevsdnode – Changes IBM Virtual Shared Disk options in the System Data Repository (SDR).

Syntax

updatevsdnode {–**n** *node_number* | **ALL** } {[–**a** *VSD_adapter*]
[–**i** *init_cache_buffer_count*]
[–**m** *max_cache_buffer_count*] [–**r** *vsd_request_count*]
[–**p** *rw_request_count*] [–**b** *min_buddy_buffer_size*]
[–**x** *max_buddy_buffer_size*] [–**s** *max_buddy_buffers*]
[–**?** *vsd_max_ip_packet_size*]}
[–**f**]

---

updatevsdtab

Purpose

updatevsdtab – Changes the IBM Virtual Shared Disk **cache**/**nocache** option in the System Data Repository (SDR).

Syntax

updatevsdtab {–**v** *vsd_names* | –**a**} –**o cache** | **nocache** [–**f**]

---

verparvsd

Purpose

verparvsd – Verifies IBM Virtual Shared Disk system partitioning.

Syntax

verparvsd [–**F**] [–**o** *output_file*] *layout_directory* [*new_partition* ...]

---

vsdatalst

Purpose

vsdatalst – Displays IBM Virtual Shared Disk system definition data from the System Data Repository (SDR).

Syntax

vsdatalst [–**G**] –**g** | –**n** | –**v**

---

vsddiag

Purpose

vsddiag – Displays information about the status of IBM Virtual Shared Disks.

Syntax

vsddiag

---

vsdelnode

Purpose

vsdelnode – Removes IBM Virtual Shared Disk information for a node or series of nodes from the System Data Repository (SDR).

Syntax

vsdelnode *node_number* ...

---

vsdelvg

Purpose

vsdelvg – Removes IBM Virtual Shared Disk global volume group information from the System Data Repository (SDR).

Syntax

vsdelvg [–**f**] *global_group_name* ...

---

vsdnode

Purpose

vsdnode – Enters IBM Virtual Shared Disk information for a node or series of nodes into the System Data Repository (SDR).

Syntax

vsdnode *node_number... adapter_name init_cache_buffer_count max_cache_buffer_count vsd_request_count rw_request_count min_buddy_buffer_size max_buddy_buffer_size max_buddy_buffers vsd_max_ip_msg_size*

---

vsdsklst

Purpose

vsdsklst – Produces output that shows you the disk resources used by the IBM Virtual Shared Disk subsystem across a system or system partition.

Syntax

vsdsklst {–**d** | –**v** | –**dv**} {[–**a**] | –**n** *node_list*]} [–**G**]

---

vsdvg

Purpose

vsdvg – Defines an IBM Virtual Shared Disk global volume group.

Syntax

vsdvg [–**g** *global_group_name*] *local_group_name primary_server_node*
[*secondary_server_node*]

---

vsdvts

Purpose

vsdvts – Verifies that the IBM Virtual Shared Disk software works.

Syntax

vsdvts [–**b** *block_size*] [–**n** *number_of_blocks*] *vsd_name* [*file*]

vsdvts

Appendix B. Single-Node Command and SMIT Interfaces for IBM Virtual Shared Disks, Data Striping Devices and the IBM Recoverable Virtual Sha red Disk Licensed Product

With Version 2 Release 2 of the SP system, commands that operated on multiple nodes at once, such as **createvsd** and **createhsd**, have been introduced to improve the usability of these components. However, some users may have scripts that use single-node commands, such as **defvsd** and **defhsd**. This appendix contains the previous version of the information on defining and managing IBM Virtual Shared Disks, Data Striping Devices, and the IBM Recoverable Virtual Shared Disk software.

Command and SMIT Interfaces for IBM Virtual Shared Disks

Entering Node Information in the SDR for One Node

You can use SMIT or the **vsdnode** command to enter IBM Virtual Shared Disk node information into the SDR.

To enter IBM Virtual Shared Disk node information using SMIT:

SELECT VSD Node Information

▲ The VSD Node Information dialog appears.

For the syntax of the **vsdnode** command, see Appendix A, “IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands” on page 77.

The following list describes the information that is needed in the SDR.

- Node number

  Each node is identified by a *node_number* that represents its position in the SP rack.

- IBM Virtual Shared Disk Adapter Name

  The name of the adapter to be used for IBM Virtual Shared Disk communications with this node.

  The IBM Virtual Shared Disk component uses this name and the related node number to get the IBM Virtual Shared Disk IP address associated with each node from the SDR **Adapter** object. Either the High Performance Switch or the Scalable Parallel (SP) Switch (both are defined as **css0**) is recommended as the IBM Virtual Shared Disk adapter for best performance.

  The set of nodes and addresses in the IBM Virtual Shared Disk configuration (the clients and servers) must define a fully-connected network, that is, all nodes can send messages directly to each other. For the best performance, however, all the nodes should be directly connected to each other over the network.

- Initial cache buffer count

© Copyright IBM Corp. 1994, 1996

The IBM Virtual Shared Disk device driver implements an optional write-through cache of pinned kernel memory with a block size of 4KB. When the first cached IBM Virtual Shared Disk is configured on a node, the cache is created, and it contains the specified number of blocks. The minimum value is 1. The recommended value is 64, which results in a 256KB cache.

- Maximum cache buffer count

  The number of buffers in the cache can be increased (using **ctlvsd**) up to this value. You cannot decrease the number of cache blocks. You can start over by unconfiguring all the IBM Virtual Shared Disks, changing the value of the this number in the SDR, and then, on the first **cfgvsd**, the initial number of IBM Virtual Shared Disk buffer cache blocks will be created. The recommended value is 256, resulting in a 1MB cache.

- Number of outstanding logical volume read/write requests

  Specifies the maximum number of outstanding requests the IBM Virtual Shared Disk device driver will allow for each underlying logical volume. Each such request uses a pbuf. A pbuf is a little bigger than a buf structure, which is approximately 104 bytes. The IBM Virtual Shared Disk device driver allocates this number of pbufs in pinned kernel memory whenever an IBM Virtual Shared Disk that has this node as its primary is configured. The minimum value is 1. The recommended value for a primary or secondary node is 48. If the node is not going to be used as a primary or secondary, this value is irrelevant.

- Minimum buddy buffer size

  The smallest buddy buffer a server will use to satisfy a remote request to an IBM Virtual Shared Disk. This value must be a power of 2 and greater than or equal to 4096. The recommended value is 4096 (4KB). For a 512-byte request, 4KB is clearly overkill but a buddy buffer is used only for the short period of time a remote request is being processed at the direct-client node.

  For more information on buddy buffers, see “Buddy Buffers” on page 53.

- Maximum buddy buffer size

  The largest buddy buffer a server will use to satisfy a remote request. This value must be a power of 2 and greater than or equal to the *min_buddy_buffer_size*. The maximum value, and the recommended value, is 65536 (64KB). This value must be the same on all nodes within a system partition.

- Number of buddy buffers

  The buddy buffer is pinned kernel memory allocated when the IBM Virtual Shared Disk device driver is loaded the first time an IBM Virtual Shared Disk is configured (and freed when the last IBM Virtual Shared Disk is unconfigured). If you do not use the switch as your IP adapter, the size of the buddy buffer affects the number of remote requests the IBM Virtual Shared Disk server node can handle at one time. Remote requests can queue waiting for a buddy buffer. The **statvsd** command and the IBM Virtual Shared Disk Statistics notebook page of the IBM Virtual Shared Disk Perspective report this queuing as buddy buffer shortages. See “Buddy Buffers” on page 53 for more information about buddy buffers and how to determine how many to define. The recommended value when the switch is used as the IP adapter is no more than 4, which results in a 256KB buddy buffer size in combination with the recommended maximum buddy buffer size of 64KB.

- Request block count

  Specifies the maximum number of outstanding IBM Virtual Shared Disk requests originating on the node. The IBM Virtual Shared Disk device driver allocates this number of request blocks in pinned kernel memory the first time an IBM Virtual Shared Disk is configured. If the number is too small, local requests will queue up waiting for a request block to become available. The recommended value is 256. The size of the block is approximately 76 bytes.

- Maximum IP message size

  The maximum IP message size for IBM Virtual Shared Disks, in bytes. The default value is 24KB (24,576). If you are using the switch as your IBM Virtual Shared Disk adapter, use a value of 60KB (61,440).

- Level of parallelism

  The IBM Virtual Shared Disk component divides I/O requests into smaller units. This allows one unit to be transmitted while another is being placed in the buffer. The default value is 9 and it should always be used.

See the description of the **vsdnode** command in Appendix A, "IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands" on page 77 for the complete syntax.

Entering Global Volume Group Information in the SDR for One Node

Global volume group information is stored in the System Data Repository (SDR) **VSD_Global_Volume_Group** object. For each volume group on which IBM Virtual Shared Disks will be defined, the following data is required:

1. Local volume group name

   The LVM volume group's name on the serving node. The length of the name must be less than or equal to 15 characters.

2. Primary server node for the volume group

   The node number of the server node. The primary server node may be identified in four ways. See the **vsdvg** command reference page for specific information.

   If you do not have the IBM Recoverable Virtual Shared Disk product installed, all server nodes are primary nodes.

3. Secondary server node for the volume group

   The secondary server node is only intended for use with the RVSD product.

   The node number of the serving node. The secondary server node may be identified in four ways. See the **vsdvg** command reference page for specific information.

4. Global volume group name (must be unique across a system partition)

   The global volume group name is the globally unique name for this IBM Virtual Shared Disk volume group. The global volume group name is usually identical to the volume group name. The length of the name must be less than or equal to 31 characters. We recommend that the global volume group names be unique across the physical SP system. It *must* be unique within a system partition.

To enter IBM Virtual Shared Disk global volume group information using SMIT:

SELECT VSD Global Volume Group Information

▲ The VSD Global Volume Group Database Information dialog appears.

You can also enter IBM Virtual Shared Disk volume group information using the **vsdvg** command:

```
vsdvg [-g global_group_name] local_group_name
primary_server_node [secondary_server_node]
```

Entering IBM Virtual Shared Disk Information in the SDR for One Node

An IBM Virtual Shared Disk is defined by creating a **VSD_Table** object in the SDR. For each IBM Virtual Shared Disk, the following data is required:

1. IBM Virtual Shared Disk name (this name must be unique across the system partition)

   The globally unique name of the IBM Virtual Shared Disk. The name must be less than or equal to 31 characters. We recommend that the name be unique across the physical SP system.

   If you choose a *vsd_name* that is already the name of another device, the **cfgvsd** command will reject the name to insure that the special device files created for the name do not overlay and destroy files of the same name representing another device type (such as a logical volume).

2. Logical volume name

   The name of the logical volume on the serving node. The length of the name must be less than or equal to 15 characters.

3. Global volume group name

   The globally unique name of the volume group on which the IBM Virtual Shared Disk resides. This field is used to access the **VSD_Global_Volume_Group** to determine the node and the volume group that contains the underlying logical volume. The length of the name must be less than or equal to 31 characters.

4. Cache or nocache option

   Use the cache option only if your application does I/O in 4K blocks aligned on 4K disk boundaries, and issues a read immediately following a write.

   Nocache is the default for new IBM Virtual Shared Disks created with the **defvsd** command.

5. Minor number

   This number is automatically allocated.

To enter IBM Virtual Shared Disk information using SMIT:

SELECT Define a Virtual Shared Disk

▲ The Define A Virtual Shared Disk dialog appears.

You can also enter IBM Virtual Shared Disk information using the **defvsd** command:

```
defvsd logical_volume_name global_group_name  vsd_name [nocache
| cache]
```

An Example of Defining an IBM Virtual Shared Disk on One Node at a Time

Figure 7 is a simple example of an IBM Virtual Shared Disk. There are three IBM Virtual Shared Disks, one on each node. We have two physical hard disks per node and we defined a logical volume on one disk per node.

BETH7

Figure 7. An Example of an IBM Virtual Shared Disk Configuration

Here's how we did it. First, we decided that three IBM Virtual Shared Disks were needed, one on each node on a separate volume group. We created a table of the information for our configuration.

Table 6. IBM Virtual Shared Disk Information

Node Number	hdisks	Volume Groups	Logical Volumes	vsd Name
1	hdisk0			
	hdisk1	vg1n1	lv1vg1n1	vsd1vg1n1
2	hdisk0			
	hdisk2	vg1n2	lv1vg1n2	vsd1vg1n2
3	hdisk0			
	hdisk5	vg1n3	lv1vg1n3	vsdvg1n3

Create Volume Groups and Logical Volumes on One Node

You must create the volume groups and logical volumes before defining the IBM Virtual Shared Disk information. We created the volume groups (**vg1n1**, **vg1n2**, **vg1n3**) on each node for each IBM Virtual Shared Disk using SMIT:

```
smit mkvg
```

and entered the volume group information requested. We took the defaults. See *IBM AIX Version 4.1 System Management Guide: Operating System and Devices* , SC23-2525, for more information.

Next we created the logical volume groups (**lv1vg1n1**, **lv1vg1n2**, **lv1vg1n3**) on each node using SMIT:

```
smit mklv
```

and entered the global volume group information. See *IBM AIX Version 4.1 System Management Guide: Operating System and Devices* , SC23-2525, for more information on tuning logical volumes.

Next we used SMIT to check the volume group and logical volume information we created:

```
smit lsvg
```

smit lslv

Define IBM Virtual Shared Disks

Now we start to define our IBM Virtual Shared Disks. We entered the IBM Virtual Shared Disk node information using SMIT. This information was entered separately for node **1**, **2**, and **3**. See "Using the IBM Virtual Shared Disk Perspective to Display and Enter Node Information in the SDR" on page 20 for more information.

smit vsd_data

and selected the VSD Node Information option.

In the table *italics* show the information we entered. A `constant width` font shows the default data. We took most of the defaults except where noted.

Table 7. IBM Virtual Shared Disk Node Information

Node Number	Adapter	Initial Cache	Max Cache	Request Blocks	pbufs	Min Buddy Buffer	Max Buddy Buffer	# Buddy Buffers	
1	*css0*	64	256	256	48	4096	24576	4	
2	*css0*	64	256	256	48	4096	24576	4	
3	*css0*	64	256	256	48	4096	24576	4	

Next we used SMIT to enter the IBM Virtual Shared Disk global volume group information for **vg1n1**, **vg1n2** and **vg1n3**. See "Entering Global Volume Group Information in the SDR for One Node" on page 91 for more information.

smit vsd_data

and selected the VSD Global Volume Information option.

Table 8. IBM Virtual Shared Disk Volume Group Information

Global Group Name	Local VG Name	Primary Server	Secondary Server
vg1n1	*vg1n1*	*1*	–
vg1n2	*vg1n2*	*2*	–
vg1n3	*vg1n3*	*3*	–

We did not enter information about secondary servers because that is only for RVSD users.

Next we entered the IBM Virtual Shared Disk definition information for **vsd1vg1n1**, **vsd1vg1n2** and **vsd1vg1n3**. See "Entering IBM Virtual Shared Disk Information in the SDR for One Node" on page 92 for more information.

Again we used the SMIT command:

smit vsd_data

and selected the Define a Virtual Shared Disk option.

Table 9. IBM Virtual Shared Disk Definition Information			
LV Name	**Global Group**	**Name**	**Option**
lv1vg1n1	*vg1n1*	*vsd1vg1n1*	nocache
lv1vg1n2	*vg1n2*	*vsd1vg1n2*	nocache
lv1vg1n3	*vg1n3*	*vsd1vg1n3*	nocache

Now we have defined our IBM Virtual Shared Disk in the SDR.

- We checked all the data base information using SMIT:

  smit
  list_vsd

- We did the following steps next to get our system running. They are described in "Managing IBM Virtual Shared Disks" on page 97.

  1. Configuring on "Configuring an IBM Virtual Shared Disk" on page 99.
  2. Starting on "Starting an IBM Virtual Shared Disk" on page 99.

Displaying IBM Virtual Shared Disk Information Stored in the SDR

Once you have defined IBM Virtual Shared Disk information in the SDR, you can display this information.

To list IBM Virtual Shared Disk definition information using SMIT (from the SP Configuration Database Management menu):

SELECT List Database Information

▲ The List Database Information menu appears.

SELECT List VSD Database Information

▲ The List VSD Database Information menu appears.

Or you can use the fastpath invocation for this menu:

smit list_vsd

At this point, you can select options for listing node, global volume group, and defined IBM Virtual Shared Disk information.

You can also list IBM Virtual Shared Disk information using the **vsdsklst** command with the appropriate flags to indicate what kind of information you want to display. The following example shows the format of the information returned by **vsdsklst** for one node from the `vsdsklst -dv -a` command.

```
k7n12.ppd.pok.ibm.com
Node Number:12; Node Name:k7n12.ppd.pok.ibm.com
    Volume group:rootvg; Partition Size:4; Total:537; Free:315
        Physical Disk:hdisk0; Total:537; Free:315
    Volume group:vsdvg; Partition Size:4; Total:537; Free:533
        Physical Disk:hdisk1; Total:537; Free:533
        VSD Name:1HsD8n12{lv1HsD8n12}; Size:2
        VSD Name:1HsD20n12{lv1HsD20n12}; Size:2
```

Listing Node Information

To view IBM Virtual Shared Disk information using SMIT:

SELECT List VSD Node Information

▲ A listing of VSD node information appears.

You can also display IBM Virtual Shared Disk node information using the **vsdatalst** command:

```
vsdatalst –n
```

Listing Global Volume Group Information

To view IBM Virtual Shared Disk global volume group information using SMIT:

SELECT List VSD Global Volume Group Information

▲ A listing of VSD global volume group information appears.

You can also display IBM Virtual Shared Disk node information using the **vsdatalst** command:

```
vsdatalst –g
```

Listing Defined IBM Virtual Shared Disks

To view defined IBM Virtual Shared Disk information using SMIT:

SELECT List Defined Virtual Shared Disks

▲ A listing of defined IBM Virtual Shared Disks appears.

You can also display defined IBM Virtual Shared Disk information using the **vsdatalst** command:

```
vsdatalst –v
```

Deleting IBM Virtual Shared Disk Information from the SDR

You can delete the IBM Virtual Shared Disk information you have stored in the SDR. Once you have defined IBM Virtual Shared Disk information in the SDR, you may want to delete or change (delete and add) information. To delete IBM Virtual Shared Disk definition information using SMIT (from the SP Configuration Database Management menu):

SELECT Delete Database Information

▲ The Delete Database Information menu appears.

SELECT Delete VSD Database Information

▲ The Delete VSD Database Information menu appears.

The fastpath invocation for this menu is:

```
smit delete_vsd
```

At this point, you can select options for deleting node and global volume group information and for undefining a IBM Virtual Shared Disk.

You can also use the **removevsd** or **undefvsd** commands to remove an IBM Virtual Shared Disk, a list of IBM Virtual Shared Disks, or all the IBM Virtual Shared

Disks in a system or system partition. See the command reference pages for **removevsd** and **undefvsd** for more information.

Managing IBM Virtual Shared Disks

Once IBM Virtual Shared Disks have been defined in the SDR, they can be configured to the system and managed to various states by using SMIT menus or commands.

Running any of the following IBM Virtual Shared Disk commands affects the IBM Virtual Shared Disk states only on the node on which the command is run:

- **varyonvg**
- **cfgvsd**
- **startvsd**
- **preparevsd**
- **resumevsd**
- **suspendvsd**
- **stopvsd**
- **varyoffvg**
- **ucfgvsd**

The SMIT panels use **dsh** to run the command on any number of nodes you select. The SMIT menus may be invoked from the control workstation. See Figure 8 on page 98 for an overview of IBM Virtual Shared Disk states and associated commands.

In order for an application to be able to read or write **vsd.***nnn*, **vsd.***nnn* must be in the active state on both the client and server. Both **cfgvsd** and **startvsd** must be run on **vsd.***nnn*. Once **vsd.***nnn* is in the active state, the application may open /**dev**/**rvsd.***nnn*, providing the permissions on the file permit access. For example, to allow an application on node **1** to access /**dev**/**rvsd1vg1n1**, **cfgvsd** and **startvsd** must have been run on node **1** for **vsd1vg1n1**. To allow access to **vsd1vg1n1** from node **2**, **cfgvsd** and **startvsd** must also be run on node **2**. You may use **cfgvsdhsd** on all nodes instead of **cfgvsd** on each node.

Figure 8 on page 98 summarizes IBM Virtual Shared Disk states, how I/O requests are treated in each state, and the commands (sometimes called methods) used to change states. New users should study Figure 8 on page 98 to understand IBM Virtual Shared Disk operation.

If you use the IBM Recoverable Virtual Shared Disk, do not issue any commands that change the state of the IBM Virtual Shared Disks (**cfgvsd**, **cfghsdvsd**, **startvsd**, **preparevsd**, **resumevsd**, **suspendvsd**, **stopvsd**, **ucfghsdvsd**, or **ucfgvsd**.) The IBM Recoverable Virtual Shared Disk issues these commands for you.

BETH2

Figure 8. IBM Virtual Shared Disk States and Associated Commands

To manage IBM Virtual Shared Disks using SMIT:

TYPE **smit**

▲ The System Management menu appears

SELECT SP System Management

▲ The SP System Management menu appears.

SELECT SP Cluster Management

▲ The SP Cluster Management menu appears.

SELECT VSD Management

▲ The VSD Management menu appears.

The fastpath invocation for this menu is:

```
smit vsd_mgmt
```

Note: If you are not using the RVSD product and you want to bring up IBM Virtual Shared Disks automatically at boot time, all **cfgvsd** and **startvsd** (or **preparevsd** and **resumevsd**) commands can be run at boot time on all IBM Virtual Shared Disk nodes. Put these commands in a script that is invoked by either **/etc/init.rc** or **/etc/inittab**. This approach allows all IBM Virtual Shared Disk configuration to occur in parallel across the system partition as each machine boots up. You must insure that each global volume group is varied on (using the **varyonvg** command) on the server node before running **startvsd** or **resumevsd** on the server node. If you subsequently install the RVSD product, make sure to undo these steps.

Configuring an IBM Virtual Shared Disk

Configuring puts a defined IBM Virtual Shared Disk in the stopped state, but does not make it available. The IBM Virtual Shared Disk configuration method (**cfgvsd**) issues **/usr/lpp/csd/bin/readSDR** to extract information from the SDR and puts it into the flat files **VSD_Global_Volume_Group**, **VSD_Table**, **VSD_ipaddr** and **Node** in the **/usr/lpp/csd/vsdfiles** directory. Other IBM Virtual Shared Disk commands access these files.

Note: Do not alter these files. Each **cfgvsd** command invocation recreates the files from the SDR information and any changes you make to the files will be lost. To make changes, update the SDR information.

Use **cfgvsd** with **dsh** to configure IBM Virtual Shared Disks on more than one node at a time.

To configure an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Configure a VSD

▲ The VSD and node selection lists appear.

You can also configure an IBM Virtual Shared Disk using the **cfgvsd** command.

To configure all IBM Virtual Shared Disks defined in the SDR to the node you are on, enter:

```
cfgvsd -a
```

To configure specific IBM Virtual Shared Disks defined in the SDR, enter:

```
cfgvsd vsd_name ...
```

Starting an IBM Virtual Shared Disk

Starting an IBM Virtual Shared Disk puts a stopped one in the active (and available) state. (This is equivalent to preparing and resuming an IBM Virtual Shared Disk.) Note that for an IBM Virtual Shared Disk to be usable, it must be in the active state on both the server and client nodes.

To start an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Start a VSD

▲ The VSD and node selection lists appear.

You can also start an IBM Virtual Shared Disk using the **startvsd** command.

To start all IBM Virtual Shared Disks defined in the SDR to the node you are on, enter:

startvsd -a

To start specific IBM Virtual Shared Disks defined in the SDR, enter:

startvsd *vsd_name* **...**

Preparing an IBM Virtual Shared Disk

Preparing an IBM Virtual Shared Disk puts a stopped one in the suspended state. In the suspend state, open and close requests are honored. Read and write requests are held until the IBM Virtual Shared Disk is brought to the active state.

To prepare an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Prepare a VSD

▲ The VSD and node selection lists appear.

You can also prepare an IBM Virtual Shared Disk using the **preparevsd** command.

To prepare all IBM Virtual Shared Disks in the SDR, enter:

preparevsd -a

To prepare specific IBM Virtual Shared Disks defined in the SDR, enter:

preparevsd *vsd_name* **...**

Resuming an IBM Virtual Shared Disk

Resuming an IBM Virtual Shared Disk puts a suspended one in the active state. The IBM Virtual Shared Disk remains available and read and write requests that were held are resumed.

To resume an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Resume a VSD

▲ The VSD and node selection lists appear.

You can also resume an IBM Virtual Shared Disk using the **resumevsd** command.

To resume all IBM Virtual Shared Disks defined in the SDR, enter:

resumevsd -a

To resume a specific IBM Virtual Shared Disk defined in the SDR, enter:

resumevsd *vsd_name* **...**

Suspending an IBM Virtual Shared Disk

Suspending an IBM Virtual Shared Disk puts an active one in the suspended state. The IBM Virtual Shared Disk remains available and read and write requests that were active are suspended and held. All read and write requests subsequent to those that were active are also held.

To suspend an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Suspend a VSD

▲ The VSD and node selection lists appear.

You can also suspend an IBM Virtual Shared Disk using the **suspendvsd** command.

To suspend all IBM Virtual Shared Disks defined in the SDR, enter:

```
suspendvsd -a
```

To suspend specific IBM Virtual Shared Disks defined in the SDR, enter:

suspendvsd *vsd_name* **...**

Stopping an IBM Virtual Shared Disk

Stopping an IBM Virtual Shared Disk puts a suspended one in the stopped state. The IBM Virtual Shared Disk becomes unavailable. All applications that have outstanding requests to a stopped IBM Virtual Shared Disk terminate in error.

To stop an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Stop a VSD

▲ The VSD and node selection lists appear.

You can also stop an IBM Virtual Shared Disk using the **stopvsd** command.

To stop all IBM Virtual Shared Disks defined in the SDR, enter:

```
stopvsd -a
```

To stop specific IBM Virtual Shared Disks defined in the SDR, enter:

stopvsd *vsd_name* **...**

Unconfiguring an IBM Virtual Shared Disk

Unconfiguring a stopped IBM Virtual Shared Disk makes it inaccessible. It does not, however, undefine or change the definition information for the IBM Virtual Shared Disk in the SDR.

To unconfigure an IBM Virtual Shared Disk using SMIT (from the VSD Management menu):

SELECT Unconfigure a VSD

▲ The VSD and node selection lists appear.

You can also unconfigure an IBM Virtual Shared Disk using the **ucfgvsd** command.

To unconfigure all IBM Virtual Shared Disks defined in the SDR, enter:

```
ucfgvsd
-a
```

To unconfigure specific IBM Virtual Shared Disks in the SDR, enter:

ucfgvsd *vsd_name* **...**

Displaying Managed IBM Virtual Shared Disk Characteristics

Configuration information for
an IBM Virtual Shared Disk includes:

- Name
- Minor number
- State
- Current server node number
- *cache*|*nocache* option
- At the server, the major and minor number of the underlying logical volume. At other nodes, this information is not available.

To display managed IBM Virtual Shared Disk characteristics for all IBM Virtual Shared Disks using SMIT (from the VSD Management menu):

SELECT Show All Managed VSD Characteristics

▲ The VSD and node selection lists appear.

You can also display managed IBM Virtual Shared Disk characteristics using the **lsvsd** command. To display configuration information for specific IBM Virtual Shared Disks in your system, enter:

lsvsd -l *vsd_name* **...**

Displaying Managed IBM Virtual Shared Disk Statistics

Usage information includes:

- Number of local logical read and write operations
- Number of remote logical read and write operations
- Number of client logical read and write operations
- Number of physical reads and writes
- Number of cache hits for read
- Number of 512KB blocks read and written

To display managed IBM Virtual Shared Disk usage statistics for all IBM Virtual Shared Disks using SMIT (from the VSD Management menu):

SELECT Show All Managed VSD Statistics

▲ The VSD and node selection lists appear.

You can also display managed IBM Virtual Shared Disk usage statistics using the **lsvsd** command.
To display usage information for IBM Virtual Shared Disks in your system, enter:

lsvsd -s *vsd_name* **...**

Note: The number of blocks read and written is cumulative. To reset this count before you measure it, issue **ctlvsd -V** command .

Displaying IBM Virtual Shared Disk Device Driver Statistics

You may wish to display usage information about the IBM Virtual Shared Disk device driver. Usage information includes:

- IBM Virtual Shared Disk parallelism value
- Nonzero sequence numbers and associated nodes
- IBM Virtual Shared Disk device driver counters

- Node outcast status
- IP message size
- Timeouts
- Retries
- Cache shortages
- Buddy buffer shortages
- Indirect I/O
- Communications buffer pool shortage

To display usage information about the IBM Virtual Shared Disk device driver using SMIT (from the VSD Management menu):

SELECT Show VSD Device Driver Statistics

▲ A node selection list appears.

You can also display IBM Virtual Shared Disk device driver usage statistics using the **statvsd** command:

```
statvsd
```

See Appendix A, “IBM Virtual Shared Disk and IBM Recoverable Virtual Shared Disk Commands” on page 77 for a description of these statistics.

Setting and Displaying IBM Virtual Shared Disk Device Driver Operational Parameters

You may wish to display and, in some cases, set IBM Virtual Shared Disk device driver operational parameters. Operational parameters include:

- Current cache buffer count
- Maximum cache buffer count
- Request block count
- pbuf count
- Level of IBM Virtual Shared Disk parallelism
- Buddy buffer configuration

To display and set IBM Virtual Shared Disk device driver operational parameters using SMIT (from the VSD Management menu):

SELECT Set/Show VSD Device Driver Operational Parameters

▲Show VSD Device Driver Operational Parameters

▲Reset Sequence Numbers

▲Increase Cache Size

▲Set Level of VSD Parallelism

▲Reset VSD Device Driver Counters

▲Reset a VSD's Statistics

▲Reset All VSDs' Statistics

You can also display and set IBM Virtual Shared Disk device driver operational parameters using the **ctlvsd** command.

Testing to Verify an IBM Virtual Shared Disk (the Verification Test Suite)

CAUTION: The **vsdvts** command writes data to the IBM Virtual Shared Disk. **Do not use the command after you have put real data on a IBM Virtual Shared Disk.** Refer to the **vsdvts** command reference page.

To test that you have successfully installed the IBM Virtual Shared Disk software and that you can successfully define and make an IBM Virtual Shared Disk active, and read and write to it, follow these steps:

1. Make sure the switch or other network adapter is up and running.
2. Create a logical volume. One physical partition will be enough space for it.
3. Define the logical volume as an IBM Virtual Shared Disk in the SDR. Then, in the following commands, substitute the name of the IBM Virtual Shared Disk you defined for *vsd.name*. Commands are in /**usr**/**lpp**/**csd**/**bin**.
4. Configure the IBM Virtual Shared Disk to your system. Enter:

   cfgvsd *vsd_name*
5. Make the IBM Virtual Shared Disk available. Enter:

   preparevsd *vsd_name*
6. Make the IBM Virtual Shared Disk active. Enter:

   resumevsd *vsd_name*
7. Display configuration information about the IBM Virtual Shared Disk. Enter:

   lsvsd
   -l *vsd_name*

   The display should show that *vsd.LV* is in the active state.
8. Confirm that the IBM Virtual Shared Disk Verification Test is successful. Enter:

   vsdvts *vsd_name*

   The command should complete within 15 seconds and you should get the message:

   ```
   VSD verification test successful!
   ```

You can use the **vsdvts** command to verify all of your IBM Virtual Shared Disks. Use it to verify both local and remote IBM Virtual Shared Disks. Run the **cfgvsd** and **startvsd** commands on all client and server nodes. Use the SMIT panels to verify that the panels work.

Command and SMIT Interfaces for Data Striping Devices

Defining HSD Device Driver Information in the SDR

The HSD must be defined before you can configure and use it. The data striping device information is stored in the System Data Repository (SDR). You can use either SMIT or a command to enter the information in the SDR.

Creating HSDs with SMIT

The SMIT panels refer to the data striping devices as “hsd”

To define the Data Striping Device information in the SDR using SMIT:

SELECT SP System Management

▲ The SP System Management menu appears.

SELECT SP Configuration Database Management

▲ The Configuration Database Management appears.

SELECT Enter Database information

▲ The Enter Database information menu appears.

SELECT IBM Virtual Shared Disk Database information

▲ The IBM Virtual Shared Disk Database information menu appears.

SELECT Define a Hashed Shared Disk

▲ The Define A Hashed Shared Disk dialog appears.

The fastpath invocation for this menu is:

smit vsd_data

An Example of Creating an HSD

Example 1

We will define our HSD, called **hsd1**, with four virtual shared disks at the same node and a stripe size of 4096 bytes. The IBM Virtual Shared Disks are named vsd1vg1n1, vsd2vg1n1, vsd3vg1n1, and vsd4vg1n1.

The HSDs were defined:

```
defhsd hsd1 4096 vsd1vg1n1 vsd2vg1n1 vsd3vg1n1 vsd4vg1n1
```

We have 16384 bytes (four 4KB blocks) of data to write. The first 4096 bytes are stored in vsd1vg1n1. The second 4096 bytes are stored in vsd2vg1n1, the third 4096 bytes are stored in vsd3vg1n1, and the last 4096 bytes are stored in vsd4vg1n1.

Example 2

Assume now we have two striped devices defined, **hsd1** and **hsd2**. Our stripe size is 4096.

We had identified the IBM Virtual Shared Disks to be included in the Data Striping Devices. Assume there are four nodes in the system and two physical hard disks per node. We wanted to stripe across all the four nodes and all the disks. We defined a logical volume on the disks and nodes. Then we defined the IBM Virtual Shared Disks:

hsd name	node	IBM Virtual Shared Disks in the HSD
hsd1	node1	vsd1vg1n1
	node2	vsd1vg1n2
	node3	vsd1vg1n3
	node4	vsd1vg1n4
hsd2	node1	vsd2vg1n1
	node2	vsd2vg1n2
	node3	vsd2vg1n3
	node4	vsd2vg1n4

We defined the HSDs with the following commands:

```
defhsd hsd1 4096 vsd1vg1n1 vsd1vg1n2 vsd1vg1n3 vsd1vg1n4
defhsd hsd2 4096 vsd2vg1n1 vsd2vg1n2 vsd2vg1n3 vsd2vg1n4
```

Displaying Data Striping Device Information in the SDR

To list HSD definition information in the SDR do one of the following:

Using SMIT

From the SP Configuration Database Management menu:

SELECT List Database Information

▲ The List Database Information menu appears.

SELECT List HSD Database Information

▲ The List IBM Virtual Shared Disk Database Information menu appears.

SELECT List Defined Hashed Shared Disks

▲ A listing of defined Hashed Shared Disks appears.

You can use the fastpath invocation for this menu :

```
smit list_vsd
```

At this point, you can select options for listing defined HSD information.

To view defined HSD information using SMIT:

SELECT List Defined Hashed Shared Disks

▲ A listing of defined Hashed Shared Disks appears.

From the Command Line

You can also display defined HSDs information using the **hsdatalst** command:

```
hsdatalst
```

Undefining HSD Information in the SDR

To delete Data Striping Device definition information

Using SMIT

From the SP Configuration Database Management menu:

SELECT Delete Database Information

▲ The Delete Database Information menu appears.

SELECT Delete IBM Virtual Shared Disk Database Information

▲ The Delete IBM Virtual Shared Disk Database Information menu appears.

The fastpath invocation for this menu is:

smit delete_vsd

At this point, you can select options for undefining an HSD.

To undefine an HSD using SMIT:

SELECT Undefine a Hashed Shared Disk

▲ The Undefine a Hashed Shared Disk selection list appears.

You can either type in the *hsd_name* or select an *hsd_name* from the list function. The *hsd_name* must be in the unconfigured state.

From the Command Line

You can also undefine a Data Striping Device using the **removehsd undefhsd** or commands:

undefhsd -v *hsd_names* [**–f**]

removehsd -v *hsd_names* [**–f**]

If the *hsd_names* are configured, you can use the –**f** parameter of **removehsd** to force the system to unconfigure and remove them.

Managing Data Striping Devices

To manage Data Striping Devices using SMIT:

TYPE **smit**

▲ The System Management menu appears

SELECT SP System Management

▲ The SP System Management menu appears.

SELECT SP Cluster Management

▲ The SP Cluster Management menu appears.

SELECT HSD Management

▲ The HSD Management menu appears.

The fastpath invocation for this menu is:

smit hsd_mgmt

Configuring an HSD

Configuration makes a defined HSD available.

Using SMIT: From the HSD Management menu,

SELECT Configure an HSD

▲ The HSD and node selection lists appear.

From the Command Line

To configure specific HSDs defined in the SDR on one node, enter:

cfghsd *hsd_name* **...**

Unconfiguring an HSD

Unconfiguring an HSD makes it unavailable. It does not, however, undefine or change the definition information for the HSD in the SDR. You must do this on each node.

Using SMIT: From the HSD Management menu,

SELECT Unconfigure an HSD

▲ The HSD and node selection lists appear.

From the Command Line

To unconfigure specific HSDs in the SDR on one node, enter:

ucfghsd *hsd_name* **...**

Displaying HSD Information at a Node

You can display managed IBM Virtual Shared Disk Data Striping Devicecharacteristics for all configured Data Striping Devices.

Using SMIT: From the HSD Management menu,

SELECT Show All Managed HSD Characteristics

▲ The HSD and node selection lists appear.

From the Command Line: You can also display managed Data Striping Device characteristics using the **lshsd** command. To display information for a specific Data Striping Device in your system, enter:

lshsd -l [*hsd_name* **...**]

Configuration information displayed for an HSD includes:

Name
Minor number
Stripe size
protect_lvcb / **not_protect_lvcb** option
Underlying IBM Virtual Shared Disks in the HSD

You can display managed IBM Virtual Shared Disk Data Striping Device usage statistics for all configured Data Striping Devices..

Using SMIT: From the HSD Management menu,

SELECT Show All Managed HSD Characteristics

▲ The HSD and node selection lists appear.

From the Command Line: You can also display managed Data Striping Device usage statistics using the **lshsd** command. To display usage information for a specific Data Striping Device in your system, enter:

lshsd -s [*hsd_name* **...**]

Usage information displayed for an HSD includes:

- Number of reads on each underlying IBM Virtual Shared Disk
- Number of writes on each underlying IBM Virtual Shared Disk

Setting and Displaying IBM Virtual Shared Disk HSD Operational Parameters

You can display and, in some cases, set IBM Virtual Shared Disk Data Striping Device parallelism level, as well as reset the device counters and statistics. Operational parameters include:

- Current level of Data Striping Device parallelism
- Number of read requests not at page boundary count
- Number of write requests not at page boundary count

Using SMIT: From the HSD Management menu,

SELECT Set/Show HSD Device Driver Operational Parameters

▲ Show HSD Device Driver Operational Parameters

▲ Set Level of HSD Parallelism

▲ Reset HSD Device Driver Counters

▲ Reset an HSD's Statistics

▲ Reset All HSDs' Statistics

You can also display and set IBM Virtual Shared Disk Data Striping Device operational parameters using the **ctlhsd** command. Refer to the *IBM Parallel System Support Programs for AIX: Command and Technical Reference* for more information.

Verifying Data Striping Device Installation

After you install the HSD software and define and configure several HSDs, you can use the verification test. Do this **before** you store data on these devices.

CAUTION:
The hsdvts command writes data to the configured *hsd_name***. Do not use the command after you have put real data on that** *hsd_name* **or the underlying IBM Virtual Shared Disks.**

Refer to the **hsdvts** command reference page. Follow these steps:

1. Enter:

   hsdvts *hsd.name*

   You should get the message:

   ```
   HSD verification test successful!
   ```

You can use the **hsdvts** command to verify all of your HSDs. Use it to verify both local and remote HSDs. Issue the **cfghsd** command on all client and server nodes.

Management and Usage Notes

Configuring an HSD

You can define, configure, unconfigure, and undefine HSDs, using the **defhsd**, **cfghsd**, **ucfghsd**, and **undefhsd** commands.

Configuring an HSD creates the special device file /**dev**/**r***hsd_name*, which loads the HSD device driver into the kernel and makes the device available for use by user applications.

Use **cfghsdvsd** and **ucfghsdvsd** to configure and unconfigure HSDs and their underlying IBM Virtual Shared Disks on multiple nodes at once.

Unconfiguring an HSD makes the device unavailable to applications.

Defining an HSD creates an HSD object in the SDR.

Undefining an HSD deletes the HSD object from the SDR and removes the special device file /**dev***hsd_name*.

As a device driver, the Data Striping Device provides interfaces to application programs through the **open**, **close**, **read**, **write**, and related system calls. A **read** or **write** first uses the offset in the data file to figure out which IBM Virtual Shared Disk the request is for. The HSD then passes the request to the IBM Virtual Shared Disk to do the I/O. Usually the **ioctl** call is device-dependent. The **ioctl** call for the HSD provides minimum information about the device.

The Data Striping Device is a character (raw) device which must be opened as /**dev**/**r***hsd_name*; all the data addresses must be at 512 block boundaries.

Command and SMIT Interfaces for the IBM Recoverable Virtual Shared Disk Licensed Program

This section tells how to maintain twin-tailed Logical Volume Manager (LVM) components shared by two nodes in an SP system running the IBM Recoverable Virtual Shared Disk software. It includes procedures for volume groups, logical volumes, and physical volumes.

Defining Twin-tailed LVM Components in the IBM Recoverable Virtual Shared Disk Environment

Some of the definitions you need to understand are:

twin-tailed volume group A volume group consisting of supported disks that are all twin-tailed to the same nodes.

twin-tailed physical volume
A disk in a twin-tailed volume group.

twin-tailed logical volume
A logical volume in a twin-tailed volume group.

Coordinating the IBM Recoverable Virtual Shared Disk Environment

The twin-tailed components must have the same definition on both twin-tailed nodes. Any change to an LVM component **must** be reflected in the Object Data Manager (ODM) definitions on both nodes.

Follow the procedures in this chapter carefully in the order given to keep the LVM ODM definitions on both nodes synchronized.

Note: Do not reboot or issue **ha_vsd reset** until you have completed all the steps in a task.

Overview of Shared Volume Groups Administrative Tasks

The overall procedure for modifying a twin-tailed LVM component is the same for all tasks. In general, you export the volume group from a secondary node, make the change on the primary node, and then reimport the volume group on the secondary node. Specific operations, however, have unique steps. They are described in the following sections.

The table below summarizes the steps you must complete on the primary node and the secondary node to change a twin-tailed LVM component in an IBM Recoverable Virtual Shared Disk system. Perform all the steps in the correct order so the data does not become corrupted.

Table 10. General Procedure for Changing a Twin-tailed LVM Component

Step Number	Primary Node	Secondary Node
Step 1	Complete prerequisite tasks	Complete prerequisite tasks
Step 2		Export a volume group
Step 3	Vary on a volume group	
Step 4	Make changes to the twin-tailed LVM component	
Step 5	Vary off the volume group	
Step 6		Import a volume group
Step 7		Change a volume group to remain dormant at start up
Step 8		Vary off the volume group
Step 9	Complete follow-up tasks	Complete follow-up tasks

Prerequisite Tasks

The prerequisite tasks, while not directly involved in modifying LVM components, must be completed before you begin to make the change.

Note: You must suspend, stop, and unconfigure all IBM Virtual Shared Disks on all nodes in the volume group to be changed before changing a twin-tailed LVM component.

The prerequisite tasks can vary for the different operations. The descriptions for each operation on the following pages have a list of specific prerequisite tasks.

Maintaining Twin-Tailed Volume Groups

Maintaining twin-tailed volume groups requires the following administrative tasks:

- "Creating a Twin-tailed Volume Group"
- "Adding Physical Volumes To a Twin-tailed Volume Group" on page 115
- "Adding or Removing Logical Volumes to or from a Twin-tailed Volume Group" on page 119
- "Reducing a Twin-tailed Volume Group" on page 121
- "Removing a Twin-tailed Volume Group" on page 126

Creating a Twin-tailed Volume Group

The table below summarizes the steps you must complete on the primary node and the secondary node to create a twin-tailed volume group. Perform all the steps in the correct order so the data does not become corrupted.

Note: Twin-tailed disks that have SCSI IDs must have a different ID on the secondary than the one on the primary.

Table 11. Steps to Create a Twin-tailed Volume Group

Step Number	Primary Node	Secondary Node
Step 1	Complete prerequisite tasks	Complete prerequisite tasks
Step 2	Create a twin-tailed volume group	
Step 3	Create logical volumes	
Step 4	Vary off the volume group	
Step 5		Import a volume group
Step 6		Change a volume group to remain dormant at startup
Step 7		Vary off the volume group
Step 8	Complete follow-up tasks	Complete follow-up tasks

Complete the following steps to create a twin-tailed volume group.

Step 1: Complete Prerequisite Tasks

1. The physical volumes (**hdisks**) should be installed, configured, named, and available.
2. Make sure the primary and the secondary nodes are active.

Step 2: Create a Twin-tailed Volume Group on the Primary Node

Use the **smit mkvg** fastpath to create a volume group.

1. As the root user at the primary node, enter:

   ```
   smit mkvg
   ```

   SMIT returns a screen similar to the following:

```
                Add a Volume Group
Type or select values in entry fields.
Press enter after making all desired changes.
                                                  [Entry Fields]
 VOLUME GROUP name                                      [ ]
 Physical partition SIZE in megabytes                   [ ]
 PHYSICAL VOLUME names                                  [ ]
 Activate volume group AUTOMATICALLY at system restart?      yes
 ACTIVATE volume group after it is created?                  yes
 Volume group MAJOR NUMBER                              [ ]
```

2. Enter field values as follows:

   VOLUME GROUP name Enter the name of the twin-tailed volume group.

   Physical partition size in megabytes
   Accept the default value, unless your site has another specific partitioning requirement.

   PHYSICAL VOLUME names
   Enter the names of all the disks to be used in the volume group. You can use the F4 key to display a list of all available physical volumes. From the list, use the arrow keys to find the correct entries, then use F7 to select them. Press Enter when you have selected all the entries you want. The physical volumes you selected are automatically entered in the PHYSICAL VOLUME names field.

   Activate volume group AUTOMATICALLY at system restart?
   Set the field to **no** so that the volume group can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts .

   ACTIVATE volume group after it is created?
   Set this field to **yes**.

   Volume Group MAJOR NUMBER
   Use the default, the next available number in the valid range.

3. Press Enter. The system asks if you are sure. Check and correct your entries, if needed. Then press Enter again.
4. Create logical volumes. Refer to the LVM reference pages.
5. After the command completes, press F10 to leave SMIT and return to the command line.

Step 3: Create Logical Volumes

Use the **smit mklv** fastpath to create logical volumes. Create the logical volumes you want for this IBM Virtual Shared Disk.

Step 4: Vary Off the Volume Group on the Primary Node

Use the **varyoffvg** command to quiesce the affected volume group.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 5: Import a Volume Group Information onto the Secondary Node

You now return to the secondary node. First, import the volume group onto the secondary node. Importing the volume group onto the secondary node synchronizes the ODM definition of the volume group on both nodes.

1. To use the **smit importvg** fastpath to import the volume group, enter:

   ```
   smit importvg
   ```

   SMIT returns a screen similar to the following:

   ```
   Press enter after making all desired changes.
                                                   [Entry Fields]
    VOLUME GROUP name                                [  ]
   *PHYSICAL VOLUME name                             [  ]
   *ACTIVATE volume group after it is imported?       yes
    Volume Group MAJOR NUMBER                        [  ]
   ```

2. Enter field values as follows:

   VOLUME GROUP name Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the primary node.

   PHYSICAL VOLUME name
   Enter the name of one of the physical volumes that resides in the volume group. Note that a disk can have a different physical name on different nodes. Make sure that you use the disk name as it is defined on the *secondary* node. Every **hdisk** has an ID. If you know that the name of a disk on the primary node is *hdisk01*, use the name that matches that ID on the secondary node. You can use the **lspv** command to determine the name.

   ACTIVATE volume group after it is imported?
   Set the field to **yes**.

   Volume Group MAJOR NUMBER
   Use the default, the next available number in the valid range.

3. Press Enter.
4. Press F10 to leave SMIT and return to the command line.

Step 6: Change a Volume Group to Remain Dormant at Start Up

By default, a volume group that was just imported is configured to automatically become active at system restart. However, an IBM Recoverable Virtual Shared Disk volume group should be varied on as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts . Therefore, after importing a volume group, use the SMIT Change a Volume Group screen to reconfigure the volume group to remain dormant at start up.

1. To use the **smit chvg** fastpath to change the characteristics of a volume group, enter:

   ```
   smit chvg
   ```

   SMIT prompts you to select the volume group.

2. Select the volume group you just imported and press Enter.

   SMIT returns a screen similar to the following: The first field contains the volume group name you specified.

```
              Change a Volume Group
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                    [ ]
*ACTIVATE volume group automatically                  [yes]   +
     at system restart?
```

3. **ACTIVATE volume group automatically at system restart?** Set the field to **no**.
4. Press Enter.
5. Press F10 to leave SMIT and return to the command line.

Step 7: Vary Off the Volume Group on the Secondary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 8: Complete Follow-up Tasks

Once you have created the volume group, do the following tasks:

- Make sure the primary and secondary nodes are active.
- Use the **vsdvg** command to define the twin-tailed volume groups.
- Use **defvsd** to define IBM Virtual Shared Disks on all the twin-tailed nodes.
- Use the **cfgvsd** command for all the IBM Virtual Shared Disks in this volume group at all the nodes.
- Use the **startvsd** command at all the IBM Virtual Shared Disk nodes to start these new IBM Virtual Shared Disks.

Adding Physical Volumes To a Twin-tailed Volume Group

The following table summarizes the steps you do on both the primary and secondary nodes to extend (add one or more physical volumes to) a twin-tailed volume group. Perform all the steps in the correct order so the data does not become corrupted.

Table 12. Procedure for Extending a Twin-tailed LVM Component		
Step Number	**Primary Node**	**Secondary Node**
Step 1	Complete prerequisite tasks	Complete prerequisite tasks
Step 2		Export a volume group
Step 3	Vary on a volume group	
Step 4	Extend a twin-tailed volume group	
Step 5	Vary off the volume group	
Step 6		Import a volume group
Step 7		Change a volume group's characteristics
Step 8		Vary off the volume group
Step 9	Complete follow-up tasks	Complete follow-up tasks

Step 1: Complete Prerequisite Tasks

1. The physical volumes (**hdisks**) should be installed, configured, named, and available.
2. Suspend, stop, and unconfigure all IBM Virtual Shared Disks in the volume group involved in the change on all nodes.

Step 2: Export Volume Group Information from the Secondary Node

Before making any changes to the LVM elements on the primary node, you must export the appropriate volume group from the secondary node. Exporting the volume group deletes the information about this volume group from the ODM.

To use the **exportvg** command to export the volume group containing the component you are going to change on the secondary node, enter:

exportvg *volume_group_name*

Step 3: Vary on a Volume Group on the Primary Node

To use the **varyonvg** command to activate the affected volume group after making the change, enter:

varyonvg *volume_group_name*

Step 4: Extend the Twin-tailed Volume Group on the Primary Node

Use the **smit extendvg** fastpath to extend a volume group.

1. As the root user at the primary node, enter:

   smit extendvg

   SMIT returns a screen similar to the following:

```
            Add a Physical Volume to a Volume Group
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                [ ]  +
*PHYSICAL VOLUME names                            [ ]  +
```

2. Enter field values as follows:

   VOLUME GROUP name Enter the name of the volume group that you are extending.

   PHYSICAL VOLUME names
   Enter the names of the physical volumes you are adding to the volume group.

3. Press Enter to extend the volume group.
4. After the command completes, press F10 to leave SMIT and return to the command line.
5. Create new logical volumes or extend existing ones.

Step 5: Vary Off the Volume Group on the Primary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 6: Import Volume Group Information onto the Secondary Node

You now return to the secondary node. First, import the volume group onto the secondary node. Importing the volume group onto the secondary node synchronizes the ODM definition of the volume group with the primary node.

1. To use the **smit importvg** fastpath to import the volume group, enter:

   smit importvg

   SMIT returns a screen similar to the following:

```
Press enter after making all desired changes.
                                                [Entry Fields]
 VOLUME GROUP name                                [  ]
*PHYSICAL VOLUME name                             [  ]
*ACTIVATE volume group after it is imported?       yes
 Volume Group MAJOR NUMBER                        [  ]
```

2. Enter field values as follows:

   VOLUME GROUP name Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the primary node.

PHYSICAL VOLUME name
Enter the name of one of the physical volumes that resides in the volume group. Note that a disk can have a different physical name on different nodes. Make sure that you use the disk name as it is defined on the *secondary* node. Every **hdisk** has an ID. If you know that the name of a disk on the primary node is *hdisk01*, use the name that matches that ID on the secondary node. You can use the **lspv** command to determine the name.

ACTIVATE volume group after it is imported?
Set the field to **yes**.

Volume Group MAJOR NUMBER
Use the default, the next available number in the valid range.

3. Press Enter.
4. Press F10 to leave SMIT and return to the command line.

Step 7: Change Volume Group Characteristics on the Secondary Node

By default, a volume group that was just imported is configured to automatically become active at system restart. However, an IBM Recoverable Virtual Shared Disk volume group should be varied on as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts . Therefore, after importing a volume group, use the SMIT Change a Volume Group screen to reconfigure the volume group to remain dormant at start up.

1. To use the **smit chvg** fastpath to change the characteristics of a volume group, enter:

   ```
   smit chvg
   ```

   SMIT prompts you to select the volume group.

2. Select the volume group you just imported and press Enter.

   SMIT returns a screen similar to the following. The first field contains the volume group name you specified.

```
                Change a Volume Group
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                     [ ]
*ACTIVATE volume group automatically                   [yes]   +
     at system restart?
```

3. **ACTIVATE volume group automatically at system restart?** Set the field to **no**.
4. Press Enter.
5. Press F10 to leave SMIT and return to the command line.

Step 8: Vary Off the Volume Group on the Secondary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 9: Complete Follow-up Tasks

Once the you have extended the volume group, do the following tasks:

- Make sure the primary and secondary nodes are active.
- Use **defvsd** to define any new IBM Virtual Shared Disks.
- Use the **cfgvsd** command for all the IBM Virtual Shared Disks in this volume group at all the nodes.
- Use the **startvsd** command at all the IBM Virtual Shared Disk nodes to start these new IBM Virtual Shared Disks.

Adding or Removing Logical Volumes to or from a Twin-tailed Volume Group

The following table summarizes the steps you do on both the primary and secondary nodes to add or remove logical volumes to or from a twin-tailed volume group. Perform all the steps in the correct order so the data does not become corrupted.

Table 13. Procedure for Extending a Twin-tailed LVM Component

Step Number	Primary Node	Secondary Node
Step 1	Complete prerequisite tasks	Complete prerequisite tasks
Step 2		Export a volume group
Step 3	Create or remove logical volumes	
Step 4	Vary off the volume group	
Step 5		Import a volume group
Step 6		Change a volume group's characteristics
Step 7		Vary off the volume group
Step 8	Vary on the volume group	
Step 9	Complete follow-up tasks	Complete follow-up tasks

Step 1: Complete Prerequisite Tasks

1. Shutdown your applications using the volume group to ensure data consistency and ODM integrity.
2. Suspend, stop, and unconfigure all IBM Virtual Shared Disks on all the IBM Virtual Shared Disk nodes.
3. Undefine IBM Virtual Shared Disks on any logical volumes being removed. Ensure all data is moved from the logical volumes before removing.

Step 2: Export Volume Group Information from the Secondary Node

Before making any changes to the LVM elements on the primary node, you must export the appropriate volume group from the secondary node. Exporting the volume group deletes the information about this volume group from the ODM.

To use the **exportvg** command to export the volume group containing the component you are going to change on the secondary node, enter:

exportvg *volume_group_name*

Step 3: Create or Remove Required Logical Volumes on the Primary Node

To create the logical volumes, use the fastpath invocation to the SMIT panel:

smit mklv

Alternatively, you can use the command line interface to create the logical volumes:

mklv

To remove the logical volumes, use the fastpath invocation to the SMIT panel:

smit rmlv

Alternatively, you can use the command line interface to remove the logical volumes:

rmlv

Step 4: Vary Off the Volume Group on the Primary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 5: Import Volume Group Information onto the Secondary Node

You now return to the secondary node. First, import the volume group onto the secondary node. Importing the volume group onto the secondary node synchronizes the ODM definition of the volume group with the primary node.

1. To use the **smit importvg** fastpath to import the volume group, enter:

   smit importvg

Step 6: Change Volume Group Characteristics on the Secondary Node

By default, a volume group that was just imported is configured to automatically become active at system restart. However, an IBM Recoverable Virtual Shared Disk volume group should be varied on as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts . Therefore, after importing a volume group, use the SMIT Change a Volume Group screen to reconfigure the volume group to remain dormant at start up.

To use the **smit chvg** fastpath to change the characteristics of a volume group, enter:

```
smit chvg
```

For more information, refer to “Step 7: Change Volume Group Characteristics on the Secondary Node” on page 118.

Alternatively, to use the command line interface to change the volume group to remain dormant at startup, enter:

chvg -a 'n' -Q'y' *volume_group_name*

Step 7: Vary Off the Volume Group on the Secondary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 8: Vary On the Volume Group on the Primary Node

Use the **varyonvg** command to quiesce the affected volume group after making the change.

To vary on the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyonvg *volume_group_name*

Step 9: Complete Follow-up Tasks

Once the you have extended the volume group, do the following tasks:

- Make sure the primary and secondary nodes are active.
- Use **defvsd** to define any new IBM Virtual Shared Disks.
- Use the **cfgvsd** command for all the IBM Virtual Shared Disks in this volume group at all the nodes.
- Use the **startvsd** command at all the IBM Virtual Shared Disk nodes to start these new IBM Virtual Shared Disks.

Reducing a Twin-tailed Volume Group

The following table summarizes the steps you do on both the primary and secondary nodes to reduce (remove one or more physical volumes from) a twin-tailed volume group:

Table 14. Procedure for Reducing a Twin-tailed Volume Group		
Step Number	**Primary Node**	**Secondary Node**
Step 1	Complete prerequisite tasks	Complete prerequisite tasks
Step 2		Export a volume group
Step 3	Vary on a volume group	
Step 4	Remove data from the physical volume	
Step 5	Reduce a twin-tailed volume group	
Step 6	Vary off the volume group	
Step 7		Import volume group
Step 8		Change volume group characteristics
Step 9		Vary off the volume group
Step 10	Complete follow-up tasks	Complete follow-up tasks

Complete the following tasks to reduce a twin-tailed volume group.

Step 1: Complete Prerequisite Tasks

1. Both the primary and secondary nodes must be up and in the active group.
2. Suspend, stop, and unconfigure IBM Virtual Shared Disks in the affected volume group on all nodes.

Step 2: Export Volume Group Information from the Secondary Node

Before making any changes to the LVM elements on the primary node, you must export the appropriate volume group from the secondary node. Exporting the volume group deletes the information about this volume group from the ODM.

To use the **exportvg** command to export the volume group containing the component you are going to change on the secondary node, enter:

exportvg *volume_group_name*

Step 3: Vary on a Volume Group on the Primary Node

To use the **varyonvg** command to verify that the affected volume group is varied on after making the change, enter:

varyonvg *volume_group_name*

Step 4: Move Data from the Volume to Be Removed

Use the **smit migratepv** fastpath to move data on the physical volume being removed from the volume group to a different physical volume. If you do not, **data will be lost**.

1. As the root user on the primary node, enter:

   smit migratepv

   SMIT returns a screen similar to the following:

```
          Move Contents of a Physical Volume
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*SOURCE physical volume name                         [ ]  +
```

2. Enter the names of the physical volumes being moved in the SOURCE physical volume name field and press Enter.

   SMIT returns a screen similar to the following. The physical volumes being migrated are entered in the SOURCE physical volume names field.

```
          Move Contents of a Physical Volume
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*SOURCE physical volume name                         [ ]
*DESTINATION physical volumes                        [ ]  +
 Move only data belonging to this                    [ ]  +
  LOGICAL VOLUME?
```

3. Enter field values as follows:

   DESTINATION physical volumes
   Enter the names of the physical volumes to which you want to move the data.

   Move only data belonging to this LOGICAL VOLUME?
   Set this field to **no**.

4. Press Enter to migrate the physical volumes.
5. After the command completes, press F10 to leave SMIT and return to the command line.

Step 5: Reduce the Shared Volume Group on the Primary Node

Use the **smit reducevg** fastpath to reduce a volume group.

1. As the root user at the primary node, enter:

   smit reducevg

   SMIT returns a screen similar to the following:

```
           Reduce a Volume Group
Move cursor to desired item and press Enter.
 Remove a Physical Volume from a Volume Group
 Remove a Volume Group
```

2. Select the **Remove a Physical Volume from a Volume Group** option and press Enter.

   SMIT returns a screen similar to the following:

```
         Remove a Physical Volume from a Volume Group
Type or select a value for the entry field.
Press enter AFTER making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                   [ ]  +
```

3. Enter the name of the volume group from which you are removing the physical volumes in the VOLUME GROUP name field. Press Enter.

   SMIT returns a screen similar to the following. The name of the volume group you specified in the preceding screen is entered.

```
          Remove a Physical Volume from a Volume Group
Type or select values in entry fields.
Press enter AFTER making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                    [ ]  +
*PHYSICAL VOLUME name                                 [ ]  +
 FORCE deallocation of all partitions on              [ ]  +
  this physical volume?
```

4. Enter field values as follows:

   PHYSICAL VOLUME names
   Enter the name of the physical volumes to remove from the volume group.

   FORCE deallocation of all partitions on this physical volume?
   Set this field to **no**.

5. Press Enter to remove the designated physical volumes from the volume group.
6. After the command completes, press F10 to leave SMIT and return to the command line.

Step 6: Vary Off the Volume Group on the Primary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

To vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts , enter:

varyoffvg *volume_group_name*

Step 7: Import Volume Group Information onto the Secondary Node

You now return to the secondary node. First, import the volume group on the secondary node. Importing the volume group onto the secondary node synchronizes the ODM definition of the volume group on the node.

1. To use the **smit importvg** fastpath to import the volume group, enter:

   smit importvg

   SMIT returns a screen similar to the following:

```
Press enter after making all desired changes.
                                                [Entry Fields]
 VOLUME GROUP name                                    [  ]
*PHYSICAL VOLUME name                                 [  ]
*ACTIVATE volume group after it is imported?           yes
 Volume Group MAJOR NUMBER                            [  ]
```

2. Enter field values as follows:

VOLUME GROUP name Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the primary node.

PHYSICAL VOLUME name
Enter the name of one of the physical volumes that resides in the volume group. Note that a disk can have a different physical name on different nodes. Make sure that you use the disk name as it is defined on the *secondary* node. Every **hdisk** has an ID. If you know that the name of a disk on the primary node is *hdisk01*, use the name that matches that ID on the secondary node. You can use the **lspv** command to determine the name.

ACTIVATE volume group after it is imported?
Set the field to **yes**.

Volume Group MAJOR NUMBER
Use the default, the next available number in the valid range.

3. Press Enter.
4. Press F10 to leave SMIT and return to the command line.

Step 8: Change Volume Group Characteristics on the Secondary Node

By default, a volume group that was just imported is configured to automatically become active at system restart. However, an IBM Recoverable Virtual Shared Disk volume group should be varied on as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts . Therefore, after importing a volume group, use the SMIT Change a Volume Group screen to reconfigure the volume group to remain dormant at startup.

1. To use the **smit chvg** fastpath to change the characteristics of a volume group, enter:

   ```
   smit chvg
   ```

   SMIT prompts you to select the volume group.

2. To select the volume group you just imported, press Enter.

   A screen similar to the following appears. The first field contains the volume group name you specified.

```
                Change a Volume Group
Type or select values in entry fields.
Press enter after making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                     [ ]
*ACTIVATE volume group automatically                   [yes]   +
     at system restart?
```

3. **ACTIVATE volume group automatically at system restart?** Set the field to **no**.
4. Press Enter.
5. Press F10 to leave SMIT and return to the command line.

Step 9: Vary Off the Volume Group on the Secondary Node

Use the **varyoffvg** command to quiesce the affected volume group after making the change.

You vary off the volume group so that it can be activated as appropriate by the IBM Recoverable Virtual Shared Disk recovery scripts . Enter:

varyoffvg *volume_group_name*

Step 10: Complete Follow-up Tasks

Once the you have reduced the volume group, complete the following tasks:

- Make sure the primary and secondary nodes are active.
- Use **undefvsd** to undefine no longer required IBM Virtual Shared Disks.
- Use the **cfgvsd** command for all the IBM Virtual Shared Disks in this volume group at all the nodes.
- Use the **startvsd** command at all the IBM Virtual Shared Disk nodes to start these new IBM Virtual Shared Disks.

Removing a Twin-tailed Volume Group

See the following table for an overview of the steps needed to remove a twin-tailed volume group.

Table 15. Steps to Remove a Shared Volume Group

Step Number	Primary Node	Secondary Node
Step 1	Complete prerequisite tasks	Complete prerequisite tasks
Step 2		Export volume group information
Step 3	Vary on a volume group	
Step 4		Delete a twin-tailed volume group

Complete the following steps to remove a twin-tailed volume group.

Step 1: Complete Prerequisite Tasks

1. Suspend, stop, and unconfigure the active IBM Virtual Shared Disks on this volume group on all nodes.
2. Undefine the volume group with **vsdelvg** at the control workstation.

Step 2: Export Volume Group Information from the Secondary Node

Before making any changes to the LVM elements on the primary node, you must export the appropriate volume group from the secondary node. Exporting the volume group deletes the information about this volume group from the ODM. To use the **exportvg** command to export the volume group containing the component you are going to change on the secondary nodes, enter:

exportvg *volume_group_name*

Step 3: Vary On a Volume Group on the Primary Node

To use the **varyonvg** command to activate the affected volume group after making the change, enter:

varyonvg *volume_group_name*

Step 4: Delete a Twin-tailed Volume Group on the Primary Node

1. As the root user at the primary node, enter:

   smit reducevg

   SMIT returns a screen similar to the following:

```
            Reduce a Volume Group
Move cursor to desired item and press Enter.
 Remove a Physical Volume from a Volume Group
 Remove a Volume Group
```

2. Select the **Remove a Volume Group** option and press Enter.

   SMIT returns a screen similar to the following:

```
          Remove a Volume Group
Type or select a value for the entry field.
Press enter AFTER making all desired changes.
                                                [Entry Fields]
*VOLUME GROUP name                                    [ ]  +
```

3. Enter the name of the volume group to be removed. Press Enter.
4. When the command completes, press 10 to leave SMIT and return to the command line.

Appendix C. Setting Up 7134 Disks as Twin-tailed Volume Groups

If you are using 7134 disks to set up twin-tailed volume groups, making and importing the volume groups is the same as on 9333 and 9334 disks, but there is some unique preparation work that needs to be done.

To install the 7134 disks as twin-tailed volume groups:

1. Connect the 7134 drawer to two nodes.
2. Reboot (or run **cfgmgr**) on both nodes.
3. On each node, run

   ```
   lsdev -C | grep "2.0 GB 16 Bit Differential SCSI Disk Drive"
   ```

   A list of the disks, similar to the following, is output:

   ```
   hdisk8  Available 00-14-01-00 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk9  Available 00-14-01-10 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk14 Available 00-14-01-20 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk15 Available 00-14-01-30 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk16 Available 00-14-01-40 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk17 Available 00-14-01-50 2.0 GB 16 Bit Differential SCSI Disk Drive
   ```

4. Rerun **lsdev** and **grep** for the slot number displayed for the disks in the previous step to get the name of the adapters for these disks. For example, the output from the previous step displayed that the disks are in slot 14:

   ```
   lsdev -C | grep "00\-14"
   ```

   The output from this command looks similar to the following:

   ```
   ascsi0       Available 00-14       Wide SCSI I/O Controller Adapter
   vscsi0       Available 00-14-00    SCSI I/O Controller Protocol Device
   vscsi1       Available 00-14-01    SCSI I/O Controller Protocol Device
   hdisk8       Available 00-14-01-00 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk9       Available 00-14-01-10 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk14      Available 00-14-01-20 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk15      Available 00-14-01-30 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk16      Available 00-14-01-40 2.0 GB 16 Bit Differential SCSI Disk Drive
   hdisk17      Available 00-14-01-50 2.0 GB 16 Bit Differential SCSI Disk Drive
   ```

5. As displayed in the above output, **vscsi1** is the controller for these disks because it is at 00-14-01; the disks are also coming from 00-14-01. Move the disks from the Available state to the Defined state by removing all the devices except **ascsi0** (but don't remove their definition from the ODM). For example:

   ```
   rmdev -l vscsi0
   rmdev -l vscsi1
   rmdev -l hdisk8
   rmdev -l hdisk9
   ⋮
   ```

   Note: Do steps 3, 4, and 5 on both nodes.

6. On one node, use SMIT to change the External SCSI ID from **7** to something else (**6**, for example).

   a. Enter **smit**

© Copyright IBM Corp. 1994, 1996

b. Select **Devices**

c. Select **SCSI Adapter**

d. Select **Change/Show Characteristics of a SCSI Adapter**

e. Select **ascsi0**

f. Change **Adapter card SCSI ID** to another ID

7. Run **cfgmgr** on both nodes.
8. Run **lspv** on both nodes. There should be a physical volume ID for all of the 7134 disks. If there is not a physical volume ID for the 7134 disks, try removing the devices and rerunning **cfgmgr**. Note that it should not say **none** in this column.
9. From here, proceed as you normally would to make volume groups, logical volumes, and so forth.

Appendix D. Glossary

C

Client. The representation of an IBM Virtual Shared Disk on a node that is not the server node. Requests for data from the IBM Virtual Shared Disk can be addressed to all clients. The data will actually be provided remotely by the server node for that IBM Virtual Shared Disk. Clients are created when an IBM Virtual Shared Disk definition is configured on a node. For example, a system partition P21S1 contains nodes 1, 2, 3, and 4 of an SP2 system. All four nodes have IBM Virtual Shared Disk setup data defined in the System Data Repository (SDR). If an IBM Virtual Shared Disk named OVSDn2 exists and the command cfgvsd OVSDn2 is issued, clients are configured and represented on all four nodes. All four nodes can read from and write to OVSDn2.

Cluster. A set of nodes that contain clients of the same IBM Virtual Shared Disk or sets of IBM Virtual Shared Disks.

D

Direct client. The IBM Virtual Shared Disk client on the server node.

M

Membership list. The list of all the nodes in the system or system partition that are active and can be reached through the network maintained by the **hatsd** daemon.

P

Primary node. When physical disks are connected to two nodes (twin-tailed), one node is designated as the primary node for each disk and the other is designated the secondary, or backup, node. The primary node is the server node for IBM Virtual Shared Disks defined on the physical disks under normal conditions. The secondary node can become the server node for the disks if the primary node is unavailable (off-line or down).

Q

Quorum. Quorum is defined as a majority of the IBM Virtual Shared Disk nodes in a system or system partition (*number of nodes* + 1 / 2). When a quorum exists, that majority of the nodes are active and responding to the network. Quorum is not applied across different software levels.

© Copyright IBM Corp. 1994, 1996

S

Secondary node. When physical disks are connected to two nodes (twin-tailed), one node is designated as the primary node for each disk and the other is designated the secondary, or backup, node. The secondary node acts as the server node for the IBM Virtual Shared disks defined on the physical disks if the primary node is unavailable (off-line or down).

Server node. The node that currently maintains and controls local access to the physical disks on which an IBM Virtual Shared Disk is created. With twin-tailed disks, where the physical disks are locally attached to two nodes, the primary node is the server node under normal conditions. If the primary node becomes unavailable, the secondary node acts as the server node. The server node satisfies requests for data from the IBM Virtual Shared Disks created on the physical disks attached to it.

Index

A

active group definition 61
adding logical volumes to a twin-tailed volume group 119
administrative tasks
  adding logical volumes to a twin-tailed volume group 119
  creating a twin-tailed volume group 112
  deleting logical volumes from a twin-tailed volume group 119
  extending a twin-tailed volume group 115
  overview 111
  reducing a twin-tailed volume group 121
  removing a twin-tailed volume group 126
administrative tasks. 110
application program recovery 46
availability services 59

B

buddy buffers
  tuning 53
buffer allocation
  tuning 52

C

cache buffer
  tuning 52
cfghsd command 77
cfghsd command considerations 64
cfghsdvsd command 77
cfgvsd command 77
cfgvsd command considerations 64
changing IBM Virtual Shared Diskstates 38
cksumvsd command 78
client 1, 61
client information
  displaying 35, 36
command considerations 64
commands 63
  cfghsd 77
  cfghsdvsd 77
  cfgvsd 77
  cksumvsd 78
  createhsd 78
  createvsd 78
  ctlhsd 78
  ctlvsd 79
  defhsd 79
  defvsd 79
  ha.vsd 79
commands *(continued)*
  hc.vsd 80
  hsdatalst 80
  hsdvts 80
  lshsd 80
  lsvsd 80
  preparevsd 81
  removehsd 81
  removevsd 81
  resumevsd 81
  spvsd 82
  startvsd 82
  statvsd 82
  stopvsd 82
  suspendvsd 83
  ucfghsd 83
  ucfghsdvsd 83
  ucfgvsd 83
  undefhsd 83
  undefvsd 84
  updatehsd 84
  updatevsdnode 84
  updatevsdtab 85
  verparvsd 85
  vsdatalst 85
  vsddiag 85
  vsdelnode 85
  vsdelvg 86
  vsdnode 86
  vsdsklst 86
  vsdvg 86
  vsdvts 87
communications failures 61
component definitions 110
configuring an IBM Virtual Shared Disk 38, 99
consideration
  data striping 66
consideration for IBM Recoverable Virtual Shared Disk environment 111
considerations
  cfghsd command 64
  cfgvsd command 64
  preparevsd command 64
  resumevsd command 64
  startvsd command 64
  stopvsd command 65
  suspendvsd command 65
  ucfgvsd command 65
  vsdelnode command 65
  vsdnode command 66
  vsdvg command 66

© Copyright IBM Corp. 1994, 1996

createhsd command 78
examples 30
examples with IBM Recoverable Virtual Shared Disk 31
createvsd command 78
examples 26
examples with IBM Recoverable Virtual Shared Disk 26
creating a twin-tailed volume group 112
creating Data Striping Devices
examples 30
examples with IBM Recoverable Virtual Shared Disk 31
creating Data Striping Devices on multiple nodes 27
creating IBM Virtual Shared Disks
examples 26
examples with IBM Recoverable Virtual Shared Disk 26
creating IBM Virtual Shared Disks on multiple nodes 22
ctlhsd command 78
ctlvsd command 79

D

daemons
group services 60
HA topology services 60
hagsd 60
hagsd daemon 60
hcd 61
heartbeatd 60
rvsdd 61
data striping considerations 66
Data Striping Device 105, 108
cfghsd command 108
configuring 108
defining 104
displaying configuration information 108
displaying statistics for 37
hsdatalst command 106
management notes 110
managing
SMIT panels for 107
overview 2
performance 56
planning 9
removehsd command 107
System Data Repository 104
deleting data 107
displaying data 106
undefining data 107
tuning performance 56
unconfiguring 108
undefhsd command 107
usage notes 110
Data Striping Device overview 1
Data Striping Devices
creating 27
displaying definition and client information 35
displaying node information 34
debug files 75
defhsd command 79
definition
active group 61
heartbeatd daemon 60
quorum 60
quorum override 61
recovery scripts 60
twin-tailed disk 4
twin-tailed volume group 5
definitions of twin-tailed components 110
defvsd command 79
deleting logical volumes from a twin-tailed volume group 119
deleting old rollback files 20
diagnosis information 75
direct client 61
disk controller failure recovery scenario 69
domain, Group Services
definition 59

E

environment, special consideration 111
error messages 75
extending a twin-tailed volume group 115

F

failure
disk controller recovery scenario 69
HA topology services 70
heartbeatd recovery scenario 70
node recovery scenario 68
recovery service scenario 70
switch recovery scenario 69
vsd recovery scenario 70

G

group services 59
group services daemon 60
Group Services domain
definition 59
Group Services subsystem
definition 59
group state data
definition 59
group state value
definition 59

group, Group Services
  definition 59

H

HA topology services daemon 60
HA topology services failure 70
HA topology services membership list 60
ha.vsd command 79
HACMP
  restrictions 73
hardware requirements 11
hc recovery subsystems 61
hc subsystem 46, 62
hc.activate script 46
hc.deactivate script 46
hc.h header file 46
hc.vsd command 80
hcdd daemon 61
header file, hc.h 46
heartbeatd daemon 60
heartbeatd recovery scenario 70
HSD considerations 66
hsdatalst command 80
hsdvts command 80

I

IBM Recoverable Virtual Shared Disk overview 1
IBM Recoverable Virtual Shared Disk recovery subsystems 61
IBM Virtual Shared Disk overview 1
IBM Virtual Shared Disk Perspective 15
  information area 17
  menu bar 15
  primary window 15
  tool bar 15
  work areas 16
IBM Virtual Shared Disks
  casting out 44
  cfgvsd command 99
  changing states 38
  configuring 38, 99
  creating 22
  ctlvsd command 103
  data integrity 43
  defvsd command 92
  displaying client information 35
  displaying configuration information 102
  displaying node information 34
  displaying operational parameters for IBM Virtual Shared Disk device driver 103
  displaying statistics for 36
  displaying usage information for an IBM Virtual Shared Disk 102
  displaying usage information for the IBM Virtual Shared Disk device driver 102
IBM Virtual Shared Disks *(continued)*
  example of defining 93
  global volume group information 91
  IBM Virtual Shared Disk definition information 92
  lsvsd command 102
  managing 97
    SMIT panels for 98
  migration 7
  node information 20
  overview 1
  performance 49
  planning 7
  preparevsd command 100
  preparing 38, 100
  protocol number 9
  resumevsd command 100
  resuming 38, 100
  sequence numbers 43
  setting operational parameters for IBM Virtual Shared Disk device driver 103
  starting 38, 99
  startvsd command 100
  states 98
  statvsd command 103
  stopping 39, 101
  stopvsd command 101
  suspending 38, 100
  suspendvsd command 101
  System Data Repository
    deleting data 96
    displaying data 95
  transmission protocol 43
  tuning performance 49
  ucfgvsd command 101
  unconfiguring 39, 101
  verification test suite 104
  verifying 33, 104
  verifying IBM Virtual Shared Disks without IBM Recoverable Virtual Shared Disk 33
  vsdsklst command 95
  vsdvg command 92
incarnation number 62
information area 17
inittab, file 59
installation
  hardware and software requirements 11
  notes 13
  planning 12
  space requirements 11
  step-by-step 12
installation notes 13

J

join request
  definition 59

joining a group
  membership list change 59

L

leaving a group
  membership list change 59
lshsd command 80
lsvsd command 80
LVM
  considerations in tuning 51

M

management, twin-tailed disks 110
maximum I/O request size
  tuning 53
mbufs
  tuning 55
membership list 60
  definition 59
  hc 62
menu bar 15
messages 75
migrating IBM Virtual Shared Disks to the IBM Recoverable Virtual Shared Disk 13

N

network failures 61
node failure recovery scenario 68
node information
  displaying 34
node statistics 36

O

overview
  Data Striping Device 1
  IBM Recoverable Virtual Shared Disk 1
  IBM Virtual Shared Disk 1
overview of administrative tasks 111

P

pbufs
  tuning 54
Perspective, IBM Virtual Shared Disk 15
  information area 17
  menu bar 15
  primary window 15
  tool bar 15
  work areas 16
Perspectives graphical user interface
  IBM Virtual Shared Disk Nodes panel 20
  primary window 20
  SDR notebook page 20
planning, installation 12
preparevsd command 81
preparevsd command considerations 64
primary node 61
primary window for IBM Virtual Shared Disks 15
programming considerations 46
programming, application 46
provider
  definition 59

Q

quorum 60, 61
quorum override 61

R

recognizing recovery 67
recoverable applications 46
recovery 67
  disk controller failure 69
  HA topology services 70
  heartbeatd failure 70
  High Performance Switch failure 69
  node failure 68
  planning for 67
  recognizing 67
  recovery service daemon failure 70
  vsd failure 70
recovery scripts 60, 63
recovery service daemon failure 70
recovery subsystems 61
reducing a twin-tailed volume group 121
removehsd command 81
removevsd command 81
removing a twin-tailed volume group 126
request blocks
  tuning 54
requirements, installation 11
restrictions
  command use 64
  with HACMP 73
resumevsd command 81
resumevsd command considerations 64
rollback files 20
RVSD installation 10
rvsd subsystem 61
rvsdd daemon 61

S

scenarios for recovery 67
secondary node 61
server 1
SMIT
  installation using 12

software requirements 11
space requirements for installation 11
spvsd command 82
startup 63
startvsd command 82
startvsd command considerations 64
states
  changing 38
statistics
  Data Striping Device 37
  IBM Virtual Shared Disks 36
statvsd command 82
stopvsd command 82
stopvsd command considerations 65
subscribe request
  definition 59
subscriber
  definition 59
subsystems
  hc 46, 62
  rvsd 61
suspendvsd command 83
suspendvsd command considerations 65
switch failure recovery scenario 69
switch parameters
  tuning 50
switch pool
  tuning 55
system partition
  domain for Group Services 59
system startup 63

T

tool bar 15
tuning
  buddy buffers 53
  buffer allocation 52
  cache buffer 52
  LVM considerations 51
   maximum I/O request size 53
  mbufs 55
  pbufs 54
  request blocks 54
  switch parameters 50
  switch pool 55
twin-tailed disk definition 4
twin-tailed disk management 110
twin-tailed volume group definition 5

U

ucfghsd command 83
ucfghsdvsd command 83
ucfgvsd command 83
ucfgvsd command considerations 65
undefhsd command 83
undefvsd command 84
updatehsd command 84
updatevsdnode command 84
updatevsdtab command 85
using the IBM Recoverable Virtual Shared Disk 59

V

verparvsd command 85
vsd recovery scenario 70
vsdatalst command 85
vsddiag command 85
vsdelnode command 85
vsdelnode command considerations 65
vsdelvg command 86
vsdnode command 86
vsdnode command considerations 66
vsds
  migrating to the IBM Recoverable Virtual Shared Disk 13
vsdsklst command 86
vsdvg command 86
vsdvg command considerations 66
vsdvts command 87

W

work areas 16