

DS4000 Best Practices and Performance Tuning Guide

Performance measurement using TPC
for Disk

ERM guidelines and
bandwidth estimator

Managing and using the
DS4000 with SVC



Bertrand Dufrasne
Bruce Allworth
Agung Indrayana
Christian Schoessler
Brian Youngs

Redbooks



International Technical Support Organization

DS4000 Best Practices and Performance Tuning Guide

March 2007

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

Third Edition (March 2007)

This edition applies to IBM TotalStorage DS4000 Storage Servers and related products that were current as of March 2007.

© Copyright International Business Machines Corporation 2007. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	ix
Trademarks	x
Preface	xi
The team that wrote this redbook.	xi
Become a published author	xii
Comments welcome.	xiii
Summary of changes	xv
March 2007, Third Edition	xv
Chapter 1. Introduction to DS4000 and SAN	1
1.1 DS4000 features and models	2
1.1.1 Telco industry standard support	4
1.1.2 DS4000 Series product comparison	4
1.2 DS4000 Storage Manager	6
1.3 Introduction to SAN	7
1.3.1 SAN components	8
1.3.2 SAN zoning	10
Chapter 2. DS4000 planning tasks	13
2.1 Planning your SAN and storage server.	14
2.1.1 SAN zoning for DS4000	15
2.2 Physical components planning	16
2.2.1 Rack considerations	16
2.2.2 Cables and connectors	18
2.2.3 Cable management and labeling	21
2.2.4 Fibre Channel adapters	23
2.2.5 Multipath driver selection	27
2.2.6 The function of ADT	29
2.2.7 Disk expansion enclosures	32
2.2.8 Selecting drives.	35
2.3 Planning your storage structure	36
2.3.1 Arrays and RAID levels.	37
2.3.2 Logical drives and controller ownership	45
2.3.3 Hot spare drive	48
2.3.4 Storage partitioning.	48
2.3.5 Media scan	51
2.3.6 Segment size	52
2.3.7 Cache parameters	53
2.4 Planning for premium features	57
2.4.1 FlashCopy.	58
2.4.2 VolumeCopy	58
2.4.3 Enhanced Remote Mirroring (ERM)	58
2.4.4 FC/SATA Intermix.	59
2.5 Additional planning considerations	60
2.5.1 Planning for systems with LVM: AIX example.	61
2.5.2 Planning for systems without LVM: Windows example.	63

Chapter 3. DS4000 configuration tasks	67
3.1 Preparing the DS4000 Storage Server	68
3.1.1 Initial setup of the DS4000 Storage Server	68
3.1.2 Installing and starting the D4000 Storage Manager Client	72
3.2 DS4000 cabling	76
3.2.1 DS4100 and DS4300 host cabling configuration	77
3.2.2 DS4100 and DS4300 drive expansion cabling	80
3.2.3 DS4200 host cabling configuration	80
3.2.4 DS4200 drive expansion cabling	82
3.2.5 DS4500 host cabling configuration	87
3.2.6 DS4500 drive expansion cabling	88
3.2.7 DS4700 host cabling configuration	90
3.2.8 DS4700 drive expansion cabling	92
3.2.9 DS4800 host cabling configuration	96
3.2.10 DS4800 drive expansion cabling	97
3.2.11 Expansion enclosures	100
3.3 Configuring the DS4000 Storage Server	104
3.3.1 Defining hot-spare drives	104
3.3.2 Creating arrays and logical drives	106
3.3.3 Configuring storage partitioning	110
3.3.4 Configuring for Copy Services functions	113
3.4 Event monitoring and alerts	113
3.4.1 ADT alert notification	114
3.4.2 Failover alert delay	115
3.4.3 DS4000 Remote Support Manager	117
3.5 Software and microcode upgrades	120
3.5.1 Staying up-to-date with your drivers and firmware using My support	120
3.5.2 Prerequisites for upgrades	121
3.5.3 Updating the controller microcode	121
3.5.4 Updating DS4000 host software	127
3.6 Capacity upgrades, system upgrades	127
3.6.1 Capacity upgrades and increased bandwidth	128
3.6.2 Storage server upgrade and disk migration procedures	128
Chapter 4. DS4000 performance tuning	133
4.1 Workload types	134
4.2 Solution-wide considerations for performance	135
4.3 Host considerations	136
4.3.1 Host based settings	136
4.3.2 Host setting examples	138
4.4 Application considerations	147
4.4.1 Application examples	148
4.5 DS4000 Storage Server considerations	148
4.5.1 Which model fits best	149
4.5.2 Storage server processes	150
4.5.3 Storage server modification functions	152
4.5.4 Storage server parameters	154
4.5.5 Disk drive types	155
4.5.6 Additional NVSRAM parameters of concern	160
4.6 Fabric considerations	161
Chapter 5. DS4000 tuning with typical applications	163
5.1 DB2 database	164
5.1.1 Data location	164

5.1.2 Database structure	164
5.1.3 Database RAID type	166
5.1.4 DB2 logs and archives	167
5.2 Oracle databases	167
5.2.1 Data types	167
5.2.2 Data location	168
5.2.3 Database RAID and disk types	168
5.2.4 Redo logs RAID type	169
5.2.5 TEMP table space	169
5.2.6 Cache memory settings	170
5.2.7 Load balancing between controllers	170
5.2.8 Volume management	170
5.2.9 Performance Monitoring	171
5.3 Microsoft SQL Server	172
5.3.1 Allocation unit size	172
5.3.2 RAID levels	173
5.3.3 File locations	173
5.3.4 User database files	173
5.3.5 Tempdb database files	173
5.3.6 Transaction logs	174
5.3.7 Maintenance plans	175
5.4 IBM Tivoli Storage Manager backup server	176
5.5 Microsoft Exchange	178
5.5.1 Exchange configuration	178
5.5.2 Calculating theoretical Exchange I/O usage	179
5.5.3 Calculating Exchange I/O usage from historical data	180
5.5.4 Path LUN assignment (RDAC/MPP)	182
5.5.5 Storage sizing for capacity and performance	183
5.5.6 Storage system settings	185
5.5.7 Aligning Exchange I/O with storage track boundaries	185
Chapter 6. Analyzing and measuring performance	187
6.1 Analyzing performance	188
6.1.1 Gathering host server data	188
6.1.2 Gathering fabric network data	189
6.1.3 Gathering DS4000 storage server data	190
6.2 Iometer	190
6.2.1 Iometer components	190
6.2.2 Configuring Iometer	191
6.2.3 Results Display	196
6.3 Xdd	197
6.3.1 Xdd components and mode of operation	197
6.3.2 Compiling and installing Xdd	199
6.3.3 Running the xdd program	201
6.4 Storage Manager Performance Monitor	204
6.4.1 Starting the Performance Monitor	204
6.4.2 Using the Performance Monitor	207
6.4.3 Using the Performance Monitor: Illustration	211
6.5 AIX utilities	216
6.5.1 Introduction to monitoring Disk I/O	217
6.5.2 Assessing disk performance with the iostat command	217
6.5.3 Assessing disk performance with the vmstat command	219
6.5.4 Assessing disk performance with the sar command	220

6.5.5 Assessing logical volume fragmentation with the lslv command.	221
6.5.6 Assessing file placement with the fileplace command	221
6.5.7 The topas command	223
6.6 QLogic SANSurfer	224
6.6.1 Using the QLogic SANSurfer diagnostic tools.	225
6.7 MPPUTIL Windows 2000/2003	227
6.8 Windows Performance Monitor	228
Chapter 7. IBM TotalStorage Productivity Center for Disk	231
7.1 IBM TotalStorage Productivity Center	232
7.1.1 TotalStorage Productivity Center structure	232
7.1.2 Standards and protocols used in IBM TotalStorage Productivity Center	234
7.2 Managing DS4000 using IBM TPC for Disk	237
7.2.1 Install CIM agent for DS4000	237
7.2.2 Registering the Engenio SMI-S provider in TPC.	242
7.2.3 Probing CIM agent	245
7.2.4 Creating a Performance Monitor job	249
7.3 TPC reporting for DS4000.	252
7.3.1 DS4000 performance report	252
7.3.2 Generating reports	253
Chapter 8. Disk Magic	265
8.1 Disk Magic overview	266
8.2 Information required for DS4000 modeling with Disk Magic	266
8.3 Disk Magic configuration example	272
Chapter 9. ERM planning and implementation.	287
9.1 Introduction to ERM	288
9.2 ERM as part of a DR solution	289
9.2.1 Planning for ERM as part of a DR solution	290
9.2.2 Implementation recommendations	294
9.2.3 Network considerations.	297
9.2.4 Application considerations	299
9.2.5 Other design considerations	301
9.3 Site readiness and installation checklist	303
9.3.1 Site readiness and installation checklist details	304
9.4 The Bandwidth Estimator Tool	307
Chapter 10. SVC guidelines for DS4000	315
10.1 IBM System Storage SAN Volume Controller overview	316
10.2 SVC components and concepts	317
10.3 SVC copy services	320
10.3.1 SVC FlashCopy	320
10.3.2 Metro mirror	321
10.3.3 Global mirror	322
10.3.4 Differences between DS4000 and SVC copy services	323
10.4 SVC maximum configuration.	325
10.5 SVC considerations.	326
10.6 SVC with DS4000 best practices	327
10.6.1 DS4000 Storage Server family and SVC configuration example	329
Chapter 11. DS4000 with AIX and HACMP	339
11.1 Configuring DS4000 in an AIX environment	340
11.1.1 DS4000 adapters and drivers in an AIX environment.	340

11.1.2	Testing attachment to the AIX host	343
11.1.3	Storage partitioning in AIX	344
11.1.4	HBA configurations	347
11.1.5	Unsupported HBA configurations	351
11.1.6	Device drivers coexistence	356
11.1.7	Setting the HBA for best performance	358
11.1.8	DS4000 series – dynamic functions	359
11.2	HACMP and DS4000	361
11.2.1	Supported environment.	363
11.2.2	General rules	364
11.2.3	Configuration limitations	365
11.2.4	Planning considerations	367
11.2.5	Cluster disks setup	368
11.2.6	Shared LVM component configuration	371
11.2.7	Fast disk takeover.	375
11.2.8	Forced varyon of volume groups.	375
11.2.9	Heartbeat over disks	376
	Appendix A. DS4000 quick guide	381
	Pre-installation checklist.	382
	Installation tasks.	383
	Rack mounting and cabling.	383
	Preparing the host server	390
	Storage Manager setup	392
	Tuning for performance.	396
	Notes	397
	Notes on Windows	397
	Notes on Novell Netware 6.x.	399
	Notes on Linux	401
	Related publications	403
	IBM Redbooks	403
	Other publications	403
	Online resources	404
	How to get IBM Redbooks	404
	Help from IBM	405
	Index	407

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

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

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

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

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

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


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

COPYRIGHT LICENSE:

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

Trademarks

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

Redbooks (logo) ™
pSeries®
z/OS®
AIX 5L™
AIX®
BladeCenter®
DB2®
DS4000™
DS6000™
DS8000™

Enterprise Storage Server®
FlashCopy®
FICON®
HACMP™
IBM®
Lotus®
Netfinity®
POWER™
Redbooks™
ServeRAID™

System i™
System p™
System x™
System z™
System Storage™
System Storage DS™
SANergy®
Tivoli®
TotalStorage®

The following terms are trademarks of other companies:

Oracle, JD Edwards, PeopleSoft, Siebel, and TopLink are registered trademarks of Oracle Corporation and/or its affiliates.

Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates.

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

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

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

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

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

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

Preface

This book represents a compilation of best practices for deploying and configuring DS4000™ Storage Servers. It gives hints and tips for an expert audience on topics such as performance measurement, analysis and tuning, troubleshooting, HACMP™ Clustering, and Enhanced Remote Mirroring.

Setting up a DS4000 Storage Server can be a complex task. There is no single configuration that will be satisfactory for every application or situation.

This book starts by providing the conceptual framework for understanding the DS4000 in a Storage Area Network and then gives recommendations, hints, and tips for the physical installation, cabling, and zoning, and we review the Storage Manager setup tasks.

Follow-on chapters focus on performance and tuning of various components and features and includes numerous recommendations. We look at performance implications for various application products such as DB2®, Oracle®, Tivoli® Storage Manager, Microsoft® SQL server, and in particular, Microsoft Exchange with a DS4000 Storage Server.

We review various tools available to simulate workloads and measure and collect performance data for the DS4000. We provide an overview and illustrate the usage of IBM TotalStorage® Productivity Center for disk and Disk Magic.

Another chapter provides guidelines for planning and implementing Enhanced Remote Mirroring.

This edition of the book also includes guidelines for managing and using the DS4000 with IBM System Storage™ SAN Volume Controller.

This book is intended for IBM technical professionals, Business Partners, and customers responsible for the planning, deployment, and maintenance of IBM System Storage DS4000 family of products.

The team that wrote this redbook

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

Bertrand Dufrasne is a Certified Consulting IT Specialist and Project Leader for IBM TotalStorage products at the International Technical Support Organization, San Jose Center. He has worked at IBM in various IT areas. Before joining the ITSO, he worked for IBM Global Services as an Application Architect. He holds a degree in Electrical Engineering.

Bruce Allworth is a Senior IT Specialist working in the storage Advanced Technical Support (ATS), Americas group. He is a Subject Matter Expert and the ATS team lead for the DS4000 product. His many years of experience on the DS4000 include management, solution design, advanced problem determination, and disaster recovery. He works closely with various IBM divisions in developing and delivering DS4000 training seminars and technical presentations to a wide range of audiences.

Agung Indrayana is a System x™ IT Specialist in IBM Indonesia. He has four years of experience in the IT field. He holds a degree in Electrical Engineering from Gadjah Mada

University Indonesia, and certifications from RedHat (RHCE), Microsoft (MCP), and Sun™ Microsystems™ (SCSA). His areas of expertise include System x servers and DS4000 Storage Subsystems in Microsoft Windows®, Linux®, and Sun Solaris™ environments.

Christian Schoessler is an IT Specialist in the EMEA Storage Advanced Technical Support Team (ATS), located in Mainz (Germany). He has eight years of experience in the IT industry and five years as an IBM storage products specialist. In his role with ATS he has especially supported the DS4000 for the last four years. He holds a degree in Physics from the TU of Darmstadt.

Brian Youngs is an Infrastructure Support Specialist for Ipswich City Council in Australia. He has worked for Ipswich City Council for over 20 years. He is a Novell CNE. Brian has extensive experience with Novell NetWare, Microsoft Windows environments, System x servers, and DS4000 Storage Servers.

Thanks to the authors of the previous edition of this book:

Alexander Watson and Michele Lunardon.

Special thanks to **Thomas M. Ruwart**, I/O Performance Inc., author of Xdd.

Special thanks to **Pier Giuseppe Corengia**, IBM Italy, who authored most of the material included in 11.1, “Configuring DS4000 in an AIX environment” on page 340.

Special thanks to **Jodi Toft**, IBM, for contributing the material included in Appendix A, “DS4000 quick guide” on page 381.

Special thanks to **Bob Lai**, LSI, for his input and advice on the Enhanced Remote Mirroring chapter.

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

Deana Polm
International Technical Support Organization, San Jose Center

Danh Le
Jay Smith
George Thomas
Regina Pope-Ford
Alexander Watson
Stanley Wu
IBM US

Become a published author

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

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- Send your comments in an e-mail to:

redbooks@us.ibm.com

- Mail your comments to:

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

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition may also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-6363-03
for DS4000 Best Practices and Performance Tuning Guide
as created or updated on January 18, 2008.

March 2007, Third Edition

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

- ▶ Measuring performance with IBM TotalStorage Productivity Center for Disk.
- ▶ Using Disk Magic with DS4000 for performance tuning and capacity planning.
- ▶ Guidelines for managing the DS4000 with IBM System Storage SAN Volume Controller (SVC).
- ▶ Guidelines for planning and implementing Enhanced Remote Mirroring.
- ▶ ERM Bandwidth Estimator tool

Changed information

- ▶ Updated cabling information for new models (DS4200 and DS4700)
- ▶ New disks and expansion enclosures.
- ▶ Updated multipath driver information.



Introduction to DS4000 and SAN

In this chapter, we introduce IBM System Storage DS4000 products with a brief description of the different models, their features, and where they fit in terms of a storage solution. We also summarize the functions of the DS4000 Storage Manager software. Finally, we include a review of some of the basic concepts and topologies of Storage Area Networks as we refer to these in other parts of the book.

Readers already familiar with the DS4000 product line and SAN concepts can skip this chapter.

1.1 DS4000 features and models

IBM has brought together into one family, known as the DS family, a broad range of disk systems to help small to large-size enterprises select the right solutions for their needs. The DS family combines the high-performance IBM System Storage DS6000™ and DS8000™ series of enterprise servers that inherit from the ESS, with the DS4000 series of mid-range systems, and other line-of-entry systems (DS3000).

The IBM System Storage DS4000 Series of disk storage systems that this book addresses are IBM solution for mid-range/departmental storage requirements. The overall positioning of the DS4000 series within IBM System Storage DS™ family is shown in Figure 1-1.

Within the DS family, the DS4000 series of servers supports both Fibre Channel (FC) and Serial ATA (SATA) disk drives. The maximum raw SATA storage capacity of this family is over 112 TB (using 500 GB SATA drives). The maximum raw FC storage capacity is over 67 TB.

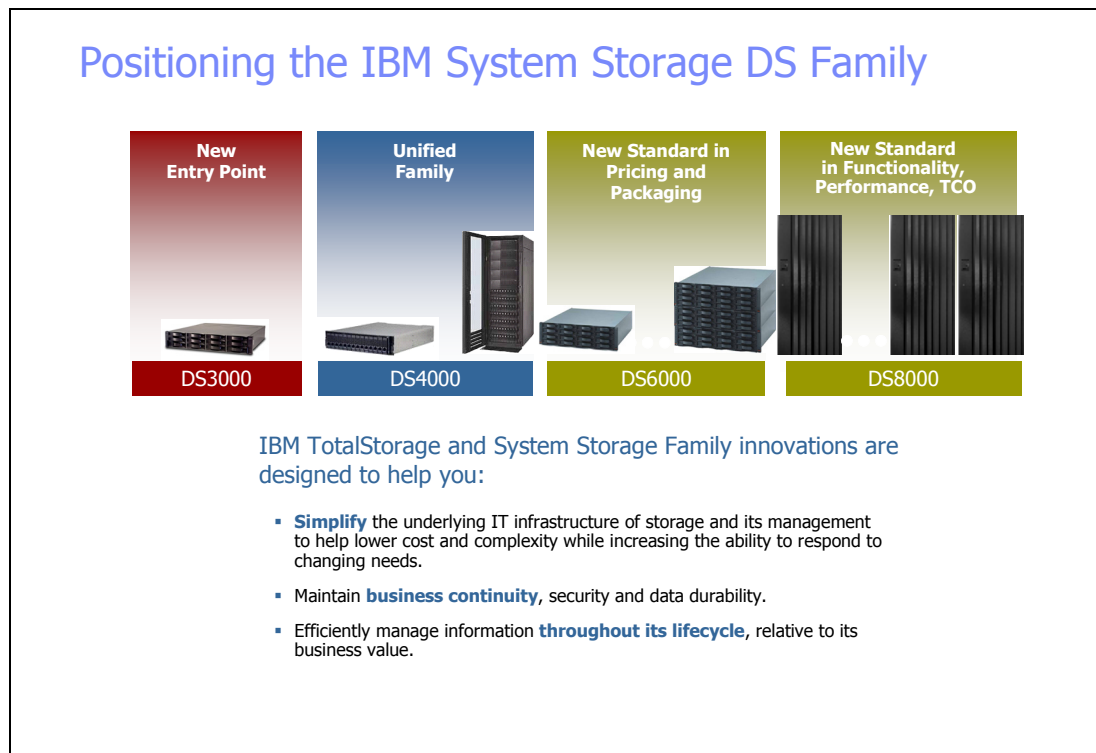


Figure 1-1 The IBM TotalStorage DS Family Overview

The DS4000 series of storage servers use Redundant Array of Independent Disks (RAID) technology. RAID technology is used to protect the user data from disk drive failures. DS4000 Storage Servers contain Fibre Channel (FC) interfaces to connect both the host systems and external disk drive enclosures.

Most of the storage servers in the DS4000 Series provide high system availability through the use of hot-swappable and redundant components. This is crucial when the storage server is placed in high-end customer environments such as server consolidation on Storage Area Networks (SANs). Most models offer autonomic functions such as Dynamic Volume Expansion and Dynamic Capacity Addition, allowing unused storage to be brought online without stopping operations.

The DS4000 Storage Servers are as follows:

- Current 4 Gbps System Storage Servers

- IBM System Storage Server DS4200 Express

The DS4200 Storage Server was, at the time of writing, the latest addition to the DS4000 Series of products. It is targeted at entry-level customers. It can hold a maximum of 16 disk drives inside the storage server enclosure and can attach up to six EXP420 Expansion Units for a total of up to 112 SATA disk drives. It is designed to deliver data throughput of up to 1600 MBps.

The DS4200 has a total of four 4 Gbps FC host ports and 2 GB of cache memory. Like other DS4000 family members, the DS4200 supports existing customer infrastructures, helping protect investments. In addition, the DS4200 is designed to efficiently handle the additional performance demands of FlashCopy®, Volume Copy, and Enhanced Remote Mirroring.

Note: The DS4200 is positioned as a replacement for the DS4100.

- IBM System Storage Server DS4700 Express

The DS4700 Storage Server is targeted at entry-level to mid-level customers. It can hold a maximum of 16 disk drives inside the storage server enclosure and can attach up to six EXP810 Expansion Units for a total of up to 112 Fibre Channel or SATA disk drives.

The DS4700 comes in two models, Model 72 and Model 70. The Model 72 has a total of eight 4 Gbps FC host ports and 4 GB of cache memory, while Model 70 has a total of four 4 Gbps FC host ports and 2 GB of cache memory. The DS4700 is a good choice for environments with intense replication requirements because it is designed to efficiently handle the additional performance demands of FlashCopy, Volume Copy, and Enhanced Remote Mirroring.

Note: The DS4700 is positioned as a replacement for the DS4300.

- IBM System Storage DS4800 Storage Server

The DS4800 Storage Server delivers breakthrough disk performance and outstanding reliability for demanding applications in compute-intensive environments. The DS4800 is a key component of IBM business continuity solutions portfolio, delivering business resilience and continuity of operations.

The DS4800 takes advantage of 4 Gbps Fibre Channel interface technology and can support up to 224 disk drives by attaching IBM System Storage EXP810, EXP710, or EXP100 disk units. It is a great choice for performance-oriented or capacity-oriented storage requirements. Four models are available: the new Model 80 with 4 GB of cache, the 82A with 4 GB of cache, the 84A with 8 GB of cache, and the Model 88A with 16 GB of cache.

Additionally, support for high-performance Fibre Channel and high-capacity Serial ATA (SATA) disk drives help enable a single DS4800 storage system to satisfy primary and secondary storage to accommodate the changing value of data over time while maintaining data availability.

The DS4800 Disk Storage System can provide enterprise-class disaster recovery strategies.

Note: The DS4800 is positioned as a replacement for the DS4500.

► Former 2 Gbps TotalStorage Servers

– IBM TotalStorage DS4100 Storage Server

The DS4100 Storage Server (formerly known as the FAStT100) is an entry-level SATA storage system that is available in a single and dual controller configuration.

– IBM TotalStorage DS4300 Storage Server

The DS4300 Storage Server (formerly known as the FAStT600) is a mid-level, highly scalable 2 Gbps Fibre Channel storage server, which is available in a single and dual controller configuration. There is also a DS4300 with Turbo feature that offers up to 65% read performance improvement and has higher Fibre Channel drive scalability over the base DS4300.

– IBM TotalStorage DS4500 Storage Server

The IBM DS4500 Storage Server (formerly known as FAStT900) delivers high disk performance and outstanding reliability for demanding applications in compute-intensive environments. The DS4500 is designed to offer investment protection with advanced functions and flexible features.

1.1.1 Telco industry standard support

Note: The DS4700 Express and the EXP810 Storage Expansion Unit offer models designed to be powered from a - 48 V dc Telco industry standard power source and are NEBS-3 compliant

1.1.2 DS4000 Series product comparison

Table 1-1 and Table 1-2 on page 5 summarize the characteristics of the DS4000 Series of products.

Table 1-1 Comparison of DS4100, DS4200, DS4300 and DS4700 models

DS model	DS4100	DS4200	DS4300 Turbo	DS4700-70	DS4700-72
Model no.	1724-100	1814-7V(H/A)	1722-60U Turbo	70A	72A
Environment	Entry level	Entry level	Midrange	Midrange	Midrange
Max Disks	112	112	112	112	112
Max raw capacity	44.8 TB	56 TB	33.6 TB FC 56 TB SATA ¹	33.6 TB FC 56 TB SATA	33.6 TB FC 56 TB SATA
Host interfaces	2 Gbps	4 Gbps	2 Gbps	4 Gbps	4 Gbps
SAN attach (max)	4 FC-SW	4 FC-SW	4 FC-SW	4 FC-SW	8 FC-SW
Direct attach (max)	4 FC-AL	4 FC-SW	4 FC-AL	4 FC-SW	4 FC-SW
Max cache memory	256 MB per controller	1 GB per controller	1 GB per controller	1 GB per controller	2 GB per controller
IOPS from cache read	70000*	120000*	77500*	120000*	120000*
IOPS from disk read	10000*	11200*	25000*	44000*	44000*

DS model	DS4100	DS4200	DS4300 Turbo	DS4700-70	DS4700-72
Throughput from disk	485 MBps*	990 MBps*	400 MBps*	990 MBps*	990 MBps*
Base/max partitions	0/16	2/64	8/64	2/64	8/64
Copy features	F	F, V, E	F, V, E	F, V, E	F, V, E
FC/SATA mix	No	No	YES	YES	YES
Available drives FC 10K rpm	N/A	N/A	36/73/146/300 GB	36/73/146/300 GB	36/73/146/300 GB
Available drives FC 15K rpm	N/A	N/A	18/36/73/146 GB	18/36/73/146 GB	18/36/73/146 GB
Available drives SATA	400 GB 7200 rpm	500 GB 7200 rpm EV-DDM Drives	400 GB 7200 rpm 500 GB 7200 rpm E-DDM ¹	400 GB 7200 rpm 500 GB 7200 rpm E-DDM	400 GB 7200 rpm 500 GB 7200 rpm E-DDM

F=FlashCopy, V=Volume Copy, E=Enhanced Remote Mirroring.

Note: * = Performance up to denoted value; may vary according to your particular environment.

¹ Intermix of EXP710, EXP100, and EXP810 allowed with 6.19 firmware. 400 GB 7200 rpm SATA drives only for EXP100 enclosures and 500 GB 7200 rpm SATA E-DDM drives only in EXP810 enclosures.

Table 1-2 Comparison of DS4500 and DS4800

DS model	DS4500	DS4800	DS4800 (4 GB cache)	DS4800 (8 GB cache)	DS4800 (16 GB cache)
Model no.	1742-900	1815-80A	1815-82A	1815-84A	1815-88A
Environment	Midrange to high end	High end	High end	High end	High end
Max disks	224	224	224	224	224
Max raw capacity	67.2 TB FC 89.6TB SATA	67.2 TB FC 112 TB SATA	67.2 TB FC 112 TB SATA	67.2 TB FC 112 TB SATA	67.2 TB FC 112 TB SATA
Host interfaces	2 Gbps	4 Gbps	4 Gbps	4 Gbps	4 Gbps
Host connections	4 (up to 8 with mini-hubs)	8	8	8	8
Drive-side interfaces	2 Gbps	4 Gbps	4 Gbps	4 Gbps	4 Gbps
Drive-side connections	4 (2 loop pairs)	8 (4 loop pairs)	8 (4 loop pairs)	8 (4 loop pairs)	8 (4 loop pairs)
SAN attach (max)	4 FC-SW ¹	4 FC-SW ¹	8 FC-SW ¹	8 FC-SW	8 FC-SW
Direct attach (max)	8 FC-AL	4 FC-SW	8 FC-SW	8 FC-SW	8 FC-SW
Max cache memory	1GB/controller	1 GB/controller	2 GB/controller	4 GB/controller	8 GB/controller

DS model	DS4500	DS4800	DS4800 (4 GB cache)	DS4800 (8 GB cache)	DS4800 (16 GB cache)
Model no.	1742-900	1815-80A	1815-82A	1815-84A	1815-88A
IOPS from cache read	148000*	575000*	575000*	575000*	575000*
IOPS from disk read	53200*	86000*	86000*	86000*	86000*
Throughput from disk	790 MBps*	1600 MBps*	1600 MBps*	1600 MBps*	1600 MBps*
Base/max partitions	16/64	8/16/64	8/16/64	8/16/64	8/16/64
Copy features	F, V, E	F, V, E	F, V, E	F, V, E	F, V, E
FC/SATA Drive Intermix	YES ⁴	YES ³	YES ³	YES ³	YES ³
Available drives FC 10k rpm	36/73/146/300 GB	36/73/146/300 GB	36/73/146/300 GB	36/73/146/300 GB	36/73/146/300 GB
Available drives FC 15k rpm	18/36/73/146 GB	36/73/146 GB	36/73/146 GB	36/73/146 GB	36/73/146 GB
Available drives SATA	400 GB 7200 rpm 500 GB 7200 rpm E-DDM ⁵	400 GB 7200 rpm 500 GB 7200 rpm E-DDM ⁵	400 GB 7200 rpm 500 GB 7200 rpm E-DDM ⁵	400 GB 7200 rpm 500 GB 7200 rpm E-DDM ⁵	400 GB 7200 rpm 500 GB 7200 rpm E-DDM ⁵

1) For more than four connections, purchase of additional mini-hubs is required.

2) F=FlashCopy, V=Volume Copy, E=Enhanced Remote Mirroring.

3) Intermix on EXP710 and EXP100 allowed on 6.15 and earlier firmware. Intermix of SATA and FC in EXP810 in separate enclosures allowed.

4) Intermix of EXP710, EXP100, and EXP810 allowed with 6.19 firmware

5) 400 GB 7200 rpm SATA drives only for EXP100 enclosures and 500 GB 7200 rpm SATA E-DDM drives only in EXP810 enclosures.

Note: * = Performance up to denoted value. May vary according to your particular environment.

1.2 DS4000 Storage Manager

The DS4000 Storage Manager software is used primarily to configure RAID arrays and logical drives, assign logical drives to hosts, replace and rebuild failed disk drives, expand the size of the arrays and logical drives, and convert from one RAID level to another. It allows troubleshooting and management tasks, like checking the status of the storage server components, updating the firmware of the RAID controllers, and managing the storage server. Finally, it offers advanced functions such as FlashCopy, Volume Copy, and Enhanced Remote Mirroring.

Note: Always consult IBM TotalStorage DS4000 Interoperability matrix for information about the latest supported Storage Manager version for your DS4000 system. It is available on the Web at:

<http://www.ibm.com/servers/storage/disk/ds4000/interop-matrix.html>

The Storage Manager software is now packaged as follows:

► *Host-based* software:

- Storage Manager 9.1x Client (SMclient):

The SMclient component provides the graphical user interface (GUI) for managing storage subsystems through the Ethernet network or from the host computer.

- Storage Manager 9.1x Runtime (SMruntime):

The SMruntime is a Java™ runtime environment that is required for the SMclient to function. It is not available on every platform as a separate package, but in those cases, it has been bundled into the SMclient package.

- Storage Manager 9.1x Agent (SMagent):

The SMagent package is an optional component that allows in-band management of the DS4000 Storage Server.

- Storage Manager 9.1x Utilities (SMutil):

The Storage Manager Utilities package contains command line tools for making logical drives available to the operating system.

- Multipath drivers:

Version 9.19 of storage manager offers a choice of multipath driver, RDAC or MPIO.

During the installation you are prompted to choose between RDAC or MPIO. Both are Fibre Channel I/O path failover drivers that are installed on host computers. These are only required if the host computer has a host bus adapter (HBA) installed.

► *Controller-based* software:

- DS4000 Storage Server controller firmware and NVSRAM:

The controller firmware and NVSRAM are always installed as a pair and provide the “brains” of the DS4000 Storage Server.

- DS4000 Storage Server Environmental Service Modules (ESM) firmware:

The ESM firmware controls the interface between the controller and the drives.

- DS4000 Storage Server Drive firmware:

The drive firmware is the software that tells the Fibre Channel (FC) drives how to behave on the FC loop.

1.3 Introduction to SAN

For businesses, data access is critical and requires performance, availability, and flexibility. In other words, there is a need for a data access network that is fast, redundant (multipath), easy to manage, and always available. That network is a Storage Area Network (SAN).

A SAN is a high-speed network that enables the establishment of direct connections between storage devices and hosts (servers) within the distance supported by Fibre Channel.

The SAN can be viewed as an extension of the storage bus concept, which enables storage devices to be interconnected using concepts similar to that of local area networks (LANs) and wide area networks (WANs). A SAN can be shared between servers or dedicated to one server, or both. It can be local or extended over geographical distances.

The diagram in Figure 1-2 shows a brief overview of a SAN connecting multiple servers to multiple storage systems.

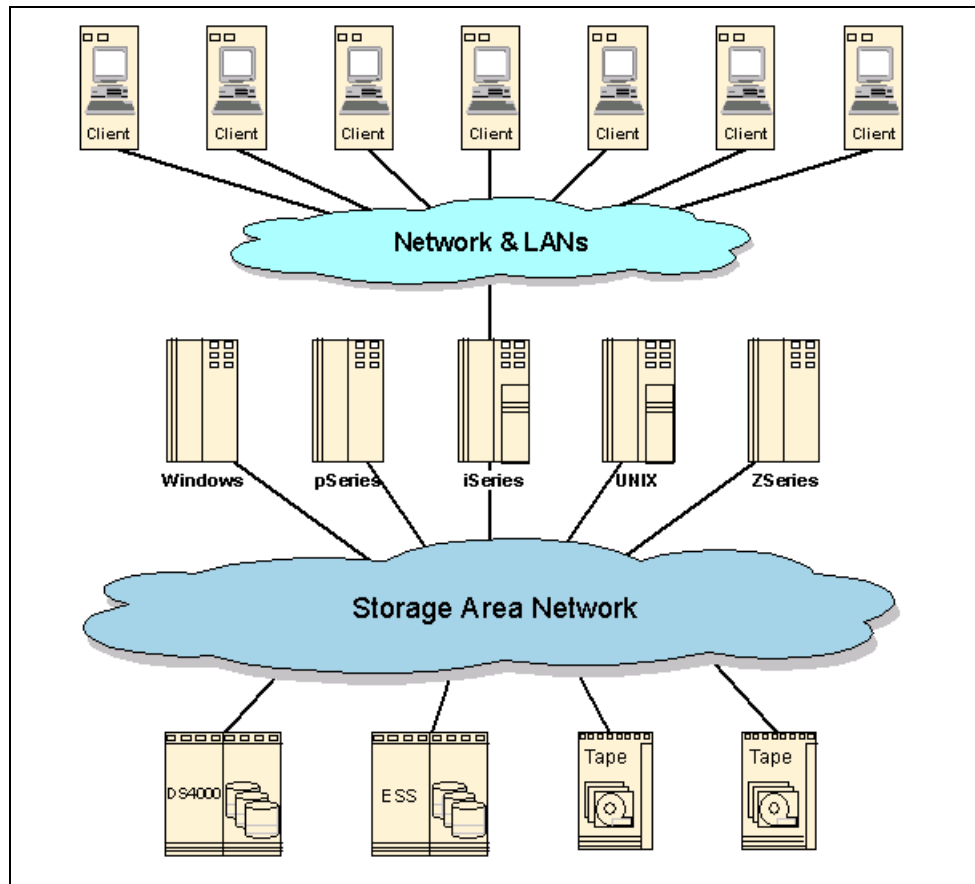


Figure 1-2 What is a SAN?

SANs create new methods of attaching storage to servers. These new methods can enable great improvements in availability, flexibility, and performance. Today's SANs are used to connect shared storage arrays and tape libraries to multiple servers, and are used by clustered servers for failover. A big advantage of SANs is the sharing of devices among heterogeneous hosts.

1.3.1 SAN components

In this section, we present a brief overview of the basic SAN storage concepts and building blocks.

SAN servers

The server infrastructure is the underlying reason for all SAN solutions. This infrastructure includes a mix of server platforms, such as Microsoft Windows, Novell NetWare, UNIX® (and its various flavors), and IBM z/OS®.

SAN storage

The storage infrastructure is the foundation on which information relies, and therefore, must support a company's business objectives and business model. In this environment, simply deploying more and faster storage devices is not enough. A SAN infrastructure provides enhanced availability, performance, scalability, data accessibility and system manageability. It is important to remember that a good SAN begins with a good design. The SAN liberates the storage device, so it is not on a particular server bus, and attaches it directly to the network. In other words, storage is externalized and can be functionally distributed across the organization. The SAN also enables the centralization of storage devices and the clustering of servers, which has the potential to make for easier and less expensive centralized administration that lowers the total cost of ownership (TCO).

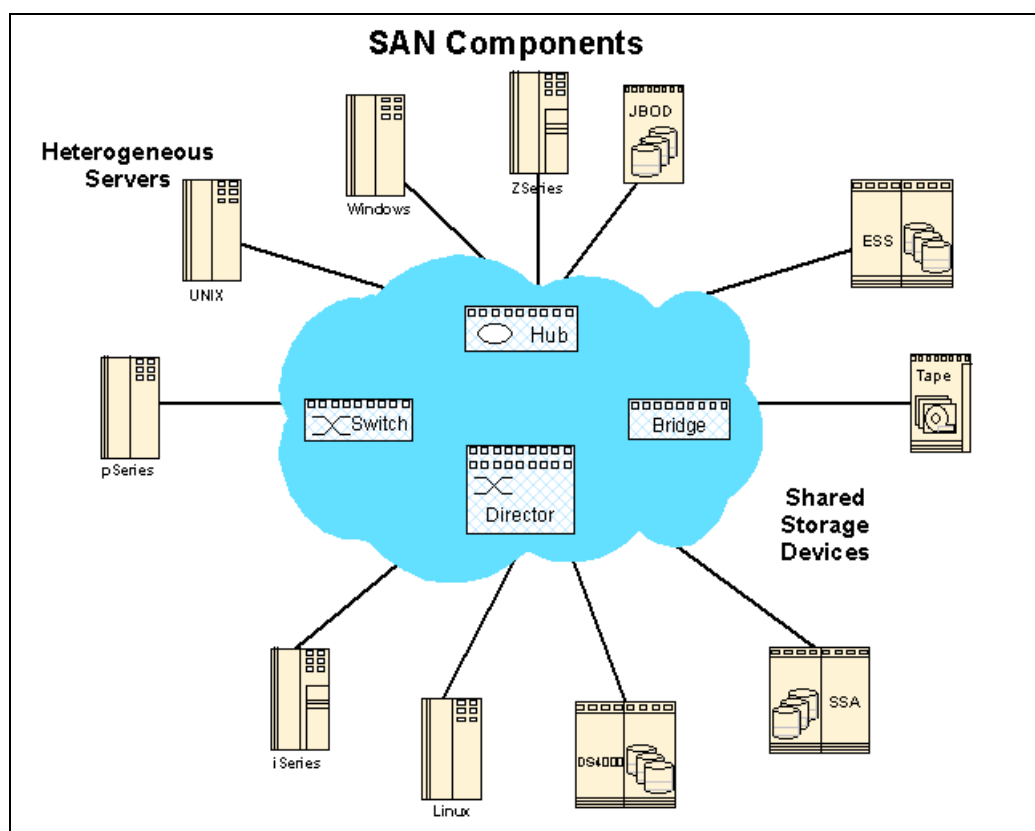


Figure 1-3 SAN components

Fibre Channel

Today, Fibre Channel (FC) is the architecture on which most SAN implementations are built. Fibre Channel is a technology standard that enables data to be transferred from one network node to another at very high speeds. Current implementations transfer data at 1 Gbps, 2 Gbps, and 4Gbps (10 Gbps data rates have already been tested).

Fibre Channel was developed through industry cooperation — unlike SCSI, which was developed by a vendor, and submitted for standardization after the fact.

Some people refer to Fibre Channel architecture as the Fibre version of SCSI. Fibre Channel is an architecture that can carry IPI traffic, IP traffic, FICON® traffic, FCP (SCSI) traffic, and possibly traffic using other protocols, all on the standard FC transport.

SAN topologies

Fibre Channel interconnects nodes using three physical topologies that can have variants. These three topologies are:

- ▶ Point-to-point: The point-to-point topology consists of a single connection between two nodes. All the bandwidth is dedicated to these two nodes.
- ▶ Loop: In the loop topology, the bandwidth is shared between all the nodes connected to the loop. The loop can be wired node-to-node; however, if a node fails or is not powered on, the loop is out of operation. This is overcome by using a hub. A hub opens the loop when a new node is connected, and closes it when a node disconnects.
- ▶ Switched or fabric: A switch enables multiple concurrent connections between nodes. There are two types of switches: circuit switches and frame switches. Circuit switches establish a dedicated connection between two nodes, while frame switches route frames between nodes and establish the connection only when needed. This is also known as switched fabric.

Note: The fabric (or switched) topology gives the most flexibility and ability to grow your installation for future needs.

SAN interconnects

Fibre Channel employs a fabric to connect devices. A fabric can be as simple as a single cable connecting two devices. However, the term is most often used to describe a more complex network using cables and interface connectors, HBAs, extenders, and switches.

Fibre Channel switches function in a manner similar to traditional network switches to provide increased bandwidth, scalable performance, an increased number of devices, and in some cases, increased redundancy. Fibre Channel switches vary from simple edge switches to enterprise-scalable core switches or Fibre Channel directors.

Inter-Switch Links (ISLs)

Switches can be linked together using either standard connections or Inter-Switch Links. Under normal circumstances, traffic moves around a SAN using the Fabric Shortest Path First (FSPF) protocol. This allows data to move around a SAN from initiator to target using the quickest of alternate routes. However, it is possible to implement a direct, high-speed path between switches in the form of ISLs.

Trunking

Inter-Switch Links can be combined into logical groups to form trunks. In IBM TotalStorage switches, trunks can be groups of up to four ports on a switch connected to four ports on a second switch. At the outset, a trunk master is defined, and subsequent trunk slaves can be added. This has the effect of aggregating the throughput across all links. Therefore, in the case of switches with 2 Gbps ports, we can trunk up to four ports, allowing for an 8 Gbps Inter-Switch Link.

1.3.2 SAN zoning

A zone is a group of fabric-connected devices arranged into a specified grouping. Zones can vary in size depending on the number of fabric-connected devices, and devices can belong to more than one zone.

Typically, you use zones to do the following tasks:

- ▶ **Provide security:** Use zones to provide controlled access to fabric segments and to establish barriers between operating environments. For example, isolate systems with different uses or protect systems in a heterogeneous environment.
- ▶ **Customize environments:** Use zones to create logical subsets of the fabric to accommodate closed user groups or to create functional areas within the fabric. For example, include selected devices within a zone for the exclusive use of zone members, or create separate test or maintenance areas within the fabric.
- ▶ **Optimize IT resources:** Use zones to consolidate equipment logically for IT efficiency, or to facilitate time-sensitive functions. For example, create a temporary zone to back up non-member devices.

Note: Utilizing zoning is always a good idea with SANs that include more than one host. With SANs that include more than one operating system, or SANs that contain both tape and disk devices, it is mandatory.

Without zoning, failing devices that are no longer following the defined rules of fabric behavior might attempt to interact with other devices in the fabric. This type of event would be similar to an Ethernet device causing broadcast storms or collisions on the whole network, instead of being restricted to one single segment or switch port. With zoning, these failing devices cannot affect devices outside of their zone.

Zone types

A zone member can be specified using one of the following zone types:

Port level zone	A zone containing members specified by switch ports (domain ID, port number) only. Port level zoning is enforced by hardware in the switch.
WWPN zone	A zone containing members specified by device World Wide Port Name (WWPN) only. WWPN zones are hardware enforced in the switch.
Mixed zone	A zone containing some members specified by WWPN and some members specified by switch port. Mixed zones are software enforced through the fabric name server.

Zones can be hardware enforced or software enforced:

- ▶ In a hardware-enforced zone, zone members can be specified by physical port number, or in recent switch models, through WWPN, but not within the same zone.
- ▶ A software-enforced zone is created when a port member and WWPN members are in the same zone.

Note: You do not explicitly specify a type of enforcement for a zone. The type of zone enforcement (hardware or software) depends on the type of member it contains (WWPNs or ports).

For more complete information regarding Storage Area Networks, refer to the following IBM Redbooks:

- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *IBM SAN Survival Guide*, SG24-6143

Zoning configuration

Zoning is not hard to understand or configure. Using your switch management software, use WWPN zoning to set up each zone so that it contains one server port, and whatever storage device ports that host port requires access to. You do not need to create a separate zone for each source/destination pair. Do not put disk and tape access in the same zone. Also avoid using the same HBA for disk and tape access.

We cannot stress enough to ensure that all zoning information is fully documented and that documentation is kept up to date. This should be kept in a safe location for reference, documentation, and planning purposes. If done correctly, the document can be used to assist in diagnosing zoning problems.

When configuring World Wide Name (WWN) based zoning, it is important to always use the World Wide Port Name (WWPN), not the World Wide Node Name (WWNN). With many systems, the WWNN is based on the Port WWN of the first adapter detected by the HBA driver. If the adapter the WWNN was based on were to fail, and you based your zoning on the WWNN, your zoning configuration would become invalid. Subsequently, the host with the failing adapter would completely lose access to the storage attached to that switch.

Keep in mind that you will need to update the zoning information, should you ever need to replace a Fibre Channel adapter in one of your servers. Most storage systems such as the DS4000, Enterprise Storage Server®, and IBM Tape Libraries have a WWN tied to the Vital Product Data of the system unit, so individual parts may usually be replaced with no effect on zoning.

For more details on configuring zoning with your particular switch, see *IBM TotalStorage: Implementing an Open IBM SAN*, SG24-6116.

Multiple fabrics

Depending on the size, levels of redundancy, and budget, you may want more than one switched fabric. Multiple fabrics increase the redundancy and resilience of your SAN by duplicating the fabric infrastructure. With multiple fabrics the hosts and the resources have simultaneous access to both fabrics, and have zoning to allow multiple paths over each fabric.

- ▶ Each server can have two or more HBAs. In a two-HBA configuration, each HBA can connect to a different fabric.
- ▶ Each DS4000 can use different host ports or mini hubs to connect to multiple fabrics. Thus giving a presence in each fabric.
- ▶ Zoning in each fabric means that the server can have many paths to its resources, this would also mean that the zoning has to be done in each fabric separately.
- ▶ The complete loss of a fabric would mean that the host could still access the resources via the other fabric.

The multiple fabric increases the complexity, resiliency, and redundancy of the SAN infrastructure. This, however, comes at a larger cost due to the duplication of switches, HBAs, zoning administration, and fibre connections. This has to be carefully examined to see whether your SAN infrastructure requirements require multiple fabrics.



DS4000 planning tasks

Careful planning is essential to any new storage installation. This chapter provides guidelines to help you in the planning process.

Choosing the right equipment and software, and also knowing what the right settings are for a particular installation, can be challenging. Every installation has to answer these questions and accommodate specific requirements, and there can be many variations in the solution.

Well-thought design and planning prior to the implementation will help you get the most out of your investment for the present and protect it for the future.

During the planning process, you need to answer numerous questions about your environment:

- ▶ What are my SAN requirements?
- ▶ What hardware do I need to buy?
- ▶ What reliability do I require?
- ▶ What redundancy do I need? (For example, do I need off-site mirroring?)
- ▶ What compatibility issues do I need to address?
- ▶ Will I use any storage virtualization product such as IBM SAN Volume controller?
- ▶ What operating system am I going to use (existing or new installation)?
- ▶ What applications will access the storage subsystem?
- ▶ What are the hardware and software requirements of these applications?
- ▶ What will be the physical layout of the installation? Only local site, or remote sites as well?
- ▶ What level of performance do I need?
- ▶ How much does it cost?

This list of questions is not exhaustive, and as you can see, some go beyond simply configuring the DS4000 Storage Server.

Some recommendations in this chapter come directly from experience with various DS4000 installations at customer sites.

2.1 Planning your SAN and storage server

When planning to set up a Storage Area Network (SAN), you want the solution to not only answer your current requirements, but also be able to fulfill future needs.

First, the SAN should be able to accommodate a growing demand in storage (it is estimated that storage need doubles every two years). Second, the SAN must be able to keep up with the constant evolution of technology and resulting hardware upgrades and improvements. It is estimated that a storage installation needs to be upgraded every two to three years.

Ensuring compatibility among different pieces of equipment is crucial when planning the installation. The important question is what device works with what, and also who has tested and certified (desirable) that equipment.

When designing a SAN storage solution, it is good practice to complete the following steps:

1. Produce a statement outlining the solution requirements that can be used to determine the type of configuration you need. It should also be used to cross-check that the solution design delivers the basic requirements. The statement should have easily defined bullet points covering the requirements, for example:
 - New installation or upgrade of existing infrastructure
 - Host Bus Adapter (HBA) selection
 - HBA driver type selection - SCSIPort or StorPort
 - Multipath Driver selection (RDAC, MPIO)
 - Types of applications accessing the SAN (are the applications I/O intensive or high throughput?)
 - Required capacity
 - Required redundancy levels
 - Type of data protection needed
 - Current data growth patterns for your environment
 - Is the current data more read or write based?
 - Backup strategies in use (Network, LAN-free or Server-less)
 - Premium Features required (FC/SATA Intermix, Partitioning, FlashCopy, Volume Copy or Enhanced Remote Mirroring)
 - Number of host connections required
 - Types of hosts and operating systems that will connect to the SAN
 - What zoning is required
 - Distances between equipment and sites (if there is there more than one site)
2. Produce a hardware checklist. It should cover such items that require you to:
 - Make an inventory of existing hardware infrastructure. Ensure that any existing hardware meets minimum hardware requirements and is supported with the DS4000.
 - Make a complete list of the planned hardware requirements.
 - Ensure that you have enough rack space for future capacity expansion.
 - Ensure that power and environmental requirements are met.
 - Ensure that your existing Fibre Channel switches and cables are properly configured.

3. Produce a software checklist to cover all the required items that need to be certified and checked. It should include such items that require you to:

- Ensure that the existing versions of firmware and storage management software are up to date.
- Ensure host operating systems are supported with the DS4000. Check IBM System Storage DS4000 interoperability matrix available at this Web site:

<http://www.ibm.com/servers/storage/disk/ds4000/interop-matrix.html>

This list is not exhaustive, but the creation of the statements is an exercise in information gathering and planning; it assists you in a greater understanding of what your needs are in your current environment and creates a clearer picture of your future requirements. The goal should be quality rather than quantity of information.

Use this planning chapter as a reference that can assist you to gather the information for the statements.

Understanding the applications is another important consideration in planning for your DS4000. Applications can typically be either be I/O intensive (high number of I/O per second or IOPS), or characterized by large I/O requests (that is, high throughput or MBps).

- ▶ Typical examples of high IOPS environments are Online Transaction Processing (OLTP), database, and Microsoft Exchange servers. These have random writes and fewer reads.
- ▶ Typical examples of high throughput applications are data mining, imaging, and backup storage pools. These have large sequential reads and writes.

4.1, “Workload types” on page 134, provides a detailed discussion and considerations for application types. The planning for each application type affects hardware purchases and configuration options.

By understanding your data and applications, you can also better understand growth patterns. Being able to estimate an expected growth is vital for the capacity planning of your DS4000 Storage Server installation. Clearly indicate the expected growth in the planning documents, to act as a guide: The actual patterns may differ from the plan but that is the dynamics of your environment.

Selecting the right DS4000 Storage Server model for your current and perceived future needs is one of the most crucial decisions that will have to be made. The good side, however, is that the DS4000 offers scalability and expansion flexibility. Premium Features can be purchased and installed at a later time to add functionality to the storage server.

In any case, it is perhaps better to purchase a higher model than one strictly dictated by your current requirements and expectations. This will allow for greater performance and scalability as your needs and data grow.

2.1.1 SAN zoning for DS4000

Zoning is an important part of integrating a DS4000 Storage Server in a SAN. When done correctly, it can eliminate many common problems.

A best practice is to create a zone for the connection between the host bus adapter (HBA1) and controller A and a separate zone that contains the other HBA2 to controller B. Then create additional zones for access to other resources. This isolates each zone down to its simplest form.

Best practice: Create separate zones for the connection between each HBA and each controller (one zone for HBA1 to controller A and one zone for HBA2 to controller B). This isolates each zone to its simplest form.

Disk and tape access should not be on the same HBA and should not be in the same zone.

Important: Disk and tape should be on separate HBAs, following the best practice for zoning; then the disk and tape access will also be in separate zones. With some UNIX systems, this is supported by the DS4000 due to hardware limitations, but generally HBA sharing is strongly *not* recommended.

For systems such as IBM BladeCenter® servers that have a limited number of FC ports available, we suggest that you perform a LAN backup instead of a LAN-free backup directly to the tape drives.

Enhanced Remote Mirroring considerations

When using Enhanced Remote Mirroring (ERM), you must create two additional zones:

- ▶ The first zone contains the ERM source DS4000 controller A and ERM target DS4000 controller A.
- ▶ The second zone contains the ERM source DS4000 controller B and ERM target DS4000 controller B.

Important: On the DS4100, DS4200, DS4300, DS4700 (Model 70), and DS4500 the ERM port is the second set of ports on controller A and controller B.

On the DS4700 (Model 72) and DS4800, the ERM port is port 4 on controller A and controller B.

2.2 Physical components planning

In this section, we review elements related to physical characteristics of an installation, such as rack considerations, fibre cables, Fibre Channel adapters, and other elements related to the structure of the storage system and disks, including enclosures, arrays, controller ownership, segment size, storage partitioning, caching, hot spare drives, and Enhanced Remote Mirroring.

2.2.1 Rack considerations

The DS4000 Storage Server and possible expansions are mounted in rack enclosures.

General planning

Consider the following general planning guidelines. Determine:

- ▶ The size of the floor area required by the equipment:
 - Floor-load capacity
 - Space needed for expansion
 - Location of columns
- ▶ The power and environmental requirements.

Create a floor plan to check for clearance problems. Be sure to include the following considerations on the layout plan:

- ▶ Service clearances required for each rack or suite of racks.
- ▶ If the equipment is on a raised floor, determine:
 - The height of the raised floor
 - Things that might obstruct cable routing
- ▶ If the equipment is not on a raised floor, determine:
 - The placement of cables to minimize obstruction
 - If the cable routing is indirectly between racks (such as along walls or suspended), the amount of additional cable needed
 - Cleanliness of floors, so that the fan units will not attract foreign material such as dust or carpet fibers
- ▶ Location of:
 - Power receptacles
 - Air conditioning equipment, placement of grilles and controls
 - File cabinets, desks, and other office equipment
 - Room emergency power-off controls
 - All entrances, exits, windows, columns, and pillars
 - Fire control systems
- ▶ Check access routes for potential clearance problems through doorways and passage ways, around corners, and in elevators for racks and additional hardware that will require installation.
- ▶ Store all spare materials that can burn in properly designed and protected areas.

Rack layout

To be sure you have enough space for the racks, create a floor plan before installing the racks. You might need to prepare and analyze several layouts before choosing the final plan.

If you are installing the racks in two or more stages, prepare a separate layout for each stage.

Consider the following things when you make a layout:

- ▶ The flow of work and personnel within the area
- ▶ Operator access to units, as required
- ▶ If the rack is on a raised floor:
 - Ensure adequate cooling and ventilation
- ▶ If the rack is not on a raised floor, determine:
 - The maximum cable lengths
 - The need for cable guards, ramps, and so on to protect equipment and personnel
- ▶ Location of any planned safety equipment.
- ▶ Future expansion.

Review the final layout to ensure that cable lengths are not too long and that the racks have enough clearance.

You need at least 152 cm (60 inches) of clearance at the front and at least 76 cm (30 inches) at the rear of the 42-U rack suites. This space is necessary for opening the front and rear doors and for installing and servicing the rack. It also allows air circulation for cooling the equipment in the rack. All vertical rack measurements are given in rack units (U). One U is

equal to 4.45 cm (1.75 inches). The U levels are marked on labels on one front mounting rail and one rear mounting rail. Figure 2-1 shows an example of the required service clearances for a 9306-900 42U rack. Check with the manufacturer of the rack for the statement on clearances.

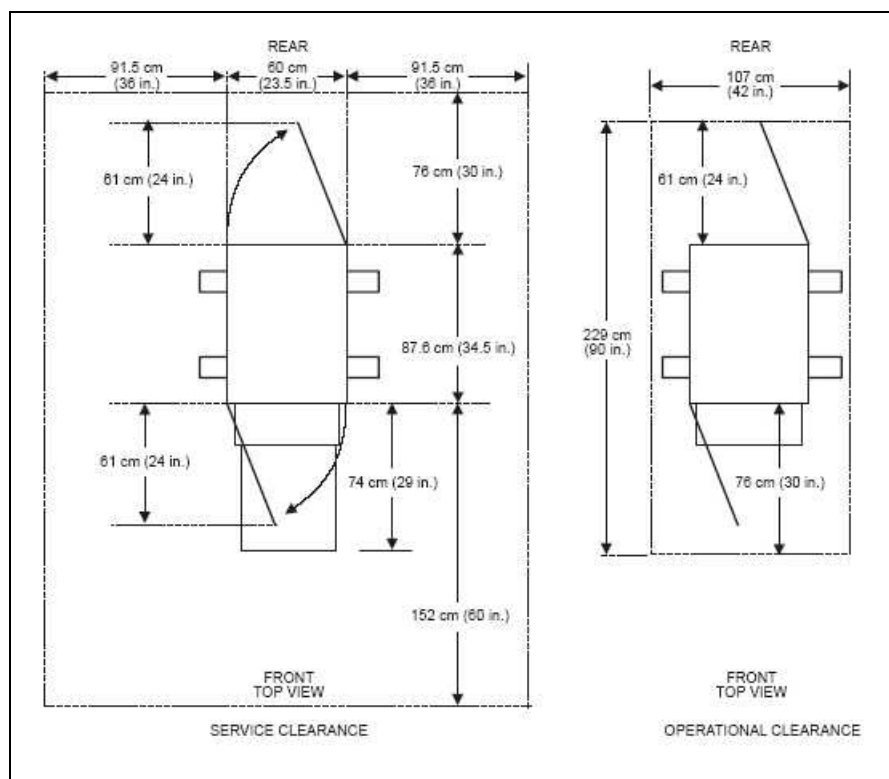


Figure 2-1 9306 Enterprise rack space requirements

2.2.2 Cables and connectors

In this section, we discuss some essential characteristics of fibre cables and connectors. This should help you understand options you have for connecting and cabling the DS4000 Storage Server.

Cable types (shortwave or longwave)

Fiber cables are basically available in multi-mode fiber (MMF) or single-mode fiber (SMF).

Multi-mode fiber allows light to disperse in the fiber so that it takes many different paths, bouncing off the edge of the fiber repeatedly to finally get to the other end (multi-mode means multiple paths for the light). The light taking these different paths gets to the other end of the cable at slightly different times (different path, different distance, different time). The receiver has to determine which signals go together as they all come flowing in.

The maximum distance is limited by how “blurry” the original signal has become. The thinner the glass, the less the signals “spread out,” and the further you can go and still determine what is what on the receiving end. This dispersion (called modal dispersion) is the critical factor in determining the maximum distance a high-speed signal can go. It is more relevant than the attenuation of the signal (from an engineering standpoint, it is easy enough to increase the power level of the transmitter or the sensitivity of your receiver, or both, but too much dispersion cannot be decoded no matter how strong the incoming signals are).

There are two different core sizes of multi-mode cabling available: 50 micron and 62.5 micron. The intermixing of the two different core sizes can produce unpredictable and unreliable operation. Therefore, core size mixing is not supported by IBM. Users with an existing optical fibre infrastructure are advised to ensure it meets Fibre Channel specifications and is a consistent size between pairs of FC transceivers.

Single-mode fiber (SMF) is so thin (9 microns) that the light can barely “squeeze” through and it tunnels through the center of the fiber using only one path (or mode). This behavior can be explained (although not simply) through the laws of optics and physics. The result is that because there is only one path that the light takes to the receiver, there is no “dispersion confusion” at the receiver. However, the concern with single mode fiber is attenuation of the signal. Table 2-1 lists the supported distances.

Table 2-1 Cable type overview

Fiber type	Speed	Maximum distance
9 micron SMF (longwave)	1 Gbps	10 km
9 micron SMF (longwave)	2 Gbps	2 km
50 micron MMF (shortwave)	1 Gbps	500 m
50 micron MMF (shortwave)	2 Gbps	300 m
50 micron MMF (shortwave)	4 Gbps	150 m
62.5 micron MMF (shortwave)	1 Gbps	175 m/300 m
62.5 micron MMF (shortwave)	2 Gbps	90 m/150 m

Note that the “maximum distance” shown in Table 2-1 is just that, a maximum. Low quality fiber, poor terminations, excessive numbers of patch panels, and so on, can cause these maximums to be far shorter. At the time of writing this book, only the 50 micron MMF (shortwave) cable is officially supported on the DS4800 for 4 Gbps connectivity.

All IBM fiber feature codes that are orderable with the DS4000 will meet the standards.

Interfaces, connectors, and adapters

In Fibre Channel technology, frames are moved from source to destination using gigabit transport, which is a requirement to achieve fast transfer rates. To communicate with gigabit transport, both sides have to support this type of communication. This is accomplished by using specially designed interfaces that can convert other communication transport into gigabit transport.

The interfaces that are used to convert the internal communication transport of gigabit transport are, depending on the DS4000 model either Small Form Factor Transceivers (SFF), also often called Small Form Pluggable (SFP) or Gigabit Interface Converters (GBIC). See Figure 2-2.



Figure 2-2 Small Form Pluggable (SFP) with LC connector Fibre Cable

Obviously, the particular connectors used to connect a fiber to a component will depend upon the receptacle into which they are being plugged.

LC connector

Connectors that plug into SFF or SFP devices are called LC connectors. The two fibers each have their own part of the connector. The connector is keyed to ensure correct polarization when connected, that is, transmit to receive and vice-versa.

The main advantage that these LC connectors have over the SC connectors is that they are of a smaller form factor, and so manufacturers of Fibre Channel components are able to provide more connections in the same amount of space.

All DS4000 Series products now use SFP transceivers and LC Fibre Cables. See Figure 2-3.

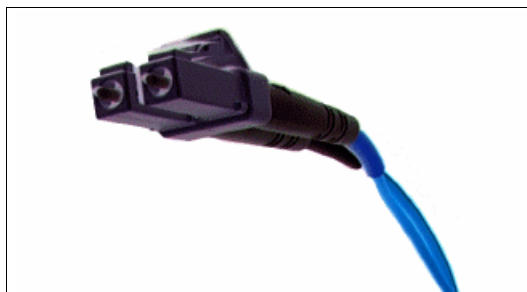


Figure 2-3 LC Fibre Cable Connector

SC connector

The duplex SC connector is a low loss, push/pull fitting connector. It is easy to configure and replace. Again, a duplex version is used so that the transmit and receive are connected in one step.

The FAStT200, FAStT500, and EXP500 use GBICs and SC connectors. See Figure 2-4.



Figure 2-4 GBIC Connector and SC Fibre Connection

Best practice: When you are not using an SFP or GBIC, it is best to remove it from the port on the DS4000 and replace it with a cover. This will help eliminate unnecessary wear and tear.

Interoperability of 1 Gbps, 2 Gbps, and 4 Gbps devices

The Fibre Channel standard specifies a procedure for speed auto-detection. Therefore, if a 2 Gbps port on a switch or device is connected to a 1 Gbps port, it should negotiate down and the link will run at 1 Gbps. If there are two 2 Gbps ports on either end of a link, the negotiation runs the link at 2 Gbps if the link is up to specifications. A link that is too long or “dirty” could end up running at 1 Gbps even with 2 Gbps ports at either end, so watch your distances and make sure your fiber is good. The same rules apply to 4 Gbps devices relative to 1 Gbps and 2 Gbps environments. The 4 Gbps devices have the ability to automatically negotiate back down to either 2 Gbps or 1 Gbps, depending upon the attached device and the link quality.

The DS4100, DS4300, DS4400, DS45000, EXP700, EXP710 and EXP100 Enclosures are 2 Gbps.

The DS4800 introduced 4Gbps functionality; There are several switches and directors which operate at this speed. At the time of writing, 4 Gbps trunking was not available.

Note that not all ports are capable of auto-negotiation. For example the ports on the DS4400 and EXP 700 must be manually set for either 1 Gbps or 2 Gbps.

2.2.3 Cable management and labeling

Cable management and labeling for solutions using racks, n-node clustering, and Fibre Channel are increasingly important in open systems solutions. Cable management and labeling needs have expanded from the traditional labeling of network connections to management and labeling of most cable connections between your servers, disk subsystems, multiple network connections, and power and video subsystems. Examples of solutions include Fibre Channel configurations, n-node cluster solutions, multiple unique solutions located in the same rack or across multiple racks, and solutions where components might not be physically located in the same room, building, or site.

Why more detailed cable management is required

The necessity for detailed cable management and labeling is due to the complexity of today's configurations, potential distances between solution components, and the increased number of cable connections required to attach additional value-add computer components. Benefits from more detailed cable management and labeling include ease of installation, ongoing solutions/systems management, and increased serviceability.

Solutions installation and ongoing management are easier to achieve when your solution is correctly and consistently labeled. Labeling helps make it possible to know what system you are installing or managing, for example, when it is necessary to access the CD-ROM of a particular system, and you are working from a centralized management console. It is also helpful to be able to visualize where each server is when completing custom configuration tasks such as node naming and assigning IP addresses.

Cable management and labeling improve service and support by reducing problem determination time, ensuring that the correct cable is disconnected when necessary. Labels will assist in quickly identifying which cable needs to be removed when connected to a device such as a hub that might have multiple connections of the same cable type. Labels also help identify which cable to remove from a component. This is especially important when a cable connects two components that are not in the same rack, room, or even the same site.

Cable planning

Successful cable management planning includes three basic activities: site planning (before your solution is installed), cable routing, and cable labeling.

Site planning

Adequate site planning completed before your solution is installed will result in a reduced chance of installation problems. Significant attributes covered by site planning are location specifications, electrical considerations, raised/non-raised floor determinations, and determination of cable lengths. Consult the documentation of your solution for special site planning considerations. IBM Netfinity® Racks document site planning information in *IBM Netfinity Rack Planning and Installation Guide*, part number 24L8055.

Cable routing

With effective cable routing, you can keep your solution's cables organized, reduce the risk of damaging cables, and allow for affective service and support. To assist with cable routing, IBM recommends the following guidelines:

- ▶ When installing cables to devices mounted on sliding rails:
 - Run the cables neatly along equipment cable-management arms and tie the cables to the arms. (Obtain the cable ties locally.)

Note: Do not use cable-management arms for fiber cables.

- Take particular care when attaching fiber optic cables to the rack. Refer to the instructions included with your fiber optic cables for guidance on minimum radius, handling, and care of fiber optic cables.
- Run the cables neatly along the rack rear corner posts.
- Use cable ties to secure the cables to the corner posts.
- Make sure the cables cannot be pinched or cut by the rack rear door.
- Run internal cables that connect devices in adjoining racks through the open rack sides.
- Run external cables through the open rack bottom.
- Leave enough slack so that the device can be fully extended without putting a strain on the cables.
- Tie the cables so that the device can be retracted without pinching or cutting the cables.

- ▶ To avoid damage to your fiber-optic cables, follow these guidelines:
 - Use great care when utilizing cable management arms.
 - When attaching to a device on slides, leave enough slack in the cable so that it does not bend to a radius smaller than 76 mm (3 in.) when extended or become pinched when retracted.
 - Route the cable away from places where it can be snagged by other devices in the rack.
 - Do not overtighten the cable straps or bend the cables to a radius smaller than 76 mm (3 in.).
 - Do not put excess weight on the cable at the connection point and be sure that it is well supported. For instance, a cable that goes from the top of the rack to the bottom *must* have some method of support other than the strain relief boots built into the cable.

Additional information for routing cables with IBM Netfinity Rack products can be found in IBM *Netfinity Rack Planning and Installation Guide*, part number 24L8055. This publication includes pictures providing more details about the recommended cable routing.

Cable labeling

When labeling your solution, follow these tips:

- ▶ As you install cables in the rack, label each cable with appropriate identification.
- ▶ Remember to attach labels to any cables you replace.
- ▶ Document deviations from the label scheme you use. Keep a copy with your Change Control Log book.

Whether using a simple or complex scheme, the label should always implement a format including these attributes:

- ▶ The function — to help identify the purpose of the cable
- ▶ Location information should be broad to specific (for example, the site/building to a specific port on a server or hub).

Other cabling mistakes

Some of the most common mistakes include these:

- ▶ Leaving cables hanging from connections with no support.
- ▶ Not using dust caps.
- ▶ Not keeping connectors clean. (Some cable manufacturers require the use of lint-free alcohol wipes in order to maintain the cable warranty.)
- ▶ Leaving cables on the floor where people might kick or trip over them.
- ▶ Not removing old cables when they are no longer needed, nor planned for future use.

2.2.4 Fibre Channel adapters

We now review topics related to Fibre Channel adapters:

- ▶ Placement on the host system bus
- ▶ Distributing the load among several adapters
- ▶ Queue depth
- ▶ Driver Selection

Host system bus

Today, there is a choice of high-speed adapters for connecting disk drives. Fast adapters can provide better performance. The HBA should be placed in the fastest supported slot available.

Important: Do not place all the high-speed Host Bus Adapters (HBAs) on a single system bus. Otherwise, the computer bus becomes the performance bottleneck.

We recommend that you distribute high-speed adapters across several busses. When you use PCI adapters, make sure you first review your system specifications. Some systems include a PCI adapter placement guide.

The number of adapters you can install depends on the number of PCI slots available on your server, but also on what traffic volume you expect on your SAN. The rationale behind multiple adapters is either redundancy (failover) or load sharing.

Failover

When multiple adapters are installed on the host system and used with a multipath driver, the multipath driver checks to see if all the available paths to the storage server are still functioning. In the event of an HBA or cabling failure, the path is changed to the other HBA, and the host continues to function without loss of data or functionality.

In general, all operating systems support two paths to the DS4000 Storage Server. Microsoft Windows 2000 and Windows 2003 and Linux support up to four paths to the storage controller. AIX® can also support four paths to the controller, provided that there are two partitions accessed within the DS4000 subsystem. You can configure up to two HBAs per partition and up to two partitions per DS4000 storage server.

Load balancing

Load balancing or load sharing means distributing I/O requests from the hosts between multiple adapters. This can be done by assigning LUNs to both the DS4000 controllers A and B alternatively (see also 2.3.2, “Logical drives and controller ownership” on page 45).

Figure 2-5 shows the principle for a load-sharing setup (Microsoft Windows environment). Microsoft Windows does a kind of forced load sharing. A multipath driver such as IBM Redundant Disk Array Controller (RDAC) checks all available paths to the controller. In Figure 2-5, that would be four paths (blue zone). RDAC now forces the data down all paths in a *round-robin* scheme. This means that it does not really check for the workload on a single path, but moves the data down in a *rotational manner* (round-robin).

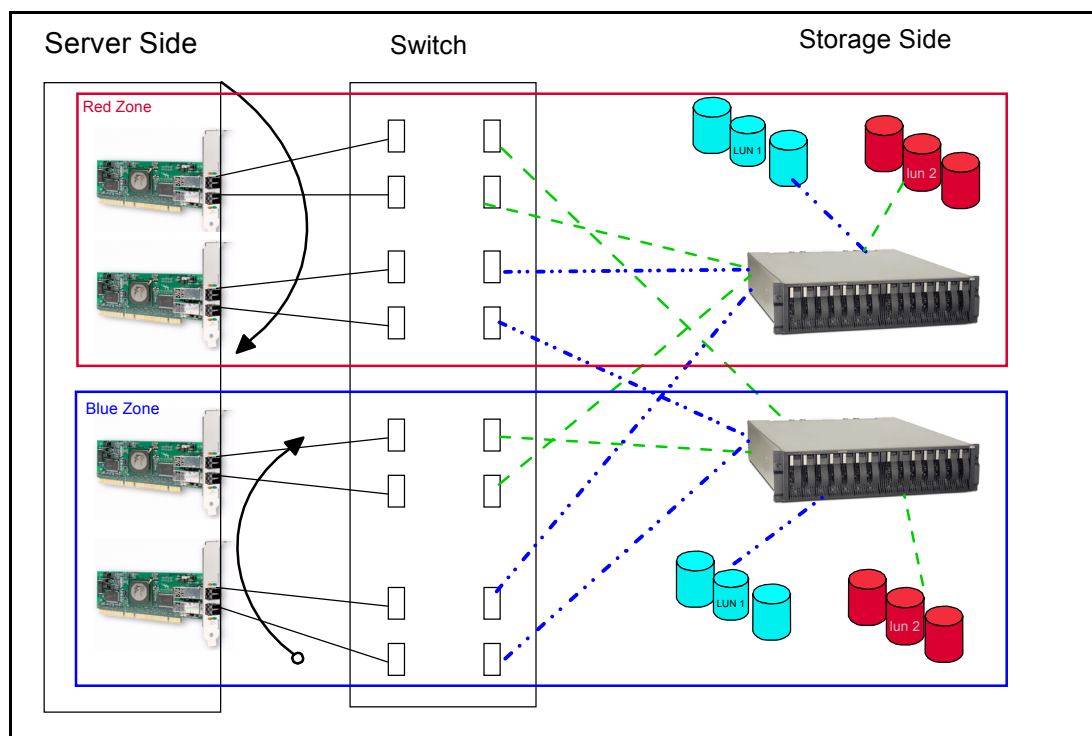


Figure 2-5 Load sharing approach for multiple HBAs

The RDAC drivers for Windows and Linux support round-robin load balancing.

Note: In a cluster environment, you need a single path to each of the controllers (A and B) of the DS4000. However, if the cluster software and host application can do persistent reservations, you can keep multiple paths and the multipath driver will route the I/O request using the appropriate path to the reserved logical drive.

In a single server environment, AIX is the other OS that allows load sharing (also called *load balancing*). However, the best practice is not to use load balancing in AIX, as it can have performance issues and cause disk thrashing.

Best practice: Do not enable load balancing for AIX.

Queue depth

The queue depth is the maximum number of commands that can be queued on the system at the same time.

The DS4000 controller firmware version 05.30.xx.xx or earlier, the queue depth is 512; for the DS4000 controller firmware Versions 06.1x.xx.xx or 05.4x.xx.xx, the queue depth is 2048. This represents 1024 per controller.

The formula for the correct queue depth on the host HBA for this level of firmware code is:
 $2048 / (\text{number of hosts} * \text{LUNs per host})$

For example, a system with four hosts, each with 32 LUNs, would have a maximum queue depth of 16: $2048 / (4 * 32) = 16$.

This setting should be set on each host bus adapter. See also 4.3.1, “Host based settings” on page 136.

For Qlogic based HBAs the queue depth is known as *execution throttle*, this can be set with either Qlogic SANSurfer or in the BIOS of the Qlogic-based HBA by pressing Ctl+Q during the boot process.

HBA driver selection

Microsoft Windows operating systems use two different types of storage interface drivers, in conjunction with vendor-written, adapter-specific drivers (miniport drivers). These are the SCSIport driver and the Storport driver.

Until recently, the HBA drivers developed by the different vendors have relied on the Microsoft SCSIport driver. The SCSIport driver, however, was designed to work optimally with the parallel SCSI interconnects used with direct attached storage, and it cannot meet the high performance standards of Fibre Channel SAN configurations and RAID storage systems, such as the DS4000.

Some of these limitations are listed below:

- ▶ Adapter I/O limits

SCSIport can support a maximum of 254 outstanding I/O requests per adapter, and thus for all the devices attached to that adapter. This is acceptable in a SCSI environment knowing that the SCSI bus will only support a maximum of 15 attached devices. In a Fibre Channel environment, and using an FC-AL topology as we have for the DS4000, the adapter can potentially support up to 126 devices, and the SCSIport limitation of 254 I/O becomes a real constraint.

- ▶ Sequential I/O processing

SCSIport cannot send an I/O request to the HBA and handle an interrupt from the storage system at the same time. This inability to handle multiple I/O requests in parallel specifically inhibits the performance of multiprocessor systems. Although up to 64 processors may be available for I/O processing, SCSIport cannot exploit the parallel processing capabilities.

- ▶ Error handling and bus resets

If an I/O operation encounters a command time out, SCSIport instructs the miniport driver to perform a bus reset. This is highly disruptive in SAN environments where many of the attached devices can be impacted. Tape backup operations can fail due to a bus reset being issued for an I/O time out on a disk device.

- ▶ Increased miniport load

The SCSIport driver was designed to do the majority of the work necessary to translate each I/O request packet into corresponding SCSI requests. However, when some non-SCSI adapters are used with SCSIport, additional translation to non-SCSI protocols is required. These additional steps must be performed by the miniport driver, resulting in a degradation of system performance.

► I/O Queue limitations

SCSIport queues I/O requests to the HBA to which the devices are connected. It does not, however, provide a mechanism for HBA miniport drivers to control how I/O is queued to their devices. Such control was not necessary with direct attached storage. In SAN environments where devices are added and removed from the network fairly regularly, I/O queues must be paused and resumed without accumulation of errors.

To overcome these limitations, Microsoft developed a new Storport device driver to supplement SCSIport on Windows Server® 2003 and beyond. Storport is a port driver using FC (SCSI-3) protocol optimization. It delivers higher I/O throughput performance, enhanced manageability, and improved miniport interface. Together, these changes help hardware vendors realize their high-performance interconnect goals.

Note: The Storport driver does not replace the multipath driver. The multipath driver communicates with the Storport driver to present the LUN to the host system.

All vendors are encouraged to use Storport where possible, rather than the SCSIport driver. Certain restrictions apply, however. Storport cannot be used with adapters or devices that do not support Plug and Play.

Best practice: In a Windows Server 2003 environment, the Storport Microsoft port driver is recommended for use with hardware RAID storage arrays and high-performance Fibre Channel interconnects.

With the latest StorPort driver for Windows 2003 (9.1.2.17), to ensure the optimal operating conditions when using the Storport Fibre Channel HBA device driver, Windows Server 2003 service pack 1 must be installed along with the Microsoft Windows Server 2003 Storport hot-fix KB916048. These updates are available for Windows Server 2003 32 bit and 64 bit, to download at the Microsoft support Web site:

<http://support.microsoft.com>

For the latest StorPort driver there are also minimum firmware levels on the DS400 Storage Server. The Fibre Channel host bus adapter STORport miniport device driver is supported with controller firmware Versions 6.12.27.xx or later only.

For compatibility information, always refer to the current DS4000 interoperability matrix or the readme file for your HBA driver.

2.2.5 Multipath driver selection

IBM offers different multipath drivers that you can use with your DS4000 Storage Server. Only one of these drivers is required. Each driver offers multipath support, I/O load balancing, and automatic path failover.

The multipath driver is a proxy for the real, physical-level HBA drivers. Each multipath driver hides from the application the fact that there are redundant connections, by creating a virtual device. The application uses this virtual device, and the multipath driver will connect the application to the correct physical path.

When you create a logical drive, you assign one of the two active controllers to own the logical drive (called *preferred controller ownership*, as described in 2.3.2, “Logical drives and controller ownership” on page 45) and to control the I/O between the logical drive and the application host along the I/O path. The preferred controller normally receives the I/O

requests from the logical drive. If a problem along the data path (such as a component failure) causes an I/O to fail, the multipath driver issues the I/O to the alternate controller.

A multipath device driver is not required when the host operating system has its own mechanism to handle multiple I/O paths.

Veritas Logical Drive Manager with Dynamic Multi-Pathing (DMP) is another example of a multipath driver. This multipath driver requires Array Support Library (ASL) software, which provides information to the Veritas Logical Drive manager for setting up the path associations for the driver.

Note: IBM now offers an SDD driver for HPUX. This SDD driver is only supported through RPQ.

For Windows hosts there is currently a choice of two multipath drivers:

- ▶ Redundant Disk Array Controller (RDAC)
- ▶ Multipath Input/Output (MPIO) Device Specific Module (MPIO)

RDAC

Prior to Storage Manager Version 9.19, RDAC was the only multipath driver available for Windows environments.

The RDAC driver is supported in the Windows 2000 and Server 2003 operating system environment. In addition, the RDAC driver is also supported with Fibre Channel host bus adapter device drivers based on either SCSIport or STORport miniport device driver models.

The current RDAC driver implementation performs the following tasks:

- ▶ Detects and claims the physical devices (LUNs) presented from the DS4000 storage subsystems (*hides* them) based on vendor/product ID strings and manages all of the paths to the physical devices
- ▶ Presents a single instance of each LUN to the rest of the Windows operating system components
- ▶ Manages all of the Plug and Play interactions
- ▶ Provides I/O routing information
- ▶ Identifies conditions requiring a request to be retried, failed, or failed over
- ▶ Automatically fails over the LUNs to their alternate controller when detecting problems in sending I/Os to the LUNs in their preferred controller and fails back the LUNs to their preferred controller when detecting the problems in the preferred path fixed
- ▶ Handles miscellaneous functions such as persistent reservation translation
- ▶ Uses round robin (load distribution or load balancing) model

RDAC is implemented between the HBA driver and the operating system disk driver, operating as a low-level filter driver. It has the following advantages:

- ▶ It is much more transparent to the OS and applications.
- ▶ I/O controls at the HBA driver level are not as tightly coupled to the OS as those at the disk driver level. Consequently, it is easier to implement I/O control functionality in the MPP-based RDAC driver for routing functions.
- ▶ As the driver is positioned at the HBA level, it has access to the SCSI command and sense data interface of the HBA driver and therefore can make more informed decisions about what to do in case of path failures.

MPIO

This multipath driver is included in the Storage Manager software package for Windows Version 9.19. MPIO is a Driver Development Kit (DDK) from Microsoft for developing code that manages multipath devices. It contains a core set of binary drivers, which are installed with the DS4000 Device Specific Module (DSM) to provide a transparent system architecture that relies on Microsoft Plug and Play to provide LUN multipath functionality while maintaining compatibility with existing Microsoft Windows device driver stacks.

The MPIO driver performs the following tasks:

- ▶ Detects and claims the physical disk devices presented by the DS4000 storage subsystems based on vendor/product ID strings and manages the logical paths to the physical devices.
- ▶ Presents a single instance of each LUN to the rest of the Windows operating system.
- ▶ Provides an optional interface via WMI for use by user-mode applications.
- ▶ Relies on the vendor's (IBM) customized Device Specific Module (DSM) for the information about the behavior of storage subsystem devices on the following:
 - I/O routing information
 - Conditions requiring a request to be retried, failed, failed over, or failed back (for example, vendor-unique errors)
 - Handles miscellaneous functions such as release/reservation commands
- ▶ Multiple Device Specific Modules (DSMs) for different disk storage subsystems can be installed in the same host server.

The MPIO driver is currently supported only with the following:

- ▶ Controller firmware Versions 6.19.xx.xx and later.
- ▶ Fibre Channel host bus adapter device drivers based on MS STORport miniport device driver models. Microsoft Windows Server 2003 with SP1 or later and with STORport hot-fix KB916048 (an updated STORport storage driver Version 5.2.3790.2723).

Coexistence of RDAC and MPIO/DSM in the same host is not supported. RDAC and MPIO/DSM drivers handle logical drives (LUNs) in fail conditions similarly because the DSM module that has code to handle these conditions are ported from RDAC. However, the MPIO/DSM driver will be the required Microsoft multipath driver for future Microsoft Windows operating systems.

See also 2.2.6, “The function of ADT” on page 29, for more discussion on multipathing and failover considerations.

2.2.6 The function of ADT

In a DS4000 Storage Server equipped with two controllers, you can provide redundant I/O paths with the host systems. There are two different components that provide this redundancy: a multipath driver and Auto Logical Drive Transfer.

Auto-Logical Drive Transfer feature (ADT)

AVT is a built-in feature of controller firmware that allows logical drive-level failover rather than controller-level failover (as is the case with RDAC). AVT is also referred to as the Auto-Logical Drive transfer feature.

Note: AVT is not a failover driver. AVT provides storage systems with the flexibility to work with some third-party failover software.

► AVT-disabled failover

The multi-path software will send a SCSI Mode Select command to cause a change in volume ownership before using the alternate path. All logical drives on the preferred controller are transferred to the alternate controller. This is the configuration setting for Microsoft Windows, IBM AIX, and Sun Solaris and Linux (when using the RDAC driver and non-failover Fibre Channel HBA driver) systems. When ADT is disabled, the I/O data path is still protected as long as you use a multi-path driver. After the I/O data path problem is corrected, the preferred controller does not automatically reestablish ownership

of the logical drive. You must open a storage management window, select **Redistribute Logical Drives** from the Advanced menu, and perform the Redistribute Logical Drives task. Figure 2-6 shows the AVT-disabled failover mode phases.

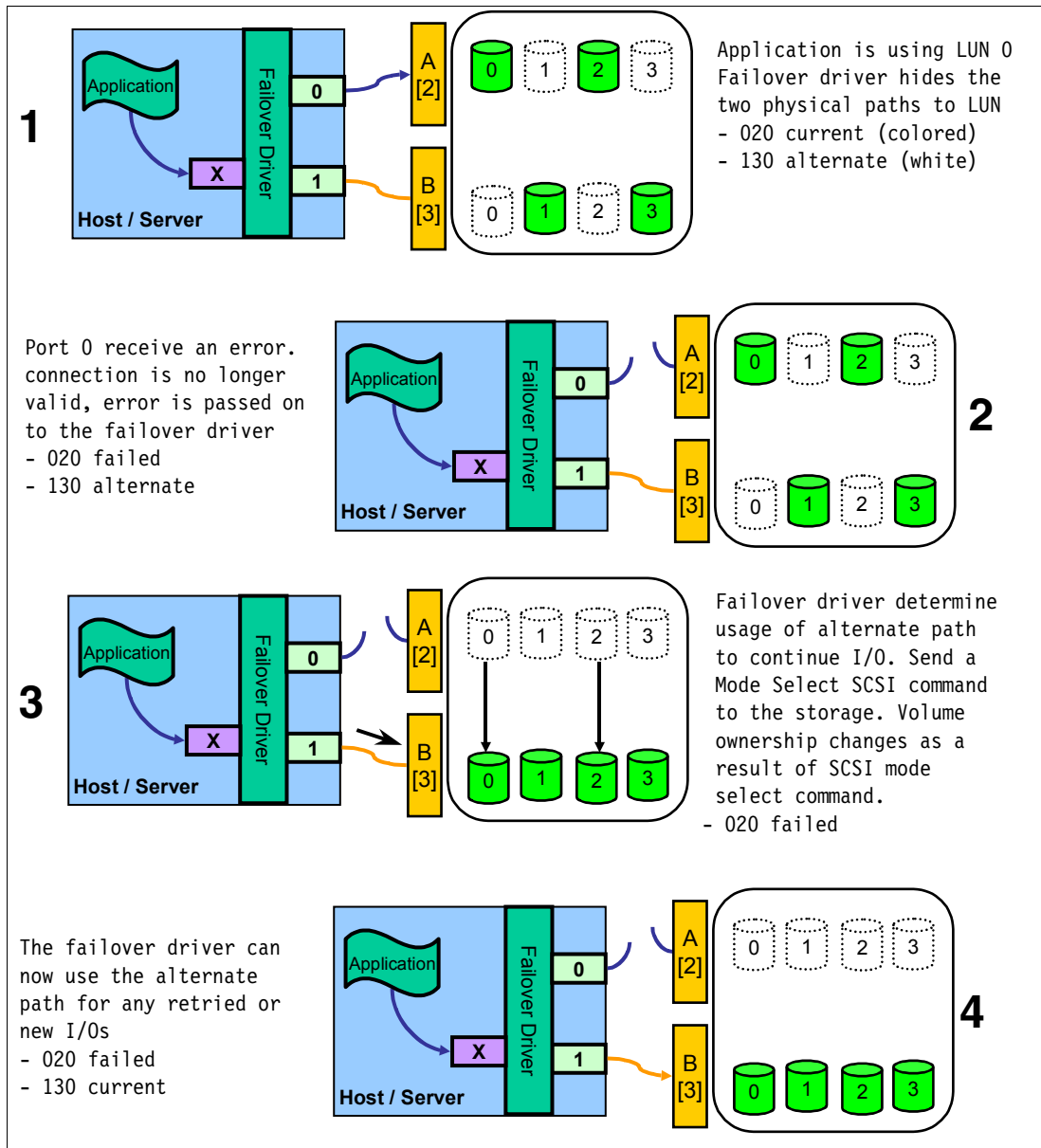


Figure 2-6 AVT-disabled mode path failover

Note: In AVT-disabled mode, you are required to issue a redistribution command manually to get the LUNs balanced across the controllers.

► AVT-enabled failover

The multi-path driver starts using the alternate path by sending the I/O down the path it chooses and lets the AVT react. This is the normal configuration setting for Novell NetWare, Linux (when using FC HBA failover driver instead of RDAC), and Hewlett Packard HP-UX systems. After the I/O data path problem is corrected, the preferred controller automatically reestablishes ownership of the logical drive as soon as the multipath driver detects that the path is normal again. Figure 2-7 shows the phases of failover in AVT-enabled case.

Note: In AVT mode, RDAC automatically redistributes the LUNs to their preferred path after the failed path is again operational.

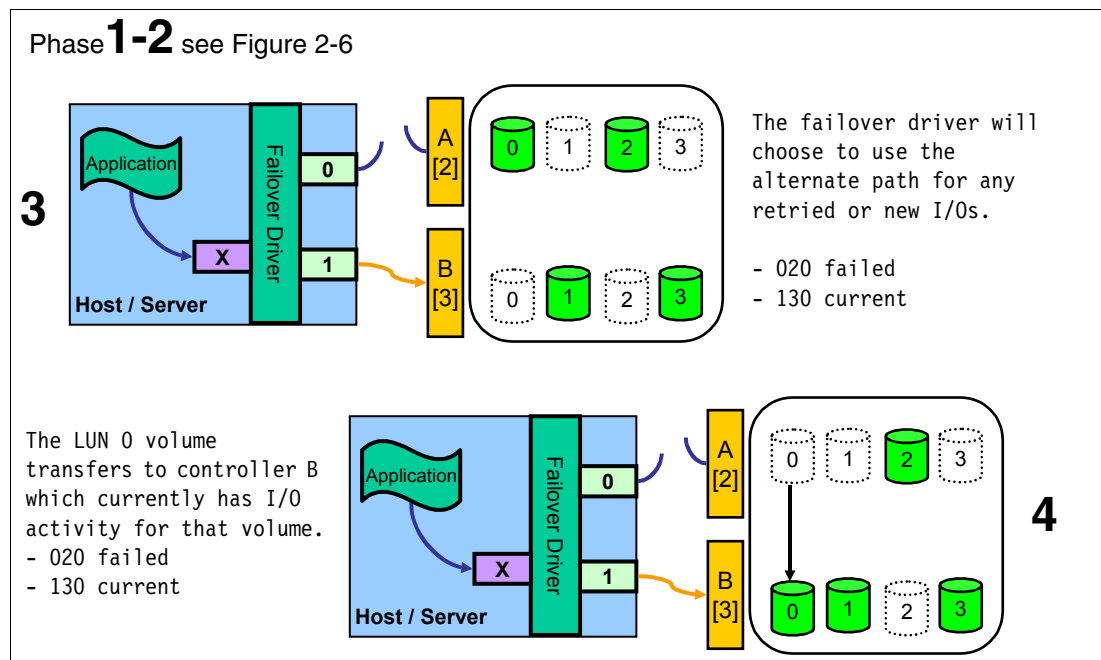


Figure 2-7 AVT-enabled mode path failover

2.2.7 Disk expansion enclosures

The DS4000 Series offers four expansion enclosures: EXP100, EXP420, EXP710, and EXP810. When planning for which enclosure to use with your DS4000, you must look at the applications and data that you will be using on the DS4000.

The EXP420 enclosure is only available with the DS4200. The EXP810 enclosure can accommodate either Fibre Channel or SATA II drives. At the time of writing, only one type of drive per enclosure was permitted. This may change in future releases of firmware. Both the EXP420 and EXP810 are 4 Gbps capable enclosures, offering high performance and value. The EXP710 and EXP100 enclosures are discontinued.

Enclosure IDs

It is very important to correctly set the tray (enclosure) IDs. They are used to differentiate multiple EXP Enclosures that are connected to the same DS4000 Storage Server. Each EXP Enclosure must use a unique value. The DS4000 Storage Manager uses the tray IDs to identify each EXP Enclosure.

For the EXP100 and EXP710, the Fibre Channel Fabric Address (EXP100/EXP710) for each disk drive is automatically set according to:

- ▶ The EXP100/EXP710 bay where the disk drive is inserted
- ▶ Tray ID setting

Two switches are available to set the tray ID:

- ▶ A switch for tens (x10)
- ▶ A switch for ones (x1)

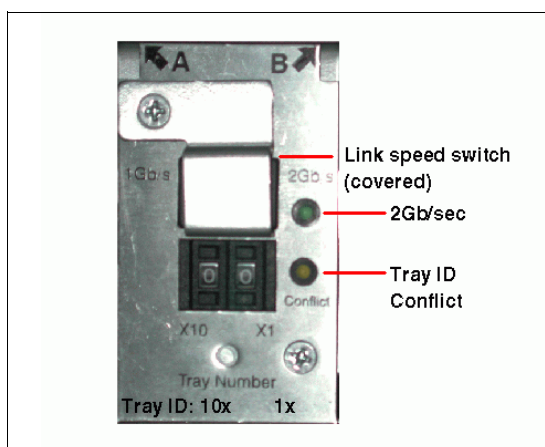


Figure 2-8 Enclosure ID for EXP100/EXP710

For the EXP810 and EXP420, the enclosure ID is indicated by a dual seven-segment LED located on the back of each ESM next to the other ESM indicator lights. The storage server firmware automatically sets the enclosure ID number. If needed, you can change the enclosure ID setting through the DS4000 storage management software only. There are no switches on the EXP810 chassis to manually set the enclosure ID. Both ESM enclosure ID numbers will be identical under normal operating conditions.

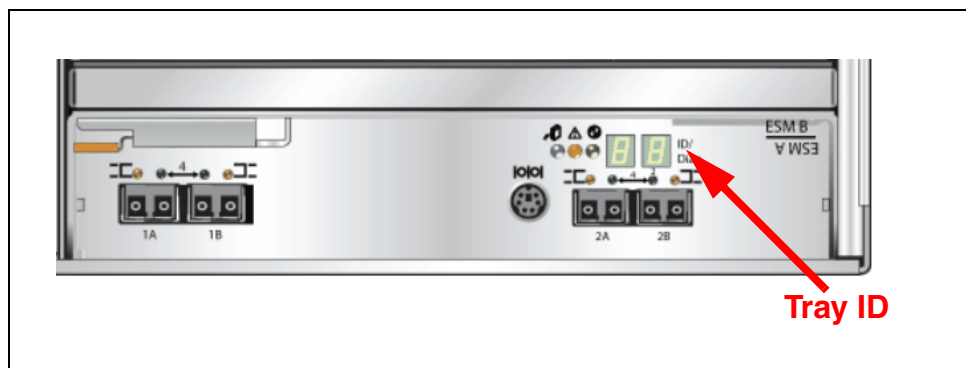


Figure 2-9 Enclosure ID LEDs for EXP810/EXP420

Enclosure guidelines

The base controller unit and each expansion enclosure has an ID number associated with it. The ID will allow each enclosure to be identified properly to the base controller unit.

Since the base and all enclosures are connected by Fibre Channel loop, it is necessary for each ID address to be distinct and unique for I/O to flow properly. An ID address is composed of two digits: a tens digit (x10) and a ones digit (x1). Enclosure IDs are typically between the values of x00 and x77.

We recommend that all enclosures in the same loop pair share the same tens digit (x10) value, but have unique ones digits (x1). This will allow you to keep better track of which enclosures belong to which loop pairs and to avoid potential problems that have been identified with the use of duplicate single digits on the same loop. See Table 2-2.

Table 2-2 Recommended enclosure ID scheme with the DS4800

	Enclosure behind first controller drive port	Enclosure behind second controller drive port	Enclosure behind third controller drive port	Enclosure behind fourth controller drive port
Enclosure 1	x00	x04	x10	x14
Enclosure 2	x01	x05	x11	x15
Enclosure 3	x02	x06	x12	x16
Enclosure 4	x03	x07	x13	x17

Note that EXP810s are automatically assigned enclosure IDs that match the values shown in Table 2-2.

Because the storage system follows a specific address assignment scheme for the drives, you should observe a few guidelines when assigning enclosure IDs to expansion units. Failure to adhere to these guidelines can cause issues with I/O error recovery, and make more difficult the troubleshooting of certain drive communication issues:

- ▶ Do not use enclosure IDs from 00 through 09.
- ▶ Ensure that the least significant digit (units) of the enclosure ID is unique within a drive loop pair. The most significant digit (tens) should be unique for each loop pair of a given storage subsystem. For instance, a loop of purely EXP710s should be numbered 10–17 for the first loop pair and 20–27 for the second loop pair.
- ▶ Whenever possible, maintain an even number of expansion units on the DS4000 server and configure for enclosure loss protection. Add EXP expansion units in pairs and (ideally) for the DS4800 in groups of four.
- ▶ Add drives into an expansion enclosure in pairs.
- ▶ Avoid fully populating all expansion units.

Once the DS4000 is fully populated with the maximum number of enclosures and drives, there is simply no more expansion possible. As the requirement increases for the numbers of drives, perhaps it is time to look at an additional DS4000 storage system or storage virtualization or even both.

Best practice: Plan on using no more than 80% of the maximum possible capacity for the DS4000 system.

2.2.8 Selecting drives

The speed and the type of the drives used will impact the performance. Typically the faster the drive, the higher the performance. This increase in performance comes at a cost, the faster drives are typically a higher cost than the lower performance drives. FC drives outperform the SATA drives.

The DS4000 supports both Fibre Channel and SATA drives. To use SATA drives in your DS4000, either an EXP100, EXP420, or EXP810 (using only SATA Drives) expansion cabinets may be required (not necessarily required for the DS4200 or DS4100, as they may incorporate SATA drives directly). If Fibre Channel drives are used as well, the FC/SATA Intermix Premium Feature is required and is an additional cost.

The following types of FC drives are currently available:

- ▶ 2 Gbps FC: 15 Krpm, 146 GB/73 GB/36 GB/ 18 GB
- ▶ 2 Gbps FC: 10 Krpm, 300 GB/146 GB/73 GB
- ▶ 4 Gbps FC: 15 Krpm, 146 GB/73 GB/36 GB

The following SATA drives are available:

- ▶ 2 Gbps SATA: 7.2K rpm, 250 GB/400 GB (for use in EXP100)
- ▶ 4 Gbps SATA: 7.2K rpm, 500 GB E-DDM (for use in the EXP810)
- ▶ 4 Gbps SATA: 7.2K rpm, 500 GB EV-DDM (for use in the EXP420 and DS4200)

At the time of writing this book, a SATA drive with a capacity of 400 GB at 7200 RPM was just announced and replaces the former 250 GB drive.

Note: The usable disk space is less than overall disk capacity. Please note the usable capacity amounts for storage capacity calculation issues:

- ▶ 18.2 GB formatted capacity is equal to 16.450 GB usable capacity.
- ▶ 36.4 GB formatted capacity is equal to 33.400 GB usable capacity.
- ▶ 73.4 GB formatted capacity is equal to 67.860 GB usable capacity.
- ▶ 146.8 GB formatted capacity is equal to 136.219 GB usable capacity.
- ▶ 300 GB formatted capacity is equal to 278.897 GB usable capacity.
- ▶ 500 GB formatted capacity is equal to 465.161 GB usable capacity

The usable capacities are what the SMclient will report as storage that can be used by the hosts. We arrive at this number by the following steps:

1. Take the listed raw disk amount (listed in decimal, as the storage industry standard dictates) and divide by 1.073742 to get a raw binary capacity (1 decimal GB = 1,000,000,000 bytes; 1 binary GB = 1,073,741,824 bytes (2^{30} bytes)).
2. Subtract out the 512 MB DACstore region (the region that holds configuration information) after converting the DACstore to binary.

This will give you the usable binary capacity that can be utilized by hosts and is what the SMclient will report to you as usable capacity.

Table 2-3 compares the Fibre Channel 10K, 15K and SATA drives (single drive).

Best practice: Use the fastest drives available for best performance.

Table 2-3 Comparison between Fibre Channel and SATA

	Fibre Channel	SATA	SATA difference
Spin Speed	10K and 15K	7.2K	
Command Queuing	Yes 16 Max	No 1 Max	
Single Disk I/O Rate (# of 512 bytes IOPS) ^a	280 & 340	88	.31 and .25
Read Bandwidth (MBps)	69 & 76	60	.86 and .78
Write Bandwidth (MBps)	68 & 71	30	.44

a. Note that the IOPS and bandwidth figures are from disk manufacturer tests in ideal lab conditions. In practice you will see lower numbers, but the ratio between SATA and FC disks still applies.

The speed of the drive is the number of revolutions per minute (RPM). A 15K drive rotates 15,000 times per minute. With the higher speeds the drives tend to be denser, as a large diameter plate driving at such speeds is likely to wobble. With the faster speeds comes the ability to have greater throughput.

Seek time is the measure of how long it takes for the drive head to move to the correct sectors on the drive to either read or write data. It is measured in thousands of a second (milliseconds or ms). The faster the seek time, the quicker data can be read from or written to the drive. The average seek time reduces when the speed of the drive increases. Typically a 7.2K will have an average seek time of around 9 ms, a 10K drive will have an average seek time of around 5.5 ms, and a 15K drive will have an average seek time of around 3.5 ms.

Command queuing allows for multiple commands to be outstanding to the disk drive at the same time. The drives have a queue where outstanding commands can be dynamically rescheduled or re-ordered, along with the necessary tracking mechanisms for outstanding and completed portions of workload. The SATA disks do not have command queuing and the Fibre Channel disks currently have a command queue depth of 16.

Avoid using the SATA drives for high IOPS operations. SATA can, however, be used for streaming and archiving applications. These are both very good uses for SATA, where good throughput rates are required, but at a lower cost.

2.3 Planning your storage structure

It is important to configure a storage system in accordance to the needs of the user. An important question and primary concern for most users or storage administrators is how to configure the storage subsystem to achieve best for the best performance. There is no simple answer, no best guideline for storage performance optimization that is valid in every environment and for every particular situation. We have dedicated a chapter of this book (see Chapter 4, “DS4000 performance tuning” on page 133) to discuss and recommend how to configure or tune the various components and features of the DS4000 to achieve the best

performance upon different circumstances. You will find some preliminary (and less detailed) performance discussion in this section.

Also, in this section, we review other aspects of the system configuration that can help optimize the storage capacity and resilience of the system. In particular, we review and discuss the RAID levels, array size, array configuration, and enclosure loss protection.

Note: Topics introduced in this section are also discussed from a performance optimization perspective in Chapter 4, “DS4000 performance tuning” on page 133.

2.3.1 Arrays and RAID levels

An array is a set of drives that the system logically groups together to provide one or more logical drives to an application host or cluster.

When defining arrays, you often have to compromise among capacity, performance, and redundancy.

RAID levels

We go through the different RAID levels and explain why we would choose this particular setting in this particular situation, and then you can draw your own conclusions. See also Figure on page 155.

RAID-0: For performance, but generally not recommended

RAID-0 (refer to Figure 2-10) is also known as *data striping*. It is well-suited for program libraries requiring rapid loading of large tables, or more generally, applications requiring fast access to read-only data or fast writing. RAID-0 is only designed to increase performance. There is no redundancy, so any disk failures require reloading from backups. Select RAID level 0 for applications that would benefit from the increased performance capabilities of this RAID level. Never use this level for critical applications that require high availability.

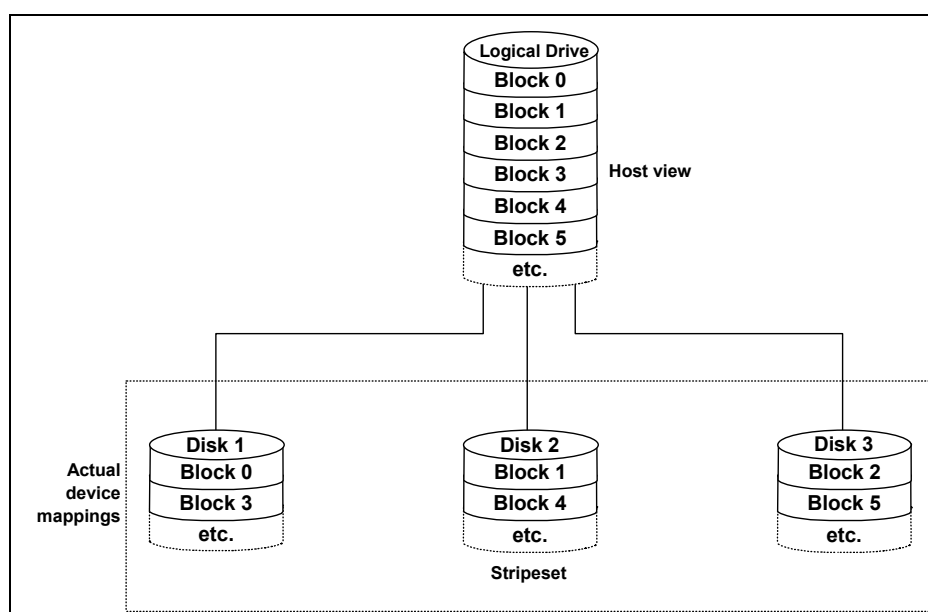


Figure 2-10 RAID 0

RAID-1: For availability/good read response time

RAID-1 (refer to Figure 2-11) is also known as *disk mirroring*. It is most suited to applications that require high data availability, good read response times, and where cost is a secondary issue. The response time for writes can be somewhat slower than for a single disk, depending on the write policy. The writes can either be executed in parallel for speed or serially for safety. Select RAID level 1 for applications with a high percentage of read operations and where the cost is not the major concern.

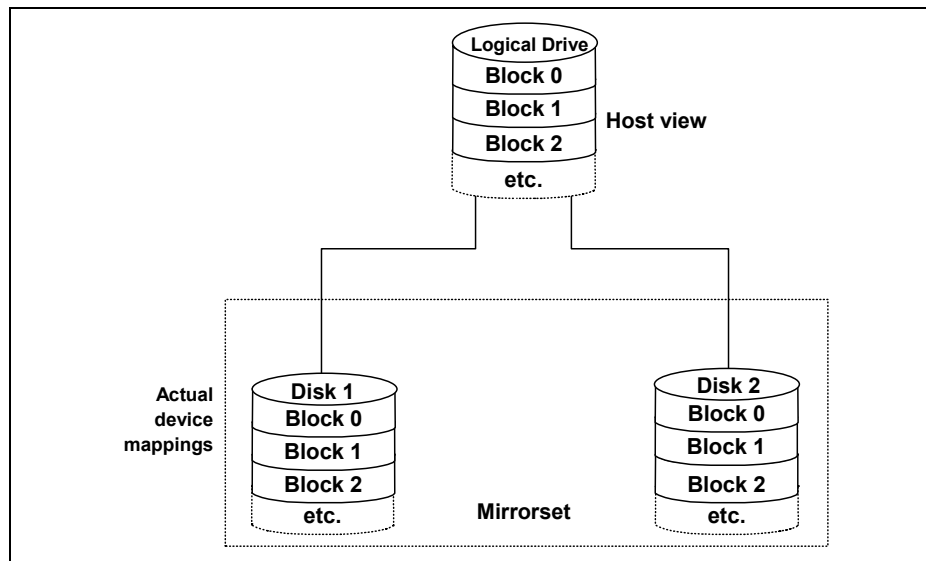


Figure 2-11 RAID 1

Because the data is mirrored, the capacity of the logical drive when assigned RAID level 1 is 50% of the array capacity.

Here are some recommendations when using RAID-1:

- ▶ Use RAID-1 for the disks that contain your operating system. It is a good choice, because the operating system can usually fit on one disk.
- ▶ Use RAID-1 for transaction logs. Typically, the database server transaction log can fit on one disk drive. In addition, the transaction log performs mostly sequential writes. Only rollback operations cause reads from the transaction logs. Therefore, we can achieve a high rate of performance by isolating the transaction log on its own RAID-1 array.
- ▶ Use write caching on RAID-1 arrays. Because a RAID-1 write will not complete until both writes have been done (two disks), performance of writes can be improved through the use of a write cache. When using a write cache, be sure it is battery-backed up.

Note: RAID 1 is actually implemented only as RAID 10 (described below) on the DS4000 products.

RAID-3: Sequential access to large files

RAID-3 is a parallel process array mechanism, where all drives in the array operate in unison. Similar to data striping, information to be written to disk is split into chunks (a fixed amount of data), and each chunk is written out to the same physical position on separate disks (in parallel). This architecture requires parity information to be written for each stripe of data.

Performance is very good for large amounts of data, but poor for small requests because every drive is always involved, and there can be no overlapped or independent operation. It is

well-suited for large data objects such as CAD/CAM or image files, or applications requiring sequential access to large data files. Select RAID-3 for applications that process large blocks of data. It provides redundancy without the high overhead incurred by mirroring in RAID-1.

RAID-5: High availability and fewer writes than reads

RAID level 5 (refer to Figure 2-12) stripes data and parity across all drives in the array. RAID level 5 offers both data protection and increased throughput. When you assign RAID-5 to an array, the capacity of the array is reduced by the capacity of one drive (for data-parity storage). RAID-5 gives you higher capacity than RAID-1, but RAID level 1 offers better performance.

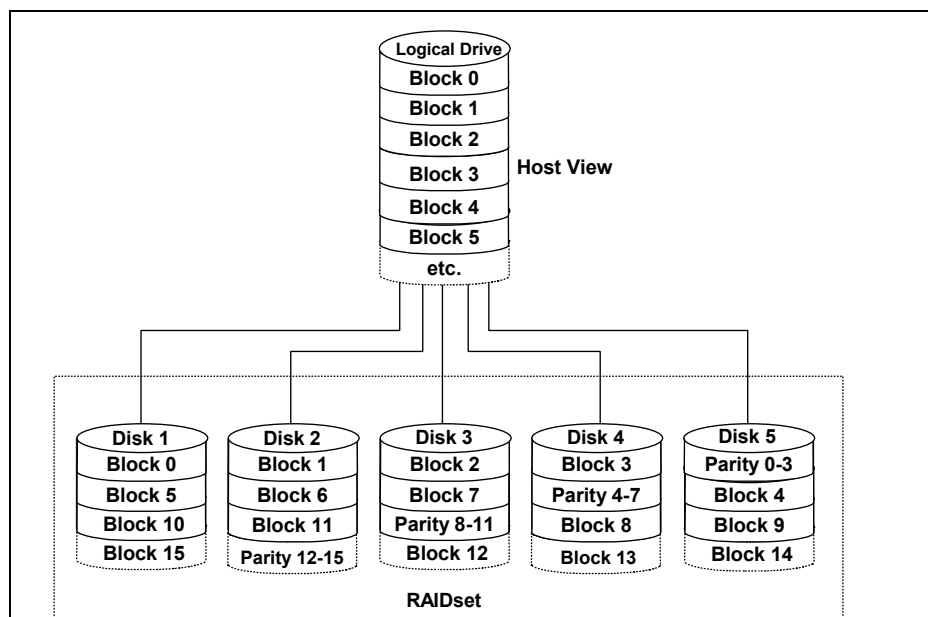


Figure 2-12 RAID 5

RAID-5 is best used in environments requiring high availability and fewer writes than reads.

RAID-5 is good for multi-user environments, such as database or file system storage, where typical I/O size is small, and there is a high proportion of read activity. Applications with a low read percentage (write-intensive) do not perform as well on RAID-5 logical drives because of the way a controller writes data and redundancy data to the drives in a RAID-5 array. If there is a low percentage of read activity relative to write activity, consider changing the RAID level of an array for faster performance.

Use write caching on RAID-5 arrays, because RAID-5 writes will not be completed until at least two reads and two writes have occurred. The response time of writes will be improved through the use of write cache (be sure it is battery-backed up). RAID-5 arrays with caching can give as good as performance as any other RAID level, and with some workloads, the striping effect gives better performance than RAID-1.

RAID-10: Higher performance than RAID-1

RAID-10 (refer to Figure 2-13), also known as RAID 1+0, implements block interleave data striping and mirroring. In RAID-10, data is striped across multiple disk drives, and then those drives are mirrored to another set of drives.

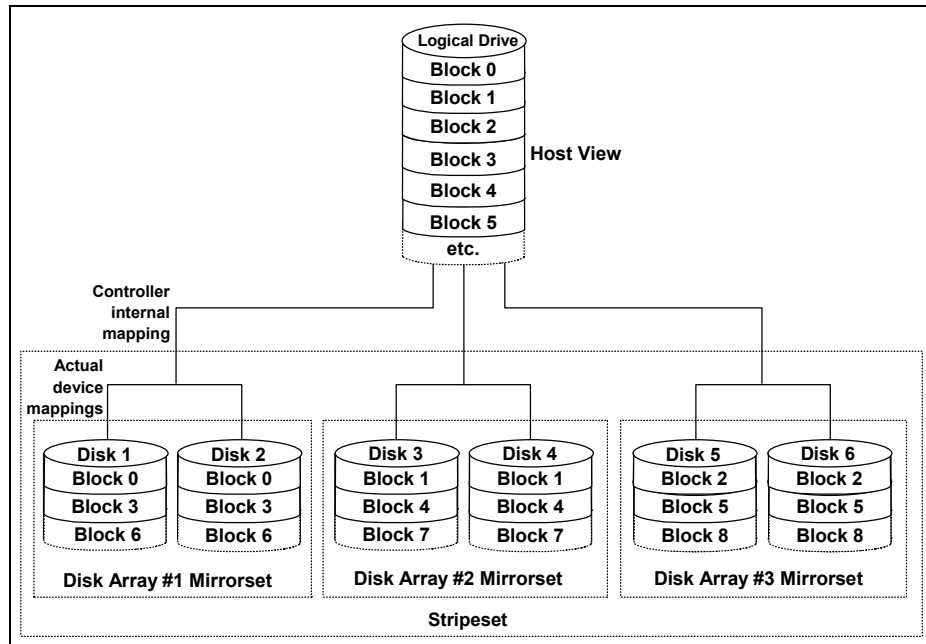


Figure 2-13 RAID 10

The performance of RAID-10 is approximately the same as RAID-0 for sequential I/Os. RAID-10 provides an enhanced feature for disk mirroring that stripes data and copies the data across all the drives of the array. The first stripe is the data stripe; the second stripe is the mirror (copy) of the first data stripe, but it is shifted over one drive. Because the data is mirrored, the capacity of the logical drive is 50% of the physical capacity of the hard disk drives in the array.

The recommendations for using RAID-10 are as follows:

- ▶ Use RAID-10 whenever the array experiences more than 10% writes. RAID-5 does not perform as well as RAID-10 with a large number of writes.
- ▶ Use RAID-10 when performance is critical. Use write caching on RAID-10. Because a RAID-10 write will not be completed until both writes have been done, write performance can be improved through the use of a write cache (be sure it is battery-backed up).

When comparing RAID-10 to RAID-5:

- ▶ RAID-10 writes a single block through two writes. RAID-5 requires two reads (read original data and parity) and two writes. Random writes are significantly faster on RAID-10.
- ▶ RAID-10 rebuilds take less time than RAID-5 rebuilds. If a real disk fails, RAID-10 rebuilds it by copying all the data on the mirrored disk to a spare. RAID-5 rebuilds a failed disk by merging the contents of the surviving disks in an array and writing the result to a spare.

RAID-10 is the best fault-tolerant solution in terms of protection and performance, but it comes at a cost. You must purchase twice the number of disks that are necessary with RAID-0.

The following note and Table 2-4 summarize this information.

Summary: Based on the respective level, RAID offers the following performance results:

- ▶ RAID-0 offers high performance, but does not provide any data redundancy.
- ▶ RAID-1 offers high performance for write-intensive applications.
- ▶ RAID-3 is good for large data transfers in applications, such as multimedia or medical imaging, that write and read large sequential chunks of data.
- ▶ RAID-5 is good for multi-user environments, such as database or file system storage, where the typical I/O size is small, and there is a high proportion of read activity.
- ▶ RAID-10 offers higher performance than RAID-1 and more reliability than RAID-5

Table 2-4 RAID levels comparison

RAID	Description	APP	Advantage	Disadvantage
0	Stripes data across multiple drives.	IOPS Mbps	Performance due to parallel operation of the access.	No redundancy. One drive fails, data is lost.
1	Disk's data is mirrored to another drive.	IOPS	Performance as multiple requests can be fulfilled simultaneously.	Storage costs are doubled.
10	Data is striped across multiple drives and mirrored to same number of disks.	IOPS	Performance as multiple requests can be fulfilled simultaneously. Most reliable RAID level on the DS4000	Storage costs are doubled.
3	Drives operated independently with data blocks distributed among all drives. Parity is written to a dedicated drive.	Mbps	High performance for large, sequentially accessed files (image, video, graphical).	Degraded performance with 8-9 I/O threads, random IOPS, smaller more numerous IOPS.
5	Drives operated independently with data and parity blocks distributed across all drives in the group.	IOPS Mbps	Good for reads, small IOPS, many concurrent IOPS and random I/Os.	Writes are particularly demanding.

RAID reliability considerations

At first glance both RAID-3 and RAID-5 would appear to provide excellent protection against drive failure. With today's high-reliability drives, it would appear unlikely that a second drive in an array would fail (causing data loss) before an initial failed drive could be replaced.

However, field experience has shown that when a RAID-3 or RAID-5 array fails, it is not usually due to two drives in the array experiencing complete failure. Instead, most failures are caused by one drive going bad, and a single block somewhere else in the array that cannot be read reliably.

This problem is exacerbated by using large arrays with RAID-5. This *stripe kill* can lead to data loss when the information to re-build the stripe is not available. The end effect of this issue will of course depend on the type of data and how sensitive it is to corruption. While

most storage subsystems (including the DS4000) have mechanisms in place to try to prevent this from happening, they cannot work 100% of the time.

Any selection of RAID type should take into account the cost of downtime. Simple math tells us that RAID-3 and RAID-5 are going to suffer from failures more often than RAID 10. (Exactly how often is subject to many variables and is beyond the scope of this book.) The money saved by economizing on drives can be easily overwhelmed by the business cost of a crucial application going down until it can be restored from backup.

Naturally, no data protection method is 100% reliable, and even if RAID were faultless, it would not protect your data from accidental corruption or deletion by program error or operator error. Therefore, all crucial data should be backed up by appropriate software, according to business needs.

Table 2-5 RAID level and performance

RAID levels	Data capacity ^a	Sequential I/O performance ^b		Random I/O performance ^b	
		Read	Write	Read	Write
Single disk	n	6	6	4	4
RAID-0	n	10	10	10	10
RAID-1	n/2	7	5	6	3
RAID-5	n-1	7	7 ^c	7	4
RAID-10	n/2	10	9	7	6

a. In the data capacity, n refers to the number of equally sized disks in the array.

b. 10 = best, 1 = worst. We should only compare values within each column. Comparisons between columns are not valid for this table.

c. With the write back setting enabled.

Array size

Maximum array size has an upper limit of 30 disks. For the DS4100, DS4300, and DS4500, the best performance for number of disks in an array is around 10 to 12 disks.

Important: The maximum size of a logical drive (LUN) is 2 TB.

Raw space means the total space available on your disk. Depending on your RAID level, the usable space will be between 50% for RAID-1 and (N - 1)* drive capacity, where N is the number of drives for RAID-5.

Tip: The first rule for the successful building of good performing storage solutions is to have enough physical space to create the arrays and logical drives as required.

DACstor

The DACstor is a reserved area on each disk of the DS4000 Storage Server. This reserved area contains information about drives and other information needed by the controller. The DACstor is approximately 512Mb in size on each disk. This size may grow as future enhancements are made to the firmware. It is always a good idea to leave some free space inside every array to cater for any future increase in the DACstor size.

Best practice: Always leave a small amount of free space on each array to allow for expansion of logical drives, changes in DACstor size, or for premium copy features.

Array configuration

Before you can start using the physical disk space, you must configure it. That is, you divide your (physical) disk drives into arrays and create one or more logical drives inside each array.

In simple configurations, you can use all of your drive capacity with just one array and create all of your logical drives in that unique array. However, this presents the following drawbacks:

- ▶ If you experience a (physical) drive failure, the rebuild process affects all logical drives, and the overall system performance goes down.
- ▶ Read/write operations to different logical drives are still being made to the same set of physical hard drives.
- ▶ There could be changes to future DACstor size.

The array configuration is crucial to performance. You must take into account all the logical drives inside the array, as all logical drives inside the array will impact on the same physical disks. If you have two logical drives inside an array and they both are high throughput, then there may be contention for access to the physical drives as large read or write requests are serviced. It is crucial to know the type of data that each logical drive is used for and try to balance the load so contention for the physical drives is minimized. Contention is impossible to eliminate unless the array only contains one logical drive.

Number of drives

The more physical drives you have per array, the shorter the access time for read and write I/O operations. See also “Number of disks per array” on page 156.

You can determine how many physical drives should be associated with a RAID controller by looking at disk transfer rates (rather than at the megabytes per second). For example, if a hard disk drive is capable of 75 nonsequential (random) I/Os per second, about 26 hard disk drives working together could, theoretically, produce 2000 nonsequential I/Os per second, or enough to hit the maximum I/O handling capacity of a single RAID controller. If the hard disk drive can sustain 150 sequential I/Os per second, it takes only about 13 hard disk drives working together to produce the same 2000 sequential I/Os per second and keep the RAID controller running at maximum throughput.

Tip: Having more physical disks for the same overall capacity gives you:

- ▶ **Performance:** By doubling the number of the physical drives, you can expect up to a 50% increase in throughput performance.
- ▶ **Flexibility:** Using more physical drives gives you more flexibility to build arrays and logical drives according to your needs.
- ▶ **Data capacity:** When using RAID-5 logical drives, more data space is available with smaller physical drives because less space (capacity of a drive) is used for parity.

Enclosure loss protection planning

Enclosure loss protection is a good way to make your system more resilient against hardware failures. Enclosure loss protection means that you spread your protection arrays across multiple enclosures rather than in one enclosure so that a failure of a single enclosure does not take a whole array offline.

By default, the automatic configuration is enabled. However, this is not the best practice as the method of creating arrays. Instead, use the manual method, as this allows for more configuration options to be available at creation time.

Best practice: Manual array configuration allows for greater control over the creation of arrays.

Figure 2-14 shows an example of the enclosure loss protection. If enclosure number 2 were to fail, the array with the enclosure loss protection would still function (in a degraded state), as the other drives are not affected by the failure.

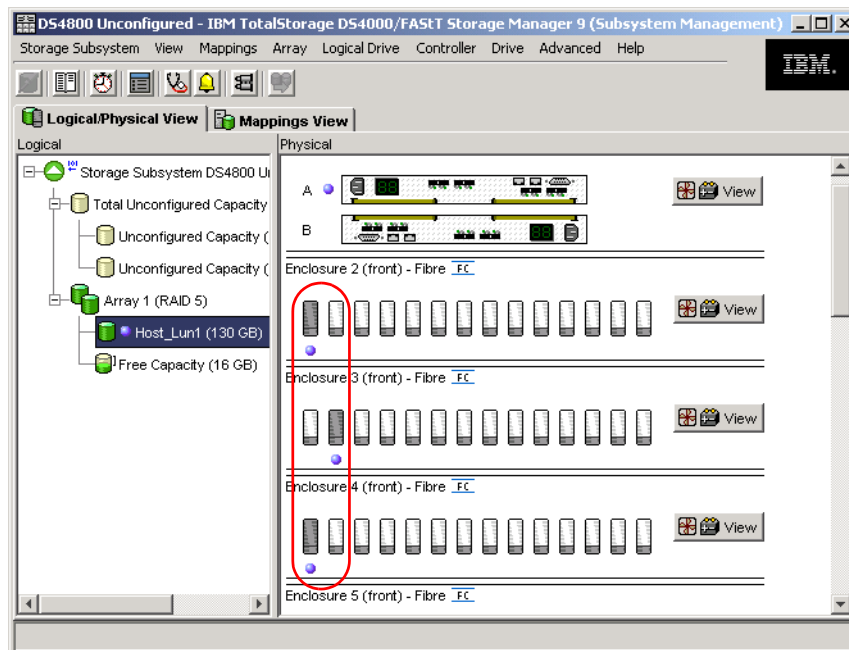


Figure 2-14 Enclosure loss protection

In the example in Figure 2-15, without enclosure loss protection, if enclosure number 2 were to fail, the entire array would become inaccessible.

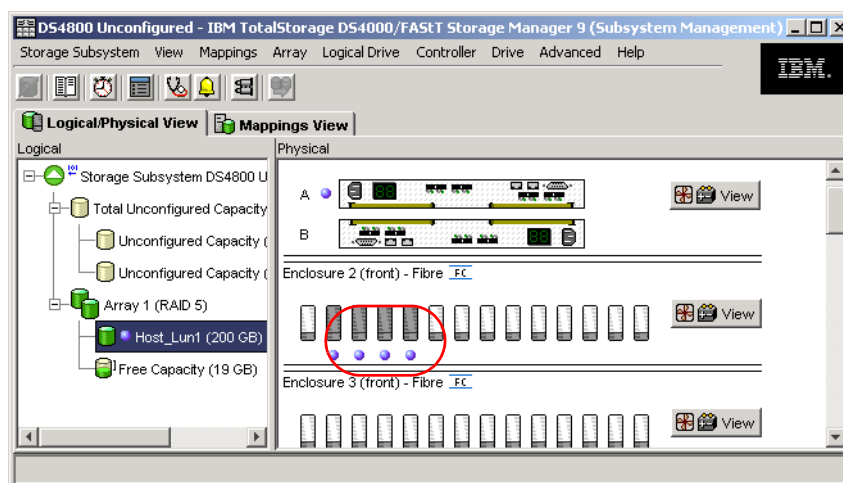


Figure 2-15 An array without enclosure loss protection

Best practice: Plan to use enclosure loss protection for your arrays.

2.3.2 Logical drives and controller ownership

Logical drives, sometimes simply referred to as volumes or LUNs (LUN stands for Logical Unit Number and represents the number a host uses to access the logical drive), are the logical segmentation of arrays. A logical drive is a logical structure you create on a storage subsystem for data storage. A logical drive is defined over a set of drives called an array and has a defined RAID level and capacity (see 2.2, “Physical components planning” on page 16). The drive boundaries of the array are hidden from the host computer.

IBM System Storage DS4000 Storage Server provides great flexibility in terms of configuring arrays and logical drives. However, when assigning logical volumes to the systems, it is very important to remember that the DS4000 Storage Server uses a preferred controller ownership approach for communicating with LUNs. This means that every LUN is owned by only one controller. It is, therefore, important at the system level to make sure that traffic is correctly balanced among controllers. This is a fundamental principle for a correct setting of the storage system. See Figure 2-16 on page 46.

Balancing traffic is unfortunately not always a trivial task. For example, if an application requires large disk space to be located and accessed in one chunk, it becomes harder to balance traffic by spreading the smaller volumes among controllers.

In addition, typically, the load across controllers and logical drives is constantly changing. The logical drives and data accessed at any given time depend on which applications and users are active during that time period, hence the importance of monitoring the system.

Best practice: Here are some guidelines for LUN assignment and storage partitioning:

- ▶ Assign LUNs across all controllers to balance controller utilization.
- ▶ Use the manual method of creating logical drives. This allows greater flexibility for configuration settings, such as enclosure loss protection and utilizing both drive loops.
- ▶ If you have highly used LUNs, where possible, move them away from other LUNs and put them on their own separate array. This will reduce disk contention for that array.
- ▶ Always leave a small amount of free space in the array after the LUNs have been created.

Assigning ownership

Ownership is assigned to an array and to a logical drive. To change the ownership of an array (see Figure 2-16), select the **Array** → **Change** → **Ownership/Preferred Path** menu option to change the preferred controller ownership for a selected array.

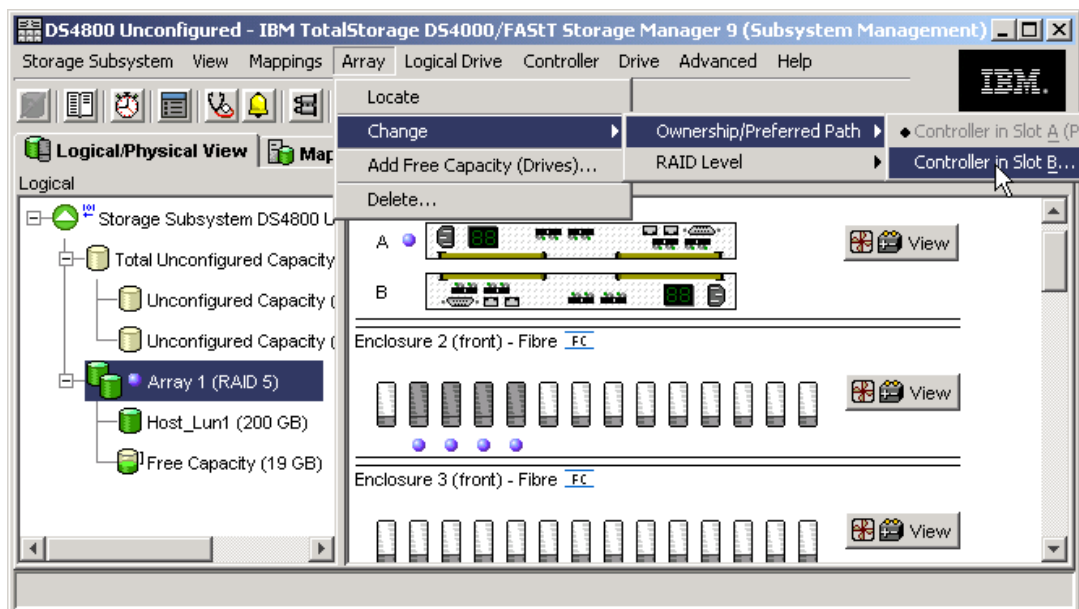


Figure 2-16 Change preferred controller ownership for an array

To change the preferred controller ownership for a Logical Drive, select the Logical Drive which you want to change, then select **Logical Drive** → **Change** → **Ownership/Preferred Path** (see Figure 2-17).

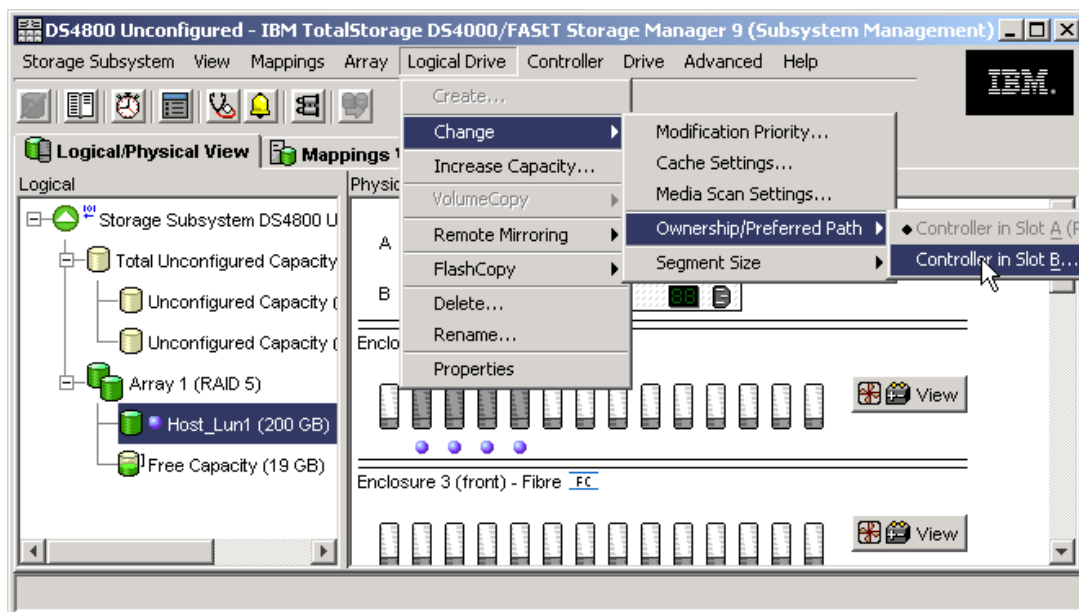


Figure 2-17 Change preferred ownership of a logical drive

The preferred controller ownership of a logical drive or array is the controller of an active-active pair that is designated to own these logical drives. The current controller owner is the controller that currently owns the logical drive or array.

If the preferred controller is being replaced or undergoing a firmware download, ownership of the logical drives is automatically shifted to the other controller, and that controller becomes the current owner of the logical drives. This is considered a routine ownership change and is reported with an informational entry in the event log.

There can also be a forced failover from the preferred controller to the other controller because of I/O path errors. This is reported with a critical entry in the event log, and will be reported by the Enterprise Management software to e-mail and SNMP alert destinations.

Important: To shift logical drives away from their current owners and back to their preferred owners, select **Advanced** → **Recovery** → **Redistribute Logical Drives**. See Figure 2-18.

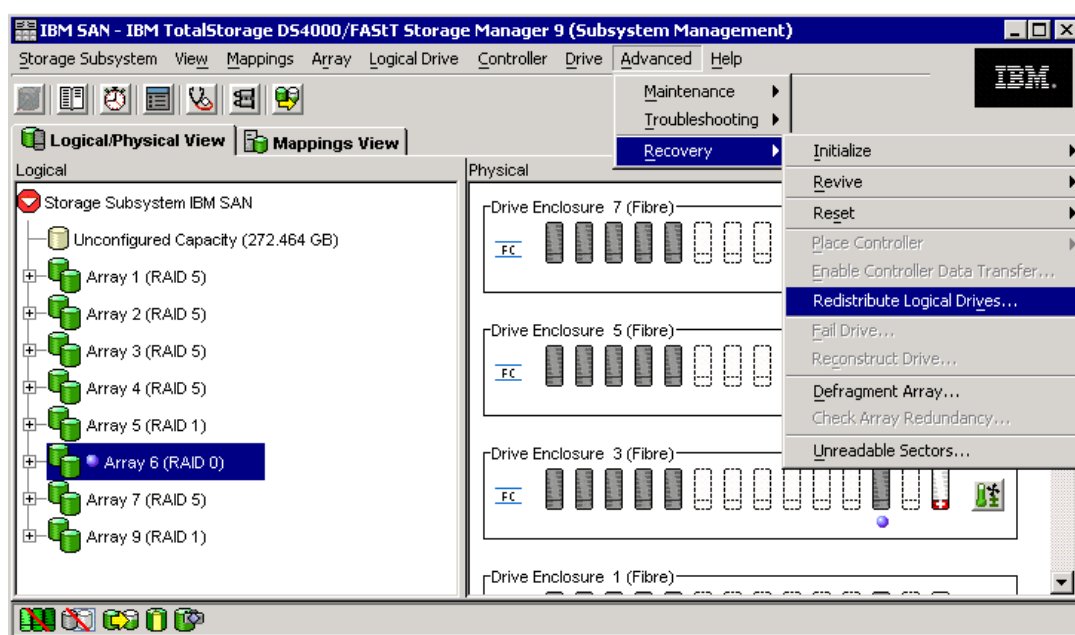


Figure 2-18 Redistribute logical drives

Best practice: Ensure that all assigned LUNs are on their preferred owner. Distribute workload evenly between the controllers with the Storage Manager. Use preferred ownership to ensure a balance between the controllers.

Enhanced Remote Mirror considerations

A secondary logical drive in a remote mirror does not have a preferred owner. Instead, the ownership of the secondary logical drive is determined by the controller owner of the associated primary logical drive. For example, if controller A owns the primary logical drive in the primary storage subsystem, controller A owns the associated secondary logical drive in the secondary storage subsystem. If controller ownership changes on the primary logical drive, then this will cause a corresponding controller ownership change of the secondary logical drive.

2.3.3 Hot spare drive

A hot spare drive is like a replacement drive installed in advance. Hot spare disk drives provide additional protection that might prove to be essential in case of a disk drive failure in a fault tolerant array.

Note: There is no definitive recommendation as to how many hot spares you should install, but it is common practice to use a ratio of one hot spare for about 28 drives.

We recommend that you also split the hot spares so that they are not on the same drive loops (see Figure 2-19).

Best practice: When assigning disks as hot spares, make sure they have enough storage capacity. If the failed disk drive is larger than the hot spare, reconstruction is not possible. Ensure that you have at least one of each size or all larger drives configured as hot spares.

If you have a mixture of 10K and 15K RPM drives of equal size, it is best practice to ensure that the hot spare is 15K RPM. This will ensure that if a hot spare is used, there is not a performance impact on the array after the rebuild is complete.

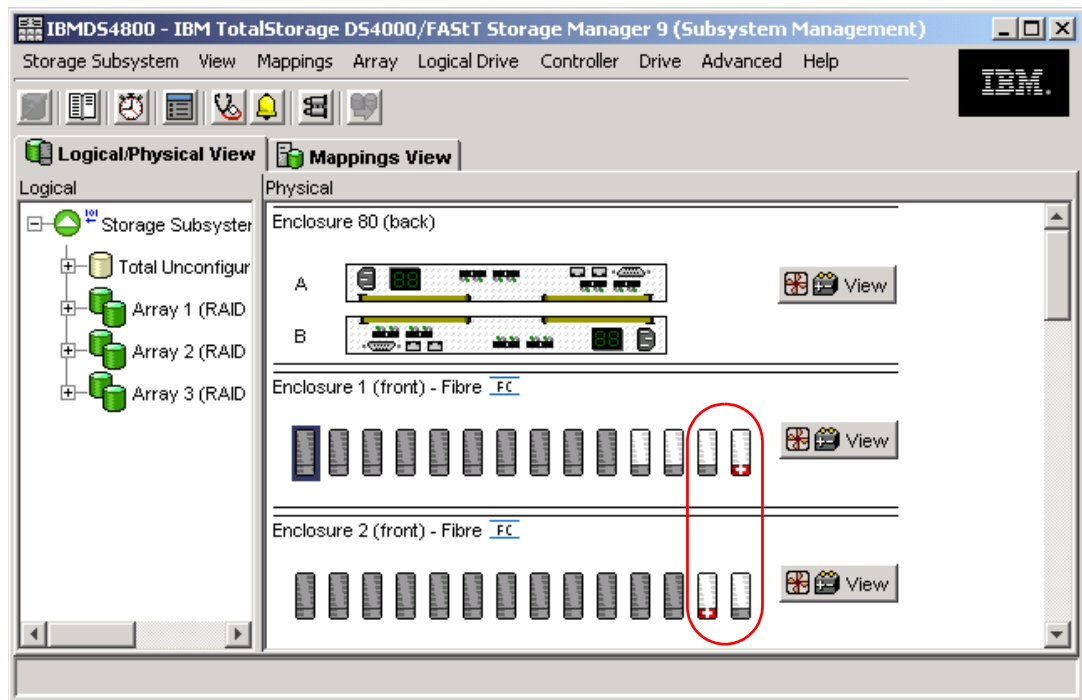


Figure 2-19 Hot spare coverage with alternating loops

2.3.4 Storage partitioning

Storage partitioning adds a high level of flexibility to the DS4000 Storage Server. It enables you to connect to the same storage server multiple and heterogeneous host systems, either in stand-alone or clustered mode. The term storage partitioning is somewhat misleading, as it actually represents a host or a group of hosts and the logical disks they access.

Without storage partitioning, the logical drives configured on a DS4000 Storage Server can only be accessed by a single host system or by a single cluster. This can lead to inefficient use of the storage server hardware.

With storage partitioning, on the other hand, you can create sets of objects containing the hosts with their host bus adapters and the logical drives. We call these sets storage partitions. Now, the host systems can only access their assigned logical drives, just as though these logical drives were locally attached to them.

Storage partitioning lets you map and mask LUNs (that is why it is also referred to as LUN masking). This means that after you assigned that LUN to a host, it is hidden to all other hosts connected to the same storage server. Therefore, the access to that LUN is exclusively reserved for that host.

Note: There are limitations as to how many logical drives you can map per host. DS4000 allows up to 256 LUNs per partition (including the access LUN) and a maximum of two partitions per host. Note that a particular OS platform (refer to the operating system documentation about restrictions) can also impose limitations to the number of LUNs they can support. Keep all these limitations in mind when planning your installation.

It is a good practice to do your storage partitioning prior to connecting to multiple hosts. Operating systems such as AIX or Windows 2000 may write their signatures to any device they can access.

Restriction: Most hosts will be able to have 256 LUNs mapped per storage partition. Solaris with RDAC, NetWare 5.x, and HP-UX 11.0 are restricted to 32 LUNs. If you try to map a logical drive to a LUN that is greater than 32 on these operating systems, the host will be unable to access it. Solaris requires use of Veritas Dynamic Multi-Pathing (DMP) for failover for 256 LUNs.

NetWare 6.x with latest support packs and latest multipath driver (LSIMPE.CDM) supports 256 LUNs. Refer to , “Notes on Novell Netware 6.x” on page 399, and the latest documentation on the Novell support Web site for further information.

Heterogeneous host support means that the host systems can run different operating systems. But be aware that all the host systems within a particular storage partition must run the same operating system, because all host systems within a particular storage partition have unlimited access to all logical drives in this partition. Therefore, file systems on these logical drives must be compatible with host systems. To ensure this, it is best to run the same operating system on all hosts within the same partition. Some operating systems might be able to mount foreign file systems. In addition, Tivoli SANergy® or IBM SAN File System can enable multiple host operating systems to mount a common file system.

Note: Heterogeneous hosts are only supported with storage partitioning enabled.

A storage partition contains several components:

- ▶ Host groups
- ▶ Hosts
- ▶ Host ports
- ▶ Logical drive mappings

A *host group* is a collection of hosts that are allowed to access certain logical drives, for example, a cluster of two systems.

A *host* is a single system that can be mapped to a *logical drive*.

A *host port* is the FC port of the host bus adapter on the host system. The host port is identified by its world-wide name (WWN). A single host can contain more than one host port. If you want redundancy then each server needs two host bus adapters. That is, it needs two host ports within the same host system.

In order to do the storage partitioning correctly, you need the WWN of your HBAs. Mapping is done on a WWN basis. Depending on your HBA, you can obtain the WWN either from the BIOS or Qlogic SANSurfer tool if you have Qlogic cards. Emulex adapters and IBM adapters for System p™ and System i™ servers have a sticker on the back of the card, as do the JNI and AMCC adapters for Solaris. The WWN is also usually printed on the adapter itself or the box the adapter was shipped in.

If you are connected to a hub or switch, check the Name Server Table of the hub or switch to identify the WWN of the HBAs.

Note: If you have to replace a host bus adapter, the WWN of the new adapter will obviously be different. Storage partitioning assignments are based on the WWN. Since the new WWN does not appear in any of the storage partitioning assignments, after replacement, this host system will have no access to any logical drives through this adapter.

When planning your partitioning, keep in mind that:

- ▶ In a cluster environment, you need to use host groups.
- ▶ You can optionally purchase partitions.

When planning for your storage partitioning, you should create a table of planned partitions and groups so that you can clearly map out and define your environment.

Best practice: If you have a single server in a host group that has one or more LUNs assigned to it, we recommend that you do the mapping to the host and not the host group. All servers having the same host type (for example, Windows servers) can be in the same group if you want, but by mapping the storage at the host level, you can define what specific server accesses which specific LUN.

However, if you have a cluster, it is good practice to assign the LUNs at the host group, so that all of the servers in the host group have access to all the LUNs.

Table 2-6 shows an example of a storage partitioning plan. This clearly shows the host groups, hosts, port names, WWN of the ports, and the operating systems used in that environment. Other columns could be added to the table for future references such as HBA BIOS levels, driver revisions, and switch ports used — all of which can then form the basis of a change control log.

Table 2-6 Sample plan for storage partitioning

Host group	Host name	Port name	WWN	OS type
Windows 2000	Windows Host	MailAdp_A	200000E08B28773C	Windows 2000 Non-Clustered
		MailAdp_B	200000E08B08773C	
Linux	Linux_Host	LinAdp_A	200100E08B27986D	Linux
		LinAdp_B	200000E08B07986D	

RS6000	AIX_Host	AIXAdp_A	20000000C926B6D2	AIX
		AIXAdp_B	20000000C926B08	

Delete the access logical drive – (LUN 31)

The DS4000 storage system will automatically create a LUN 31 for each host attached. This is used for in-band management, so if you do not plan to manage the DS4000 storage subsystem from that host, you can delete LUN 31, which will give you one more LUN to use per host.

If you attached a Linux or AIX to the DS4000 storage server, you need to delete the mapping of the access LUN.

2.3.5 Media scan

Media scan is a background process that checks the physical disks for defects by reading the raw data from the disk and writing it back. This detects possible problems caused by bad sectors of the physical disks before they disrupt normal data reads or writes. This is sometimes known as *data scrubbing*.

Media scan continuously runs in the background, using spare cycles to complete its work. The default media scan is for a scan every 30 days, that is, the maximum time media scan will have to complete the task. During the scan process, the DS4000 calculates how much longer the scan process will take to complete, and adjusts the priority of the scan to ensure that the scan completes within the time setting allocated. Once the media scan has completed, it will start over again and reset its time for completion to the current setting. This media scan setting can be reduced, however if the setting is too low, priority will be given to media scan over host activity to ensure that the scan completes in the allocated time. This scan can impact on performance, but improve data integrity. See Figure 2-20 on page 52.

Media scan should be enabled for the entire storage subsystem. The system wide enabling specifies the duration over which the media scan will run. The logical drive enabling specifies whether or not to do a redundancy check as well as media scan.

A media scan can be considered a surface scan of the hard drives while a redundancy check scans the blocks of a RAID 3 or 5 logical drive and compares it against the redundancy data. In the case of a RAID 1 logical drive, then the redundancy scan compares blocks between copies on mirrored drives.

We have seen no effect on I/O with a 30 day setting unless the processor is utilized in excess of 95%. The length of time that it will take to scan the LUNs depends on the capacity of all the LUNs on the system and the utilization of the controller. See Figure 2-20.

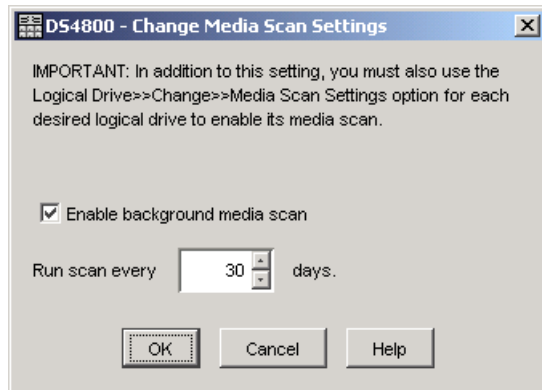


Figure 2-20 Default media scan settings at storage server

An example of logical drive changes to the media scan settings is shown in Figure 2-21.

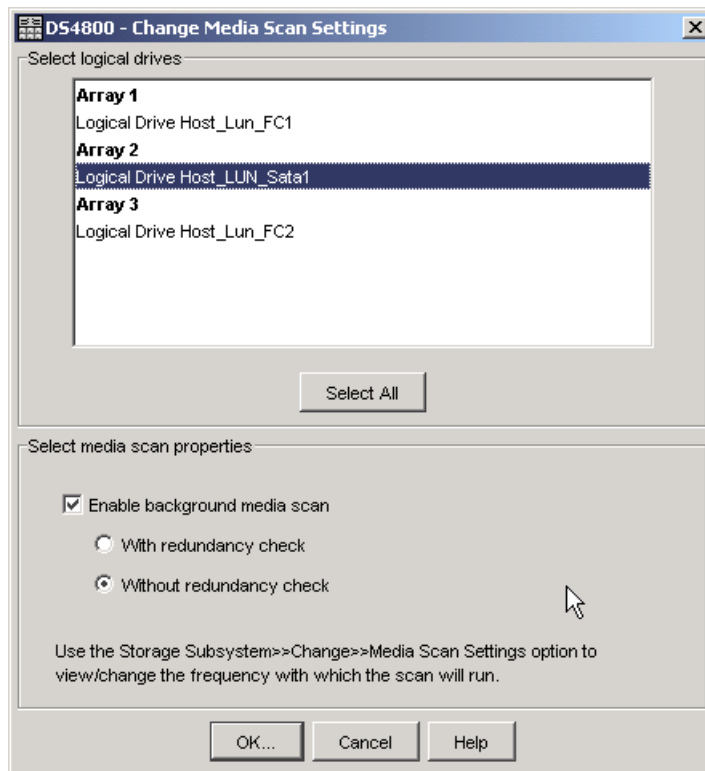


Figure 2-21 Logical drive changes to media scan settings

See also “Media Scan” on page 152.

2.3.6 Segment size

A segment, in a logical drive, is the amount of data, in kilobytes, that the controller writes on a single physical drive before writing data on the next physical drive.

The choice of a segment size can have a major influence on performance in both IOPS and throughput. Smaller segment sizes increase the request rate (IOPS) by allowing multiple disk drives to respond to multiple requests. Large segment sizes increase the data transfer rate (Mbps) by allowing multiple disk drives to participate in one I/O request.

Refer to “Logical drive segments” on page 158 for a more detailed discussion.

2.3.7 Cache parameters

Cache memory is an area of temporary volatile storage (RAM) on the controller that has a faster access time than the drive media. This cache memory is shared for read and write operations.

Efficient use of the RAID controller cache is essential for good performance of the DS4000 storage server.

In this section we define the different concepts and elements that come into play for setting cache parameters on a DS4000. Additional performance related information can be found in 4.5, “DS4000 Storage Server considerations” on page 148.

The diagram shown in Figure 2-22 is a schematic model of the major elements of a disk storage system, elements through which data moves (as opposed to other elements such as power supplies). In the model, these elements are organized into eight vertical layers: four layers of electronic components shown inside the dotted ovals and four layers of paths (that is, wires) connecting adjacent layers of components to each other. Starting at the top in this model, there are some number of host computers (not shown) that connect (over some number of paths) to host adapters. The host adapters connect to cache components. The cache components, in turn, connect to disk adapters that, in turn, connect to disk drives.

Here is how a read I/O request is handled in this model. A host issues a read I/O request that is sent over a path (such as a Fibre Channel) to the disk system. The request is received by a disk system host adapter, which checks whether the requested data is already in cache, in which case, it is immediately sent back to the host. If the data is not in cache, the request is forwarded to a disk adapter that reads the data from the appropriate disk and copies the data into cache. The host adapter sends the data from cache to the requesting host.

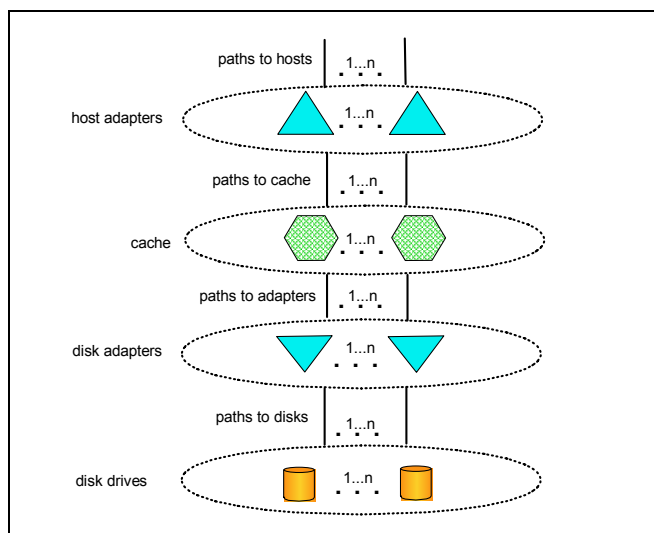


Figure 2-22 Conceptual model of disk caching

Most (hardware) RAID controllers have some form of read or write caching, or both. You should plan to take advantage of this caching capabilities, because they enhance the effective I/O capacity of the disk subsystem. The principle of these controller-based caching mechanisms is to gather smaller and potentially nonsequential I/O requests coming in from the host server (for example, SQL Server) and try to batch them with other I/O requests. Consequently, the I/O requests are sent as larger (32 KB to 128 KB) and possibly sequential requests to the hard disk drives. The RAID controller cache arranges incoming I/O requests by making the best use of the hard disks underlying I/O processing ability. This increases the disk I/O throughput.

There are many different settings (related to caching). When implementing a DS4000 Storage Server as part of a whole solution, you should plan at least one week of performance testing and monitoring to adjust the settings.

The DS4000 Storage Manager utility enables you to configure various cache settings:

Set at the DS4000 system wide settings:

- ▶ Start and stop cache flushing levels (this setting will affect all arrays and Logical drives created on the system)
- ▶ Cache Block size

Settings per Logical Drive.

- ▶ Read caching
- ▶ Cache read-ahead multiplier
- ▶ Write caching or write-through mode (write caching disabled)
- ▶ Enable or disable write cache mirroring

Figure 2-23 shows the typical values when using the Create Logical Drive Wizard. With the Storage Manager, you can specify cache settings for each logical drive independently for more flexibility.

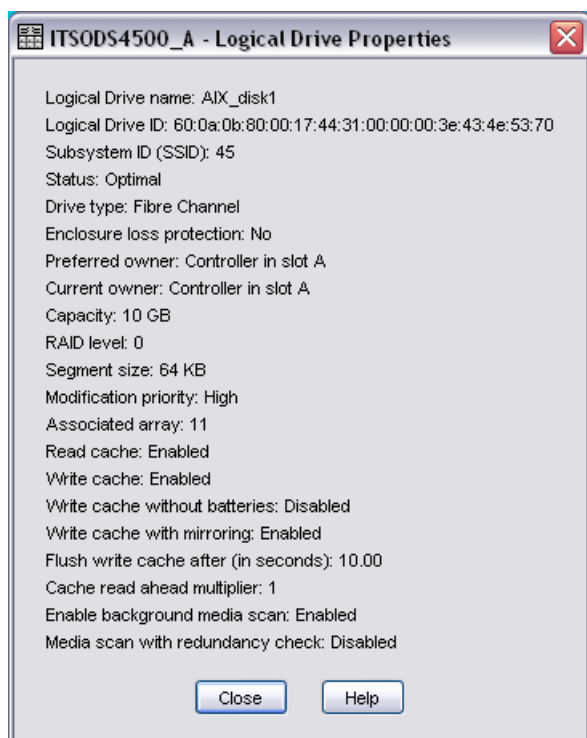


Figure 2-23 Typical values used by the Create Logical Drive Wizard

Note: We recommend that you manually set the values during creation to suit the performance needs of the logical drive. These settings can be changed after Logical drive creation for tuning.

These settings can have a large impact on the performance of the DS4000 Storage Server and on the availability of data. Be aware that performance and availability often conflict with each other. If you want to achieve maximum performance, in most cases, you must sacrifice system availability and vice versa.

The default settings are read and write cache for all logical drives, with cache mirroring to the alternate controller for all write data. The write cache is only used if the battery for the controller is fully charged. Read ahead is not normally used on the logical drives.

Read caching

The read caching parameter can be safely enabled without risking data loss. There are only rare conditions when it is useful to disable this parameter, which then provides more cache for the other logical drives.

Read-ahead multiplier

The cache read-ahead multiplier, or prefetch, allows the controller, while it is reading and copying host-requested data blocks from disk into the cache, to copy additional data blocks into the cache. This increases the chance that a future request for data will be fulfilled from the cache. Cache read-ahead is important for multimedia applications that use sequential I/O.

There is a new automatic pre-fetching for the cache read ahead multiplier introduced with Storage Manager 9.1x. This feature is implemented in the firmware and is enabled by specifying any non-zero value for the cache read-ahead multiplier. This will turn on monitoring of the I/O to the logical drive and enable the new algorithm to dynamically choose how much to read ahead. This simplifies the process for the administrator as there is no need to manually set a specific value for the read ahead multiplier, just change the value from zero.

Write caching

The write caching parameter enables the storage subsystem to cache write data instead of writing it directly to the disks. This can improve performance significantly, especially for environments with random writes such as databases. For sequential writes, the performance gain varies with the size of the data written. If the logical drive is only used for read access, it might improve overall performance to disable the write cache for this logical drive. Then, no cache memory is reserved for this logical drive.

Write cache mirroring

DS4000 write cache mirroring provides the integrity of cached data if a RAID controller fails. This is excellent from a high availability perspective, but it decreases performance. The data is mirrored between controllers across the drive-side FC loop. This competes with normal data transfers on the loop. We recommend that you keep the controller write cache mirroring enabled for data integrity reasons in case of a controller failure.

By default, a write cache is always mirrored to the other controller to ensure proper contents, even if the logical drive moves to the other controller. Otherwise, the data of the logical drive can be corrupted if the logical drive is shifted to the other controller and the cache still contains unwritten data. If you turn off this parameter, you risk data loss in the case of a controller failover, which might also be caused by a path failure in your fabric.

The cache of the DS4000 Storage Server is protected, by a battery, against power loss. If the batteries are not fully charged, for example, just after powering on, the controllers automatically disable the write cache. If you enable the parameter, the write cache is used, even if no battery backup is available, resulting in a higher risk of data loss.

Write caching or write-through

Write-through means that writing operations do not use cache at all. The data is always going to be written directly to the disk drives. Disabling write caching frees up cache for reading (because the cache is shared for read and write operations).

Write caching can increase the performance of write operations. The data is not written straight to the disk drives; it is only written to the cache. From an application perspective, this is much faster than waiting for the disk write operation to complete. Therefore, you can expect a significant gain in application writing performance. It is the responsibility of the cache controller to eventually flush the unwritten cache entries to the disk drives.

Write cache mode appears to be faster than write-through mode, because it increases the performance of both reads and writes. But this is not always true, because it depends on the disk access pattern and workload.

A lightly loaded disk subsystem usually works faster in write-back mode, but when the workload is high, the write cache can become inefficient. As soon as the data is written to the cache, it has to be flushed to the disks in order to make room for new data arriving into cache. The controller would perform faster if the data went directly to the disks. In this case, writing the data to the cache is an unnecessary step that decreases throughput.

Starting and stopping cache flushing levels

These two settings affect the way the cache controller handles unwritten cache entries. They are only effective when you configure the write-back cache policy. Writing the unwritten cache entries to the disk drives is called *flushing*. You can configure the start and stop flushing level values. They are expressed as percentages of the entire cache capacity. When the number of unwritten cache entries reaches the start flushing value, the controller begins to flush the cache (write the entries to the disk drives). The flushing stops when the number of unwritten entries drops below the stop flush value. The controller always flushes the oldest cache entries first. Unwritten cache entries older than 20 seconds are flushed automatically.

The default is the start flushing level and the stop flushing level set to 80%. This means the cache controller does not allow more than 80% of the entire cache size for write-back cache, but it also tries to keep as much of it as possible for this purpose. If you use such settings, you can expect a high number of unwritten entries in the cache. This is good for writing performance, but be aware that it offers less data protection.

If the stop level value is significantly lower than the start value, this causes a high amount of disk traffic when flushing the cache. If the values are similar, the controller only flushes the amount needed to stay within limits.

Refer to “Cache flush control settings” on page 154 for further information.

Cache block size

This is the size of the cache memory allocation unit and can be either 4 K or 16 K. By selecting the proper value for your particular situation, you can significantly improve the caching efficiency and performance. For example, if applications mostly access the data in small blocks up to 8 K, but you use 16 K for the cache block size, each cache entry block is only partially populated. You always occupy 16 K in cache to store 8 K (or less) of data. This means that only up to 50% of the cache capacity is effectively used to store the data. You can expect lower performance. For random workloads and small data transfer sizes, 4 K is better.

On the other hand, if the workload is sequential, and you use large segment sizes, it is a good idea to use a larger cache block size of 16 K. A larger block size means a lower number of cache blocks and reduces cache overhead delays. In addition, a larger cache block size requires fewer cache data transfers to handle the same amount of data.

Refer to “Cache blocksize selection” on page 154 for further information.

2.4 Planning for premium features

When planning for any of the premium features it is a good idea to document what are the goals and rationale for purchasing the feature. This clearly defines from the outset what you want to achieve and why.

Some things that should be considered include:

- ▶ Which premium feature to use FlashCopy, VolumeCopy or Enhanced Remote Mirroring
- ▶ The data size to copy
- ▶ Additional arrays required
- ▶ Amount of free space
- ▶ Number of copies
- ▶ Retention of copies
- ▶ Automated or manual copies
- ▶ Disaster recovery or backup operations

All of the needs and requirements should be documented.

2.4.1 FlashCopy

A FlashCopy logical drive is a point-in-time image of a logical drive. It is the logical equivalent of a complete physical copy, but you create it much more quickly than a physical copy. Additionally, it requires less disk space. In DS4000 Storage Manager, the logical drive from which you are basing the FlashCopy, called the base logical drive, must be a standard logical drive in the storage subsystem. Typically, you create a FlashCopy so that an application (for example, an application to take backups) can access the FlashCopy and read the data while the base logical drive remains online and user-accessible.

FlashCopy takes only a small amount of space compared to the base image

For further information regarding FlashCopy refer to *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

2.4.2 VolumeCopy

The VolumeCopy feature is a firmware-based mechanism for replicating logical drive data within a storage subsystem. This feature is designed as a system management tool for tasks such as relocating data to other drives for hardware upgrades or performance management, data backup, and restoring snapshot logical drive data.

A VolumeCopy creates a complete physical replication of one logical drive (source) to another (target) within the same storage subsystem. The target logical drive is an exact copy or clone of the source logical drive. VolumeCopy can be used to clone Logical Drives to other arrays inside the DS4000 storage server. Careful planning should be considered with regard to space available to make the FlashCopy of a logical drive.

The VolumeCopy premium feature must be enabled by purchasing a Feature Key. For efficient use of VolumeCopy, FlashCopy must be installed as well.

For further information regarding VolumeCopy refer to *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

2.4.3 Enhanced Remote Mirroring (ERM)

The Enhanced Remote Mirroring option is a premium feature that comes with the DS4000 Storage Manager Version 9.x software and is enabled by purchasing a premium feature key. The Enhanced Remote Mirroring option is used for online, real-time replication of data between storage subsystems over a remote distance.

Note: Refer to Chapter 9, “ERM planning and implementation” on page 287, for a detailed discussion and advice on ERM planning and implementation.

The Enhanced Remote Mirroring is a redesign of the former Remote Volume Mirroring and offers three different operating modes:

- Metro mirroring

Metro mirroring is a synchronous mirroring mode. Any host write requests are written to the primary (local) storage subsystem and then transferred to the secondary (remote) storage subsystem. The remote storage controller reports the result of the write request operation to the local storage controller which reports it to the host. This mode is called

synchronous, because the host application does not get the write request result until the write request has been executed on both (local and remote) storage controllers.

- ▶ Global copy

Global copy is an asynchronous write mode. All write requests from a host are written to the primary (local) storage subsystem and immediately reported as completed to the host system. Regardless of when data was copied to the remote storage subsystem, the application does not wait for the I/O commit from the remote site. However, global mirror does not ensure that write requests performed to multiple drives on the primary site are later processed in the same order on the remote site.

- ▶ Global mirroring

Global mirroring ensures that the dependent write requests are carried out in the same order at the remote site.

The Enhanced Remote Mirroring has also been equipped with new functions for better business continuance solution design and maintenance tasks.

A minimum of two storage subsystems is required. One storage subsystem can have primary volumes being mirrored to arrays on other storage subsystems and hold secondary volumes from other storage subsystems. Also note that because replication is managed on a per-logical drive basis, you can mirror individual logical drives in a primary storage subsystem to appropriate secondary logical drives in several different remote storage subsystems.

For further information regarding Enhanced Remote Mirroring refer to *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

Planning considerations for Enhanced Remote Mirroring

Here are some planning considerations:

- ▶ DS4000 storage servers (minimum of two)
- ▶ Fibre links between sites
- ▶ Distances between sites (ensure that it is supported)
- ▶ Switches or directors used
- ▶ Redundancy
- ▶ Additional storage space requirements

Note: ERM requires a dedicated *switched fabric* connection per controller to be attached and zoned specific for its use. The dedication is at the following levels:

- ▶ DS4100 and DS4300 - Host port 2 on each controller (must be dual controller models).
- ▶ DS4200 and DS4700 (Model 70) - Host port 2 on each controller.
- ▶ DS4400 and DS4500 - Host-side minihubs 3 and 4 for A and B controllers with 1 port attached to the switched fabric.
- ▶ DS4800 and DS4700 (Model 72) Host port 4 on both A and B controllers.

This dedication is required at both ends of the ERM solution.

For further information regarding planning considerations refer to Chapter 9, “ERM planning and implementation” on page 287.

2.4.4 FC/SATA Intermix

It is possible to intermix Fibre Channel (FC) and SATA drives expansions attached to a storage server. The flexibility to intermix different drive types (FC and SATA) with one storage

server gives you the ability to use the advantages of both drive technologies. For example, it is now possible to have the primary (production) storage on FC drives and the secondary (near line or archiving) storage on SATA drives without the need of having different, separate storage servers. Or, the FlashCopy or VolumeCopy of an array made of FC drives can be created on cheaper SATA drives.

There are other considerations when implementing both FC and SATA expansion units. The main one is to ensure that placement of the expansion units and the cabling is done correctly. The incorrect cabling of the EXP100 SATA enclosure could impact the performance of your DS4000 storage server, mainly when EXP710 expansion units are involved. The EXP710 will degrade itself to an EXP700 specification. Figure 2-24 shows the grouping recommended when intermixing EXP710 and EXP100 enclosures.

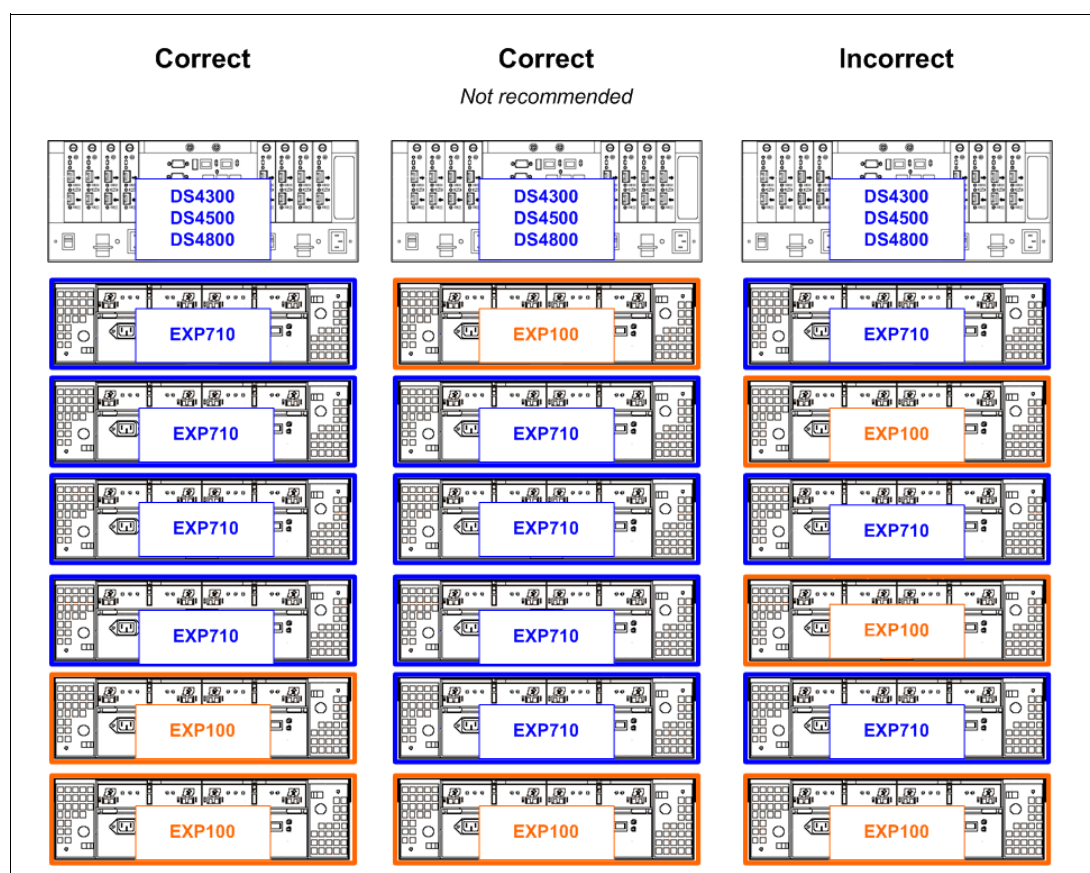


Figure 2-24 Enclosure planning for Intermix Feature with EXP710 and EXP100 Expansion Units

With the release of controller firmware 6.19, it is possible to intermix EXP100, EXP710 and EXP810 enclosures behind a DS4300 or DS4500. Grouping these enclosures in a similar manner is highly desirable to assist in troubleshooting problems.

2.5 Additional planning considerations

In this section, we review additional elements to consider when planning your DS4000 storage subsystems. These considerations include whether using a Logical Volume Manager or not, multipath drivers, failover alert delay, and others.

2.5.1 Planning for systems with LVM: AIX example

Many modern operating systems implement the concept of a Logical Volume Manager (LVM) that can be used to manage the distribution of data on physical disk devices.

The Logical Volume Manager controls disk resources by mapping data between a simple and flexible logical view of storage space and the actual physical disks. The Logical Volume Manager does this by using a layer of device driver code that runs above the traditional physical device drivers. This logical view of the disk storage is provided to applications and is independent of the underlying physical disk structure. Figure 2-25 illustrates the layout of those components in the case of the AIX Logical Volume Manager.

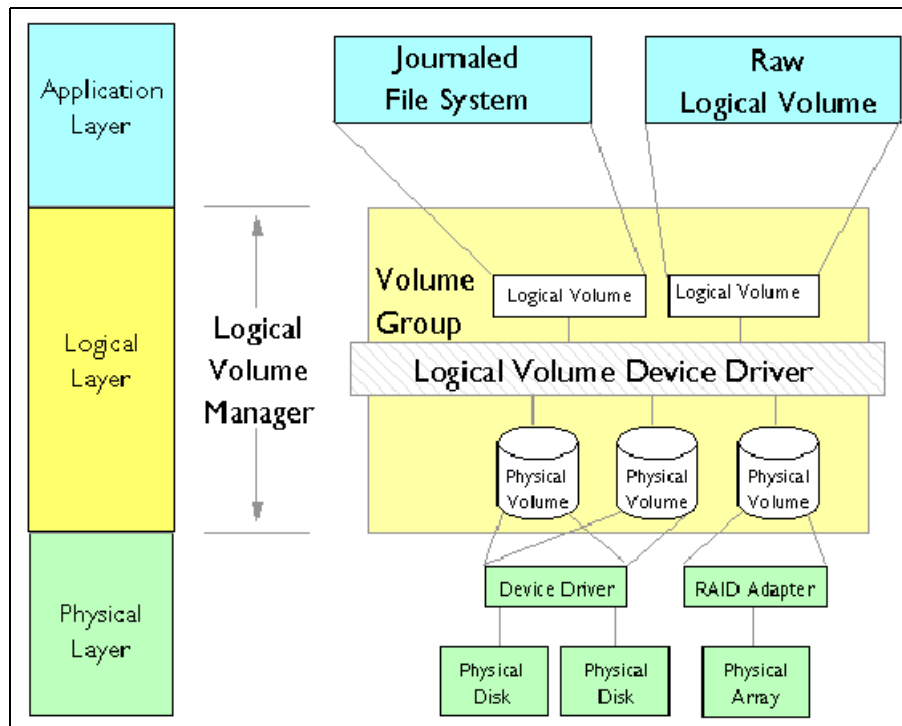


Figure 2-25 AIX Logical Volume Manager

A hierarchy of structures is used to manage the actual disk storage, and there is a well defined relationship among these structures.

In AIX, each individual disk drive is called a physical volume (PV) and has a name, usually /dev/hdiskx (where x is a unique integer on the system). In the case of the DS4000 such physical volumes correspond to a LUN.

- ▶ Every physical volume in use belongs to a volume group (VG) unless it is being used as a raw storage device.
- ▶ Each physical volume is divided into physical partitions (PPs) of a fixed size for that physical volume.
- ▶ Within each volume group, one or more logical volumes (LVs) are defined. Logical volumes are groups of information located on physical volumes. Data on logical volumes appear contiguous to the user, but can be spread (striped) on multiple physical volumes.
- ▶ Each logical volume consists of one or more logical partitions (LPs). Each logical partition corresponds to at least one physical partition (see Figure 2-26 on page 62). If mirroring is specified for the logical volume, additional physical partitions are allocated to store the

additional copies of each logical partition (with DS4000, this is not recommended, because DS4000 can do the mirroring).

- Logical volumes can serve a number of system purposes (paging, for example), but each logical volume that holds ordinary systems, user data, or programs, contains a single journaled filesystem (JFS or JFS2). Each filesystem consists of a pool of page-size blocks. In AIX Version 4.1 and later, a given filesystem can be defined as having a fragment size of less than 4 KB (512 bytes, 1 KB, 2 KB).

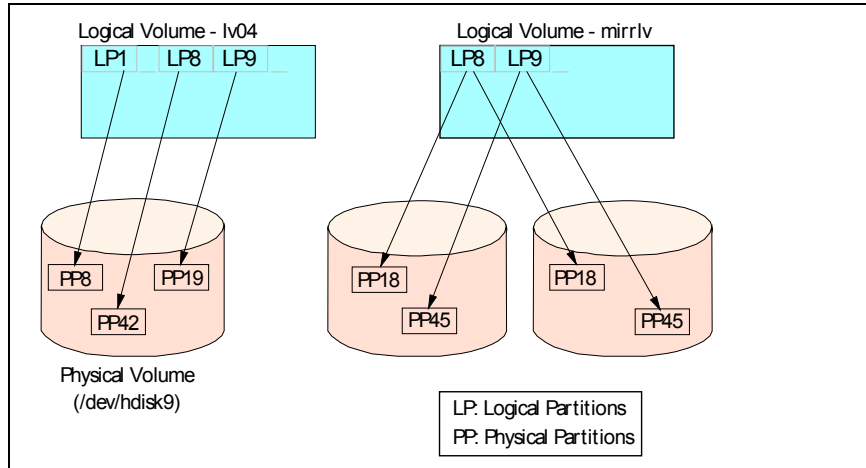


Figure 2-26 Relationships between LP and PP

The Logical Volume Manager controls disk resources by mapping data between a simple and flexible logical view of storage space and the actual physical disks. The Logical Volume Manager does this by using a layer of device driver code that runs above the traditional physical device drivers. This logical view of the disk storage is provided to applications and is independent of the underlying physical disk structure (Figure 2-27).

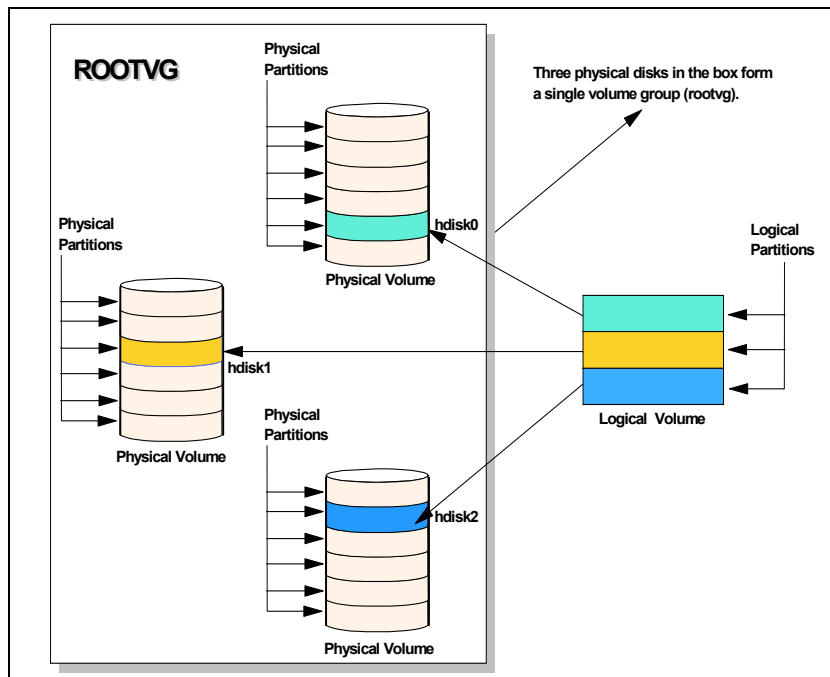


Figure 2-27 AIX LVM conceptual view

Best practice: When using DS4000 with operating systems that have a built-in LVM, or if a LVM is available, you should make use of the LVM.

The AIX LVM provides a number of facilities or policies for managing both the performance and availability characteristics of logical volumes. The policies that have the greatest impact on performance in general disk environment are the intra-disk allocation, inter-disk allocation, write scheduling, and write-verify policies.

Because DS4000 systems has its own RAID arrays and logical volumes, we do not work with real physical disks in the system. Functions, such as intra-disk allocation, write scheduling, and write-verify policies, do not help much, and it is hard to determine the performance benefits when using them. They should only be used after additional testing, and it is not unusual that trying to use these functions will lead to worse results.

On the other hand, we should not forget about the important inter-disk allocation policy.

Inter-disk allocation policy

The inter-disk allocation policy is used to specify the number of disks how the logical partitions (LPs) are placed on specified physical volumes. This is also referred to as *range of physical volumes* in the smitty mklv panel:

- ▶ With an inter-disk allocation policy of minimum, LPs are placed on the first PV until it is full, then on the second PV, and so on.
- ▶ With an inter-disk allocation policy of maximum, the first LP is placed on the first PV listed, the second LP is placed on the second PV listed and so on, in a round robin fashion.

By setting the inter-physical volume allocation policy to maximum, you also ensure that the reads and writes are shared among PVs, and in systems like DS4000, also among controllers and communication paths.

Best practice: For random I/O, the best practice is to create arrays of the same type and size. For applications that don't spread I/Os equally across containers, create VGs comprised of one LUN from every array, use a maximum inter-disk allocation policy for all LVs in the VG, and use a random disk order for each LV. Applications which spread their I/Os equally across containers such as DB2 use a different layout.

If systems are using only one big volume, it is owned by one controller, and all the traffic goes through one path only. This happens because of the static load balancing that DS4000 controllers use.

2.5.2 Planning for systems without LVM: Windows example

Today, the Microsoft Windows operating system does not have a powerful LVM like some of the UNIX systems. Distributing the traffic among controllers in such an environment might be a little bit harder. Actually, Windows systems have an integrated reduced version of Veritas Volume Manager (also known as Veritas Foundation Suite) called Logical Disk Manager (LDM), but it does not offer the same flexibility as regular LVM products. The integrated LDM version in Windows that is used for the creation and use of *dynamic disks*.

With Windows 2000 and Windows 2003, there are two types of disks, basic disks and dynamic disks. By default, when a Windows system is installed, the basic disk system is used.

Basic disks and basic volumes are the storage types most often used with Microsoft Windows operating systems. A basic disk refers to a disk that contains basic volumes, such as primary partitions and logical drives. A basic volume refers to a partition on a basic disk. These partitions in Windows 2000 are set to the size they were created. For Windows 2003, a primary partition on a basic disk can be extended using the **extend** command in the **diskpart.exe** utility. This is explained in more detail in “Using diskpart to extend a basic disk” on page 141.

Dynamic disks were first introduced with Windows 2000 and provide features that basic disks do not, such as the ability to create volumes that span multiple disks (spanned and striped volumes), as well as the ability to create software level fault tolerant volumes (mirrored and RAID-5 volumes). All volumes on dynamic disks are known as dynamic volumes.

With the DS4000 storage server you can use either basic or dynamic disks, depending upon your needs and requirements (some features might not be supported when using dynamic disks). There are cases for both disk types; this depends on your individual circumstances. In certain large installations where you may have the requirement to span or stripe logical drives and controllers to balance the work load, then dynamic disk may be your only choice. For smaller to mid-size installations, you may be able to simplify and just use basic disks. This is entirely dependent upon your environment and your choice of disk system should be made on those circumstances.

When using the DS4000 as the storage system, the use of software mirroring and software RAID 5 is not required. Instead, configure the storage on the DS4000 storage server for the redundancy level required.

If you need greater performance and more balanced systems, you have two options:

- ▶ If you wish to have the UNIX-like capabilities of LVM, you could purchase and use the Veritas Storage Foundation (from Symantec) suite or a similar product. With this product, you get several features that go beyond LDM. Volume Manager does not just replace the Microsoft Management Console (MMC) snap-in, it adds a much more sophisticated set of storage services to Windows 2000 and 2003. After Windows is upgraded with Volume Manager, you are able to manage better multidisk direct server-attached (DAS) storage, JBODs (just a bunch of disks), Storage Area Networks (SANs), and RAID.

The main benefit that you get is the ability to define sub-disks and disk groups. You can divide a dynamic disk into one or more sub-disks. A sub-disk is a set of contiguous disk blocks that represent a specific portion of a dynamic disk, which is mapped to a specific region of a physical disk. A sub-disk is a portion of a dynamic disk's public region.

A sub-disk is the smallest unit of storage in Volume Manager. Therefore, sub-disks are the building blocks for Volume Manager arrays. A sub-disk can be compared to a physical partition. With disk groups, you can organize disks into logical collections.

You assign disks to disk groups for management purposes, such as to hold the data for a specific application or set of applications. A disk group can be compared to a volume group. By using these concepts, you can make a disk group with more LUNs that are spread among the controllers.

Using Veritas Volume Manager and tuning the databases and applications go beyond the scope of this guide. You should look for more information about the application vendor sites or refer to the vendor documentation.

For Veritas Volume Manager (VxVM), go to:

http://www.symantec.com/enterprise/products/overview.jsp?pcid=1020&pvid=203_1

Note that Veritas (Symantec) also offers VxVM also for other platforms, not just Windows.

- ▶ You could use the DS4000 and Windows dynamic disks to spread the workload between multiple logical drives and controllers. This can be achieved with spanned, striped, mirrored or RAID 5:
 - Spanned volumes combine areas of un-allocated space from multiple disks into one logical volume. The areas of un-allocated space can be different sizes. Spanned volumes require two disks, and you can use up to 32 disks. If one of the disks containing a spanned volume fails, the entire volume fails, and all data on the spanned volume becomes inaccessible.
 - Striped volumes can be used to distribute I/O requests across multiple disks. Striped volumes are composed of stripes of data of equal size written across each disk in the volume. They are created from equally sized, un-allocated areas on two or more disks. The size of each stripe is 64 KB and cannot be changed. Striped volumes cannot be extended and do not offer fault tolerance. If one of the disks containing a striped volume fails, the entire volume fails, and all data on the striped volume becomes inaccessible.
 - Mirrored and RAID 5 options are software implementations that add an additional overhead on top of the existing underlying fault tolerance level configured on the DS4000. They could be employed to spread the workload between multiple disks, but there would be two lots of redundancy happening at two different levels.

These possibilities must be tested in your environment to ensure that the solution chosen suits your needs and requirements.

Operating systems and applications

There are big differences among operating systems when it comes to tuning. While Windows 2000 or 2003 does not offer many options to tune the operating system itself, the different flavors of UNIX, such as AIX or Linux, give the user a greater variety of parameters that can be set. These details are beyond the scope of this section. Refer to Chapter 4, “DS4000 performance tuning” on page 133, or consult the specific operating system vendor Web site for further information.

The same is true for tuning specific applications or database systems. There is a large variety of systems and vendors, and you should refer to the documentation provided by those vendors for how to best configure your DS4000 Storage Server.



DS4000 configuration tasks

In this chapter we recommend a sequence of tasks to set up, install, and configure IBM System Storage DS4000 Storage Server, including these:

- ▶ Initial setup of DS4000 Storage Server
- ▶ Setting up the IP addresses on the DS4000 Storage Server
- ▶ Installing and starting the D4000 Storage Manager Client
- ▶ Cabling the DS4000 Storage Server
- ▶ Setting up expansion enclosures
- ▶ Configuring the DS4000 with the Storage Manager Client
- ▶ Defining logical drives and hot-spares
- ▶ Setting up storage partitioning
- ▶ Event monitoring and alerts
- ▶ Software and firmware upgrades
- ▶ Capacity upgrades

3.1 Preparing the DS4000 Storage Server

We assume that you have installed the operating system on the host server; and have all the necessary device drivers and host software installed and configured. We also assume that you have a good understanding and working knowledge of the DS4000 Storage Server product. If you require detailed information about how to perform the installation, setup, and configuration of this product refer to *IBM System Storage IBM System Storage DS4000 Series and Storage Manager*, SG24-7010, at:

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

3.1.1 Initial setup of the DS4000 Storage Server

When installing a DS4000 Storage Server, you must first ensure that the unit has at least two disks that are attached for available DACstore regions to be used. The DS4000 Storage Server uses these regions to store configuration data. When initially powered on, these drives must be available for the system to store information about its state. With the DS4100, DS4200, DS4700, and DS4300, these drives can reside in the main chassis along with the RAID controllers. However, with the DS4500 and DS4800, these drives must reside in an expansion chassis (expansion enclosure).

Attention: With the DS4800, failure to attach an expansion with drives prior to powering it on will result in the loss of the default partition key, and will require it to be regenerated.

Therefore, you must first connect at least one expansion chassis to the main controller chassis. If migrating the expansions from a previously installed system, you must ensure that the ESM firmware is at 9326 or higher, before migrating. For details on attaching expansion units for the different models see 3.2, “DS4000 cabling” on page 76.

Network setup of the controllers

Tip: With Version 8.4x and higher levels of the DS4000 Storage Manager Client, and assuming you have the appropriate firmware level for the controllers to support, it is also possible to set the network settings using the SM Client graphical user interface (GUI).

By default, the DS4000 tries to use the bootstrap protocol (BOOTP) to request an IP address. If no BOOTP server can be contacted, the controllers fall back to the fixed IP addresses. These fixed addresses, by default, are:

- ▶ Controller A: 192.168.128.101
- ▶ Controller B: 192.168.128.102

The DS4100, DS4300 and DS4500 all have a single Ethernet network port on each controller for connecting to the management network. The DS4200, DS4700, and DS4800 have two separate Ethernet ports per controller. One port is for connecting to the management network. The other is for connecting the controllers to the private service/support network for isolation.

Important: On the DS4200, DS4700 and DS4800, each controller has *two* Ethernet ports. The default IP addresses of the additional Ethernet ports are:

- ▶ Controller A: 192.168.129.101
- ▶ Controller B: 192.168.129.102

To use the management network ports of the controllers, you need to attach both controllers to an Ethernet switch or hub. The built-in Ethernet controller supports either 100 Mbps or 10 Mbps.

Tip: To manage storage subsystems through a firewall, configure the firewall to open port 2463 for TCP and UDP data.

To change the default network setting (BOOTP with fallback to a fixed IP address), you can either use the Client GUI controller change function to access the Network settings, or use a serial connection to the controllers in the DS4000 Storage Server.

Changing network setup with the SM Client GUI

To set up the controllers through the GUI, you must be able to connect your Storage Manager Client console to the default IP addresses. This may require you to use a private network or crossover cables at first. When connected, select the storage server you wish to manage. This will bring up the Subsystem Management window for the server you wish to work with. From the Subsystem Management window, highlight the controller you are connected to, and select **Controller** → **Change** → **Network Configuration**.

Enter the network parameters for the controllers. In addition to setting the IP addresses for access, you can also define the option of enabling the use of **rlogin** to remotely access the controller shell commands. This can be done through the GUI's Advanced button, and selecting **Enable remote login to Port # (Controller in slot A/B)** and clicking **OK**. Then click **OK** for the Change Network Configuration. For examples of the windows used in this process see Figure 3-1, Figure 3-2 on page 70, and Figure 3-3 on page 70.

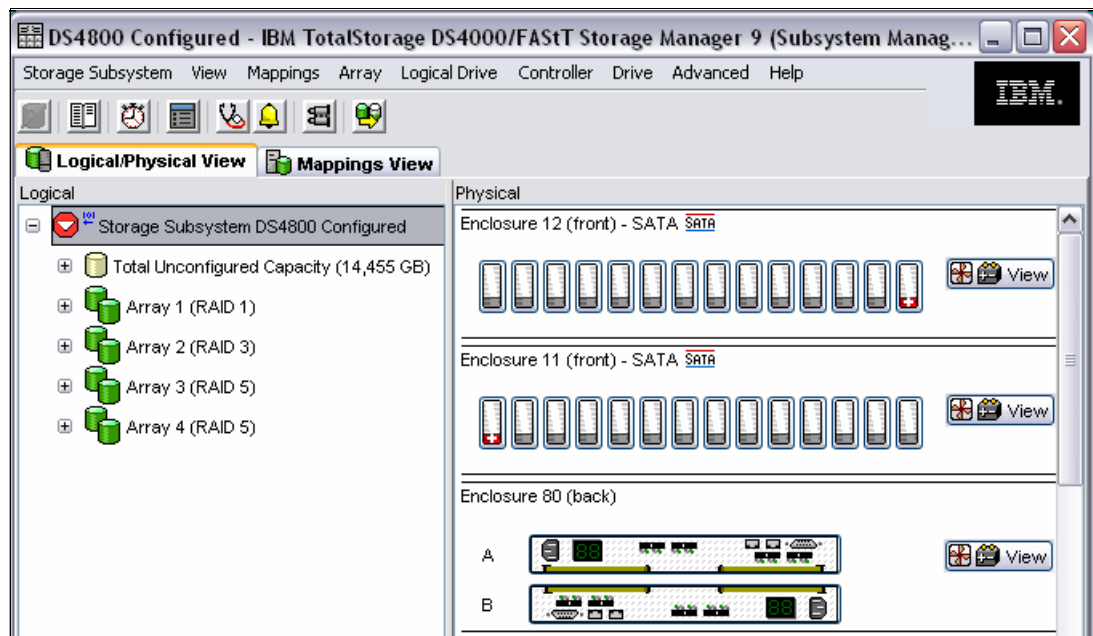


Figure 3-1 Example IBM TotalStorage DS4000 Subsystem Management window

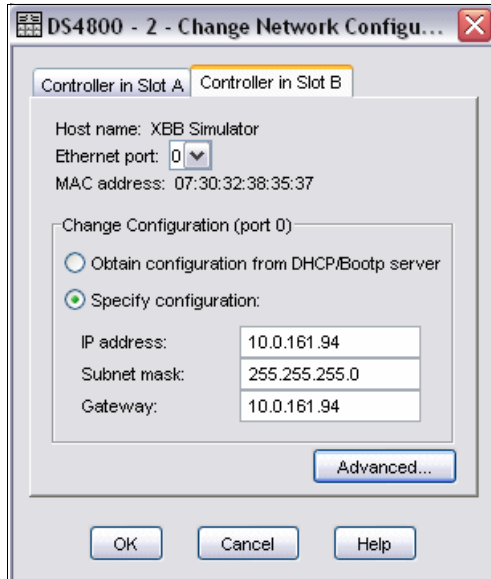


Figure 3-2 Example Change Network Configuration window

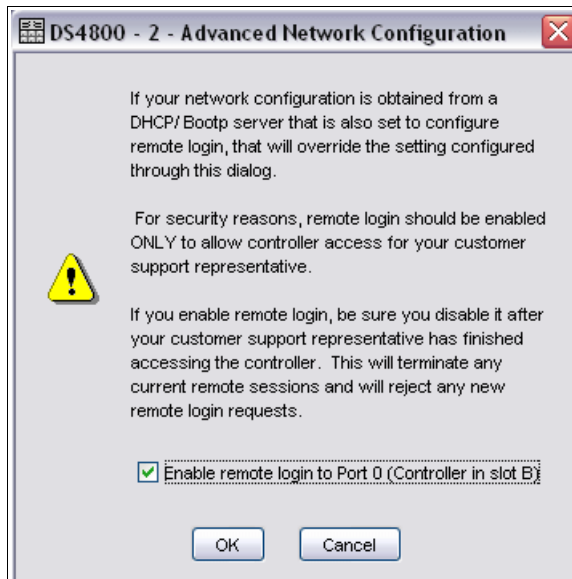


Figure 3-3 Example Advanced Network Configuration window

Using the rlogin capability

The **rlogin** shell capability is useful to enable you to check the network configuration settings for the two controllers without the use of the serial port. To do so, you need a server capable of running an **rlogin** process to the DS4000 Storage Server controllers.

Tip: For many servers without **rlogin** capability, you can use a freeware utility like “PuTTY” for running a secure shell over the network to the DS4000 Storage Server. For greater details on PuTTY see:

<http://www.chiark.greenend.org.uk/~sgtatham/putty/>

You will be prompted for the password when you connect Enter `infiniti`. Then you will want to enter the **netCfgShow** command to see the values that are set for the controller which you logged into. To change these values, enter the **netCfgSet** command.

Do not attempt to use any other shell commands. Some commands have destructive effects (causing loss of data or even affecting the functionality of your DS4000).

Best practice: Unless needed for continued support purposes, we strongly recommend having the **rlogin** function disabled once you have completed all your configuration work, and access is no longer necessary.

Network setup using the serial port

Attention: If using the serial cable, follow the procedure outlined below exactly as it is presented.

To set up the controllers through the serial port:

1. Connect to the DS4000 Storage Server with a null modem cable to the serial port of your system. For the serial connection, choose the correct port and the following settings:
 - 57600 Baud
 - 8 Data Bits
 - 1 Stop Bit
 - No Parity
 - Xon/Xoff Flow Control
2. Send a break signal to the controller. This varies depending on the terminal emulation. For most terminal emulations, such as HyperTerm, which is included in Microsoft Windows products, press Ctrl+Break.
3. If you only receive unreadable characters, press Ctrl+Break again, until the following message appears:

Press <SPACE> for baud rate within 5 seconds.
4. Press the Space bar to ensure the correct baud rate setting. If the baud rate was set, a confirmation appears.
5. Press Ctrl+Break to log on to the controller. The following message appears:

Press within 5 seconds: <ESC> for SHELL, <BREAK> for baud rate.
6. Press the Esc key to access the controller shell. The password you are prompted for is `infiniti`.
7. Run the **netCfgShow** command to see the current network configuration.
8. To change these values, enter the **netCfgSet** command. For each entry, you are asked to keep, clear, or change the value. After you assign a fixed IP address to controller A, disconnect from controller A and repeat the procedure for controller B. Remember to assign a different IP address.
9. Because the configuration changed, the network driver is reset and uses the new network configuration.

In addition to setting the IP addresses for access, you can also define usage options. One that is frequently desired is the option to be able to access the server's login shell via an **rlogin** command. This can be done through the GUI's **Advanced** button, and selecting **Enable remote login to Port # (Controller in slot A/B)**. When using the serial connection, it is set by changing the value of the *Network Init Flags* to include a 1 in the bit 5 position.

3.1.2 Installing and starting the D4000 Storage Manager Client

You can install the DS4000 Storage Manager Client (SM Client) for either in-band management or out-of-band management. It is possible to use both access methods on the same machine if you have a TCP/IP connection and a Fibre Channel connection to the DS4000 Storage Server.

In-band management uses the Fibre Channel to communicate with the DS4000 Storage Server, and requires you install the Storage Manager Agent software, and create a UTM LUN for access for the SM Client. This method is not supported on all host server OS types. Use of this method is outlined for your reference in *IBM System Storage IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

Out-of-band management uses the TCP/IP network to communicate with the DS4000 Storage Server. In our example, we use the out-of-band management and install the Storage Manager Client on a machine that only has a TCP/IP connection to the DS4000 Storage Server.

Tip: For ease of management and security, we recommend installing a management workstation on a separate IP network.

If you are unable to use a separate network, ensure that you have at least an adequate password set on your DS4000 Storage Server.

There are some advantages for doing out of band management using a separate network. First, it makes the storage more secure and limits the number of people that have access to the storage management functions. Second, it provides more flexibility, because it eliminates the need for the storage administrator to access the server console for administration tasks. In addition, the Storage Manager agent and software do not take up resources on the host server.

Note: If you install DS4000 Storage Manager Client on a stand-alone host and manage storage subsystems through the Fibre Channel I/O path, rather than through the network, you must still install the TCP/IP software on the host and assign an IP address to the host.

Installing the SMclient

We assume for this illustration that the SM Client is to be installed on a Microsoft Windows workstation, as is commonly the case. However, the SM Client is available for other OS platforms, such as AIX.

Installing with InstallAnywhere

The host software for Windows includes the following components:

- ▶ SMclient
- ▶ Multipath driver (RDAC or MPIO)
- ▶ SMagent
- ▶ SMutil

InstallAnywhere makes sure that the proper selection of these components is installed.

Because you are installing new software, including new drivers, you need to log on with administrator rights.

Important: If you want to use the Event Monitor with SNMP, you have to install the Microsoft SNMP service first, since the Event Monitor uses its functionality.

Locate and run the installation executable file, either in the appropriate CD-ROM directory, or the file that you have downloaded from IBM support Web site. After the introduction and copyright statement windows, you will be asked to accept the terms of the License Agreement. This is required to proceed with the installation.

The next step is selection of the installation target directory. The default installation path is C:\Program Files\IBM_DS4000, but you can select another directory.

Now you need to select the installation type, as shown in Figure 3-4.



Figure 3-4 InstallAnywhere - Select Installation Type

The installation type you select defines the components that will be installed. For example, if you select Management Station, then the multipath driver and Agent components will not be installed, because they are not required on the management computer. In most cases, you would select either the Management Station or Host installation type.

Since having only these two options could be a limitation, two additional choices are offered: typical (full installation) and custom. As the name suggests, full installation installs all components, where as custom installation lets you choose the components.

With Version 9.19 of Storage Manager there is a choice of multipath driver. When you select typical, host, or custom (and select the Storage Manager 9 Fail-Over Driver), you will be prompted for which type of multipath driver you wish to install. The choices are RDAC Multi-Pathing Driver or MPIO Device Specific Module (DSM) (see Figure 3-5). Note that MPIO is only supported with Version 6.19 of the firmware.

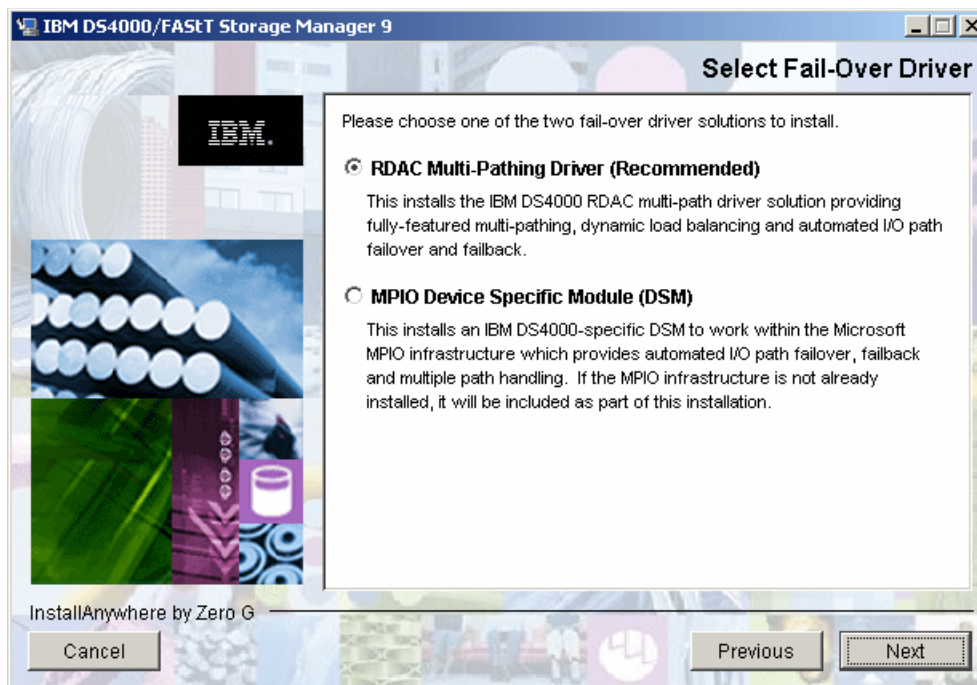


Figure 3-5 Multipath driver installation selection

Which driver is best for your environment becomes a matter of choice. To assist in the choice process you must consider the following points:

- ▶ If you have existing hosts using RDAC, then perhaps standardization is required with your environment.
- ▶ Microsoft is moving towards the MPIO model for future versions of Windows Server operating system for multipath drivers. If this is your first host for your SAN environment, it is perhaps worthwhile using MPIO, as it will allow for a standard driver in the future operating system releases.

Each of these multipath drivers has been previously discussed in 2.2.5, “Multipath driver selection” on page 27.

The next installation panel asks you whether you wish to automatically start the Storage Manager Event Monitor. This depends on your particular management setup: In case there are several management machines, the Event Monitor should only run on one.

Finally, you are presented with the Pre-Installation Summary window, just to verify that you have selected correct installation options. Click the **Install** button and the actual installation process starts. When the installation process has completed, you need to restart the computer.

Starting the SMclient

When you start the DS4000 Storage Manager Client, it launches the Enterprise Management window. With the new SM Client 9.1x release; you will be presented with the new configuring feature, Task Assistant, to aid you in the discovery and adding of storage servers that are detected. (If you do not want to use the Task Assistant, you can disable it by marking the appropriate check box. In case you want to use it again, you can invoke it from the Enterprise Management window). The first time you start the client, you are prompted to select whether you want an initial discovery of available storage subsystems (see Figure 3-6).

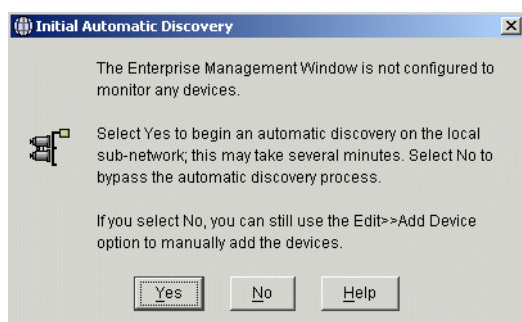


Figure 3-6 Initial Automatic Discovery

The client software sends out broadcasts through Fibre Channel and the subnet of your IP network if it finds directly attached storage subsystems or other hosts running the DS4000 Storage Manager host agent with an attached storage subsystem.

You have to invoke the Automatic Discovery every time you add a new DS4000 Storage Server in your network or install new host agents on already attached systems. To have them detected in your Enterprise Management window, click **Tools** → **Rescan**. Then, all DS4000 Storage Servers are listed in the Enterprise Management window, as shown in Figure 3-7.

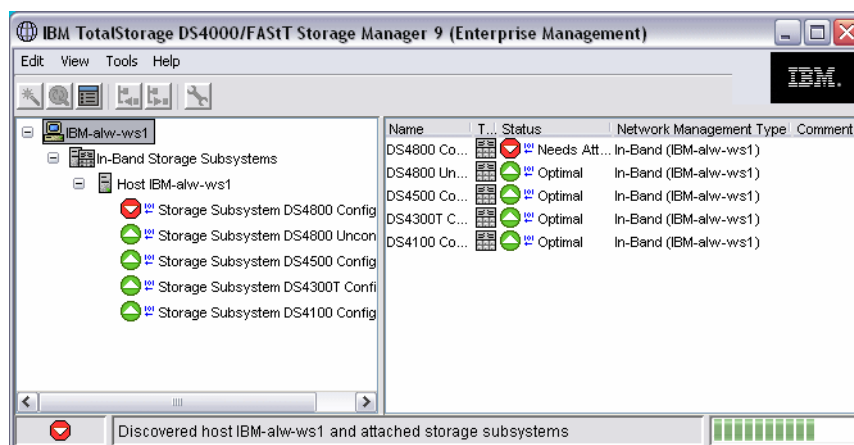


Figure 3-7 Example DS4000 Enterprise Management window

If you are connected through FC and TCP/IP, you will see the same DS4000 Storage Server twice.

The DS4000 Storage Server can be connected through Ethernet, or you might want to manage it through the host agent of another host, which is not in the same broadcast segment as your management station. In either case, you have to add the devices manually. Click **Edit** → **Add device** and enter the host name or the IP address you want to attach. To

choose the storage subsystem you want to manage, right-click and select **Manage Device** for the attached storage subsystem. This launches the Subsystem Management window (Figure 3-8).

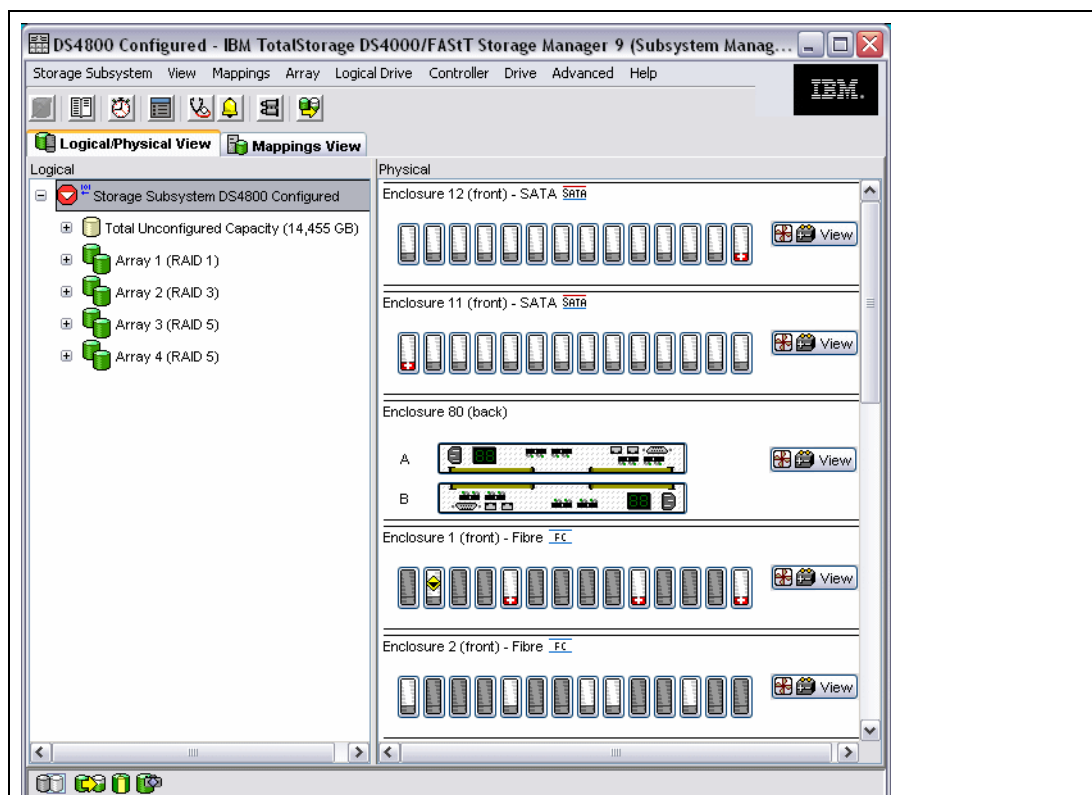


Figure 3-8 First launch of the Subsystem Management window

If you add a DS4000 Storage Server that is directly managed, be sure to enter both IP addresses, one per controller. You receive a warning message from Storage Manager if you only assign an IP address to one controller.

Verify that the enclosures in the right side of the window reflects your physical layout. If the enclosures are listed in an incorrect order, select **Storage Subsystem** → **Change** → **Enclosure Order** and sort the enclosures according to your site setup.

3.2 DS4000 cabling

In the following sections we explain the typical recommended cabling configuration for the DS4100, DS4200, DS4300, DS4400, DS4500, DS4700, and DS4800, respectively.

A basic design point of a DS4000 Storage Server is to enable hosts to directly attach it. However, the *best practice* for attaching host systems to your DS4000 storage is through a switched fabric (SAN attached).

Best practice: Attach the DS4000 to hosts systems through a SAN fabric.

Both methods are explained in this section for each DS4000 Storage Server type.

Tip: We recommend that you remove all unused small form factor plug (SFP) modules.

3.2.1 DS4100 and DS4300 host cabling configuration

Here we describe the various cabling configurations that you can use.

Direct attached

Both the DS4100, and the DS4300 offer fault tolerance by the use of two host HBAs connected to two RAID controllers in the DS4000 storage servers. At the same time, you can get higher performance, because the dual active controllers allow for distribution of the load across the two paths. See the left side of Figure 3-9 for an example of the DS4300.

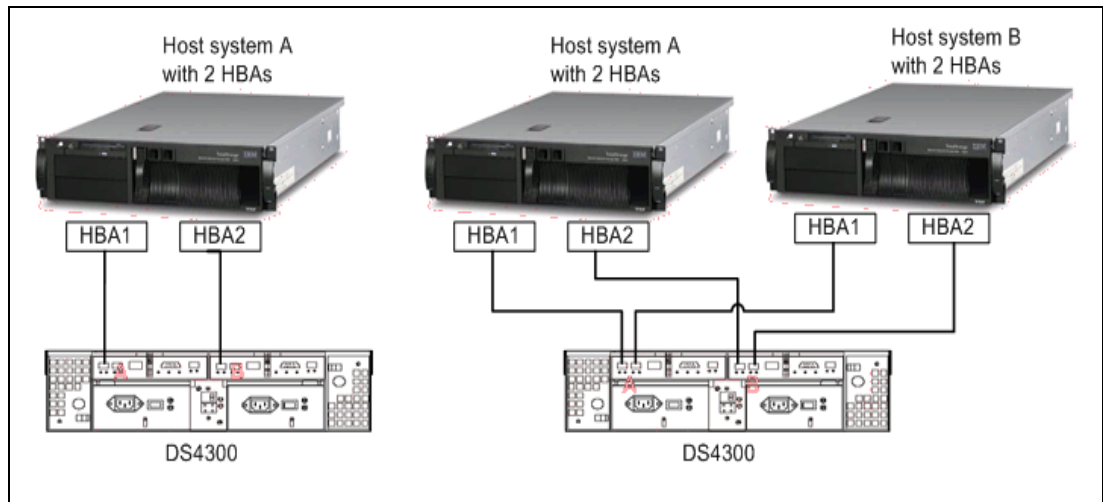


Figure 3-9 DS4300 cabling configuration

These two DS4000 Storage servers can support a dual-node cluster configuration without using a switch, shown on the right side of Figure 3-9. This provides the lowest priced solution for a 2-node cluster environment due to four Fibre Channel host ports on the storage servers.

SAN attached

The recommended configuration is to connect the DS4100, and DS4300 to managed hubs, or Fibre switches to expand their connections for multiple servers, as shown in Figure 3-10.

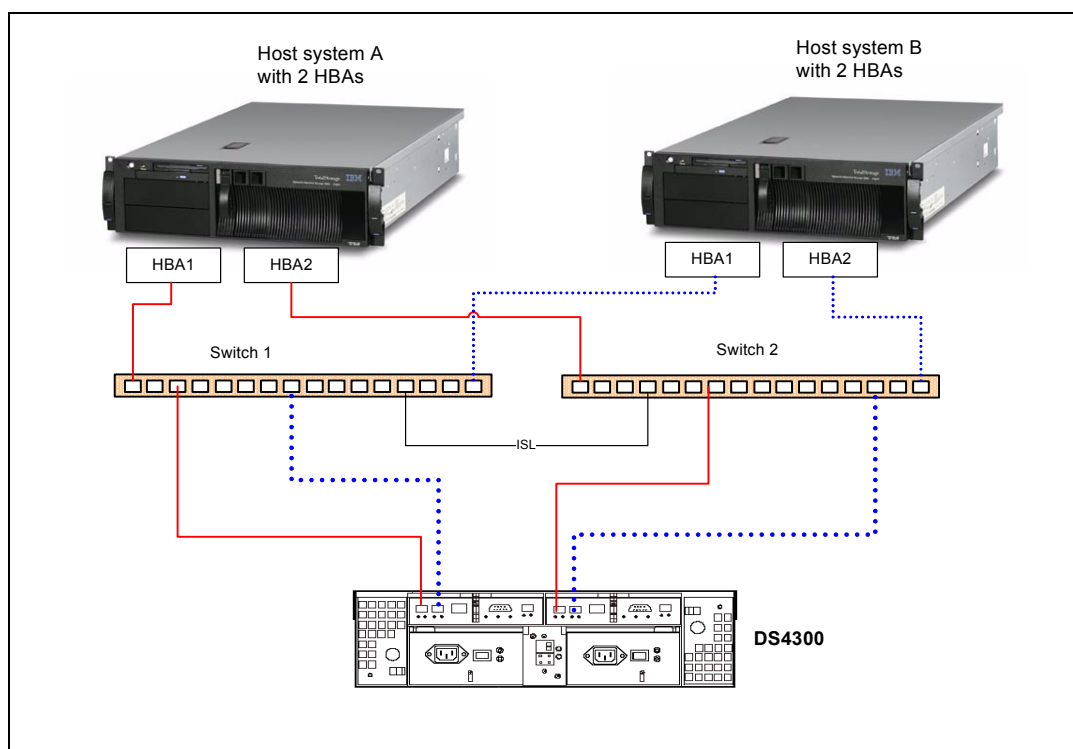


Figure 3-10 DS4300 connected through managed hub or Fibre switches

Multiple hosts can access a single DS4000 system, but also have the capability of accessing data on any DS4000 subsystem within the SAN. This configuration allows more flexibility and growth capability within the SAN: The attachment of new systems is made easier when adopting such structured cabling techniques. This method of connection with Fibre switches is required for support of the Enhanced Remote Mirroring (ERM) feature if used. When using this method, you must ensure that you properly zone all storage and host HBAs to avoid fabric noise, or undesired data transfer interruptions from other members of the fabric. Best practices for zoning recommends zoning a single initiator, and single target per zone.

This means that a zone would consist of two members: an HBA from a host server, and a controller port from a storage server. Shared mappings of the initiator or the storage controller to other zones is supported, and recommended when storage is shared. In the case of ERM, a separate pair of zones for each of the second two ports with their remote ends is required. The ERM ports cannot be shared with host server access. Managed hubs cannot be used for ERM support. More details on ERM are discussed in a later section.

Figure 3-11 shows an example of a dual DS4300 configuration in a SAN fabric.

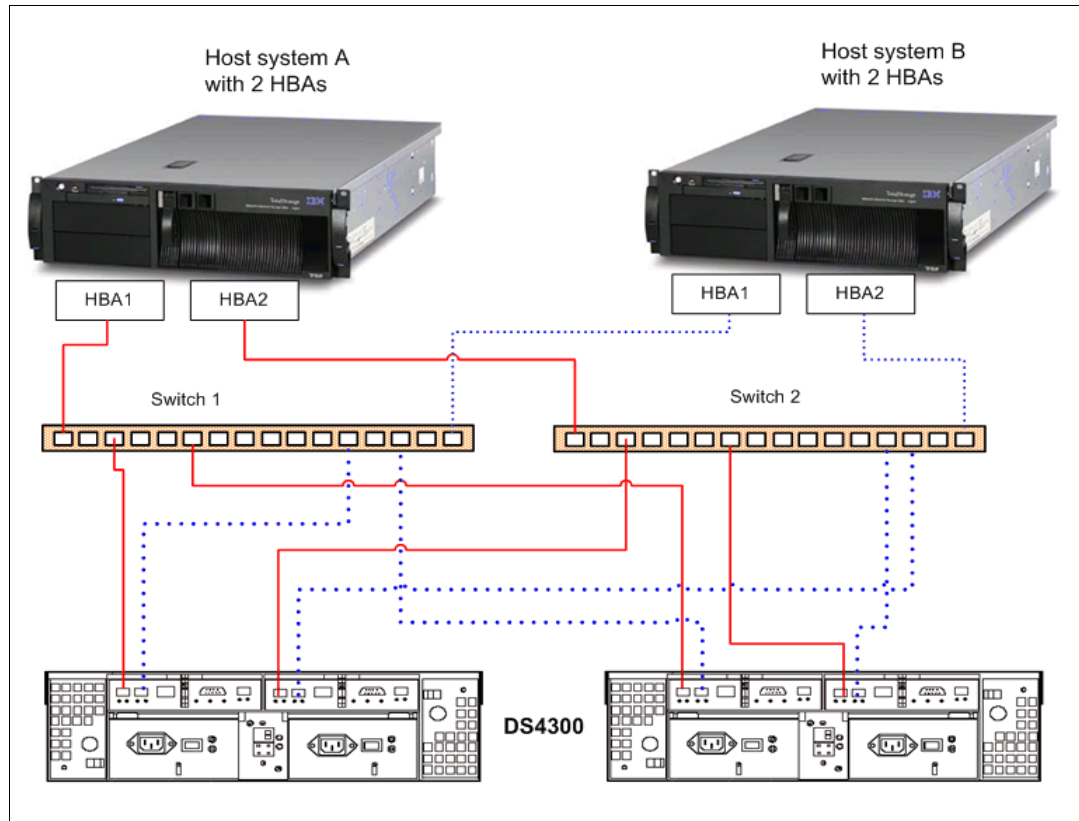


Figure 3-11 Dual DS4300 connected through Fibre switches

3.2.2 DS4100 and DS4300 drive expansion cabling

The DS4100 supports up to seven DS4000 EXP100 units. These expansion units provide Fibre Channel Arbitrated Loop connection to the DS4100 drive expansion ports; and hold up to 14 Serial ATA (SATA) drives, for a maximum of 112 SATA drives. The DS4300 Turbo supports up to *eight* exp100, or seven EXP710 units, for a maximum of 112 drives. On the base DS4300, you can have up to three EXP710 units for 56 fibre drives, or up to eight EXP100 units for 112 SATA drives. With firmware Version 6.19 and up, the EXP810 (up to six) can also be attached to the DS4300. The cabling of both the DS4100 and the DS4300 is done in the same manner regardless of the expansion used. The example diagram in Figure 3-12 shows a DS4300 connection scheme with two EXP710 expansion units.

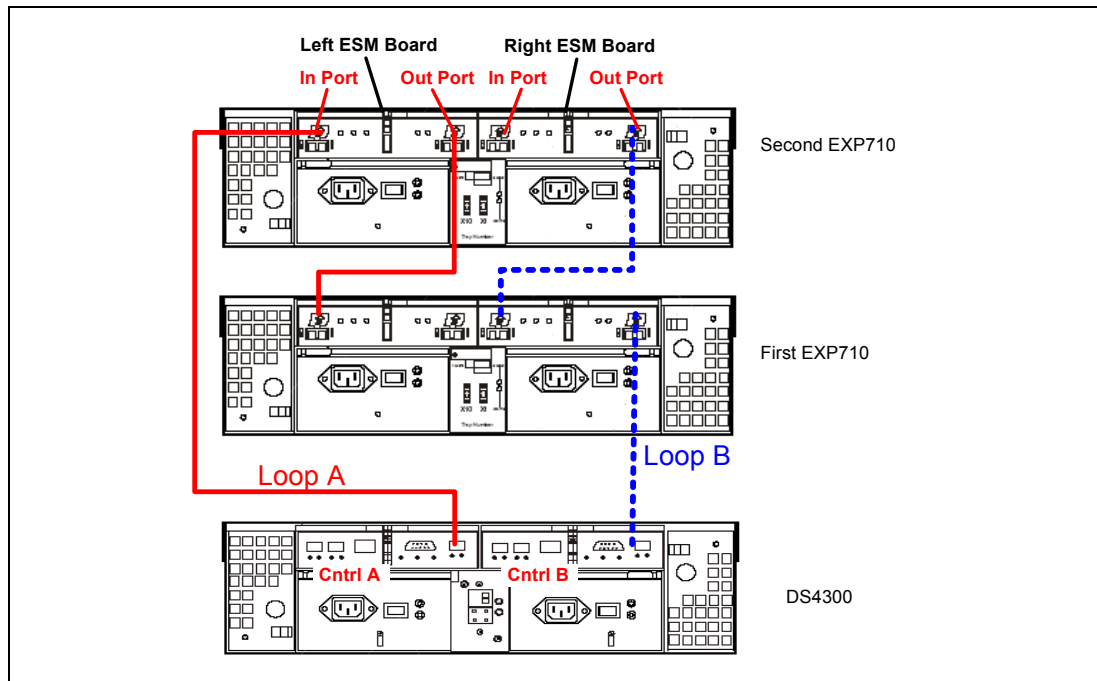


Figure 3-12 Example of DS4300 with two expansion units Fibre Channel cabling

Please note that in order to have path redundancy, you need to connect a multipath loop to the DS4300 from the EXP710. As shown in Figure 3-12, Loop A is connected to controller A, and Loop B is connected to controller B. If there was a break in one of the fiber cables, the system would still have a path for communication with the EXP710, thus providing continuous uptime and availability.

Note: Although storage remains accessible, Storage Manager will report a path failure and request that you check for a faulty cable connection to the DS4000.

3.2.3 DS4200 host cabling configuration

The DS4200 has four 4 Gbps host connections (two per controller). They Fibre Channel attachment through SAN switches or direct connections.

It is important to match up host or fabric connections to the DS4200 by attaching one connection to each controller. In doing so, you take advantage of the DS4200's ability to fail over and distribute the workload among the two controllers. For any given host, make sure to connect to the same host port number on each controller.

Direct attached hosts

The layout for a single direct connect host with two HBAs to the DS4200 is such that HBA 1 is connected to controller A on drive loop 2 and HBA 2 is connected to controller B on drive loop 1.

The layout for two hosts, each with 2 HBAs, is such that host 1 is connected to controller A and controller B and host 2 is also connected to controller A and controller B. Host 1 is using both drive loops, as is host 2. This gives fault tolerance in the unlikely event of a controller or a loop failure.

Fibre attached

The recommended configuration is to connect the DS4200 to fibre switches to expand its connection for multiple servers, as shown in Figure 3-13.

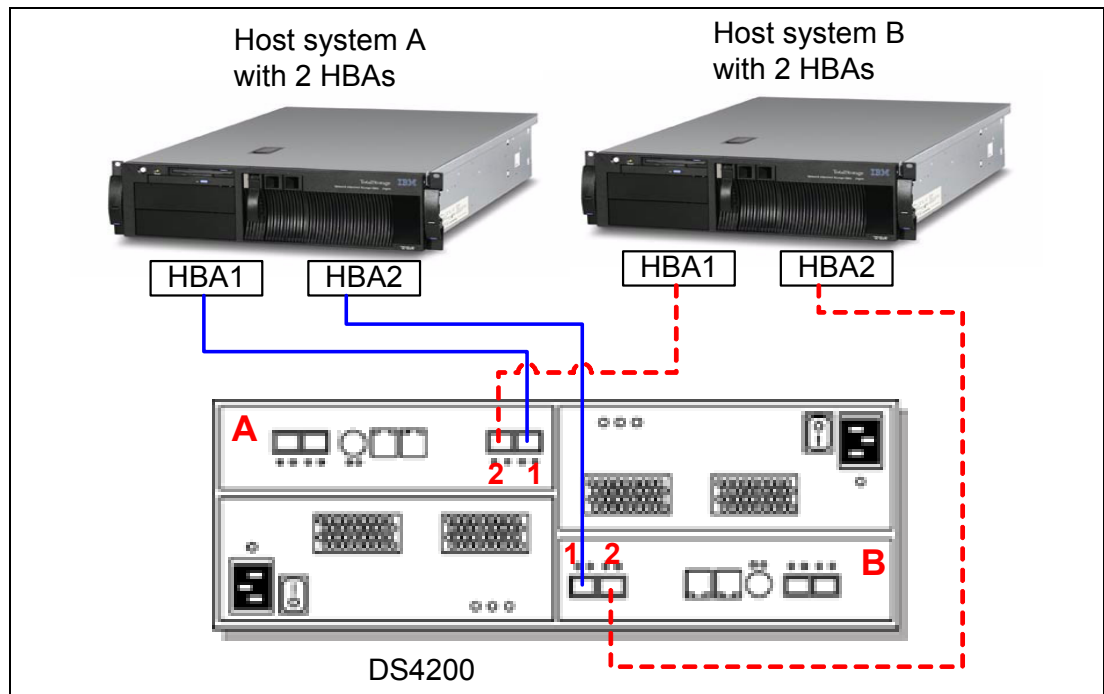


Figure 3-13 Two direct connected hosts to the DS4200

Multiple hosts can access a single DS4000 system, but also have the capability of accessing data on any DS4000 subsystem within the SAN. This configuration allows more flexibility and growth capability within the SAN. The attachment of new systems is made easier when adopting such structured cabling techniques. This method of connection with Fibre switches is required for support of the Enhanced Remote Mirroring (ERM) feature if used. When using this method, you must ensure that you properly zone all storage and host HBAs to avoid fabric noise, or undesired data transfer interruptions from other members of the fabric. Best practices for zoning recommend zoning a single initiator, and single target per zone. See Figure 3-14.

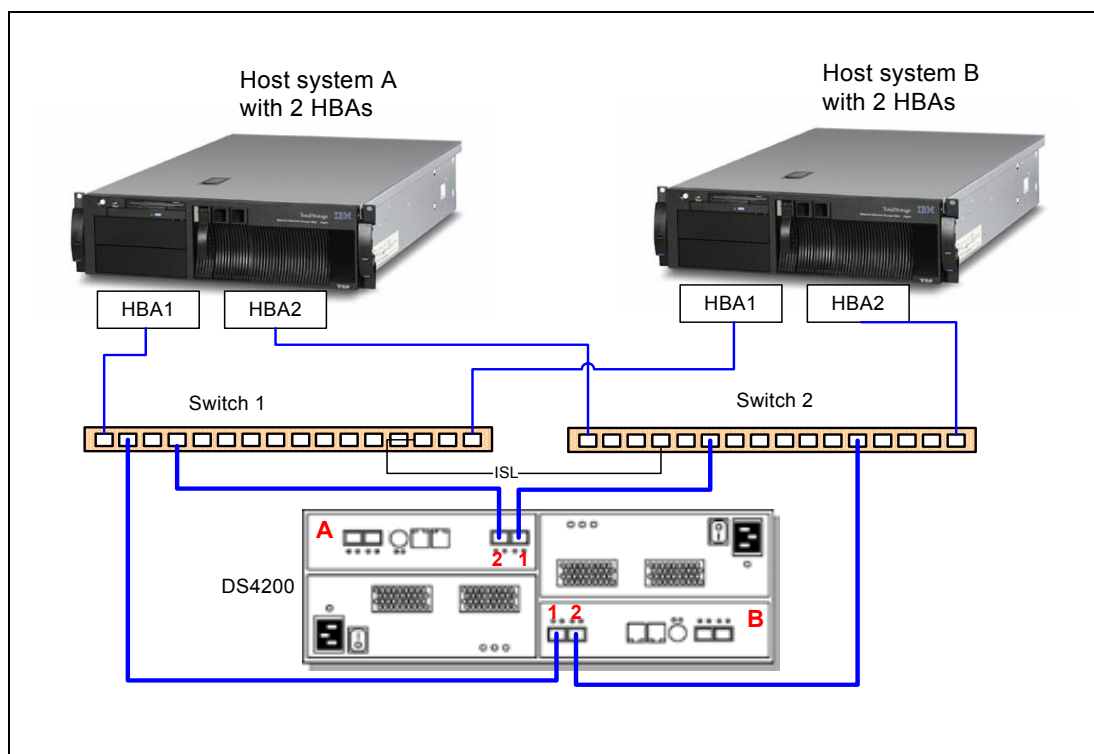


Figure 3-14 DS4200 SAN attached

3.2.4 DS4200 drive expansion cabling

The DS4200 allows for two loop pairs (each controller has a drive channel with two ports) for drive-side attachment.

Each loop pair (or drive channel) can contain a maximum of 112 drives. In other words, each loop pair can attach a maximum of six EXP420 (each enclosure contains 16 drives, and including 16 drives in the DS4200, this gives the maximum of 112 drives). However, the loop pair from each controller must combine to create one set of redundant loop pairs, supporting a maximum of six EXP420s.

Important: The DS4200 supports the connection of a maximum of six EXP420.

The EXP420 enclosure can only use the new 500 GB SATA EV-DDM drives. These drives are 4 Gbps capable drives.

Note: There are three rules for the EXP420 cabling:

- ▶ With the DS4200 you should only connect a maximum of three EXP420 enclosures per controller drive port.
- ▶ The DS4200 controller drive port must always be connected to the EXP420 port labelled 1B. Because the left and right EXP420 ESMs (ESMs A and B) are inserted in the ESM bays in different orientations, ensure that you use the port labeled 1B before making the Fibre Channel connection to the DS4200 storage server. Refer to Figure 3-15.
- ▶ Spread expansion enclosures among the two loops pairs. For example, if you attach four EXP420 enclosures, it is better to have two EXP420s behind each drive port, rather than three and one.

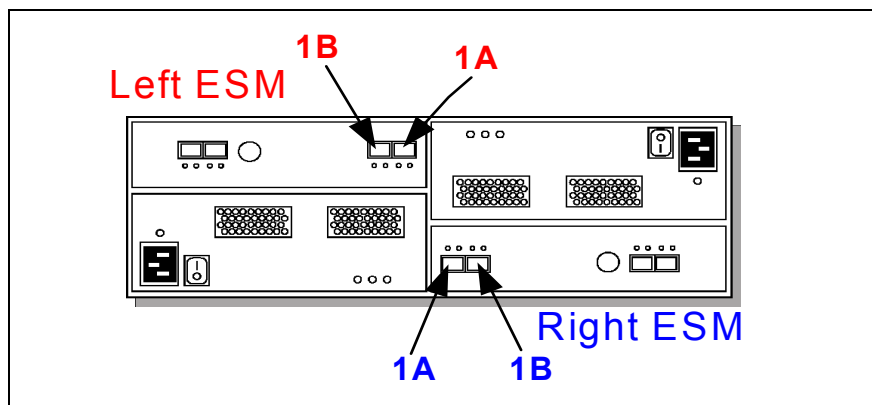


Figure 3-15 Port labels on EXP420

The drive-side cabling for the DS4200 depends on how many EXP420s you need to attach:

- If you attach only one enclosure, make sure that you have one connection to each of the controllers, thus using one of the two ports on each controller, as shown in Figure 3-16.

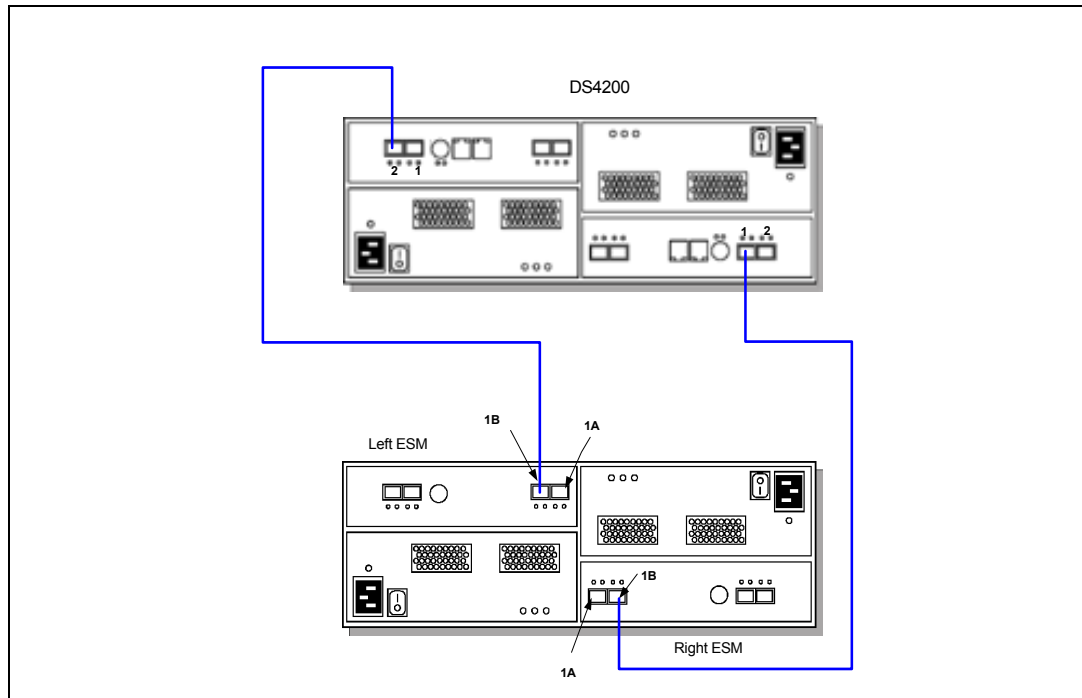


Figure 3-16 DS4200 drive cabling with one EXP420 enclosure

- If you attach a second EXP420, connect it by using the second port on the controller, as shown in Figure 3-17.

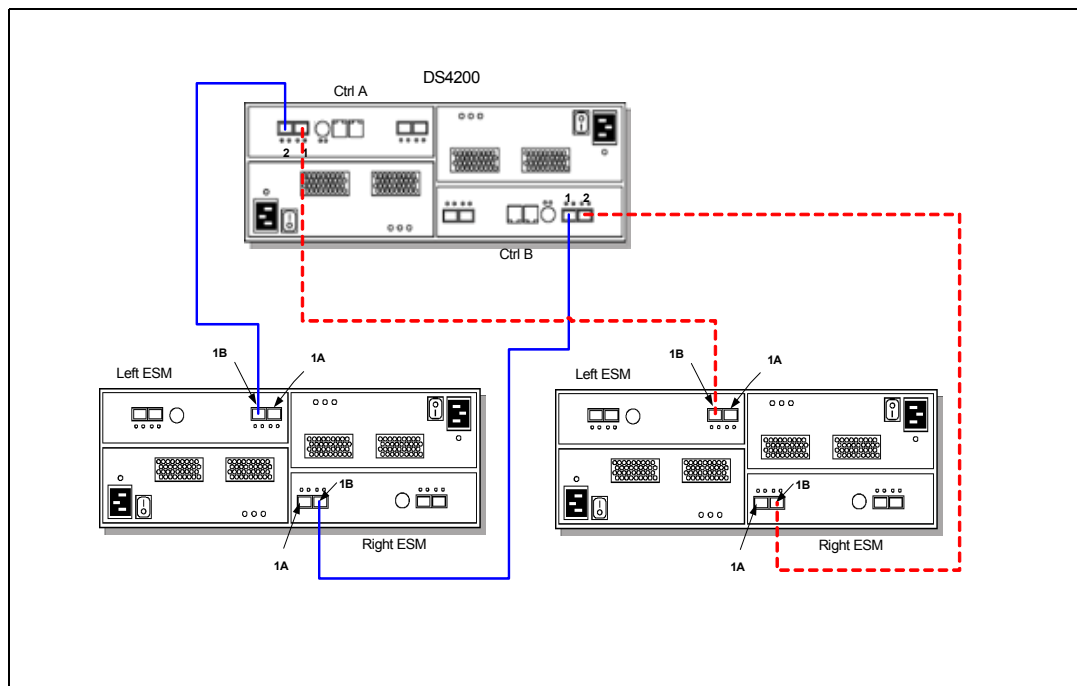


Figure 3-17 DS4200 drive cabling with two EXP420 enclosures

- Beyond two enclosures (up to a maximum of six, just make sure that you equally distribute the enclosures among the redundant loop pairs (see Figure 3-18 and Figure 3-19 on page 85).

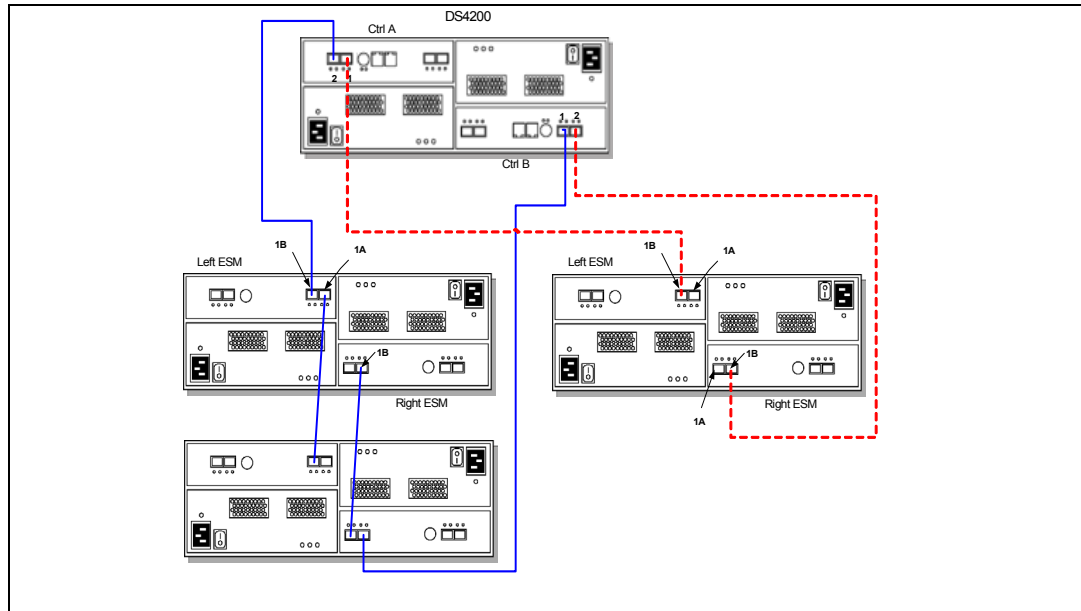


Figure 3-18 DS4200 drive cabling with three EXP420 enclosures

When six enclosures are required, the same method is employed again, maintaining loop redundancy and using both controllers. Refer to Figure 3-19.

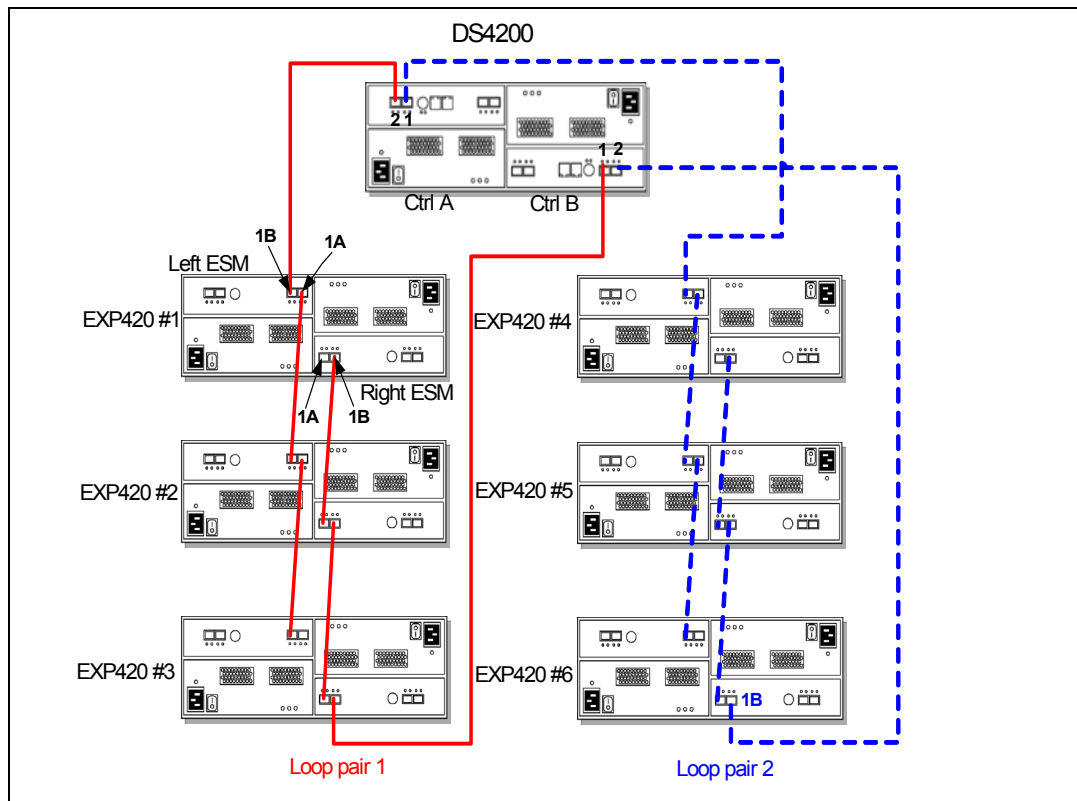


Figure 3-19 DS4200 drive cabling with six enclosures

The best sequence to populate loop pairs is as follows:

1. Controller A, port 2/controller B, port 1 (loop pair 1)
2. Controller A, port 1/controller B, port 2 (loop pair 2)

Drive-side cabling sequence

In the example Figure 3-19, the DS4200 is cabled using all two loop pairs, assuming that there are six total expansion enclosures evenly spread out across the loop pairs (three each).

Each EXP420 ESM only has one pair of ports, labelled 1A and 1B, that can be used to connect FC cables (the other pair of ports is reserved for future use). Proceed as follows:

1. Start with the first expansion enclosure, which we will attach to loop pair #1. Cable controller A, port 2 to the leftmost port (1B) of the first pair of ports of the left ESM of the first EXP420 unit.
2. Cable the rightmost port (1A) of the first pair of ports of the left ESM of the first EXP420 unit to the leftmost port (1B) of the first pair of ports on the left ESM of the second EXP420 unit.
3. Cable the rightmost port (1A) of the first pair of ports of the left ESM of the second EXP420 unit to the leftmost port (1B) of the first pair of ports on the left ESM of the third EXP420 unit.
4. Cable the rightmost port (1B) of the first pair of ports of the right ESM of the first EXP420 unit to the leftmost port (1A) of the first pair of ports of the right ESM of the second EXP420 unit.
5. Cable the rightmost port (1B) of the first pair of ports of the right ESM of the second EXP420 unit to the leftmost port (1A) of the first pair of ports of the right ESM of the third EXP420 unit.
6. Cable controller B, port 1 to the rightmost port (1B) of the first pair of ports of the right ESM of the third EXP420 unit located on the first loop pair. This is the last step of the first loop pair.
7. Repeat steps 1–6 (using the next drive-side loop pair ports) for the second loop pairs (three EXP420 units each).

3.2.5 DS4500 host cabling configuration

The DS4500 has up to four host mini-hubs. The mini-hubs numbered 1 and 3 correspond to the top controller (controller A), and mini-hubs 2 and 4 correspond to the bottom controller (controller B). These are illustrated in Figure 3-20.

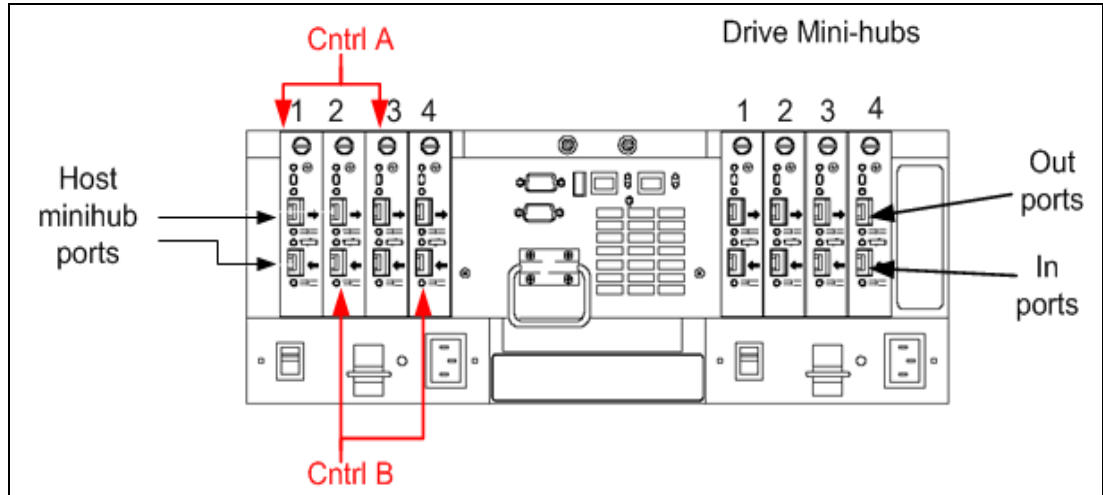


Figure 3-20 Rear view of the DS4500 Storage Server

Direct attached

To ensure redundancy, you must connect each host to both RAID controllers (A and B).

Figure 3-21 illustrates a direct connection of hosts (each host must be equipped with two host adapters).

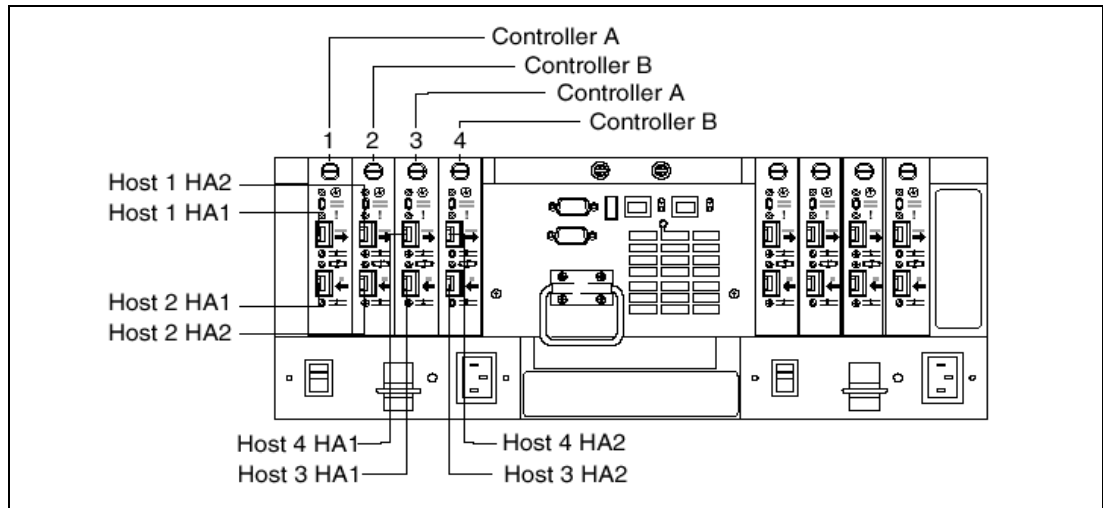


Figure 3-21 Connecting hosts directly to the controller

With AIX, only two hosts servers are supported in a direct attached configuration. These two hosts must be attached to their own separate pairs of mini-hubs and would therefore connect as Host 1 and Host 3 as shown above (for supported AIX configurations see Chapter 11, “DS4000 with AIX and HACMP” on page 339).

SAN attached

Figure 3-22 illustrates the recommended dual path configuration using Fibre Channel switches (rather than direct attachment). As stated earlier, this is the preferred best practice. Host 1 contains two HBAs that are connected through two switches to two separate host mini-hubs. So to configure a host with dual path redundancy, connect the first host bus adapter (HA1) to SW1, and HA2 to SW2. Then, connect SW1 to host mini-hub 1 and SW2 to host mini-hub 2 as shown.

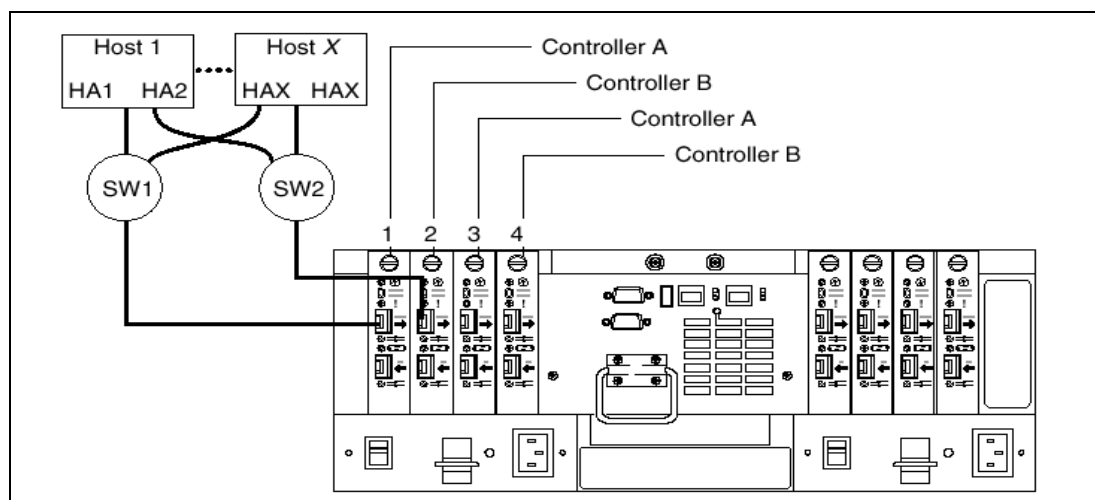


Figure 3-22 Using two Fibre Channel switches to connect a host

3.2.6 DS4500 drive expansion cabling

In this section we discuss the device cabling of the DS4500.

Devices can be dynamically added to the device mini-hubs. The DS4500 can support the following expansion types: EXP710, EXP100 (SATA), and the EXP810. Intermix of these expansions is supported with Version 6.19 of the firmware. For configuration details and limitations see “Intermixing drive expansion types” on page 101.

On the drive-side mini-hub, one SFP module port is marked as IN, the other one as OUT.

With the DS4500 it is important to ensure that the cabling rules apply.

- ▶ Drive-side mini hubs 1 and 3 are used for group one.
- ▶ Drive-side mini hubs 2 and 4 are used for group two.

This is shown in Figure 3-23.

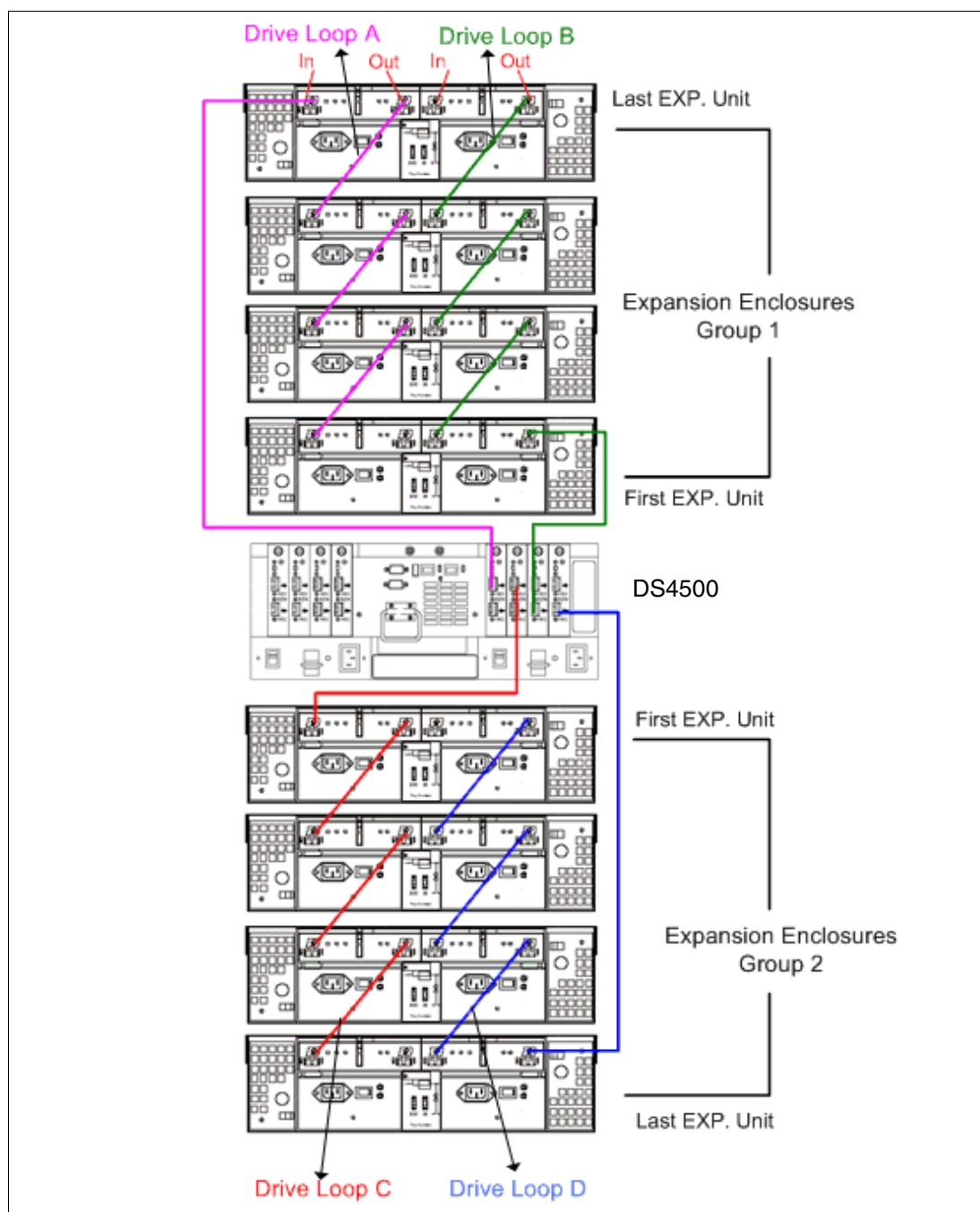


Figure 3-23 DS4500 drive-side Fibre Channel cabling

The steps are:

1. Starting with the first expansion unit of drive enclosures group 1 and connect the In port on the left ESM board to the Out port on the left ESM board of the second (next) expansion unit.
2. Connect the In port on the right ESM board to the Out port on the right ESM board of the second (next) expansion unit.
3. If you are cabling more expansion units to this group, repeat steps 1 and 2, starting with the second expansion unit.

4. If you are cabling a second group, repeat step 1 to step 3 and reverse the cabling order; connect from the Out ports on the ESM boards to the In ports on successive expansion units according to the illustration on the left. See Figure 3-20 on page 87.
5. Connect the Out port of drive-side mini-hub 4 (far left drive side) to the In port on the left ESM board of the last expansion unit in the drive enclosures group 1.
6. Connect the In port of drive-side mini-hub 2 to the Out port on the right ESM board of the first expansion unit in the drive enclosures group 1.
7. If you are cabling a second group, connect the Out port of the drive-side mini-hub 3 to the In port on the left ESM board of the first expansion unit in drive enclosures group 2. Then, connect the In port of the drive-side mini-hub 1 (far right drive side) to the Out port on the right ESM board of the last expansion unit in Drive enclosures group 2.
8. Ensure that each expansion unit has a unique ID (switch setting) and that the left and right ESM board switch settings on each expansion unit are identical.

3.2.7 DS4700 host cabling configuration

There is a total of eight host connections (four per controller) on Model 72, and a total of four host connections (two per controller) on Model 70. Each connection can operate at 4 Gbps but will auto-negotiate to support 2 Gbps and 1 Gbps connections as well. Host connections support Fibre Channel attachment through SAN switches and direct connections (Figure 3-24).

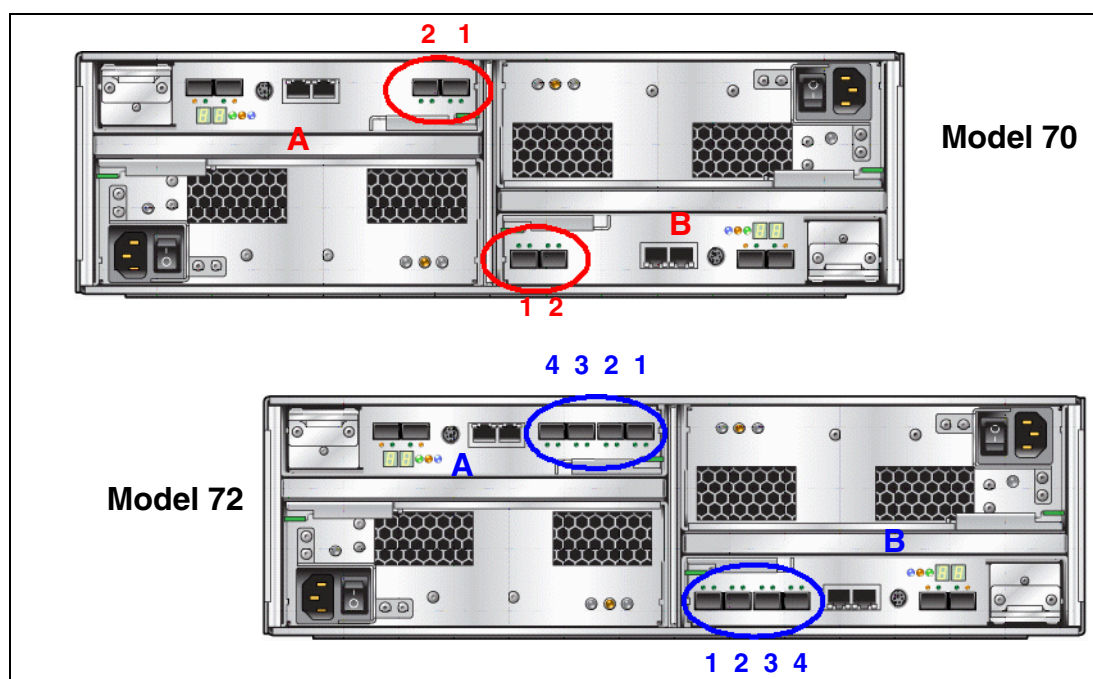


Figure 3-24 Model 70 - 2 host ports and Model 72 - 4 host ports

On Model 72, the host ports are labeled sequentially from 1 through 4, from left to right on controller B (bottom right). Conversely, they are labeled in reverse order, from 4 to 1, from the left to the right on controller A (top left).

On Model 70, the host ports are labeled sequentially from 1 through 2, from left to right, on controller B (bottom). Conversely, they are labeled in reverse order, from 2 to 1, from the left to the right on controller A (top).

Direct attached

Having eight independent host ports (Model 72) allows us to establish fully redundant direct connections to up to four hosts.

It is important to match up host or fabric connections to the DS4700 by attaching one connection to each controller. In doing so, you take advantage of the DS4700's ability to fail over and distribute the workload among the two controllers. For any given host, make sure to connect to the same host port number on each controller.

In Figure 3-25, a host is directly connected to each controller on the DS4700 using host port 1. By utilizing its four pairs of host ports, the DS4700 can support up to four directly connected hosts and maintain high availability.

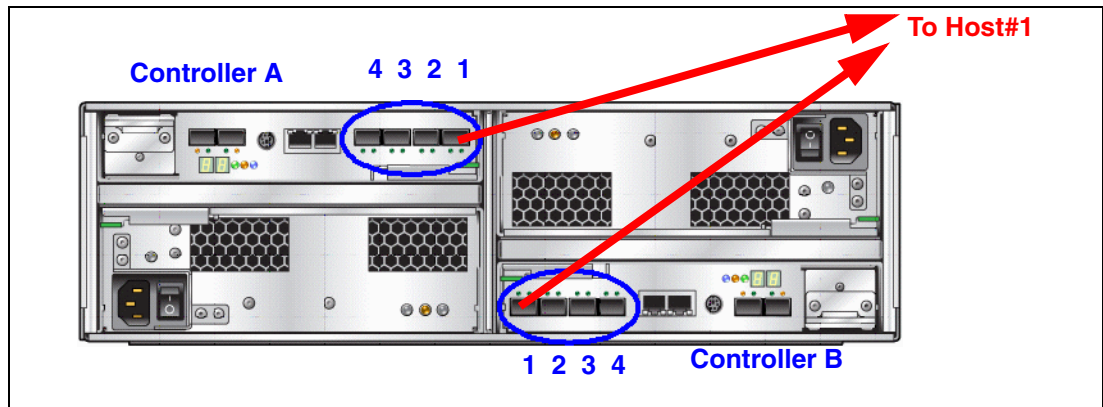


Figure 3-25 Directly connected host to DS4700 (Model 72)

SAN attached

The DS4700 also fully supports switched connections. Figure 3-26 depicts how the DS4700 would be connected into dual-redundant fabrics. Note how the same host port (1) is used on each controller to connect to the fabrics.

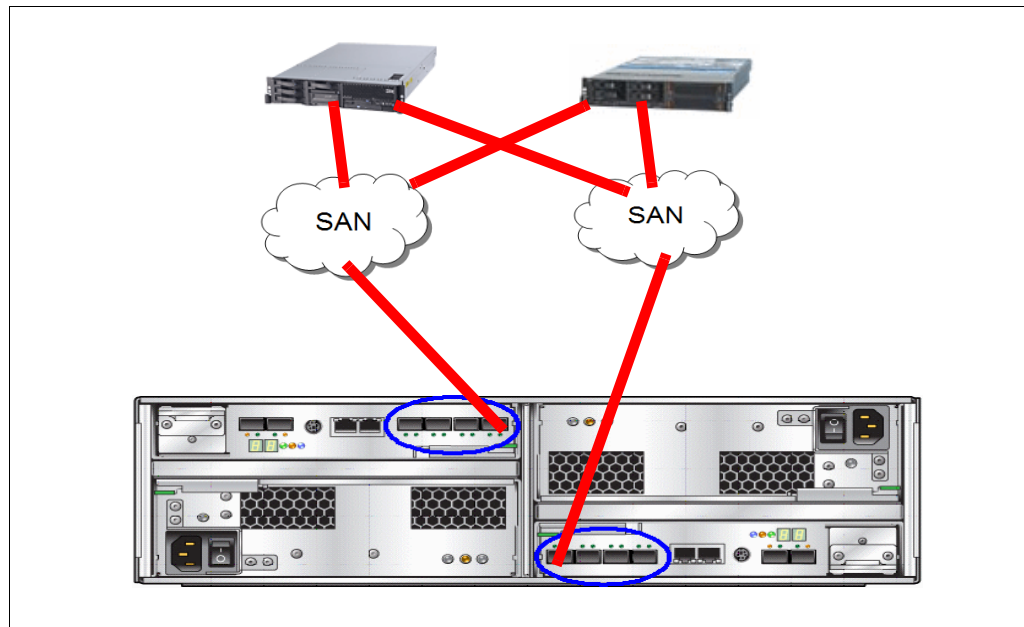


Figure 3-26 SAN connected hosts to DS4700 (Model 72)

3.2.8 DS4700 drive expansion cabling

There are four total drive-side connections (two on each controller). The drive-side connections operate at up to 4 Gbps (2 Gbps or 4 Gbps) and allow connection of disk expansion enclosures to the base controller unit. This is negotiated between the enclosures and the controller.

Since controller A is upside-down, the left-to-right numbering of the drive connection ports is reversed. This means that controller A is numbered left to right, 2 through 1, in reverse sequential order. controller B is numbered left to right, 1 through 2, in forward sequential order, as shown on Figure 3-27.

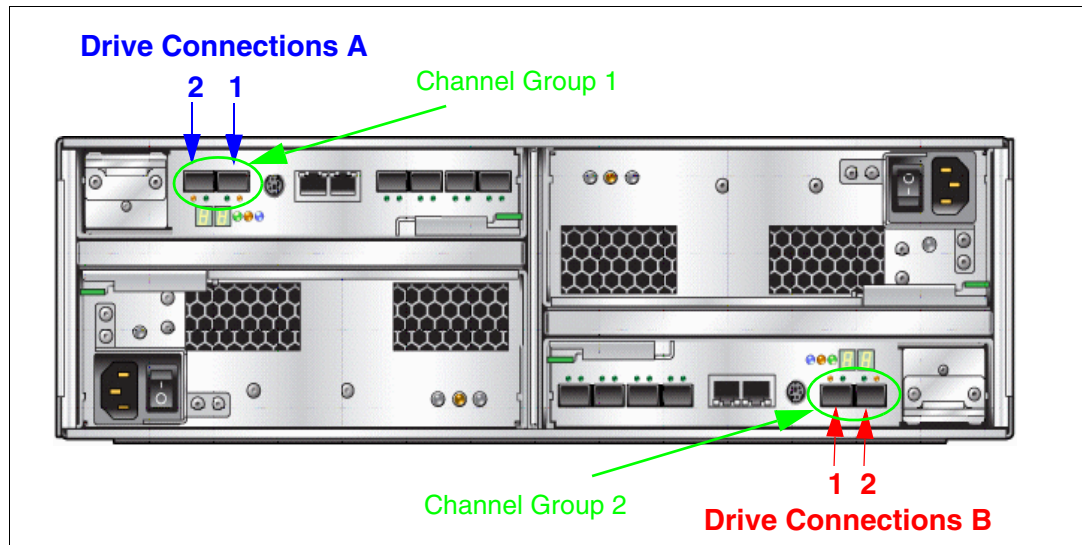


Figure 3-27 Drive-side connection of DS4700

The DS4700 can attach up to six EXP810 enclosures. It is generally best to spread enclosures evenly among the two loop pairs as you scale up your DS4700 in storage capacity. This allows you to fully utilize the maximum drive-side bandwidth. A fully configured DS4700 should have three expansion enclosures on each drive-side loop pair.

The drive-side cabling for the DS4700 depends on how many expansion units (usually EXP810) you need to attach (refer to the examples given for the DS4200 and EXP420 in Figure 3-16 on page 84 through Figure 3-18 on page 85).

Figure 3-28 shows a fully configured DS4700 with six EXP810s attached.

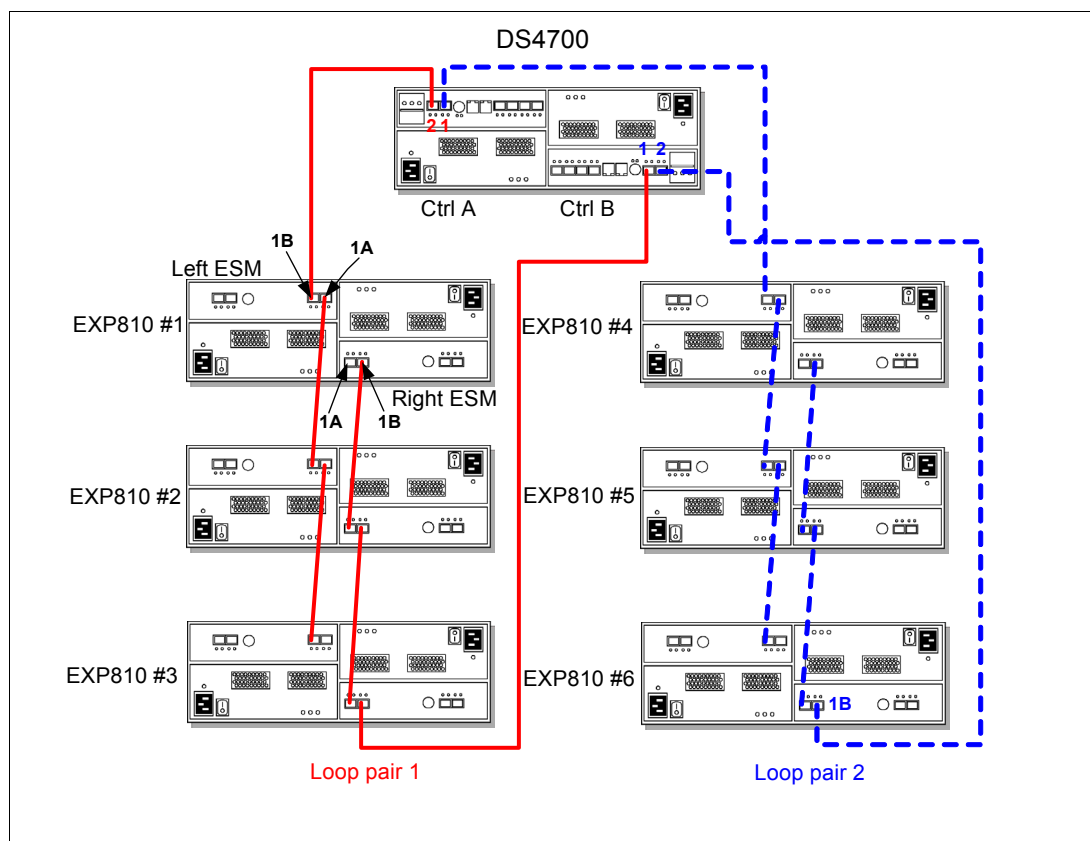


Figure 3-28 DS4700 with 6 EXP810s attached

If you were to implement a homogeneous environment (all 2 Gbps or all 4 Gbps), you should scale by adding enclosures one-by-one across all two loop pairs in an even distribution. Since all your drive-side loop pairs will operate at the same speed in a balanced configuration, scaling out in this manner is the most efficient method for maintaining high availability and performance.

The best sequence in which to populate loop pairs would be as follows:

1. Controller A, port 2/controller B, port 1 (loop pair 1)
2. Controller A, port 1/controller B, port 2 (loop pair 2)

Again, this sequence spreads out the workload among drive loop pairs and spreads the enclosures across drive channel groups in the most efficient fashion.

Drive-side cabling sequence

The DS4700 can use either the EXP710 or the EXP810 enclosures. In the example in Figure 3-28, the DS4700 is cabled using all two loop pairs, assuming that there are six total EXP810 expansion enclosures evenly spread out across the loop pairs (three each).

Each EXP810 ESM only has one pair of ports, labelled 1A and 1B, that can be used to connect FC cables (the other pair of ports is reserved for future use). Proceed as follows:

1. Start with the first expansion enclosure, which we will attach to loop pair #1. Cable controller A, port 2 to the leftmost port (1B) of the first pair of ports of the left ESM of the first EXP810 unit.

2. Cable the rightmost port (1A) of the first pair of ports of the left ESM of the first EXP810 unit to the leftmost port (1B) of the first pair of ports on the left ESM of the second EXP810 unit.
3. Cable the rightmost port (1A) of the first pair of ports of the left ESM of the second EXP810 unit to the leftmost port (1B) of the first pair of ports on the left ESM of the third EXP810 unit.
4. Cable the rightmost port (1B) of the first pair of ports of the right ESM of the first EXP810 unit to the leftmost port (1A) of the first pair of ports of the right ESM of the second EXP810 unit.
5. Cable the rightmost port (1B) of the first pair of ports of the right ESM of the second EXP810 unit to the leftmost port (1A) of the first pair of ports of the right ESM of the third EXP810 unit.
6. Cable controller B, port 1 to the rightmost port (1B) of the first pair of ports of the right ESM of the third EXP810 unit located on the first loop pair. This is the last step of the first loop pair.
7. Repeat steps 1–6 (using the next drive-side loop pair ports) for the second loop pairs (three EXP810 units each).

Note: There are three rules for the EXP810 cabling:

- ▶ Connect a maximum of three EXP810 enclosures per DS4700 controller drive port.
- ▶ The DS4700 controller drive port must always be connected to the EXP 810 port labelled 1B. Because the left and right EXP 810 ESMs (ESMs A and B) are inserted in the ESM bays in different orientations, ensure that you use the port labeled 1B before making the Fibre Channel connection to the DS4000 storage server. Refer to Figure 3-29.
- ▶ Also, as previously stated, spread expansion enclosures among the two loops pairs. For example, if you attach a maximum of six EXP810 enclosures, it is better to have three EXP810s behind each drive port rather than four and two.

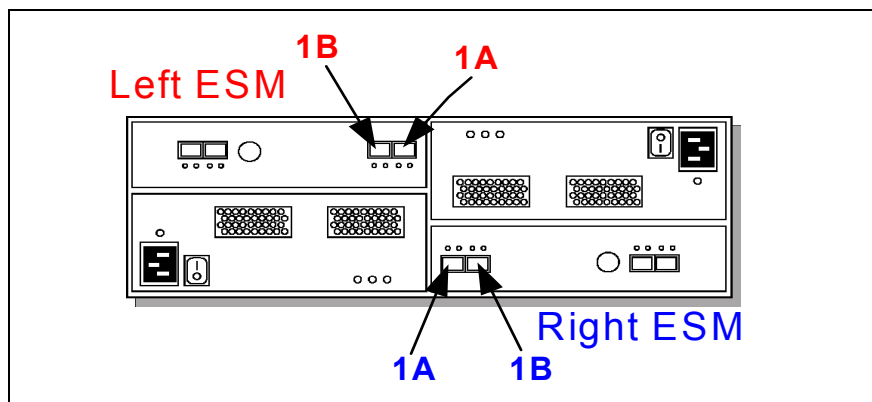


Figure 3-29 Port labels on EXP810

In addition, we also recommend that you do not intermix expansion enclosure types (EXP710 and EXP810) behind the same controller drive port. If you have a mix of EXP810 and EXP710, use a configuration similar to that shown in Figure 3-30, where EXP710 and EXP810 are attached to different drive ports.

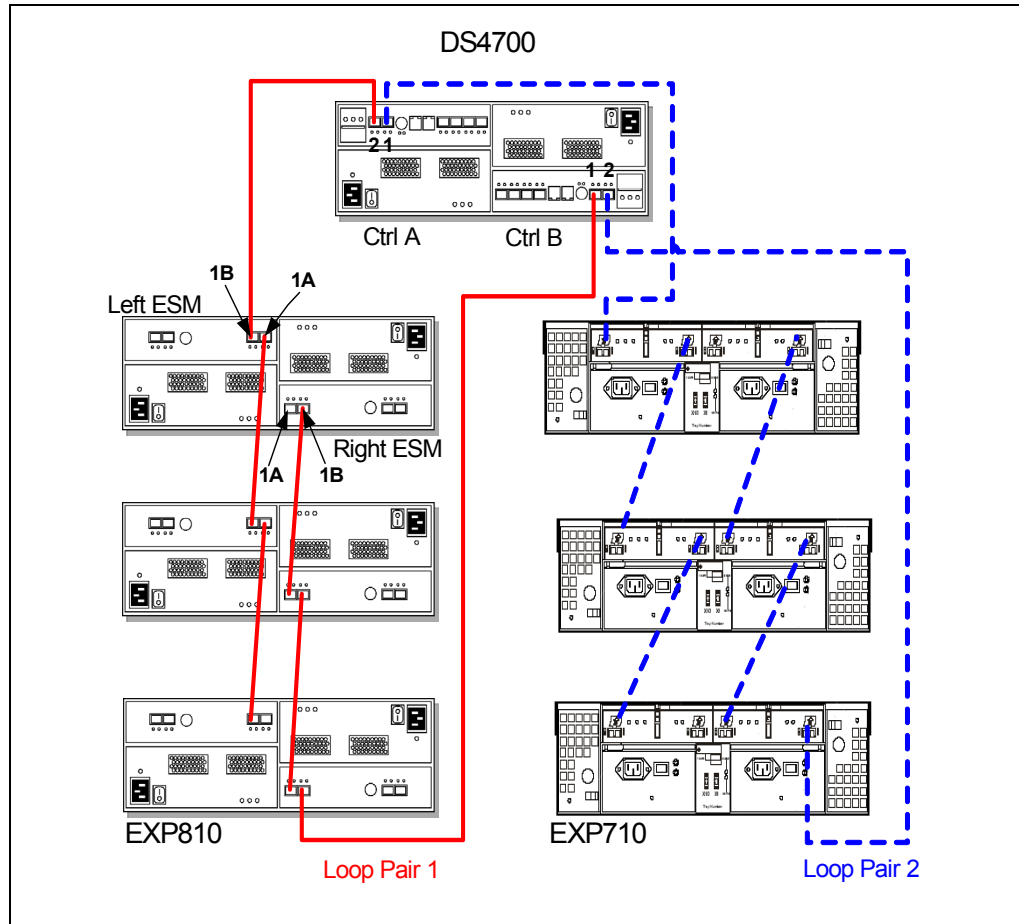


Figure 3-30 Connect EXP810 and EXP710 to different drive ports

Additional materials regarding cabling and setting up the DS4700 can be found on IBM Support Web site:

<http://www.ibm.com/servers/storage/support/disk/ds4700>

3.2.9 DS4800 host cabling configuration

The new DS4800 Storage Server supports a maximum of four independent host connection ports per storage controller. This enables it to support up to four dual-pathed hosts. As with previous DS4000 systems, the DS4800 supports direct-attached hosts. However, we recommend connecting using a SAN fabric (switch) environment. The direct-attached configuration is shown in Figure 3-31.

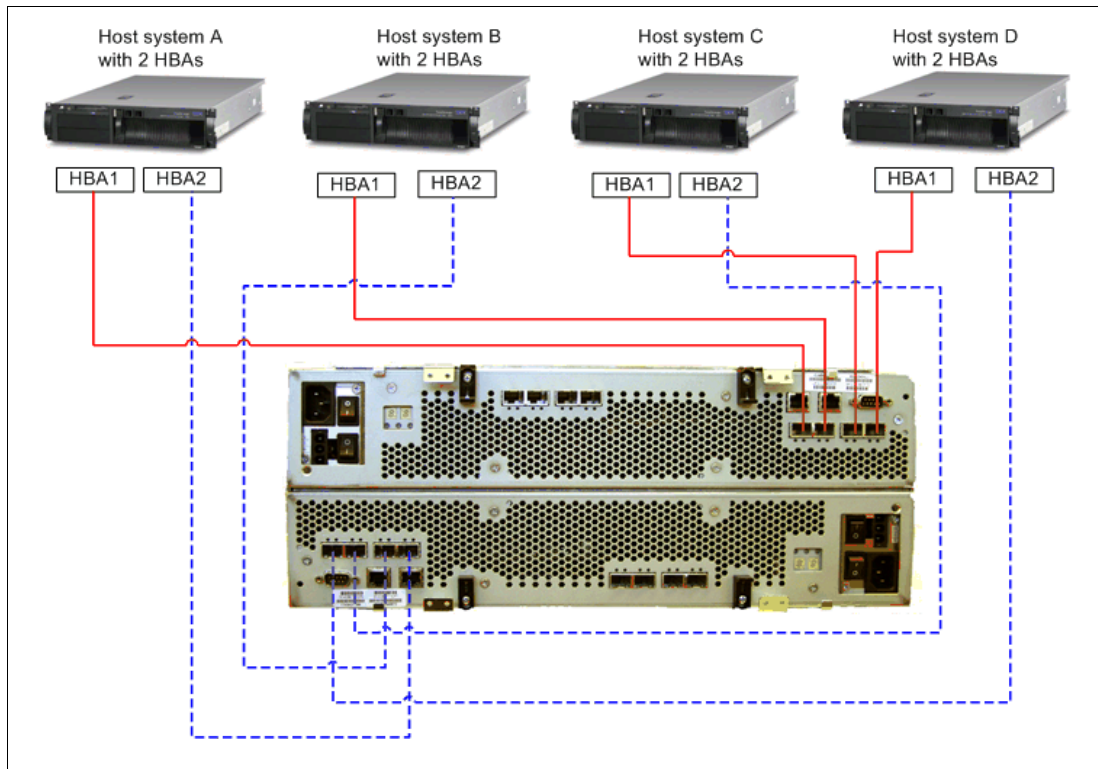


Figure 3-31 Basic DS4800 host side direct connect cabling

In the switched fabric environment, the best practice is to connect to dual fabrics by attaching the two DS4800 controllers to two independent switches (one to each controller), and then attaching each switch to an HBA in each of your host servers. If supported by the host and the operating system; additional ports may be connected for redundancy and load balancing. When adding anything greater than one initiator and one target to a fabric zoning of the switches, we recommend, as a best practice, to ensure that the fabric's reliability and ultimately performance is not impacted.

It should also be noted that port 4 of each controller is defined as the Enhanced Remote Mirroring (ERM) port, and cannot be used for host access if an ERM network is planned to be used. A basic switch attached configuration is shown in Figure 3-32.

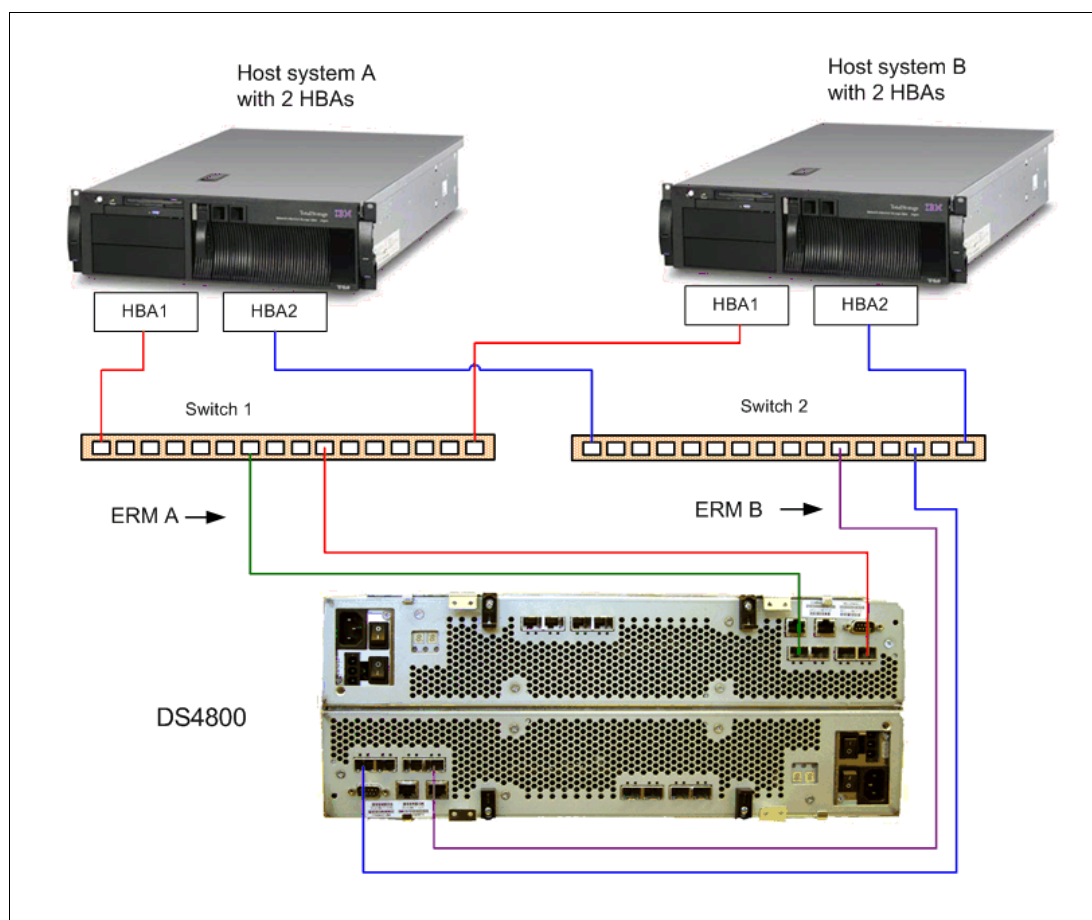


Figure 3-32 Basic DS4800 host-side switch-attached cabling

3.2.10 DS4800 drive expansion cabling

In the initial installation of a DS4800, you can add only new storage expansion enclosure and drives to the DS4800 Storage Subsystem. This means that there must be no existing configuration information on the storage expansion enclosures that you want to install.

If the storage expansion enclosures that you want to install currently contain logical drives or configured hot-spares, and you want them to be part of the DS4800 Storage Subsystem configuration, refer to *IBM DS4000 Hard Drive and Storage Expansion Enclosure Installation and Migration Guide*. Improper drive migration might cause loss of configuration and other storage subsystem problems. Contact your IBM support representative for additional information.

With the DS4800 Storage Servers, the recommended drive expansion cabling method is to connect drive channel 1 of controller A (ports 4 and 3) and drive channel 3 of controller B (ports 1 and 2) to form a DS4800 Storage Subsystem redundant drive loop/channel pair (loop pairs 1 and 2 in Figure 3-33 on page 98). If any component of drive channel 1 fails, the RAID controllers can still access the storage expansion enclosures in redundant drive loop 1 through drive channel 3. Similarly, drive channel 2 of controller A (ports 2 and 1) and drive channel 4 of controller B (port 4 and 3) combine to form the second set of redundant drive

loop/channel pairs (loop pairs 3 and 4 in Figure 3-35 on page 99). If any component of drive channel 2 fails, the RAID controllers can still access the storage expansion enclosures in redundant drive loop pairs 3 and 4, through drive channel 4.

Figure 3-33 shows the storage expansion enclosures in each drive loop pairs connected to only one drive port in the two-ported drive channel. For example, in drive channel/loop pair 1, only port 4 of channel 1 and port 1 of channel 3 are used. This results in only half of the storage expansion enclosures in the redundant drive loop pair being connected to the first port of the dual-ported drive channel. The other half of the enclosures are connected to the second port of the dual-ported drive channels.

Figure 3-33 through Figure 3-35 on page 99 are examples of different quantities of drive expansion being installed with the DS4700 or DS4800 Server.

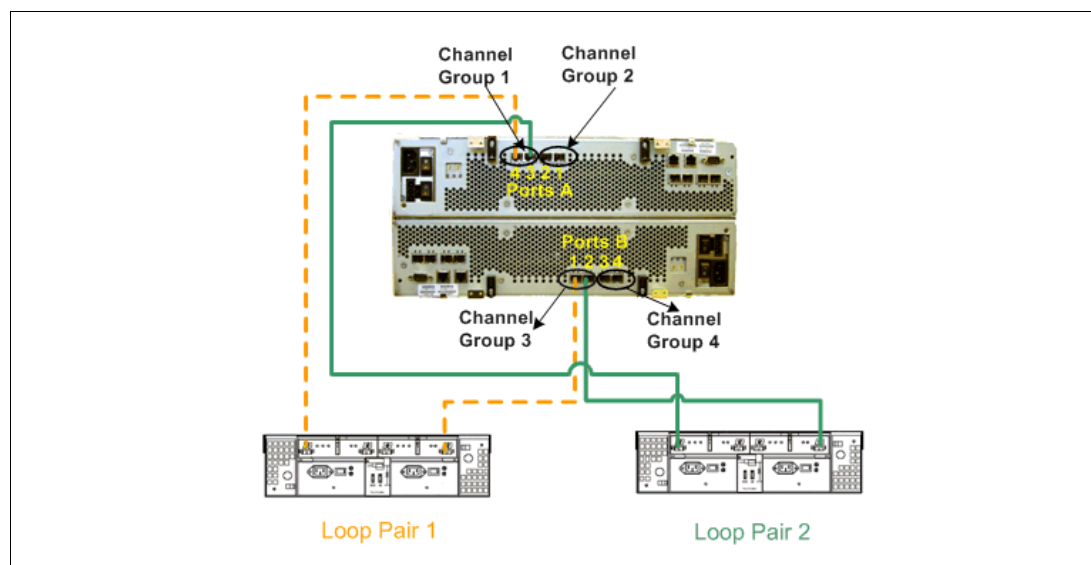


Figure 3-33 DS4800 with two EXP710 drive expansion enclosures

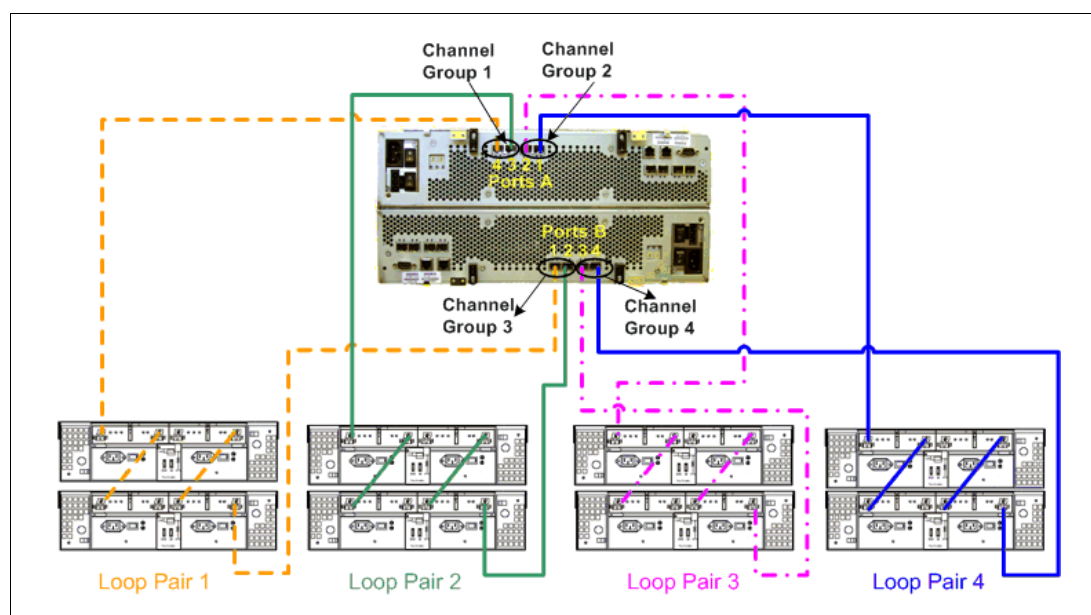


Figure 3-34 DS4800 with eight EXP710 drive expansion trays

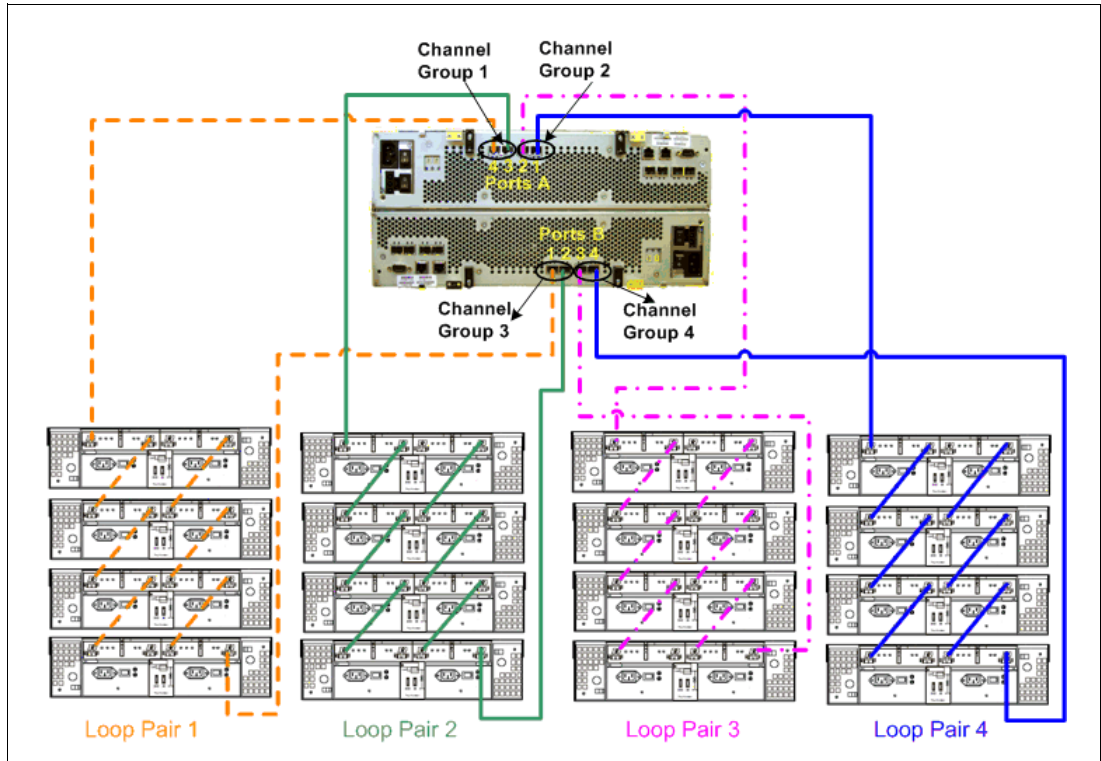


Figure 3-35 DS4800 with maximum (16) EXP710 drive expansion trays

To have a fully 4 Gbps SAN solution, the DS4800 must use the EXP810 4 Gbps enclosures. The EXP810 enclosure takes up to 16 drives. This means that a maximum of 14 enclosures may be connected to the DS4800.

There are three rules for the EXP810 cabling (Figure 3-36 on page 100):

- ▶ Connect a maximum of four EXP810 enclosures per DS4800 controller drive port.
- ▶ The DS4000 controller drive port must always be connected to the EXP810 port labelled 1B. Because the left and right EXP810 ESMs (ESMs A and B) are inserted into the ESM bays in different orientations, ensure that the port is labeled 1B before making the Fibre Channel connection to the DS4000 storage server.

- Also, as previously stated, spread expansion enclosures among the four loops pairs. Using only the EXP810 enclosures, a maximum of 14 enclosures can be connected to a DS4800. See Figure 3-36.

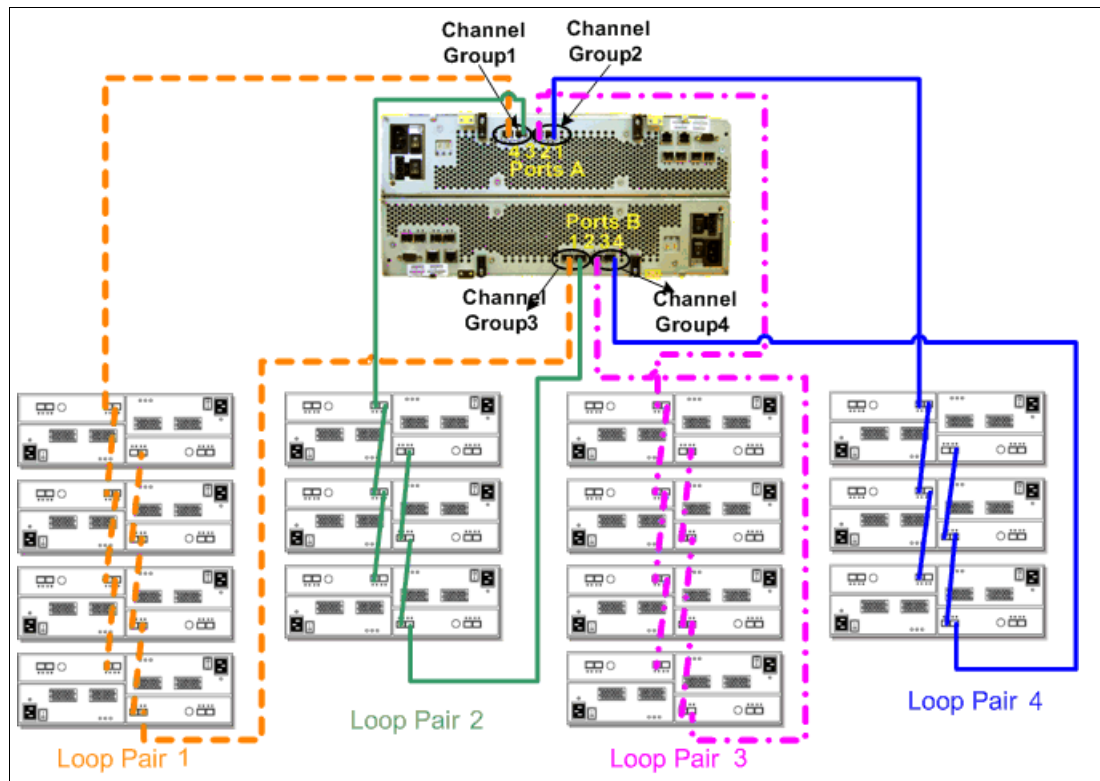


Figure 3-36 Cabling with the DS4800 and EXP810 enclosures

3.2.11 Expansion enclosures

The DS4000 Storage Server offers disk expansions of both Fibre Channel drives and SATA drive types. Support of expansions may be limited on certain models of the DS4000 Storage Server. Table 3-1 shows the DS4000 Storage Servers and their supported expansions.

Table 3-1 DS4000 Storage Servers and expansions support

DS4000 Storage Server	EXP100	EXP420	EXP700	EXP710	EXP810
DS4100	Supported	Not supported	Not supported	Not supported	Not supported
DS4200	Not supported	Supported	Not supported	Not supported	Not supported
DS4300	Supported	Not supported	Supported	Supported	Supported with 6.19 firmware
DS4500	Supported	Not supported	Supported	Supported	Supported with 6.19 firmware
DS4700	Not supported	Not supported	Not supported	Supported	Supported

DS4000 Storage Server	EXP100	EXP420	EXP700	EXP710	EXP810
DS4800	Supported with 6.15 firmware only	Not supported	Not supported	Supported	Supported with 6.16 firmware or later

Intermixing drive expansion types

When intermixing expansion types on a DS4000 Storage Server, you need to follow special rules with your configuration.

Best practice: When intermixing Fibre Channel and SATA expansions on the same drive loop be conscious of the performance characteristics of your application. Balancing the I/Os across the total drive-side bandwidth can be considered only when the different application workloads have similar characteristics.

For support of the Fibre Channel and SATA intermix, the FC/SATA Intermix feature key is required. This feature is purchased separately and installed as a premium feature key (see also 2.4.4, “FC/SATA Intermix” on page 59).

For detailed instructions for any intermix configuration refer to the *Installation and Migration Guide - IBM TotalStorage DS4000 hard drive and Storage Expansion Enclosure*, GC26-7849, available at:

<http://www-1.ibm.com/support/docview.wss?uid=psg1MIGR-57818>

Limit the number of drives installed in the DS4000 to 80% of maximum capacity. In an intermix environment, with a maximum number allowed of 224, 180 drives would be the maximum number recommended.

Dependant upon firmware version, EXP710, EXP100, and EXP810 enclosures can be intermixed. Below is a description of the intermixing details:

- ▶ With Version 6.15 firmware, the intermix of EXP710 and EXP100 is supported with the DS4800.
- ▶ With the Version 6.16 firmware, the intermix of EXP710 and EXP810 is supported on the DS4800, but the EXP100 was no longer supported with this firmware version.
- ▶ With the Version 6.16 firmware, the intermix of the EXP710 and EXP810 is supported on the DS4700.
- ▶ With Version 6.19 firmware, the intermix of EXP710, EXP100, and EXP810 is supported with the DS4300 and DS4500.

EXP810, EXP700, EXP 710, and EXP100 can all be intermixed on the DS4300 and DS4500. However, our recommendation is that when intermixing EXP700 and EXP710 on the same loops, group the EXP710s together. The best practice is to upgrade the EXP700 ESM to the EXP710 ESM.

When using the EXP810 with the DS4500 and the DS4800, the maximum number of drives cannot exceed 112 per redundant drive channel/loop pair.

Below is a series of tables that contain the intermix drive configurations. Refer to Table 3-2 for the DS4800 and DS4500. Refer to Table 3-3 on page 102 for the DS4700, and Table 3-4 on page 103 for the DS4300.

Table 3-2 Supported EXP810 and EXP710 enclosures per drive loop on DS4800 and DS4500

Number of EXP810s	Total number of drives in EXP810	Number of EXP710s	Total number of drives in EXP710s	Total number of drives in a mixed EXP710 and EXP810 drive loop
0	0	8	112	112
0	0	7	98	98
1	16	6	84	100
2	32	5	70	102
3	48	4	56	104
4	64	3	42	106
5	80	2	28	108
6	96	1	14	110
7	112	0	0	112

It should be noted that when mixing EXP710 and EXP810 enclosures, the maximum number of drives per redundant drive channel/loop pair varies. With the DS4500 you can intermix either the EXP710 or EXP100.

Intermixing the EXP100 and EXP810 on the DS4800 is not yet supported.

Table 3-3 Supported EXP810 and EXP710 enclosures with the DS4700

Number of EXP810s	Total Number of drives in EXP810	Number of EXP710s	Total number of drives in EXP710s	Drives in DS4700	Total number of drives in mixed EXP810 and EXP710
0	0	6	84	16	100
1	16	5	70	16	102
2	32	4	56	16	104
3	48	3	42	16	106
4	64	2	28	16	108
5	80	1	14	16	110
6	96	0	0	16	112

It should be noted that when mixing EXP710 and EXP810 enclosures, the maximum number of drives per redundant drive channel/loop pair varies. With the DS4700 we recommend that the EXP710s are on their own drive loops.

Table 3-4 Supported EXP810 and EXP710/EXP100 enclosures with the DS4300 Turbo

Number of EXP810s	Total number of drives in EXP810	Number of EXP710s or EXP100	Total number of drives in EXP710s and or EXP100s	Drives in DS4300	Total number of drives in mixed EXP810 and EXP710 and or EXP100s
0	0	7	98	14	112
1	16	5	70	14	100
2	32	4	56	14	102
3	48	3	42	14	104
4	64	2	28	14	106
5	80	1	14	14	108
6	96	0	0	14	110

Note: We recommend that you use EXP810 with the DS4300, DS4500, DS4700, and DS4800 storage servers.

Using the EXP810s with the DS4700 and DS4800 will allow the storage server to make full use of the 4 Gbps technology.

Expansion enclosure ID addressing

It is very important to correctly set the expansion enclosure ID switches on the EXPs. The IDs are used to differentiate multiple EXP enclosures that are connected to the same DS4000 Storage Server. Each EXP must use a unique value. The DS4000 Storage Manager uses the expansion enclosure IDs to identify each DS4000 EXP enclosure.

Additionally, the Fibre Channel loop ID for each disk drive is automatically set according to the following items:

- The EXP bay where the disk drive is inserted
- EXP ID setting

It is important to avoid hard ID contention (two disks having the same ID on the loop). Such contention can occur when the units digit of two drive expansions on the same drive side loop are identical. For example, expansions with IDs 0 and 10, 4 and 14, and 23 and 73, could all have hard ID contention between devices.

For additional details refer to 2.2.7, “Disk expansion enclosures” on page 32.

3.3 Configuring the DS4000 Storage Server

Now that you have set up the storage server and it is connected to a server or the SAN, you can proceed with additional configuration and storage setting tasks. If there is previous configuration data on the DS4000 storage server that you wish to be able to reference, then first save a copy of the *storage subsystem profile* to a file. Once you have completed your changes, you should save the profile to a (different) file as well. This will be of great value when you are discussing questions or problems with support; or reviewing your configuration with your performance data.

Best practice: You should save a new profile each time you change the configuration of the DS4000 storage subsystem. This applies to all changes regardless of how minor they might be. The profile should be stored in a location where it is available even after a complete configuration loss, for example, after a site loss.

The configuration changes you desire can be done using the new storage subsystem level Task Assistant capabilities with the SM Client 9.1x code, or by following the steps outlined here.

Before defining arrays or logical drives, you must perform some basic configuration steps. This also applies when you reset the configuration of your DS4000 Storage Server:

1. If you install more than one DS4000 Storage Server, it is important to give them literal names. To name or rename the DS4000 Storage Server, open the Subsystem Management window. Right-click the subsystem, and click **Storage Subsystem** → **Rename**.
2. Because the DS4000 Storage Server stores its own event log, synchronize the controller clocks with the time of the host system used to manage the DS4000 units. If you have not already set the clocks on the storage servers, set them now. Be sure that your local system is working using the correct time. Then, click **Storage Subsystem** → **Set Controller Clock**.

Note: Make sure the time of the controllers and the attached systems are synchronized. This simplifies error determination when you start comparing the different event logs. A network time server can be useful for this purpose.

3. For security reasons, especially if the DS4000 Storage Server is directly attached to the network, you should set a password. This password is required for all actions on the DS4000 Storage Server that change or update the configuration in any way.

To set a password, highlight the storage subsystem, right-click, and click **Set Password**. This password is then stored on the DS4000 Storage Server. It is used if you connect through another DS4000 client. It does not matter whether you are using in-band or out-of-band management.

3.3.1 Defining hot-spare drives

Hot-spare drives are special, reserved drives that are *not* normally used to store data. When a drive in a RAID array with redundancy, such as 1, 3, 5, or 10, fails, the hot-spare drive takes on the function of the failed drive and the data is rebuilt on the hot-spare drive, which becomes part of the array. After this rebuild procedure, your data is again fully protected. A hot-spare drive is like a replacement drive installed in advance.

If the failed drive is replaced with a new drive, the data stored on the hot-spare drive is copied back to the replaced drive, and the original hot-spare drive that is now in use becomes a free hot-spare drive again. The location of a hot-spare drive is fixed and does not wander if it is used.

A hot-spare drive defined on the DS4000 Storage Server is always used as a so-called *global hot-spare*. That is, a hot spare drive can always be used for a failed drive. The expansion or storage server enclosure in which it is located is not important.

A hot-spare drive must be of the same type (FC or SATA), and at least of the capacity of the configured space on the failed drive. The DS4000 Storage Server can use a larger drive to recover a smaller failed drive to it. It will not use smaller drives to recover a larger failed drive. If a larger drive is used, the remaining excess capacity is blocked from use.

Best practice: We recommend that you use a ratio of one hot-spare for every 28 drives, or one for every two fully populated chassis (controller or enclosure). A pool of up to 15 hot-spare drives can be defined for a given storage subsystem.

When a drive failure occurs on a storage server configured with multiple hot-spare drives, the DS4000 Storage Server will attempt to find a hot spare drive in the enclosure with the failed drive first. It will find a drive that is at least the same size as the failed drive, but not necessarily giving preference to one the exact same size as the failed drive. If a match does not exist in the same enclosure, it will look for spares in the other enclosures that will contain sufficient capacity to handle the task.

The controller uses a free hot-spare drive as soon as it finds one, even if there is another one that might be closer to the failed drive.

To define a hot-spare drive, highlight the drive you want to use. From the Subsystem Management window, click **Drive** → **Hot Spare Coverage** → **Manually assign individual drives**.

If there are larger drives defined in any array on the DS4000 Storage Server than the drive you chose, a warning message appears and notifies you that not all arrays are protected by the hot-spare drive. The newly defined hot-spare drive then has a small red cross in the lower part of the drive icon.

Especially in large configurations with arrays containing numerous drives, we recommend the definition of multiple hot-spare drives, because the reconstruction of a failed drive to a hot spare drive and back to a replaced drive can take a long time. See also 2.3.3, “Hot spare drive” on page 48.

Best practice: In large configurations with many drives (more than 30), define multiple hot-spare drives.

Ensure that at least one hot spare is as large as the largest size drive in use in the storage server.

If different speed (10K or 15K) drives are used, the fastest hot spares are used so as to not slow down an array if the hot spare is required.

To unassign a hot-spare drive and have it available again as a free drive, highlight the hot-spare drive and select **Drive** → **Hot Spare Coverage** → **Manually unassign individual drives**.

3.3.2 Creating arrays and logical drives

At this stage, the storage subsystem has been installed and upgraded to the newest microcode level. You can now configure the arrays and logical drives according to your requirements. With the SM Client you can use an automatic default configuration mode to configure the arrays and LUNs for the sizes you want and the RAID type you select. If you have planned your configuration for maximum performance and reliability as discussed and recommended in 2.3.1, “Arrays and RAID levels” on page 37, and also in “Arrays and logical drives” on page 155, you will need to define them manually to enter your specific configuration and needs.

Best practice: When defining the arrays and logical drives for your configuration, we recommend that you plan your layout and use the manual configuration method to select the desired drives and specify your settings.

If you need any further assistance with how to define the available drives into the arrays you want, or logical drive definitions, or which restrictions apply to avoid improper or inefficient configurations of the DS4000 Storage Server see *IBM System Storage IBM System Storage DS4000 Series and Storage Manager*, SG24-7010, available at:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247010.pdf>

To create arrays and logical drives, you can define logical drives from unconfigured-capacity or free-capacity nodes on the storage subsystem.

The main difference is that you have to decide whether to use unconfigured capacity on free disks or free capacity in an already existing array:

- ▶ When you create a logical drive from unconfigured capacity, you create an array and the logical drive at the same time. Note that the unconfigured capacity for Fibre Channel and SATA disks are grouped separately.
- ▶ When you create a logical drive from free capacity, you create an additional logical drive on an already existing array from free unconfigured space that is available.

While creating your logical drives, select **Customize settings** to be able to set specific values for the cache settings, and segment size for the logical drive options.

1. In the Subsystem Management window, right-click the unconfigured capacity and select **Create Logical Drive**.

This action starts the wizard for creating the logical drives. The first window of the wizard is an introduction to the process, as shown in Figure 3-37. Read the introduction and then click **Next** in order to proceed.

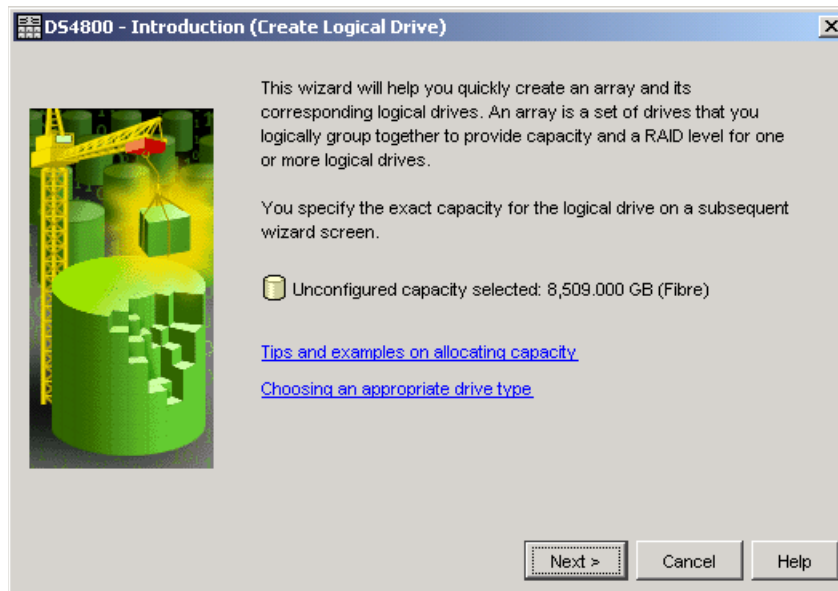


Figure 3-37 Create Logical Drive

2. You have two choices for specifying the array details: automatic and manual. The default is the automatic method. In automatic mode, the RAID level is used to create a list of available array sizes. The Storage Manager software selects a combination of available drives which it believes will provide you with the optimal configuration (Figure 3-38).

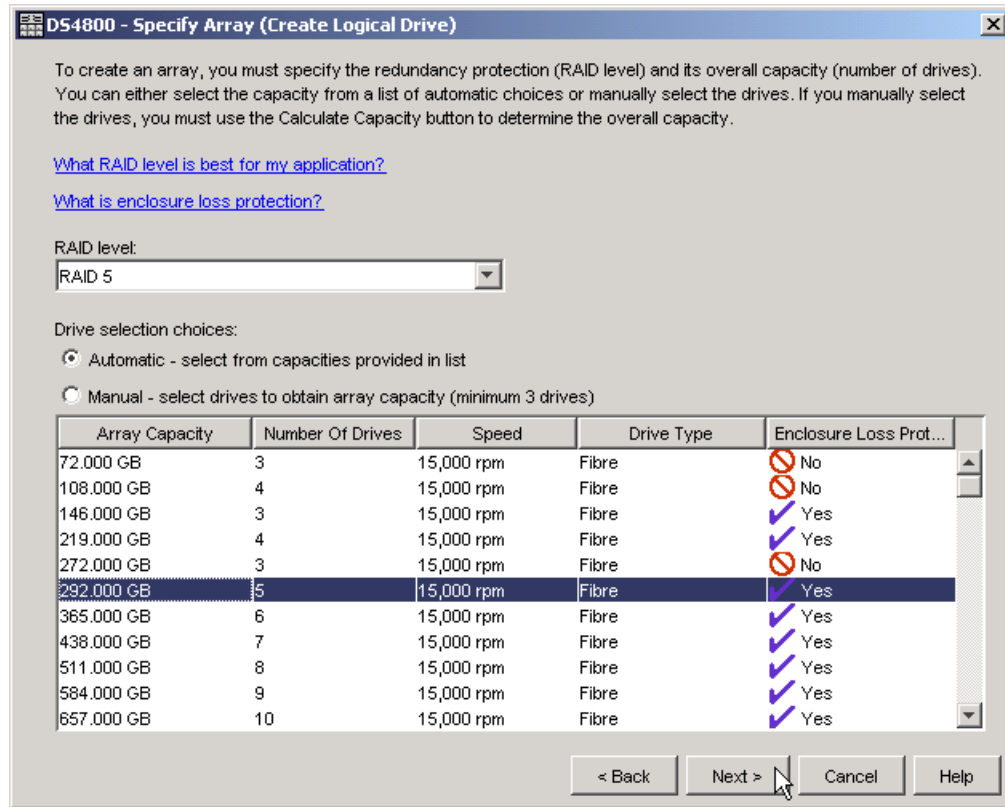


Figure 3-38 Enclosure loss protection with Automatic configuration

To define your specific layout as planned to meet performance and availability requirements, we recommend that you use the manual method.

Best practice: When defining the arrays and logical drives for your configuration, we recommend that you plan your layout and use the manual configuration method to select the desired drives and specify your settings.

Where possible, always ensure that enclosure loss protection is used for your arrays.

The manual configuration method (Figure 3-39) allows for more configuration options to be available at creation time, as discussed in “Enclosure loss protection planning” on page 43.

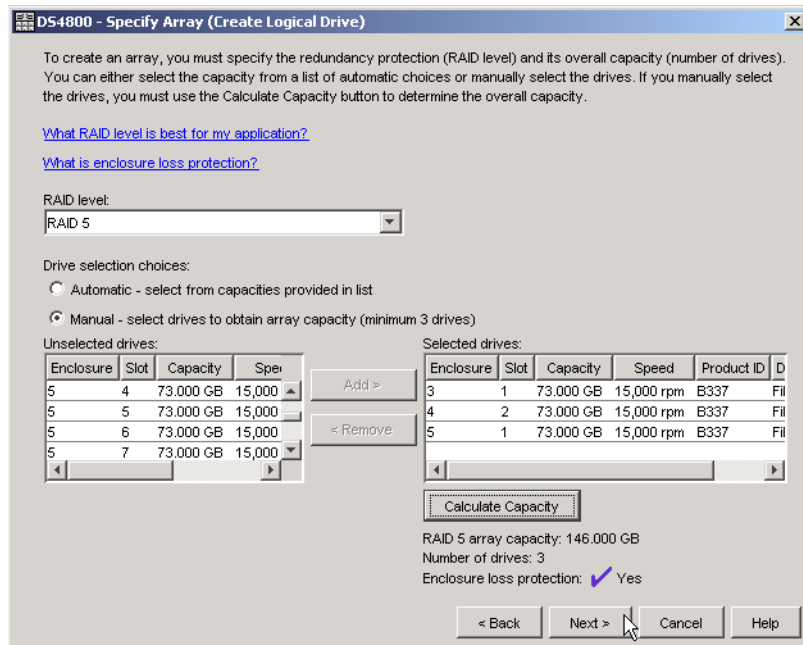


Figure 3-39 Manual configuration for enclosure loss protection

Select the drives for the array. Ensure that the drives are staggered between the enclosures so that the drives are evenly distributed on the loops. To select drives, hold down the Ctrl key and select the desired unselected drives, then click **Add**. In order to proceed, you must click **Calculate Capacity**. When enclosure loss protection is achieved, then a checkmark and the word Yes will appear in the bottom right of window, otherwise it will have a circle with a line through it and the word No.

3. The Specify Capacity dialog appears. By default, all available space in the array is configured as one logical drive, so:
 - a. If you want to define more than one logical drive in this array, enter the desired size.
 - b. Assign a name to the logical drive.
 - c. If you want to change advanced logical drive settings such as the segment size or cache settings, select the **Customize settings** option and click **Next**.

Best practice: We strongly recommend that you leave some free space in the array even if you only need a single **logical drive** in that array.

4. The Customize Advanced Logical Drive Parameters dialog appears. You can customize the logical drive using predefined defaults, or manually change the cache read ahead multiplier, segment size, and controller ownership.
 - a. For logical drive I/O characteristics, you can specify file system, database, or multimedia defaults. The Custom option allows you to manually select the cache read ahead multiplier and segment size.
 - b. The segment size is chosen according to the usage pattern. For custom settings, you can directly define the segment size.

- c. As discussed earlier the cache read ahead setting is really an on or off decision since the read ahead multiplier is dynamically adjusted in the firmware. It should be set to any non-zero value to enable the dynamic cache read-ahead function.
- d. The preferred controller handles the logical drive normally if both controllers and I/O paths are online. You can distribute your logical drives between both controllers to provide better load balancing between them. The default is to alternate the logical drives on the two controllers.
- e. Obviously it is better to spread the logical drives by the load they cause on the controller. It is possible to monitor the load of each logical drive on the controllers with the Performance Monitor and change the preferred controller in case of need.

When you have completed setting your values as desired (either default or custom), click **Next** to complete the creation of the logical drive.

5. The Specify Logical Drive-to-LUN Mapping dialog appears. This step allows you to choose between default mapping and storage partitioning. Storage partitioning is a separate licensing option. In the case of storage partitioning being used, you must select **Map later using the Mappings View**.

If you choose **Default mapping**, then the physical volume will be mapped to the default host group. If there are host groups or hosts defined under the default host group, they will all be able to access the logical drive.

If the logical drive is smaller than the total capacity of the array, a window opens and asks whether you want to define another logical drive on the array. The alternative is to leave the space as unconfigured capacity. After you define all logical drives on the array, the array is now initialized and immediately accessible.

If you left unconfigured capacity inside the array, you can define another logical drive later in this array. Simply highlight this capacity, right-click, and choose **Create Logical Drive**. Simply follow the steps that we outlined in this section, except for the selection of drives and RAID level. Because you already defined arrays that contain free capacity, you can choose where to store the new logical drive, on an existing array or on a new one.

As stated earlier, it is always good to leave a small portion of space unused on an array for emergency use. This is also a good practice for future enhancements and code releases, as there can be slight changes in the available size. If moving from a 5.x based storage server firmware to a 6.1x based, you will encounter this issue and need to expand the array to be able to handle a full sized LUN on the 6.1x server. This can impact VolumeCopy, and Enhanced Remote Mirroring efforts especially. If planning to use the FlashCopy feature, this may be a good place to plan to have repositories placed.

3.3.3 Configuring storage partitioning

As the DS4000 Storage Server is capable of having heterogeneous hosts attached, a way is needed to define which hosts are able to access which logical drives on the DS4000 Storage Server. Therefore, you need to configure storage partitioning, for two reasons:

- Each host operating system has slightly different settings required for proper operation on the DS4000 Storage Server. For that reason, you need to tell the storage subsystem the host type that is attached.
- There is interference between the hosts if every host has access to every logical drive. By using storage partitioning, you mask logical drives from hosts which are not to use them (also known as LUN masking), and you ensure that each host or host group only has access to its assigned logical drives. You can have a maximum of 256 logical drives

assigned to a single storage partition. You may have a maximum of 2048 logical drives (LUNs) per storage server depending on the model.

Restriction: The maximum logical drives per partition can exceed some host limits. Check to be sure the host can support the number of logical drives that you are configuring for the partition. In some cases you may need to split the logical drives across two separate partitions, with a second set of host side HBAs.

The process of defining the storage partitions is as follows:

1. Define host groups.
2. Define hosts.
3. Define host ports for each host.
4. Define storage partitions by assigning logical drives to the hosts or host groups.

Storage Manager Version 9.12 introduced the **Task Assistant wizards**. You can now use two new wizards that assist you with setting up the storage partitioning:

- ▶ Define Host wizard
- ▶ Storage Partitioning wizard

The first step is to select the Mappings View in the Subsystem Management window. All functions that are performed with regards to partitioning are performed from within this view.

If you have not defined any storage partitions yet, the Mapping Start-Up Help window pops up. The information in the window advises you to only create the host groups you intend to use. For example, if you want to attach a cluster of host servers, then you surely need to create a host group for them. On the other hand, if you want to attach a host that is not a part of the cluster, it is not necessary put it into a particular host group. However, as requirements may change, we recommend that you create a host group anyway.

Best practice: We recommend that all hosts be mapped to a **host group** for their specific purpose. This is not required, but can prevent confusion and mistakes.

For detailed information for each of the process steps see *IBM System Storage IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

Here are some additional points to be aware of when performing your partition mapping:

1. All information, such as host ports and logical drive mappings, is shown and configured in the **Mappings View**. The right side of the window lists all mappings that are owned by the object you choose in the left side.
2. If you highlight the storage subsystem, you see a list of all defined mappings. If you highlight a specific host group or host, only its mappings are listed.
3. If you accidentally assigned a host to the wrong host group, you can move the host to another group. Simply right-click the host name and select **Move**. A pop-up window opens and asks you to specify the host group name.
4. Storage partitioning of the DS4000 Storage Server is based on the World Wide Names of the host ports. The definitions for the host groups and the hosts only represent a view of the physical and logical setup of your fabric. Having this structure available makes it much easier to identify which host ports are allowed to see the same logical drives, and which are in different storage partitions.

5. Storage partitioning is not the only function of the storage server that uses the definition of the host port. When you define the host port, the host type of the attached host is defined as well. Through this information, the DS4000 determines what NVSRAM settings AVT, and RDAC it should expect to use with the host.

It is important to carefully choose the correct host type from the list of available types, because this is the part of the configuration dealing with heterogeneous host support. Each operating system expects slightly different settings and can handle SCSI commands a little differently. Incorrect selections can result in failure to boot, or loss of path failover function when attached to the storage server.

6. As the host port is identified by the World Wide Name of the host bus adapter, you may need to change it when an HBA failure results in replacement. This can be done by first highlighting the old host port, right-clicking, and selecting **Replace**. In the drop-down box, you see only the World Wide Names that are currently active. If you want to enter a host port that is not currently active, type the World Wide Name in the field. Be sure to check for typing errors. If the WWN does not appear in the drop-down box, you need to verify your zoning for accuracy or missing changes.
7. If you have a single server in a host group that has one or more logical drive assigned to it, we recommend that you assign the mapping to the host, and not the host group. Numerous servers can share a common host group; but may not necessarily share drives. Only place drives in the host group mapping that you truly want *all* hosts in the group to be able to share.
8. If you have a cluster, you will want to assign the logical drives that are to be shared across all, to the host group, so that all of the host nodes in the host group have access to them.

Note: If you create a new mapping or change an existing mapping of a logical drive, the change happens immediately. Therefore, make sure that this logical drive is not in use or even assigned by any of the machines attached to the storage subsystem.

9. If you attached a host server that is not configured to use the in-band management capabilities; we recommend that you ensure that the access LUNs are deleted or unconfigured from that host server's mapping list.

Highlight the host or host group containing the system in the Mappings View. In the right side of the window, you see the list of all logical drives mapped to this host or host group. To delete the mapping of the access logical drive, right-click it and select **Remove**. The mapping of that access logical drive is deleted immediately. If you need to use the access LUN with another server, you will have the opportunity to do so when you create the host mapping. An access LUN is created whenever a host server partition is created.

Now all logical drives and their mappings are defined and are now accessible by their mapped host systems.

To make the logical drives available to the host systems without rebooting, the DS4000 Utilities package provides a **hot_add** command line tool for some operating systems. You simply run **hot_add**, and all host bus adapters are re-scanned for new devices, and the devices should be accessible to the operating system.

You will need to take appropriate steps to enable the use of the storage inside the operating system, or by the volume manager software.

3.3.4 Configuring for Copy Services functions

The DS4000 Storage Server has a complete set of Copy Services functions that can be added to it. These features are all enabled by premium feature keys which come in the following types:

- ▶ FlashCopy

FlashCopy is used to create a point-in-time image copy of the *base* LUN for use by other system applications while the base LUN remains available to the base host application. The secondary applications can be read-only, such as a backup application, or they might also be read/write, for example, such as a test system or analysis application. For more in-depth application uses, we recommend that you use your FlashCopy image to create a VolumeCopy, which will be a complete image drive and fully independent of the base LUN image.

- ▶ VolumeCopy

VolumeCopy creates a complete physical replication of one logical drive (source) to another (target) within the same storage subsystem. The target logical drive is an exact copy or *clone* of the source logical drive. This feature is designed as a system management tool for tasks such as relocating data to other drives for hardware upgrades or performance management, data backup, and restoring snapshot logical drive data. Because VolumeCopy is a full replication of a point-in-time image, it allows for analysis, mining, and testing without any degradation of the production logical drive performance. It also brings improvements to backup and restore operations, making them faster and eliminating I/O contention on the primary (source) logical drive. The use of the FlashCopy → Volume copy combined process is recommended as a best practice when used with ERM for Business Continuity and Disaster Recovery (BCDR) solutions with short *recovery time objectives* (RTOs).

- ▶ Enhanced Remote Mirroring (ERM)

ERM is used to allow mirroring to another DS4000 Storage Server either co-located, or situated at another site. The main usage of this premium feature is for to enable business continuity in the event of a disaster or unrecoverable error at the primary storage server. It achieves this by maintaining two copies of a data set in two different locations, on two or more different storage servers and enabling a second storage subsystem to take over responsibility. Methods available for use with this feature are: synchronous, asynchronous, and asynchronous with write order consistency (WOC).

The configuration of all of these features are documented in great detail in *IBM System Storage IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

3.4 Event monitoring and alerts

Included in the DS4000 Client package is the Event Monitor service. It enables the workstation running this monitor to send out alerts by e-mail (SMTP) or traps (SNMP). The Event Monitor can be used to alert you of problems in any of the DS4000 Storage Servers in your environment. It is also used for the Remote Support Manager Service described in 3.4.3, “DS4000 Remote Support Manager” on page 117.

Tip: The Event Monitor service should be installed and configured on at least two systems that are attached to the storage subsystem and allow in-band management, running 24 hours a day. This practice ensures proper alerting, even if one server is down.

Depending on the setup you choose, different storage subsystems are monitored by the Event Monitor. If you right-click your local system in the Enterprise Management window (at the top of the tree) and select **Configure Alerts**, this applies to all storage subsystems listed in the Enterprise Management window. Also, if you see the same storage subsystem through different paths, directly attached and through different hosts running the host agent, you receive multiple alerts. If you right-click a specific storage subsystem, you only define the alerting for this particular DS4000 Storage Server.

An icon in the lower-left corner of the Enterprise Management window indicates that the Event Monitor is running on this host.

If you want to send e-mail alerts, you have to define an SMTP server first. Click **Edit** → **Configure Alerts**. Ensure that you are on the Mail Server tab. Enter the IP address or the name of your mail server and the sender address.

On the Mail tab, define the e-mail addresses to which alerts are sent. If you do not define an address, no SMTP alerts are sent. You also can validate the e-mail addresses to ensure a correct delivery and test your setup.

If you choose the SNMP tab, you can define the settings for SNMP alerts: the IP address of your SNMP console and the community name. As with the e-mail addresses, you can define several trap destinations.

You need an SNMP console for receiving and handling the traps sent by the service. There is an MIB file included in the Storage Manager software, which should be compiled into the SNMP console to allow proper display of the traps. Refer to the documentation of the SNMP console you are using to learn how to compile a new MIB.

3.4.1 ADT alert notification

ADT alert notification is provided with Storage Manager. This accomplishes three things:

- ▶ It provides notifications for persistent “Logical drive not on preferred controller” conditions that resulted from ADT.
- ▶ It guards against spurious alerts by giving the host a “delay period” after a preferred controller change, so it can get reoriented to the new preferred controller.
- ▶ It minimizes the potential for the user or administrator to receive a flood of alerts when many logical drives failover at nearly the same point in time due to a single upstream event, such as an HBA failure.

Upon an ADT event or an induced logical drive ownership change, the DS4000 controller firmware waits for a configurable time interval, called the *alert delay period*, after which it reassesses the logical drives distribution among the arrays.

If, after the delay period, some logical drives are not on their preferred controllers, the controller that owns the not-on-preferred-logical drive logs a critical Major Event Log (MEL) event. This event triggers an alert notification, called the *logical drive transfer alert*. The critical event logged on behalf of this feature is in addition to any informational or critical events that are already logged in the RDAC. This can be seen in Figure 3-40.

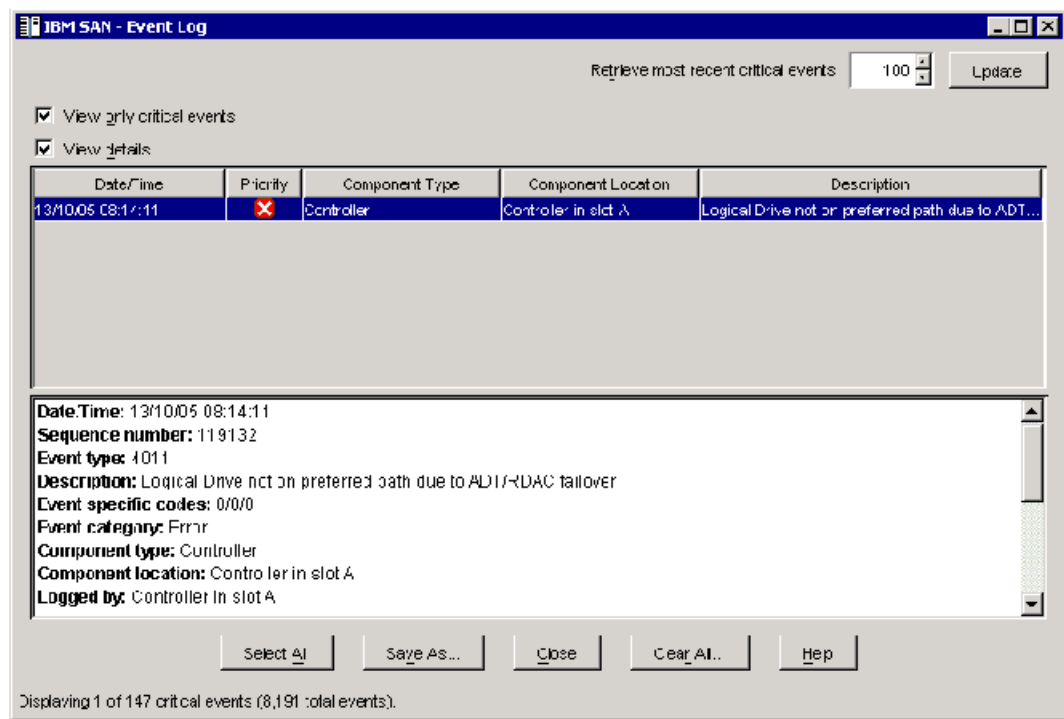


Figure 3-40 Example of alert notification in MEL of an ADT/RDAC logical drive failover

Note: Logical drive controller ownership changes occur as a normal part of a controller firmware download. However, the logical-drive-not-on-preferred-controller events that occur in this situation will *not* result in an alert notification.

3.4.2 Failover alert delay

The failover alert delay lets you delay the logging of a critical event if the multipath driver transfers logical drives to the non-preferred controller. If the multipath driver transfers the logical drives back to the preferred controller within the specified delay period, no critical event is logged. If the transfer exceeds this delay period, a logical drive-not-on-preferred-path alert is issued as a critical event. This option also can be used to minimize multiple alerts when many logical drives failover because of a system error, such as a failed host adapter.

The logical drive-not-on-preferred-path alert is issued for any instance of a logical drive owned by a non-preferred controller and is in addition to any other informational or critical failover events. Whenever a logical drive-not-on-preferred-path condition occurs, only the alert notification is delayed; a needs attention condition is raised immediately.

To make the best use of this feature, set the failover alert delay period such that the host driver fallback monitor runs at least once during the alert delay period. Note that a logical drive ownership change might persist through the alert delay period, but correct itself before you can inspect the situation. In such a case, a logical drive-not-on-preferred-path alert is issued as a critical event, but the array will no longer be in a needs-attention state. If a logical

drive ownership change persists through the failover alert delay period, refer to the Recovery Guru for recovery procedures.

Important: Here are several considerations regarding failover alerts:

- ▶ The failover alert delay option operates at the storage subsystem level, so one setting applies to all logical drives.
- ▶ The failover alert delay option is reported in minutes in the storage subsystem profile as a storage subsystem property.
- ▶ The default failover alert delay interval is five minutes. The delay period can be set within a range of 0 to 60 minutes. Setting the alert delay to a value of zero results in instant notification of a logical drive not on the preferred path. A value of zero does not mean alert notification is disabled.
- ▶ The failover alert delay is activated after controller start-of-day completes to determine if all logical drives were restored during the start-of-day operation. Thus, the earliest that the not-on-preferred path alert will be generated is after boot up and the configured failover alert delay.

Changing the failover alert delay

To change the failover alert delay:

1. Select the storage subsystem from the Subsystem Management window, and then select either the **Storage Subsystem** → **Change** → **Failover Alert Delay** menu option, or right-click and select **Change** → **Failover Alert Delay**. See Figure 3-41.

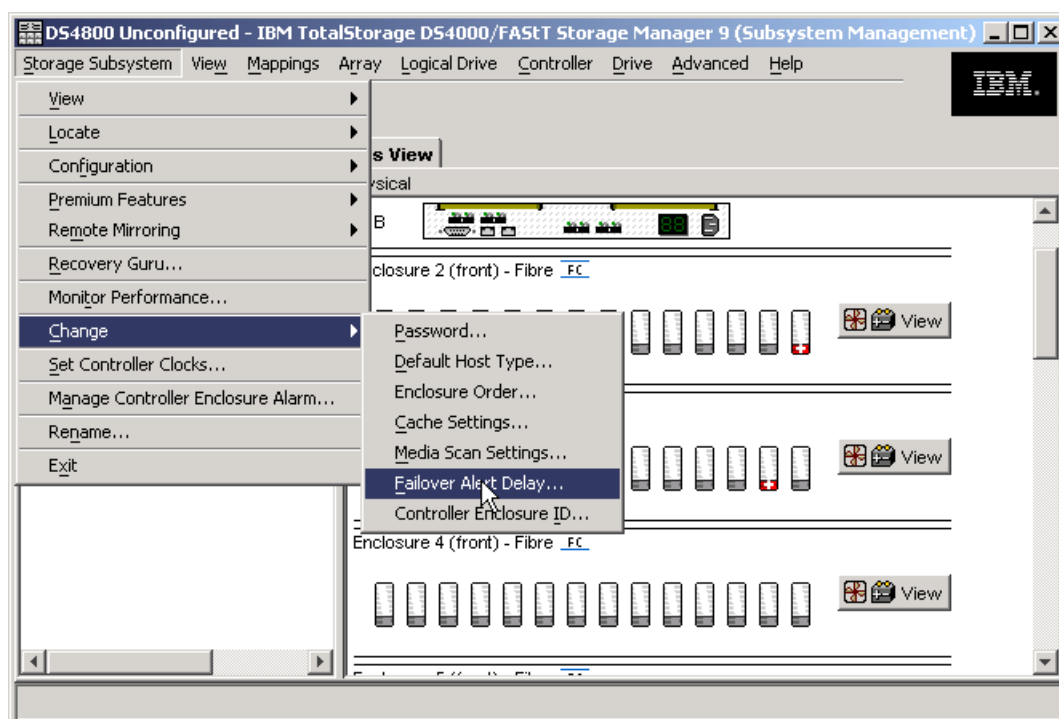


Figure 3-41 Changing the failover alert delay

The Failover Alert Delay dialog box opens, as seen in Figure 3-42.

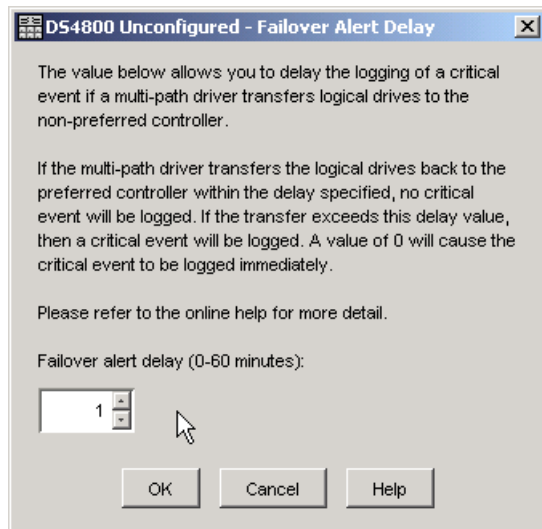


Figure 3-42 Failover Alert Delay dialog box

Enter the desired delay interval in minutes and click **OK**.

3.4.3 DS4000 Remote Support Manager

The IBM Remote Support Manager for Storage (RSM™ for Storage) is an application that installs on an IBM System x server running Novell SUSE Linux Enterprise Server 9. It provides problem reporting and remote access for IBM Service for the DS4000 family.

RSM relies on the Event Monitor for problem reporting. The problem reporting provided by the RSM application automatically creates an entry in IBM call management system for each storage subsystem that reports a problem. Monitoring of storage subsystems is performed by your existing IBM Storage Manager application, which is configured to send SNMP traps to the Remote Support Manager when critical events are detected.

RSM must be installed on a dedicated workstation. IBM service personnel can remotely access the RSM server through a modem connection and a command line interface, but only after access has been authorized. Files and logs needed for problem determination are sent to IBM using e-mail or an FTP connection using the RSM server Ethernet interface.

Once installed, the server should be considered a single purpose appliance for problem reporting and remote access support for your DS4000 storage subsystems.

Isolation of remote and local users of the system from other devices on your intranet is performed by an internal firewall that is managed by the RSM for Storage application. Remote users do not have the ability to change any security features of the application.

One RSM for Storage system can support up to 50 subsystems. The RSM for Storage system must have IP connectivity to the Ethernet management ports of the subsystems.

RSM hardware requirements

The RSM for Storage application is designed to run on a System x server. It has been tested with and is supported on the following System x servers:

- ▶ x306m 8849
- ▶ x100 8486

With these options:

- ▶ 512 MB memory.
- ▶ 73 GB hard disk drive.
- ▶ If your SAN devices are on a private management LAN, a second Ethernet port for accessing your company's SMTP server and the Internet will be required if your selected server has only a single Ethernet port.
- ▶ IBM 10/100/1000 Base-TX Ethernet PCI-X Adapter (52P8642) or equivalent.

The RSM for Storage application is designed to work with an external modem attached to the first serial port.

RSM requirements

The RSM for Storage requires the following prerequisite software:

- ▶ IBM Storage Manager Application 9.16 or later (the latest version is recommended) with Event Monitor installed in a management station in a different server.
- ▶ Storage subsystems with controller firmware supported by the Storage Manager 9.16 or later. The latest supported firmware version is recommended.
- ▶ Novell SLES 9 (Service Pack 3) to install the RSM for Storage application in the RSM server.

RSM for Storage uses an Ethernet connection for problem reporting and a modem for remote access by IBM Service, as shown in Figure 3-43.

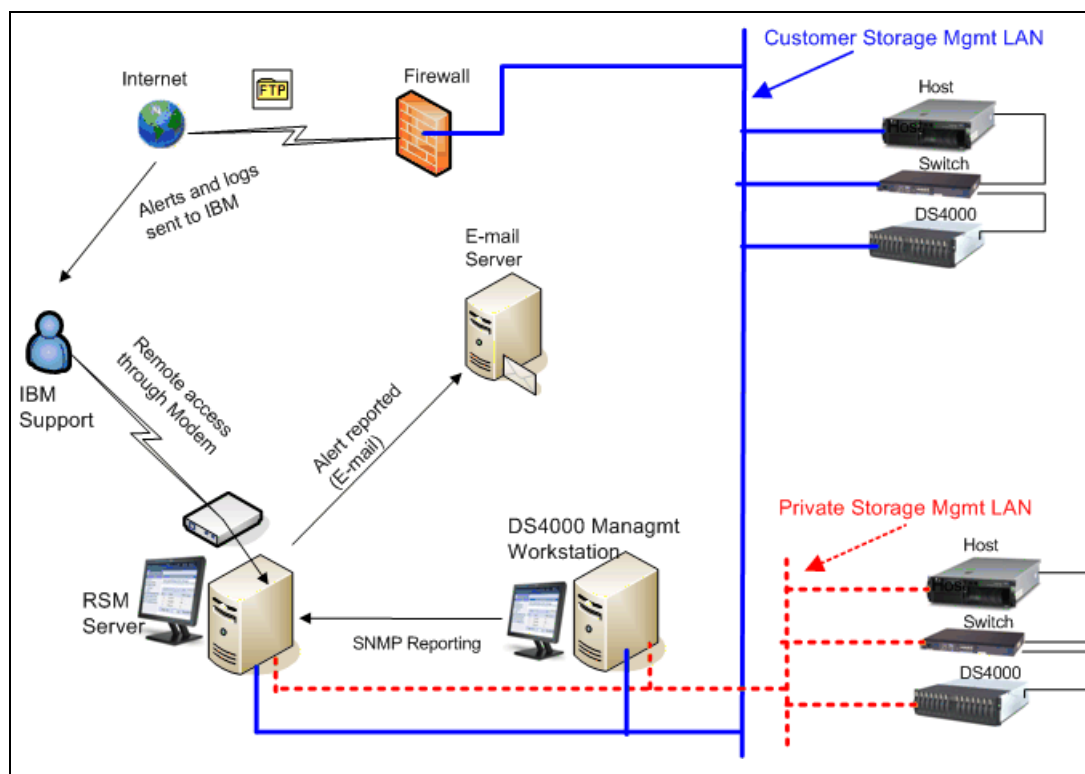


Figure 3-43 RSM for Storage connection diagram

The RSM for Storage server must have IP connectivity to the Ethernet management ports of the DS4000 storage subsystems, as well a management station running Storage Manager client and Event Monitor.

It is also required that all storage subsystems, the management station running IBM Storage Manager, the e-mail server, and Internet gateway are accessible from the RSM for Storage server without requiring authorization through a firewall.

Note: Refer to *IBM Remote Support Manager for Storage Compatibility Guide* for the latest update of supported servers, modem, and operating systems. The document can be downloaded from the following Web page:

<http://www.ibm.com/support/docview.wss?uid=psg1MIGR-66062&rs=594>

Security considerations

Adding a modem to one of your systems creates a potential entry point for unauthorized access to your network. The RSM for Storage application modifies many characteristics and behaviors of the system it is installed on to protect this entry point and to maximize the amount of control you have in managing remote access. To ensure the integrity of these controls, you should consider the server the RSM for Storage application is installed on to be a single-purpose appliance.

There are several security features in RSM for Storage to maximize protection from unauthorized access:

- ▶ Remote Access security

The modem used for remote access will not answer a call from service unless one of the storage subsystems has an active alert or Remote Access has manually been enabled.

Usually, Remote Access is enabled automatically when an alert is sent to IBM, but you can choose to wait for IBM Service to contact you when an alert is received and manually enable Remote Access at that time.

Remote Access also has configurable time-out from 12 to 96 hours. You can manually disable remote access when service is complete or allow it to time out, thereby guaranteeing that the system will return to a secure state without intervention.

The user ID for Remote Access is only valid when Remote Access is enabled, and the initial password is changed daily at midnight UTC. In addition, remote users (IBM Service) are also presented with a challenge string.

- ▶ Internal firewall

RSM for Storage includes an internal firewall to limit the scope of access a remote user has to your network. It limits the IP destinations that can be accessed by local and remote users of the system. The rules for inbound and outbound IP traffic that control the internal firewall are managed dynamically by the RSM for Storage application.

To maintain the integrity of the firewall and other security controls, remotely connected IBM Service users cannot change any security-related settings.

- ▶ DS4000 subsystem security

Storage Manager has the ability to require an administrative password in order to make changes to the subsystem configuration. We recommend that this password be configured.

DS4000 also has a controller shell environment, which is accessible using a remote login (RLOGIN) client. IBM Storage Manager for DS4000 has an option to disable RLOGIN, and we recommend that RLOGIN usually be disabled.

The configuration of all of these features are documented in great detail in *IBM System Storage IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

3.5 Software and microcode upgrades

Every so often IBM will release new firmware (it is posted on the support the Web site) that will need to be installed. Occasionally, IBM may remove old firmware versions from support. Upgrades from unsupported levels are mandatory to receive warranty support.

This section reviews the required steps to upgrade your IBM TotalStorage DS4000 Storage Server when firmware updates become available. Upgrades to the DS4000 Storage Server firmware should generally be preceded by an upgrade to the latest available version of the Storage Manager client software as this may be required to access the DS4000 Storage Server when the firmware upgrade completes. In most cases, it is possible to manage a DS4000 Storage Server running down-level firmware with the latest SMclient, but not possible to manage a storage server running the latest version of firmware with a down-level client. In some cases the only management capability provided may be to upgrade the firmware to newer level; which is the desired goal.

Note: The version number of the Storage Manager firmware and the Storage Manager client are not completely connected. For example; Storage Manager 9.15x can manage storage servers which are running storage server firmware 5.3x on them.

Always check the readme file for details of latest storage manager release and special usage.

3.5.1 Staying up-to-date with your drivers and firmware using My support

My support registration provides E-mail notification when new firmware levels have been updated and are available for download and installation. To register for My support, visit:

<http://www.ibm.com/support/mysupport/us/en>

Following is the registration process:

1. The Sign In window displays:
 - If you have a valid IBM ID and Password, then sign in.
 - If you are not currently registered with the site, click Register now and register your details.
2. The My Support window opens. Click **Add Products** to add products to your profile.
3. Use the pull-down menus to choose the appropriate DS4000 storage server and expansion enclosures that you want to add to your profile.
4. To add the product to your profile, select the appropriate box or boxes next to the product names and click **Add Product**.
5. Once the product or products are added to your profile, click the **Subscribe to Email** folder tab.
6. Select **Storage** in the pull-down menu. Select **Please send these documents by weekly email** and select **Downloads and drivers and Flashes** to receive important information about product updates. Click **Updates**.
7. Click **Sign Out** to log out of My Support.

You will be notified whenever there is new firmware available for the products you selected during registration.

We also suggest that you explore and customize to your needs the other options available under My support.

3.5.2 Prerequisites for upgrades

Upgrading the firmware and management software for the DS4000 Storage Server is a relatively simple procedure. Before you start, you should make sure that you have an adequate maintenance window to do the procedure, because on large configurations it can be a little time consuming. The times for upgrading all the associated firmware and software are in Table 3-5. These times are only approximate and can vary from system to system.

Table 3-5 Upgrade times

Element being upgraded	Approximate time of upgrade
Storage Manager software and associated drivers and software	35 minutes
DS4000 Storage Server firmware	5 minutes
DS4000 ESM firmware	5 minutes per ESM
Hard drives	3 minutes per drive type

It is critical that if you update one part of the firmware, you update all the firmware and software to the same level. You must *not* run a mismatched set.

All the necessary files for performing this upgrade are available at:

<http://www-304.ibm.com/jct01004c/systems/support/storage/disk>

Look for your specific DS4000 storage system.

3.5.3 Updating the controller microcode

We recommend that your DS4000 Storage Server always be at a recent level of microcode. Occasionally, IBM will withdraw older levels of microcode from support. In this case, an upgrade to the microcode is mandatory. In general, you should plan on upgrading all drivers, microcode, and management software in your SAN on a periodic basis. New code levels may contain important fixes to problems you may not have encountered yet.

Important: Before upgrading the storage server firmware and NVSRAM, make sure that the system is in an optimal state. If not, run the Recovery Guru to diagnose and fix the problem before you proceed with the upgrade.

The microcode of the DS4000 Storage Server consists of two packages:

- ▶ The firmware
- ▶ The NVSRAM package, including the settings for booting the DS4000 Storage Server

The NVSRAM is similar to the settings in the BIOS of a host system. The firmware and the NVSRAM are closely tied to each other and are therefore *not* independent. Be sure to install the correct combination of the two packages.

The upgrade procedure needs two independent connections to the DS4000 Storage Server, one for each controller. It is not possible to perform a microcode update with only one controller connected. Therefore, both controllers must be accessible either via Fibre Channel or Ethernet. Both controllers must also be in the active state.

If you plan to upgrade via Fibre Channel, make sure that you have a multipath I/O driver installed on your management host. In most cases, this would be the RDAC. This is necessary since access logical drive moves from one controller to the other during this procedure and the DS4000 Storage Server must be manageable during the entire time.

Important: Here are some considerations for your upgrade:

- ▶ Refer to the *readme* file to find out which between the ESM or the controllers must be upgraded first. In some cases, the expansion enclosure ESMs must be updated to the latest firmware level before starting the controller update (outdated ESM firmware could make your expansion enclosures inaccessible after the DS4000 Storage Server firmware update). In some cases it is just the opposite.
- ▶ Update the controller firmware and then the NVSRAM.
- ▶ Ensure that all hosts attached to the DS4000 Storage Server have a multipath I/O driver installed.
- ▶ Any power or network/SAN interruption during the update process may lead to configuration corruption. Therefore, do not power off the DS4000 Storage Server or the management station during the update. If you are using in-band management and have Fibre Channel hubs or managed hubs, then make sure no SAN connected devices are powered up during the update. Otherwise, this can cause a loop initialization process and interrupt the process.

Staged microcode upgrade

Staged microcode upgrade was introduced with the Storage Manager 9.10 and firmware 6.10. You can load the controller firmware and NVSRAM to a designated flash area on the DS4000 controllers and activate it at a later time. Of course, you can still transfer the controller microcode to the storage subsystem and activate it in one step if necessary.

The firmware is transferred to one of the controllers. This controller copies the image to the other controller. The image is verified through a CRC check on both controllers. If the checksum is OK, the uploaded firmware is marked ready and available for activation. If one of the two controllers fails to validate the CRC, the image is marked invalid on both controllers and not available for activation. An error is returned to the management station as well.

The activation procedure is similar to previous firmware versions. The first controller moves all logical drives to the second one. Then it reboots and activates new firmware. After that it takes ownership of all logical drives, and the second controller is rebooted in order to have its new firmware activated. When both controllers are up again, the logical drives are redistributed to the preferred paths. Because the logical drives move between the controllers during the procedure and are all handled by just one controller at a certain point, we recommend activating the new firmware when the disk I/O activity is low.

A normal reboot of a controller or a power cycle of the DS4000 does not activate the new firmware. It is only activated after the user has chosen to activate the firmware.

Note: Single controller DS4000 models do not support staged firmware download.

To perform the staged firmware and NVSRAM update, follow these steps:

Important: Before conducting any upgrades of firmware or NVSRAM, you *must* read the readme file for the version you are upgrading to see what restrictions and limitations exist.

1. Open the Subsystem Management window for the DS4000 Storage Server you want to upgrade. To download the firmware select **Advanced** → **Maintenance** → **Download** → **Controller Firmware**, as shown in Figure 3-44.

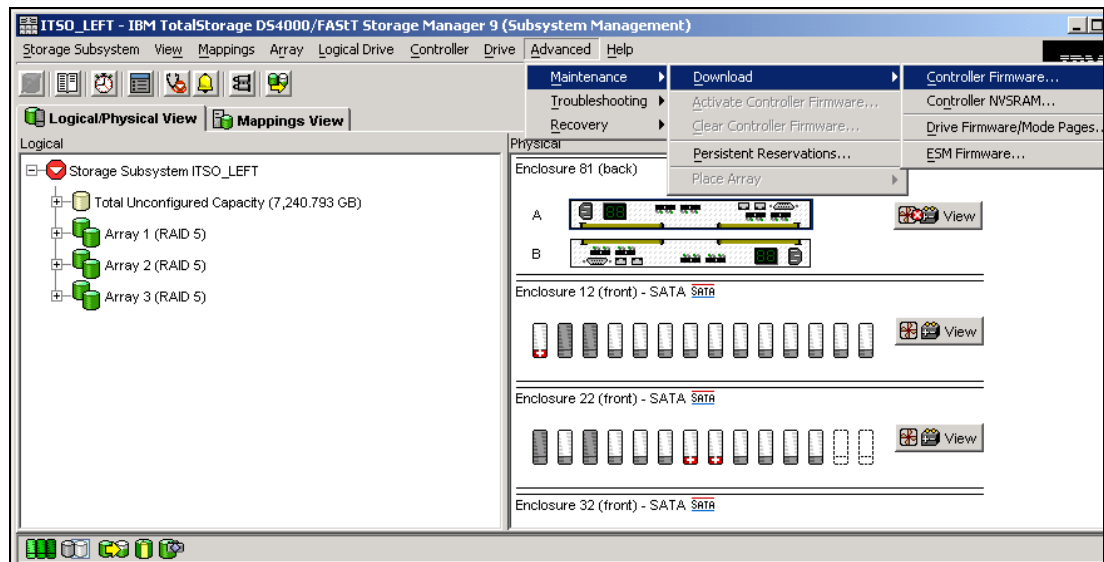


Figure 3-44 Subsystem Management window - Controller firmware update

2. The Download Firmware window opens, showing the current firmware and NVSRAM versions. Select the correct firmware and NVSRAM files, as shown in Figure 3-45. Do not forget to mark the check box to download the NVSRAM file as well.

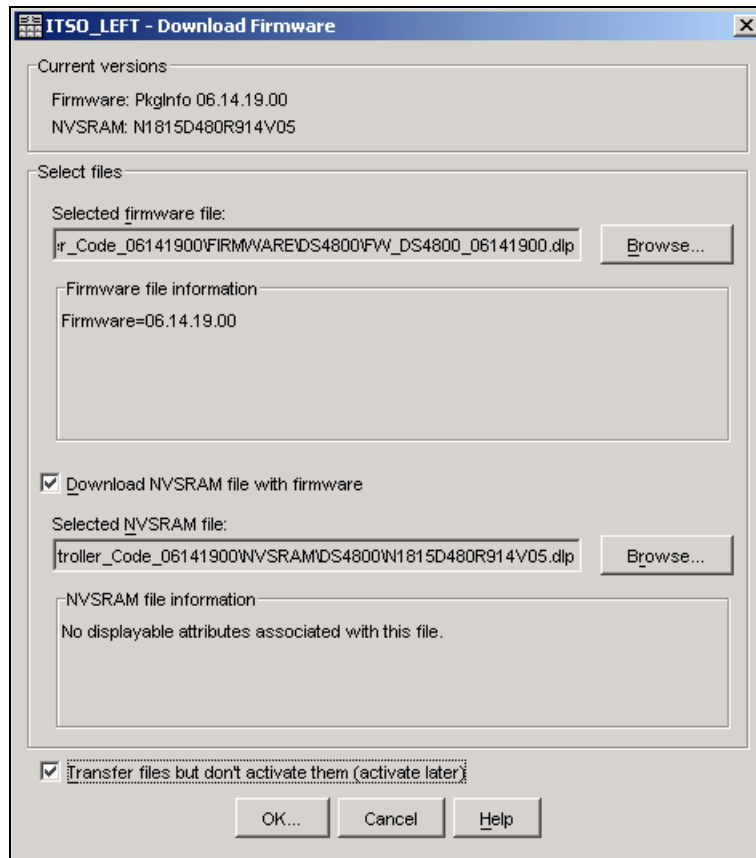


Figure 3-45 Download Firmware window

There is another check box at the bottom of the window (Transfer files but do not activate them). Mark this check box if you want to activate the new firmware at a later time. Then click **OK** to continue.

3. The next window instructs you to confirm the firmware and NVSRAM download (because the process cannot be cancelled once it begins).

Confirm by clicking **Yes**. The firmware/NVSRAM transfer begins and you can watch the progress. When the process finishes, the Transfer Successful message is displayed, as shown in Figure 3-46.

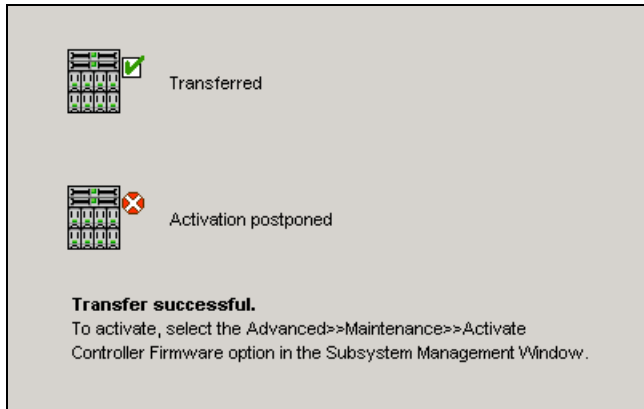


Figure 3-46 Firmware/NVSRAM download finished

4. After clicking **Close**, you are back in the Subsystem Management window. Because this is a staged firmware upgrade, the new firmware is now ready for activation. This is indicated by an icon (blue 101) next to the storage subsystem name (as shown in Figure 3-47).

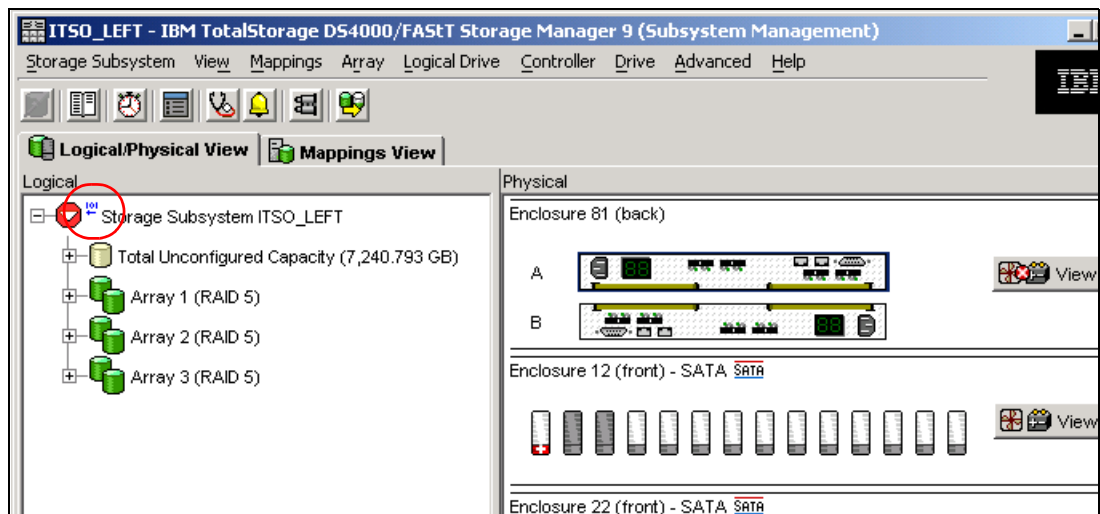


Figure 3-47 Subsystem Management window - Firmware ready for activation

- To activate the new firmware, select **Advanced** → **Maintenance** → **Activate Controller Firmware**, as shown in Figure 3-48.

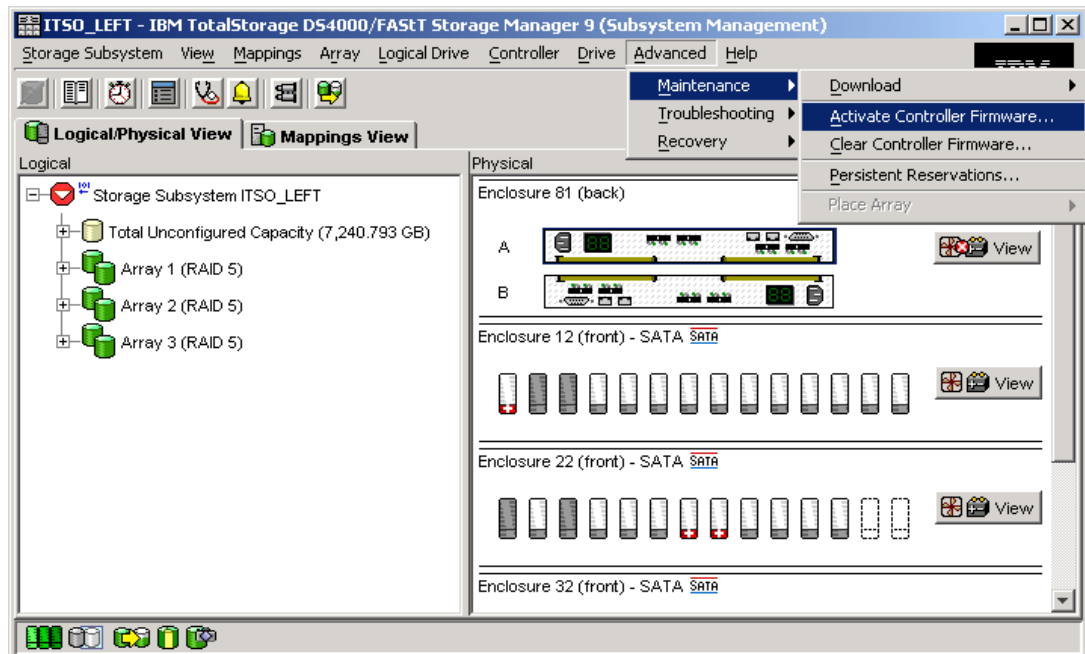


Figure 3-48 Subsystem Management window - Firmware activation

The Activate Firmware window opens and asks you for confirmation that you want to continue. After you click **Yes**, the activation process starts. You can monitor the progress in the Activation window, as shown in Figure 3-49.

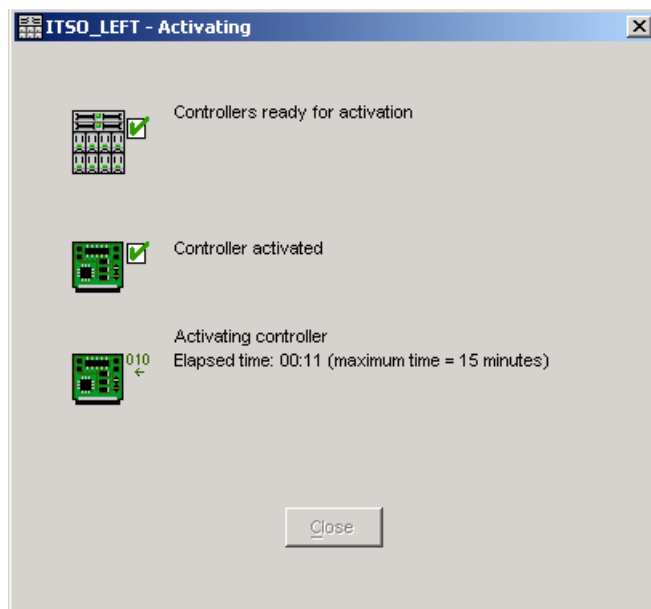


Figure 3-49 Activating the firmware

When the new firmware is activated on both controllers, you will see the Activation successful message. Click **Close** to return to the Subsystem Management window.

Note: If you applied any changes to the NVSRAM settings, for example, running a script, you must re-apply them after the download of the new NVSRAM completes. The NVSRAM update resets all settings stored in the NVSRAM to their defaults.

3.5.4 Updating DS4000 host software

This section describes how to update the DS4000 software in Windows and Linux environments.

Updating in a Windows environment

To update the host software in a Windows environment:

1. Uninstall the storage management components.
From Add/Remove Programs in the control panel, select **Complete Uninstall**. This will remove the SMagent, SMutil, Multipath driver, and SMclient.
2. Verify that IBM HBA firmware and device driver versions are current.
If they are not current, download the latest versions and refer to the readme file located with the device driver and then upgrade the device drivers. IBM offers two different HBA device drivers for the Windows environment, StorPort and SCSIPort. We recommend the newer StorPort drivers.
3. Install the storage manager from the Storage Manager InstallAnywhere (SMIA) and select **Host**.
4. Choose the type of Multipath I/O driver (RDAC or MPIO).

Updating in a Linux environment

To update the host software in a Linux environment:

1. Uninstall the storage manager components in the following order:
 - a. DS4000 Runtime environment
 - b. SMutil
 - c. RDAC
 - d. SMclient

Note: In pre-9.1x versions the SMagent was not used, and so did not need to be uninstalled.

2. Verify that IBM host adapter device driver versions are current. If they are not current, refer to the *readme* file located with the device driver and then upgrade the device drivers.
3. Install the storage manager components in the following order:
 - a. SMagent
 - b. DS4000 Runtime environment\
 - c. RDAC
 - d. SMutil
 - e. SMclient

3.6 Capacity upgrades, system upgrades

The DS4000 has the ability to accept new disks or EXP units dynamically, with no downtime to the DS4000 unit. In fact, the DS4000 *must* be powered on when adding new hardware.

3.6.1 Capacity upgrades and increased bandwidth

With the DS4000 Storage Server you can add capacity by adding expansion enclosures or disks to the enclosure being used. Care must be taken when performing these tasks to avoid damaging the configuration currently in place. For this reason you must follow the detailed steps laid out for each part.

Important: Prior to physically installing new hardware, refer to the instructions in IBM *TotalStorage DS4000 hard drive and Storage Expansion Enclosure Installation and Migration Guide*, GC26-7849, available at:

<http://www.ibm.com/support/docview.wss?rs=594&uid=psg1MIGR-57818>

Failure to consult this documentation may result in data loss, corruption, or loss of availability to your storage.

For further recommendations on cabling and layout, see the specific section for the DS4000 Storage Server you are working with.

After physical installation, use Storage Manager to create new arrays/LUNs, or extend existing arrays/LUNs. (Note that some operating systems may not support dynamic LUN expansion.)

3.6.2 Storage server upgrade and disk migration procedures

The procedures to migrate disks and enclosures or upgrade to a newer DS4000 controller are not particularly difficult, but care must be taken to ensure that data is not lost. The checklist for ensuring data integrity and the complete procedure for performing capacity upgrades or disk migration is beyond the scope of this book.

Here we explain the DS4000 feature that makes it easy for upgrading subsystems and moving disk enclosures. This feature is known as DACstore.

DACstore

DACstore is an area on each drive in a DS4000 storage subsystem or expansion enclosure where configuration information is stored. This 512 MB reservation (as pictured in Figure 3-50) is invisible to a user and contains information about the DS4000 configuration.

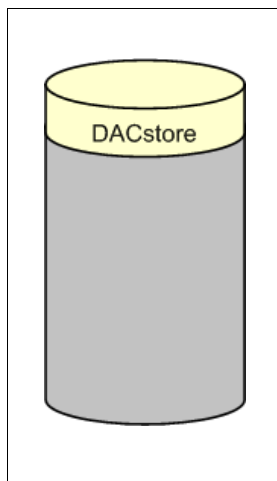


Figure 3-50 The DACstore area of a DS4000 disk drive

The standard DACstore on every drive stores:

- ▶ Drive state and status
- ▶ WWN of the DS4000 controller (A or B) behind which the disk resides
- ▶ Logical drives contained on the disk

Some drives also store extra global controller and subsystem level information, these are called sundry drives. The DS4000 controllers will assign one drive in each array as a sundry drive, although there will always be a minimum of three sundry drives even if only one or two arrays exist.

Additional information stored in the DACstore region of the sundry drive:

- ▶ Failed drive information
- ▶ Global Hot Spare state/status
- ▶ Storage subsystem identifier (SAI or SA Identifier)
- ▶ SAFE Premium Feature Identifier (SAFE ID)
- ▶ Storage subsystem password
- ▶ Media scan rate
- ▶ Cache configuration of the storage subsystem
- ▶ Storage user label
- ▶ MEL logs
- ▶ LUN mappings, host types, and so on.
- ▶ Copy of the controller NVSRAM

Why DACstore

This unique feature of DS4000 storage servers offers a number of benefits:

- ▶ Storage system level reconfiguration: Drives can be rearranged within a storage system to maximize performance and availability through channel optimization.
- ▶ Low risk maintenance: If drives or disk expansion units are relocated, there is no risk of data being lost. Even if a whole DS4000 subsystem needed to be replaced, all of the data and the subsystem configuration could be imported from the disks.
- ▶ Data intact upgrades and migrations: All DS400 subsystem recognize configuration and data from other DS400 subsystems so that migrations can be for the entire disk subsystem as shown in Figure 3-51, or for array-group physical relocation as illustrated in Figure 3-52 on page 130.

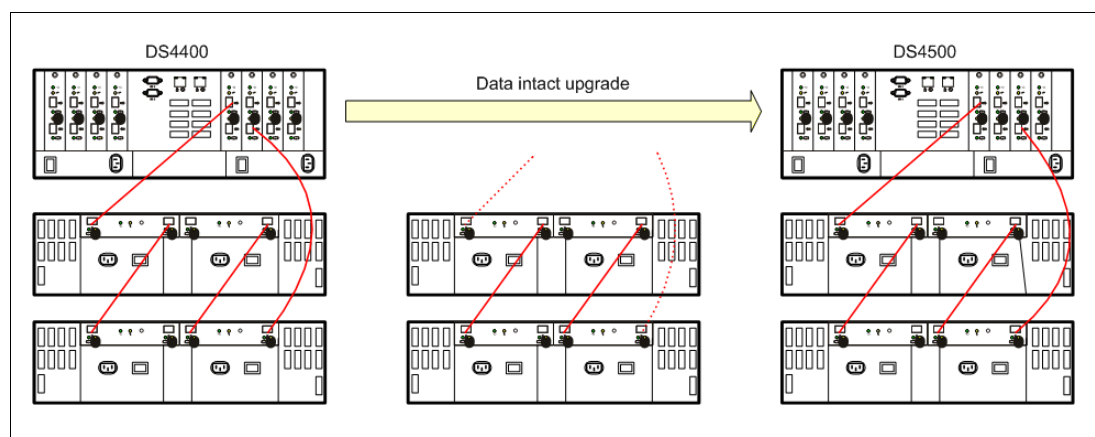


Figure 3-51 Upgrading DS4000 controllers

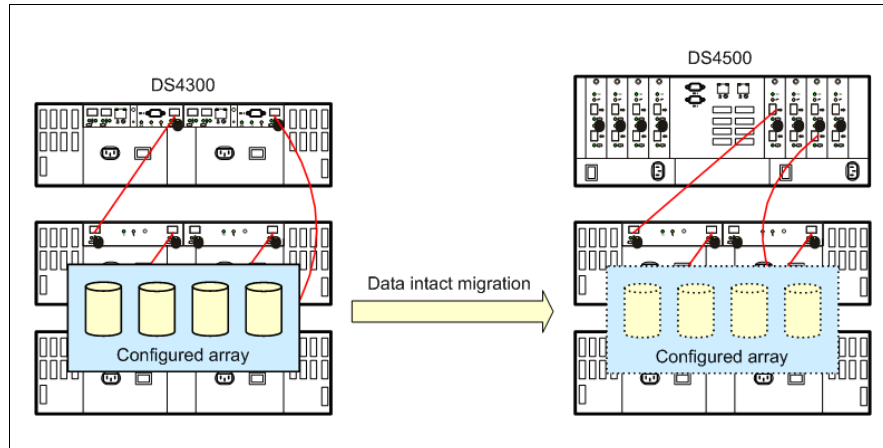


Figure 3-52 Relocating arrays

Other considerations when adding expansion enclosures and drives

These are some recommendations to keep in mind:

- ▶ If new enclosures have been added to the DS4000, we recommend that you eventually schedule some downtime to re-distribute the physical drives in the array.
- ▶ Utilizing the DACstore function available with the DS4000, drives can be moved from one slot (or enclosure, or loop) to another with no effect on the data contained on the drive. This must, however, be done while a subsystem is offline.
- ▶ When adding drives to an expansion unit, do not add more than two drives at a time.
- ▶ For maximum resiliency in the case of failure, arrays should be spread out among as many EXP units as possible. If you merely create a 14-drive array in a new drawer every time you add an EXP710 full of disk, all of the traffic for that array will be going to that one tray. This can affect performance and redundancy (see also “Enclosure loss protection planning” on page 43).
- ▶ For best balance of LUNs and I/O traffic, drives should be added into expansion units in pairs. In other words, every EXP should contain an even number of drives, not an odd number such as 5.
- ▶ If you are utilizing two drive loop pairs, approximately half of the drives in a given array should be on each loop pair. In addition, for performance reasons, half of the drives in an array should be in even numbered slots, and half in odd-numbered slots within the EXP units. (The slot number affects the default loop for traffic to a drive.)
- ▶ To balance load among the two power supplies in an EXP, there should also be a roughly equal number of drives on the left and right hand halves of any given EXP. In other words, when adding pairs of drives to an EXP, add one drive to each end of the EXP.

The complete procedure for drive migration is given in the *Fibre Channel Hard Drive and Storage Expansion Enclosure Installation and Migration Guide*, GC26-7639.

Increasing bandwidth

You can increase bandwidth by moving expansion enclosures to a new or unused mini-hub pair (this doubles the drive-side bandwidth).

This reconfiguration can also be accomplished with no disruption to data availability or interruption of I/O.

Let us assume that the initial configuration is the one depicted on the left in Figure 3-53. We are going to move EXP2 to the unused mini-hub pair on the DS4500.

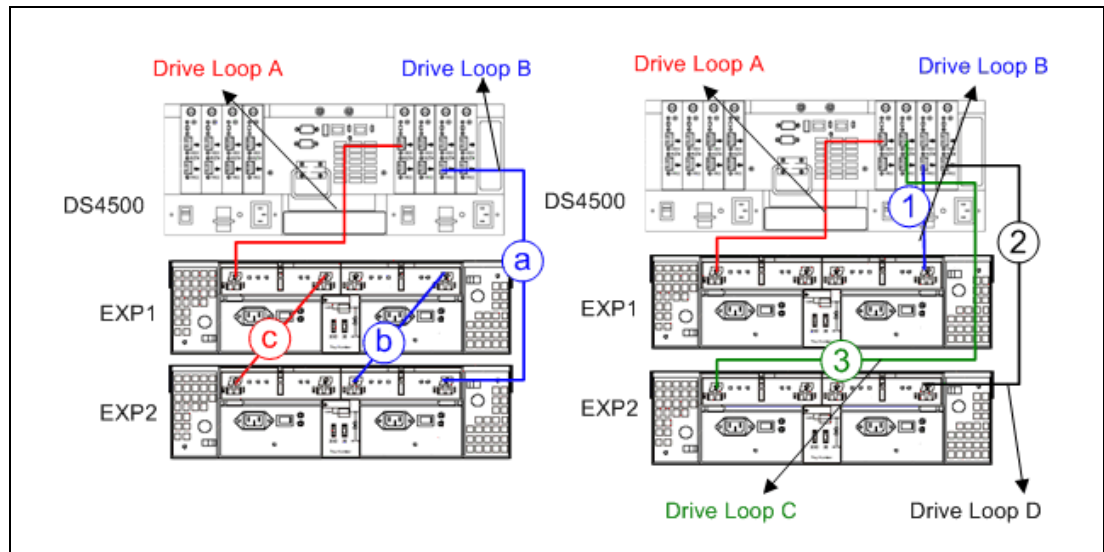


Figure 3-53 Increasing bandwidth

To move EXP2 to the unused mini-hub pair, proceed as follows:

1. Remove the drive loop B cable between the second mini-hub and EXP2 (cable labeled a). Move the cable from EXP2 going to EXP1 (cable labeled b, from loop B) and connect to second mini-hub from EXP1 (cable labeled 1).
2. Connect a cable from the fourth mini-hub to EXP2, establishing drive loop D (represented by cable labeled 2).
3. Remove the drive loop A cable between EXP1 and EXP2 (cable labeled c) and connect a cable from the third mini-hub to EXP2, establishing drive loop C (represented by the cable, labeled 3).



DS4000 performance tuning

In this chapter we describe and discuss performance topics as they apply to the DS4000 Storage Servers.

It must be understood that the storage server itself is but a piece of the performance puzzle. However, we are focused here on the DS4000 Storage Server's role. We discuss the different workload types generally observed in storage, their impact on performance and how it can be addressed by the configuration and parameters.

With all DS4000 Storage Servers, good planning and data layout can make the difference between having excellent workload and application performance, and having poor workload, with high response times resulting in poor application performance. It is therefore not surprising that first-time DS4000 clients ask for advice on how to optimally layout their DS4000 Storage Servers.

This chapter covers the following areas:

- ▶ Understanding workload
- ▶ Solution wide considerations
- ▶ Host considerations
- ▶ Application considerations
- ▶ DS4000 Storage Server considerations
- ▶ Performance data analysis
- ▶ DS4000 tuning

4.1 Workload types

In general, there are two types of data workload (processing) that can be seen:

- ▶ transaction based
- ▶ throughput based

These two workloads are very different in their nature, and must be planned for in quite different ways. Knowing and understanding how your host servers and applications handle their workload is an important part of being successful with your storage configuration efforts, and the resulting performance of the DS4000 Storage Server.

To best understand what is meant by transaction based and throughput based, we must first define a workload. The workload is the total amount of work that is performed at the storage server, and is measured through the following formula:

$$\text{Workload} = [\text{transactions (number of host IOPS)}] * [\text{throughput (amount of data sent in one IO)}]$$

Knowing that a storage server can sustain a given maximum workload (see Figure 4-3 on page 149 through Figure 4-5 on page 150), we can see with the above formula that if the number of host transactions increases, then the throughput must decrease. Conversely, if the host is sending large volumes of data with each I/O, the number of transactions must decrease.

A workload characterized by a high number of transactions (IOPS) is called a *transaction based* workload. A workload characterized by large I/Os is called *throughput based* workload.

These two workload types are conflicting in nature, and consequently will require very different configuration settings across all the pieces of the storage solution. Generally, I/O (and therefore application) performance will be best when the I/O activity is evenly spread across the entire I/O subsystem.

But first, let us describe each type in greater detail, and explain what you can expect to encounter in each case.

Transaction based processes (IOPS)

High performance in transaction based environments cannot be created with a low cost model (with a small number of physical drives) of a storage server. Indeed, transaction process rates are heavily dependent on the number of back-end drives that are available for the controller to use for parallel processing of the hosts I/Os. This frequently results in a decision to be made: Just how many drives will be good enough?

Generally, transaction intense applications also use a small *random* data block pattern to transfer data. With this type of data pattern, having more back-end drives enables more host I/Os to be processed simultaneously, as read cache is far less effective, and the misses need to be retrieved from disk.

In many cases, slow transaction performance problems can be traced directly to “hot” files that cause a bottleneck on some critical component (such as a single physical disk). This situation can occur even when the overall storage server is seeing a fairly light workload. When bottlenecks occur, they can present a very difficult and frustrating task to resolve. As workload content can be continually changing throughout the course of the day, these bottlenecks can be very mysterious in nature and appear and disappear or move over time from one location to another.

Generally, I/O (and therefore application) performance will be best when the I/O activity is evenly spread across the entire I/O subsystem.

Throughput based processes (MBps)

Throughput based workloads are seen with applications or processes that require massive amounts of data sent, and generally use large *sequential* blocks to reduce disk latency. Generally, only a small number of drives (20–28) are needed to reach maximum throughput rates with the DS4000 Storage Servers. In this environment, read operations make use of the cache to stage greater chunks of data at a time, to improve the overall performance. Throughput rates are heavily dependent on the storage server's internal bandwidth. Newer storage servers with broader bandwidths are able to reach higher numbers and bring higher rates to bear.

Why we should care

With the DS4000 Storage Server, these two workload types have different parameter settings that are used to optimize their specific workload environments. These settings are not limited to strictly the storage server, but span the entire solution being used. With care and consideration, it is possible to create an environment of very good performance with both workload types and sharing the same DS4000 Storage Server. However, it must be understood that portions of the storage server configuration will be tuned to better serve one workload or the other.

For maximum performance of both workloads, we recommend considering two separate smaller storage servers each tuned for their specific workload, rather than one large server being shared, but this model is not always financially feasible.

4.2 Solution-wide considerations for performance

Considering the different pieces of the solution that can impact performance, we first look at the host and the operating systems settings, as well as how volume managers come into play. Then we look at the applications; what their workload is like, as well as how many different types of data patterns they may have to use and that you must plan for.

Of course, we also look at the DS4000 Storage Server and the many parameter settings that should be considered, according to the environment where the storage server is deployed. And finally, we look at specific SAN settings that can affect the storage environment as well.

When looking at performance, one must first consider the location. This is a three phase consideration, consisting of:

1. Looking at the location of the drive/logical drive with regards to the path the host uses to access the logical drive. This encompasses the host volume manager, HBA, SAN fabric, and the storage server controller used in accessing the logical drive. Many performance issues stem from mis-configured settings in these areas.
2. Looking at the location of the data within the storage server on its configured array and logical drive. Check that the array has been laid out to give best performance, and the RAID type is the most optimum for the type of workload. Also, if multiple logical drives reside in the same array, check for interference from other logical drive members when doing their workload.
3. Looking at the location of the data on the back-end drives which make up the array, and how the data is carved up (segmented and striped) across all the members of the array. This includes the number of drives being used, as well as their size, and speed. This area

can have a great deal of impact that is very application specific, and usually requires tuning to get to the best results.

We consider each of these areas separately in the following sections.

4.3 Host considerations

When discussing performance, we need to consider far more than just the performance of the I/O workload itself. Many settings within the host frequently affect the overall performance of the system and its applications. All areas should be checked to ensure we are not focusing on a result rather than the cause. However, in this book we are focusing on the I/O subsystem part of the performance puzzle; so we will discuss items that affect its operation.

Some of the settings and parameters discussed in this section are defined and must match both for the host operating system and for the HBAs which are being used as well. Many operating systems have built-in definitions that can be changed to enable the HBAs to be set to the new values. In this section we try to cover these with AIX, and Windows operating systems as an illustration.

4.3.1 Host based settings

Some host operating systems can set values for the DS4000 Storage server logical drives assigned to them. For instance, some hosts can change the *write cache* and the *cache read-ahead* values through attribute settings. These settings can affect both the transaction and throughput workloads. Settings that affect cache usage can have a great impact in most environments.

Other host device attributes that can affect high transaction environments are those affecting the *blocksize*, and *queue depth* capabilities of the logical drives.

- ▶ The blocksize value used by the host I/O helps to determine the best *segment size* choice. We recommend that the segment size be set to at least twice the size of the I/O blocksize being used by the host for the high transaction workload.
- ▶ The queue depth value cannot exceed the storage server maximum of 2048 for DS4000 storage servers running firmware 6.1x and later; and a maximum of 512 for firmware 5.3x and 5.4x. All logical drives on the storage server must share these queue limits. Some hosts define the queue depth only at the HBA level, while others may also define this limit at the storage device level, which ties to the logical drive. The following formulas can be used to determine a good starting point for your queue depth value on a per logical drive basis:

For firmware level 6.1x and higher: $2048 / (\text{number-of-hosts} * \text{logical drives-per-host})$, and for firmware level 5.3 and 5.4: $512 / (\text{number-of-hosts} * \text{logical drives-per-host})$

As an example: A storage server with 4 hosts with 12, 14, 16 and 32 logical drives attached respectively would be calculated as follows:

$2048 / 4 * 32$ (largest number of logical drives per host) = 16
 $512 / 4 * 32$ (largest number of logical drives per host) = 4

If configuring only at the HBA level, you can use the formula: $2048 / (\text{total number-of-HBAs})$, and $512 / (\text{total number-of-HBAs})$ for the respective firmware levels.

Important: Setting queue depth too high can result in loss of data and possible file corruption; therefore, being conservative with these settings is better.

In the high throughput environments, you should try to use a host I/O blocksize that is equal to, or an even multiple of the stripe width of the logical drive being used.

Also, we are interested in settings that affect the large I/O blocksize, and settings mentioned earlier that may force a cache read-ahead value.

Additionally, you will want to ensure that the cache read-ahead value is enabled. This function is discussed in detail later in this chapter; but some operating system environments may have a variable value for changing this through device settings. Using the DS4000 to change this setting is the recommended method.

Finally, there are settings that may also impact performance with some servers and HBA types that enhance FC tape support. This setting should not be used with FC disks attached to the HBA.

Best practice: Though it is supported in theory, we strongly recommend that you keep *Fibre Channel tape* and *Fibre Channel disks* on separate *HBAs*. These devices have two very different data patterns when operating in their optimum mode, and the switching between them can cause undesired overhead and performance slowdown for the applications.

Host data layout

Ensure the host operating system aligns its device data partitions or slices, with those of the logical drive. Misalignment can result in numerous boundary crossings which are responsible for unnecessary multiple drive I/Os. Some operating systems do this automatically, and you just need to know the alignment boundary they use. Others however, may require manual intervention to set their start point to a value which would align them.

Understanding how your host based volume manager (if used) defines, and makes use of the logical drives once they are presented is also an important part of the data layout. As an example, the AIX Logical Volume Manager (LVM) is discussed in 2.5.1, “Planning for systems with LVM: AIX example” on page 61).

Volume managers are generally setup to place logical drives into usage groups for their use. The volume manager then creates volumes by carving up the logical drives into *partitions* (sometimes referred to as a *slice*); and then building a volume from them by either striping, or concatenating them to form the volume size desired. How the partitions are selected for use, and laid out may vary from system to system. In all cases, you need to ensure that spreading of the partitions is done in a manner to achieve maximum I/Os available to the logical drives in the group. Generally, large volumes are built across a number of different logical drives to bring more resources to bear. The selection of logical drives when doing this should be made carefully as not to use logical drives which will compete for resources, and degrade performance. The following are some general basic rules to apply.

- ▶ In a RAID 1 (Or RAID 10) environment, these logical drives can be from the same array, but must be *preferred path through different controllers* for greater bandwidth and resource utilization.
- ▶ For RAID 5, these logical drives should be on *separate arrays*, and *preferred path through different controllers*. This will ensure that the logical drives will not be in conflict with each other when the volume is used and both slices are accessed.

Best practice: For best performance, when building (host) volumes from logical drives, use logical drives from different arrays, with the preferred paths evenly spread among the two controllers of the DS4000 Storage Server.

- If striping is used, ensure that the stripe size chosen is a value that is complementing the size of the underlying *stripe width* defined for the logical drives (see “Logical drive segments” on page 158). The value used here will be dependent on the application and host I/O workload that will be using the volume. If the stripe width can be configured to sizes that complement the logical drives stripe; then benefits can be seen with using it. In most cases this model requires larger stripe values and careful planning to properly implement.

The exception to this would be for single threaded application processes with sequential I/O streams that have a high throughput requirement. In this case a small LVM stripe size of 64K or 128K can allow for the throughput to be spread across multiple logical drives on multiple arrays and controllers, spreading that single I/O thread out, and potentially giving you better performance. This model is generally not recommended for most workloads as the small stripe size of LVM can impair the DS4000 ability to detect and optimize for high sequential I/O workloads.

With file systems you again will want to revisit the need to ensure that they are aligned with the volume manager or the underlying RAID. Frequently, an offset is needed to ensure that the alignment is proper. Focus here should be to avoid involving multiple drives for a small I/O request due to poor data layout. Additionally, with some operating systems you can encounter interspersing of file index (also known as an inode) data with the user data, which can have a negative impact if you have implemented a *full stripe write* model. To avoid this issue you may wish to use raw devices (volumes) for full stripe write implementations.

4.3.2 Host setting examples

The following are example settings that may be used to start off your configuration in the specific workload environment. Those settings are suggested, they are not guaranteed to be the answer to all configurations. You should always try to setup a test of your data with your configuration to see if there is some further tuning that may better help (see recommended methods and tools in Chapter 6, “Analyzing and measuring performance” on page 187). Again, knowledge of your specific data I/O pattern is extremely helpful.

AIX operating system settings

The following section outlines the settings that can affect performance on an AIX host. We look at these in relations to how they impact the two different workload types.

Transaction settings

Early AIX driver releases allowed for the changing of the *cache read-ahead* through the logical drive attribute settings; this has been discontinued with current releases, and is now set by the DS4000 Storage Server value, and can now only report the value from the operating system.

All attribute values that are changeable can be changed using the **chdev** command for AIX. See the AIX man pages for details on the usage of **chdev**.

For the logical drive known as the hdisk in AIX, the setting is the attribute `queue_depth`:

```
# chdev -l hdiskX -a queue_depth=Y -P
```

In the above example “X” is the hdisk number, and “Y” is the value for `queue_depth` you are setting it to.

For the HBA settings, the attribute `num_cmd_elem` for the fcs device is used. This value should not exceed 512:

```
chdev -l fcsX -a num_cmd_elem=256 -P
```

Best practice: For high transactions on AIX, we recommend that you set `num_cmd_elem` to 256 for the fcs devices being used.

Throughput based settings

In the throughput based environment, you would want to decrease the queue depth setting to a smaller value such as 16. In a mixed application environment, you would not want to lower the “`num_cmd_elem`” setting, as other logical drives may need this higher value to perform. In a pure high throughput workload, this value will have no effect.

AIX settings which can directly affect throughput performance with large I/O blocksize are the `lg_term_dma`, and `max_xfer_size` parameters for the fcs device.

Best practice: The recommended start values for high throughput sequential I/O environments are `lg_term_dma = 0x800000` and `max_xfr_size = 0x200000`.

Note that setting the `max_xfer_size` affects the size of a memory area used for data transfer by the adapter. With the default value of `max_xfer_size=0x100000`, the area is 16 MB in size, and for other allowable values of `max_xfer_size`, the memory area is 128 MB in size.

See also 11.1.7, “Setting the HBA for best performance” on page 358.

AIX LVM impact

AIX uses Logical Volume Manager (LVM) to manage the logical drives and physical partitions. By default, with standard and big VGs, LVM reserves the first 512 bytes of the volume for the *Logical Volume Control Block*. Therefore, the first data block will start at an offset of 512 bytes into the volume. Care should be taken when laying out the segment size of the logical drive to enable the best alignment. You can eliminate the Logical Volume Control Block on the LV by using a scalable VG, or by using the `-T 0` option for big VGs.

Additionally, in AIX, file systems are aligned on a 16K boundary. Remembering these two items helps when planning for AIX to fit well with the DS4000 segment size. JFS and JFS2 file systems intersperse inode data with the actual user data, and can potentially disrupt the *full stripe write* activity. To avoid this issue, you can place files with heavy sequential writes on raw logical volumes. See also the recommendations defined in “Logical drive segments” on page 158.

With AIX LVM, it is generally recommended to spread high transaction logical volumes across the multiple logical drives that you have chosen, using the maximum interpolicy setting (also known as maximum range of physical volumes) with a random ordering of PVs for each LV. Ensure that your logical drive selection is done as recommended above, and is appropriate for the RAID type selected.

In environments with very high rate, sequentially accessed structures and a large I/O size, try to make the segment size times the (N-1 for RAID 5, or N/2 for RAID 10) to be equal to the application I/O size. And keep the number of sequential I/O streams per array to be less than the number of disks in the array.

Windows operating system settings

In this section we discuss settings for performance with the Windows operating system and the DS4000 Storage Server. Topics include:

- ▶ Fabric settings
- ▶ Disk types
- ▶ Disk alignment
- ▶ Allocation unit size

Fabric settings

With Windows operating systems, the queue depth settings are the responsibility of the host adapters, and configured through the BIOS setting. This varies from vendor to vendor: Refer to your manufacturer's instructions on how to configure your specific cards.

For IBM FASTT FC2-133 (and Qlogic based HBAs), the queue depth is known as *execution throttle*, which can be set with either the Qlogic SANSurfer tool, or in the BIOS of the Qlogic based HBA, by pressing CTL+Q during the boot process.

Disk types

With Windows 2000 and Windows 2003, there are two types of disks, basic disks and dynamic disks. By default, when a Windows system is installed, the basic disk system is used. Disks can be changed from basic to dynamic at anytime without impact on system or data.

Basic disks use partitions. These partitions in Windows 2000 are set to the size they were created. In Windows 2003, a primary partition on a basic disk can be extended using the **extend** command in the diskpart.exe utility.

In Windows 2000 and 2003, dynamic disks allow for expansion, spanning, striping, software mirroring, and software RAID 5.

With the DS4000 Storage Server, you can use either basic or dynamic disks. The appropriate type depends on your individual circumstances:

- ▶ In certain large installations where you may have the requirement to span or stripe logical drives and controllers to balance the workload, then dynamic disk may be your only choice.
- ▶ For smaller to mid-size installations, you may be able to simplify and just use basic disks.

When using the DS4000 as the storage system, the use of software mirroring and software RAID 5 is not required. Instead, configure the storage on the DS4000 storage server for the redundancy level required.

Basic disks

Basic disks and basic volumes are the storage types most often used with Windows operating systems. Basic disk is the default disk type during initial installation. A basic disk refers to a disk that contains basic volumes, such as primary partitions and logical drives. A basic volume refers to a partition on a basic disk. Basic disks are used in both x86-based and Itanium-based computers.

Basic disks support clustered disks. Basic disks do not support spanning, striping, mirroring and software level RAID 5. To use this functionality, you must convert the basic disk to a dynamic disk. If you want to add more space to existing primary partitions and logical drives, you can extend the volume using the **extend** command in the diskpart utility.

Here is the syntax for the **diskpart** utility to extend the disk:

Extend [size=n] [disk=n] noerr

Where:

size=n

The space in megabytes to add to the current partition. If you do not specify one it will take up all the unallocated space of that disk.

disk=n

The dynamic disk on which to extend the volume. Space equal to size=n is allocated on the disk. If no disk is specified the volume is extended on the current disk.

noerr

For scripting purposes. When an error is encountered the script will continue as if no error had occurred.

Using diskpart to extend a basic disk

In the following example, we have a Windows 2003 system with a basic disk partition of 20 GB. The partition has data on it, the partition is disk 3 and its drive letter is F. We have used the DS4000 Dynamic Volume Expansion (DVE) to expand the logical drive to 50 GB. This leaves the operating system with a disk of 50 GB, with a partition of 20 GB and free space of 30 GB. See Figure 4-1.

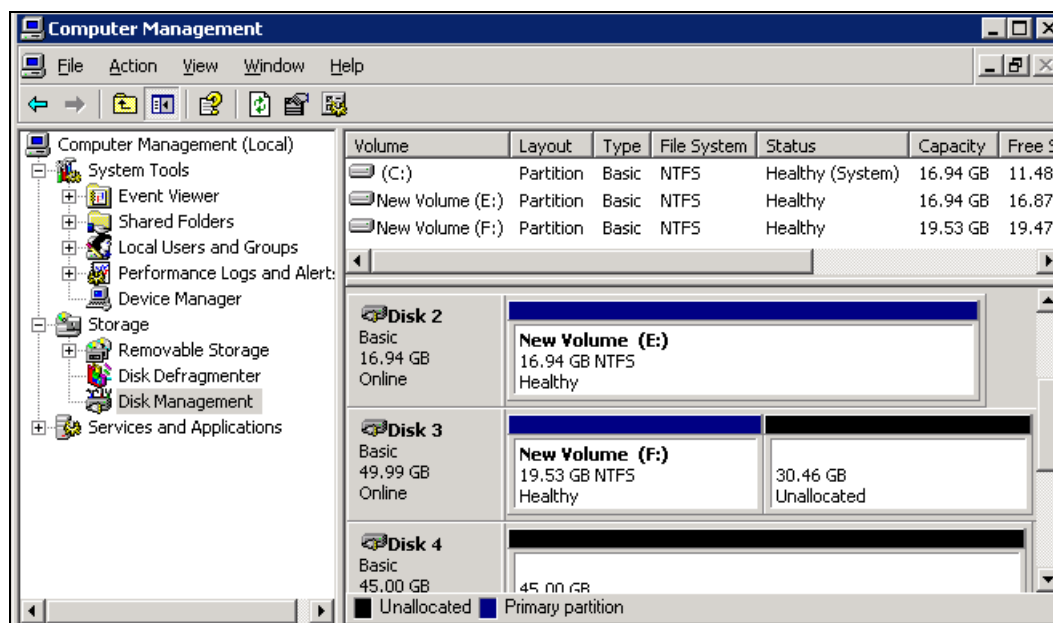


Figure 4-1 The Windows 2003 basic disk with free space

We use the Windows 2003 command line utility diskpart.exe to extend the 20 GB partition to the full size of the disk (Example 4-1).

Example 4-1 The diskpart utility to extend the basic disk in a command window

C:\>diskpart.exe

Microsoft DiskPart version 5.2.3790.1830
Copyright (C) 1999-2001 Microsoft Corporation.
On computer: RADON
DISKPART> list volume

Volume ###	Ltr	Label	Fs	Type	Size	Status	Info
Volume 0	D			CD-ROM	0 B	Healthy	
Volume 1	C		NTFS	Partition	17 GB	Healthy	System
Volume 2	E	New Volume	NTFS	Partition	17 GB	Healthy	
Volume 3	F	New Volume	NTFS	Partition	20 GB	Healthy	

DISKPART> select volume 3

Volume 3 is the selected volume.

DISKPART> extend

DiskPart successfully extended the volume.

DISKPART> list volume

Volume ###	Ltr	Label	Fs	Type	Size	Status	Info
Volume 0	D			CD-ROM	0 B	Healthy	
Volume 1	C		NTFS	Partition	17 GB	Healthy	System
Volume 2	E	New Volume	NTFS	Partition	17 GB	Healthy	
* Volume 3	F	New Volume	NTFS	Partition	50 GB	Healthy	

DISKPART>exit

C:\>

After diskpart.exe has extended the disk, the partition is now 50 GB. All the data is still intact and usable. See Figure 4-2.

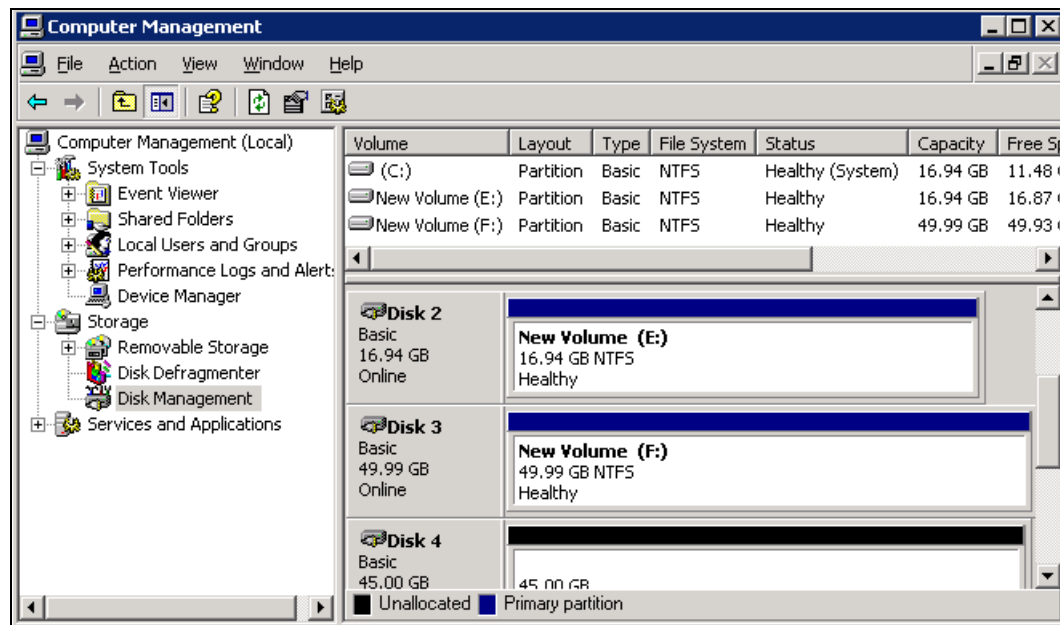


Figure 4-2 Disk Management after diskpart has extended the partition

Notes:

- ▶ The **extend** operation is dynamic.
- ▶ The **extend** command only works on NTFS formatted volumes.
- ▶ Officially, you do not need to stop I/O operations to the disk before you extend. However, keeping I/O operations at a minimum just makes good sense.

With dynamic disks, you use the Disk Management GUI utility to expand logical drives.

Dynamic disks

Dynamic disks were first introduced with Windows 2000 and provide some features that basic disks do not. These features are the ability to create volumes that span multiple disks (spanned and striped volumes), and the ability to create software based fault tolerant volumes (mirrored and RAID-5 volumes).

Dynamic disks can use the Master Boot Record (MBR) or GUID partition table (GPT) partitioning scheme; this depends on the version of operating system and hardware. The x86 platform uses MBR and the itanium based 64 bit versions use GPT or MBR.

All volumes on dynamic disks are known as dynamic volumes. There are five types of dynamic volumes that are currently supported:

- ▶ **Simple Volume:** A simple volume is a volume created on a single dynamic disk.
- ▶ **Spanned Volume:** Spanned volumes combine areas of unallocated space from multiple disks into one logical volume. The areas of unallocated space can be different sizes. Spanned volumes require two disks, and you can use up to 32 disks. If one of the disks containing a spanned volume fails, the entire volume fails, and all data on the spanned volume becomes inaccessible.

- **Striped Volume:** Striped volumes improve disk I/O performance by distributing I/O requests across multiple disks. Striped volumes are composed of stripes of data of equal size written across each disk in the volume. They are created from equally sized, unallocated areas on two or more disks. The size of each stripe is 64 KB and cannot be changed.

Striped volumes cannot be extended or mirrored and do not offer fault tolerance. If one of the disks containing a striped volume fails, the entire volume fails, and all data on the striped volume becomes inaccessible. The reliability for the striped volume is only as good as the least reliable disk in the set.

- **Mirrored Disk:** A mirrored volume is a software level fault tolerant volume that provides a copy of a volume on another disk. Mirrored volumes provide data redundancy by duplicating the information contained on the volume. Each disk in the mirror is always located on a different disk. If one of the disks fails, the data on the failed disk becomes unavailable, but the system continues to operate by using the available disk.
- **RAID 5 Volume:** A software RAID 5 volume is a fault tolerant volume that stripes data and parity across three or more disks. Parity is a calculated value that is used to reconstruct data if one disk fails. When a disk fails, the server continues to operate by recreating the data that was on the failed disk from the remaining data and parity. This is software level RAID and is not to be confused with hardware level RAID, this has a performance impact to the operating system. We do not recommend using this volume type with a DS4000 Storage Server logical drives included in it.

Dynamic disks offer greater flexibility for volume management because they use a database to track information about dynamic volumes on the disk, they also store information about other dynamic disks in the computer. Because each dynamic disk in a computer stores a replica of the dynamic disk database, you can repair a corrupted database on one dynamic disk by using the database from another dynamic disk in the computer.

The location of the database is determined by the partition style of the disk:

- On MBR disks, the database is contained in the last 1 megabyte of the disk.
- On GPT disks, the database is contained in a 1-MB reserved (hidden) partition known as the Logical Disk Manager (LDM) Metadata partition.

All online dynamic disks in a computer must be members of the same disk group. A disk group is a collection of dynamic disks. A computer can have only one dynamic disk group, this is called the primary disk group. Each disk in a disk group stores a replica of the dynamic disk database. A disk group usually has a name consisting of the computer name plus a suffix of Dg0.

Disk alignment

With the DS4000 logical drives as with physical disks that maintain 64 sectors per track, the Windows operating system always creates the partition starting with the sixty-fourth sector. This results in the partition data layout being misaligned with the segment size layout of the DS4000 logical drives. To ensure that these are both aligned, you should use the **diskpar.exe** or **diskpart.exe** command, to define the start location of the partition.

The **diskpar.exe** command is part of the Microsoft Windows Server 2000 Resource Kit, it is for Windows 2000 and Windows 2003. The **diskpar.exe** functionality was put into **diskpart.exe** with Windows Server 2003 Service Pack 1.

Using this tool, you can set the starting offset in the Master Boot Record (MBR) by selecting 64, or 128 (sectors). By setting this value to 64, you will skip the first 32K before the start of first partition. If you set this value to 128, you will skip the first full 64K segment (where the

MBR resides) before the start of the partition. The setting that you define depends on the allocation unit size of the formatted volume.

Doing so ensures track alignment and improves the performance. Other values can be defined, but these two offer the best chance to start out with the best alignment values. At a bare minimum, you should ensure that you aligned to at least a 4K boundary. Failure to do so may cause a single I/O operation to require the DS4000 to perform multiple I/O operations on its internal processing, causing extra work for a small host I/O, and resulting in performance degradation.

Important: The use of diskpart is a data destructive process. The diskpart utility is used to create partitions with the proper alignment. When used against a disk that contains data, all the data and the partitions on that disk must be wiped out, before the partition can be recreated with the storage track boundary alignment. Therefore, if the disk on which you will run diskpart contains data, you should back up the disk before performing the following procedure.

For more information, please use the following Web site:

<http://www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/0e24eb22-fbd5-4536-9cb4-2bd8e98806e7.mspx>.

In Example 4-2, we align disk 4 to the 128th sector and format it with 64K allocation unit size.

Example 4-2 Using diskpart in Windows 2003 Service Pack 1 to align the disks

From the command line run the following command:

diskpart.exe

```
DISKPART> select disk 4
Disk 4 is now the selected disk.
DISKPART> create partition primary align=128
DiskPart succeeded in creating the specified partition.
```

```
Exit diskpart utility and start the Disk Management snap-in.
Select the disk 4
Select format
Assign volume label
Leave file system as NTFS
Change Allocation unit size from default to 64K.
Click Ok.
```

In Example 4-3 we align Disk 4 to the 64th sector and format it with the default 4K allocation unit size.

Example 4-3 Using diskpart in Windows 2003 Service Pack 1 to align the disks

From the command line run the following command:

diskpart.exe

```
DISKPART> select disk 4
Disk 4 is now the selected disk.
DISKPART> create partition primary align=64
DiskPart succeeded in creating the specified partition.
```

Exit diskpart utility and start the Disk Management snap-in.
Select the disk 4
Select format
Assign volume label
Leave file system as NTFS
Leave Allocation unit size at default (4K).
Click Ok.

Allocation unit size

An allocation unit (or cluster) is the smallest amount of disk space that can be allocated to hold a file. All file systems used by Windows 2000, and Windows 2003 organize hard disks based on an allocation unit size, which is determined by the number of sectors that the cluster contains. For example, on a disk that uses 512-byte sectors, a 512-byte cluster contains one sector, while a 4 KB cluster contains eight sectors. See Table 4-1.

Table 4-1 Default cluster sizes for volumes

Volume size	Default NTFS allocation unit size
7 Mb to 512 Mb	512 bytes
513 Mb to 1024 Mb	1 Kb
1025 Mb to 2 Gb	2 Kb
2 Gb to 2 Tb	4 Kb

In the Disk Management snap-in, you can specify an allocation unit size of up to 64 KB when you format a volume. If you use the format command to format a volume, but do not specify an allocation unit size by using the /a:size parameter, the default values shown in Table 4-1 are used. If you want to change the cluster size after the volume is formatted, you must reformat the volume. This must be done with care as all data is lost when a volume is formatted. The available allocation unit sizes when formatting are 512 bytes, 1K, 2K, 4K, 8K, 16K, 32K, and 64K.

Important: The allocation unit size is set during a format of a volume. This procedure is data destructive, so if the volume on which you will run a format contains data, you should back up the volume before performing the format procedure.

Restriction: In Windows 2000 and Windows 2003, setting an allocation unit size larger than 4K will disable file or folder compression on that partition.

In Windows 2000, the disk defragmenter ceased to function with an allocation size greater than 4K. In Windows 2003, the disk defragmenter functions correctly.

In your environment, always test to ensure that all the functionality remains after any changes.

For more information about formatting an NTFS disk, see the Windows 2000 and Windows 2003 documentation.

4.4 Application considerations

When gathering data for planning from the application side, it is again important to first consider the workload type for the application.

If multiple applications or workload types will be sharing the system, you need to know the type of workloads each have; and if mixed (transaction and throughput based) which will be the most critical. Many environments have a mix of transaction and throughput workloads; with generally the transaction performance being considered the most critical.

However, in dedicated environments (for example, a TSM backup server with a dedicated DS4000 Storage Server attached), the streaming high throughput workload of the backup itself would be the critical part of the operation; and the backup database, though a transaction centered workload, is the less critical piece.

Transaction environments

Applications that use high transaction workloads are OnLine Transaction Processing (OLTP), mostly databases, mail servers, Web servers, and file servers.

If you have a database, you would tune the server type parameters as well as the database's logical drives to meet the needs of the database application. If the host server has a secondary role of performing nightly backups for the business, you would need another set of logical drives that are tuned for high throughput for the best backup performance you can get within the limitations of the mixed storage server's parameters.

So, what are the traits of a transaction based application? In the following sections we explain this in more detail.

As mentioned earlier, you can expect to see a high number of transactions and a fairly small block size. Different databases use different I/O sizes for their logs (see examples below), these vary from vendor to vendor. In all cases the logs are generally high write workloads. For table spaces most databases use between a 4 KB and 16 KB blocksize. In some applications larger chunks (for example, 64 KB) will be moved to host application cache memory for processing. Understanding how your application is going to handle its I/O is critical to laying out the data properly on the storage server.

In many cases the table space is generally a large file made up of small blocks of data records. The records are normally accessed using small I/Os of a *random* nature, which can result in about a 50% cache miss ratio.

For this reason, and to not waste space with unused data, plan for the DS4000 to read and write data into cache in small chunks. Avoid also doing any cache read-ahead with the logical

drives, due to the random nature of the I/Os (Web servers and file servers frequently use 8KB as well, and would generally follow these rules as well).

Another point to consider is whether the typical I/O is read, or write? In most Online Transaction Processing (OLTP) environments this is generally seen to be a mix of about 70% reads, and 30% writes. However, the transaction logs of a database application have a much high write ratio, and as such perform better in a different RAID array. This reason also adds to the need to place the logs on a separate logical drive which for best performance should be located on a different array that is defined to better support the heavy write need. Mail servers also frequently have a higher write ratio than read. Use the RAID array configuration for your specific usage model. This is covered in detail in Figure on page 155.

Best practice: Database table spaces, journals and logs should never be co-located on the same logical drive or RAID array. See further recommendations in 4.5, “DS4000 Storage Server considerations” on page 148 for RAID types to use.

Throughput environments

With throughput workloads you have fewer transactions, but much greater size I/Os. I/O sizes of 128K or greater are normally seen; and these I/Os are generally of a sequential nature. Applications that typify this type of workload are imaging, video servers, seismic processing, high performance computing (HPC) and backup servers.

With large size I/O, it is better to use large cache blocks to be able to write larger chunks into cache with each operation. Ensure the storage server is configured for this type of I/O load. This is covered in detail in “Cache blocksize selection” on page 154.

These environments best work when defining the I/O layout to be equal to or an even multiple of the storage servers *stripe width*. There are advantages with writes in the RAID 5 configurations that make setting the I/O size to equal that of the *full stripe* performant. Generally, the desire here is to make the sequential I/Os take as few back-end I/Os as possible, and to get maximum throughput from them. So, care should be taken when deciding how the logical drive will be defined. We discuss these choices in greater detail in “Logical drive segments” on page 158.

4.4.1 Application examples

For general suggestions and tips to consider when implementing certain applications with the DS4000 Storage Servers, refer to Chapter 5, “DS4000 tuning with typical applications” on page 163.

4.5 DS4000 Storage Server considerations

In this section we look at the specific details surrounding the DS4000 Storage Server itself when considering performance planning and configuring. Topics covered in this section are:

- ▶ Which model fits best
- ▶ Storage server processes
- ▶ Storage server modification functions
- ▶ Storage server parameters
- ▶ Disk drive types
- ▶ Arrays and logical drives
- ▶ Additional NVSRAM parameters of concern

4.5.1 Which model fits best

When planning for a DS4000 Storage Server, the first thing to consider is the choice of an appropriate model for your environment and the type of workload it will handle.

You want to be sure, if your workload is going to require a high number of disks, that the system chosen will support them, and the I/O workload they will be processing. If the workload you have requires a high storage server bandwidth, then you want your storage server to have a data bus bandwidth in line with the throughput and workload needs. See Figure 4-3, Figure 4-4 on page 150, and Figure 4-5 on page 150 for details on the DS4000 Storage Server family.

DS4000 Specification Comparison						
	DS4800 (80,82,84,88)	DS4500	DS4700 (70,72)	DS4300 Turbo	DS4200	DS4100
Host Interfaces	4 Gbps FC	2 Gbps FC	2 Gbps FC	2 Gbps FC	2Gp/s SFP	2 Gbps FC
SAN attachments	8 FC-SW	4 FC -SW	4 FC-SW	4 FC-SW	4 FC-SW	4 FC-SW
Direct attachments	8 FC-SW	8 FC-AL	4 FC-SW	4 FC-AL	4 FC-SW	4 FC-AL
Processor	Intel Xeon 2.4Ghz	Intel Pentium III 850 MHz	Intel xScale 667 MHz	Intel xScale 600MHz	Intel xScale 667MHz	Intel xScale 600MHz
Data Bus Bandwidth	Up to 1,6GB/s	Up to 790MB/s	Up to 990MB/s	Up to 400MB/s	Up to 1,6GB/s	Up to 485MB/s
XOR technology	ASIC	ASIC	ASIC	Integrated	ASIC	Integrated
memory	2/4/8/16 GB	2 GB	2/4 GB	2 GB	2 GB	512 MB

Figure 4-3 DS4000 Storage Server Family Comparison

DS4000 Specification Comparison– Cont'd

	DS4800	DS4500	DS4700	DS4300 Turbo	DS4200	DS4100
Redundant drive channels	Eight 4 Gb FC	Four 2 Gb FC	Two 4 Gb FC	Two 2 Gb FC	Two 4 Gb FC	Two 2 Gb FC 14 SATA
Drive types supported	FC or SATA	FC or SATA	FC or SATA	FC or SATA	EV-DDM SATA	SATA
Max drives	224	224	112	112	112 SATA	56 SATA
Max capacity with FC	67.2 TB**	67.2 TB**	33.6TB**	33.6 TB**	---	---
Max capacity with SATA	112 TB**	112TB**	48 TB**	48 TB**	56TB**	28TB**

** Physical capacity; usable capacity may be less.

Figure 4-4 DS4000 Storage Server Family Comparison - part 2

Maximum IOPS	DS4800	DS4500	DS4700	DS4300	DS4200	DS4100
Cache reads	575,000	148,000	120,000	77,500	120,000	70,000
Disk reads	86,000	53,200	44,000	25,000	11,200	10,000
Disk writes	22,000	10,900	9,000	5,200	1,800	2,000

Figure 4-5 DS4000 Family Transaction (IOPS) Comparison

In addition to doing your primary processing work, you may also have some storage server background processes you want to run. All work being performed requires resources, and you need to understand how they all will impact your DS4000 storage server.

4.5.2 Storage server processes

When planning for the system, remember to take into consideration any additional premium features and background management utilities you are planning to implement.

DS4000 copy services

With the DS4000 Storage Server, there is a complete suite of copy services features that are available. All of these features run internally on the DS4000 Storage Server, and therefore use processing resources. It is important that you understand the amount of overhead that these features require, and how they can impact your primary I/O processing performance.

Enhanced Remote Mirroring (ERM)

ERM is a critical back-end process to consider, as in most cases, it is expected to be constantly running with all applications while it is mirroring the primary logical drive to the secondary location. After the initial synchronization is complete, we have continuous mirroring updates that will run. These updates can be synchronous or asynchronous. For further details on these methods see 2.4.3, “Enhanced Remote Mirroring (ERM)” on page 58. When planning to use ERM with any application, you must carefully review the data workload model, as well as the networking path followed by the remote data.

The DS4800 is the best choice when ERM is a strategic part of your storage plan.

Consider also the impact of the initial synchronization overhead. When a storage subsystem logical drive is a primary logical drive and a full synchronization is necessary, the controller owner performs the full synchronization in the background while processing local I/O writes to the primary logical drive and associated remote writes to the secondary logical drive. Because the full synchronization diverts controller processing resources from I/O activity, it will impact performance on the host application. The *synchronization priority* allows you to define how much processing time is allocated for synchronization activities relative to other system work so that you can maintain accepted performance.

The synchronization priority rates are lowest, low, medium, high, and highest.

Note: The lowest priority rate favors system performance, but the full synchronization takes longer. The highest priority rate favors full synchronization, but system performance can be compromised.

The following guidelines roughly approximate the differences between the five priorities. Logical drive size and host I/O rate loads affect the synchronization time comparisons:

- ▶ A full synchronization at the *lowest* synchronization priority rate takes approximately eight times as long as a full synchronization at the highest synchronization priority rate.
- ▶ A full synchronization at the *low* synchronization priority rate takes approximately six times as long as a full synchronization at the highest synchronization priority rate.
- ▶ A full synchronization at the *medium* synchronization priority rate takes approximately three and a half times as long as a full synchronization at the highest synchronization priority rate.
- ▶ A full synchronization at the *high* synchronization priority rate takes approximately twice as long as a full synchronization at the highest synchronization priority rate.

The synchronization progress bar at the bottom of the Mirroring tab of the logical drive Properties dialog box displays the progress of a full synchronization.

VolumeCopy function

With VolumeCopy, several factors contribute to system performance, including I/O activity, logical drive RAID level, logical drive configuration (number of drives in the array or cache parameters), and logical drive type. For instance, copying from a FlashCopy logical drives might take more time to copy than standard logical drives. A major point to consider is whether you want to be able to perform the function while the host server applications are functioning, or during an outage. If the outage is not desired, using the FlashCopy volume as the source will allow you to perform your VolumeCopy in the background while normal processing continues. Like the ERM functions, VolumeCopy does have a background process penalty which you need to decide how much you want it to affect your front-end host. With the FlashCopy image you can use lower priority and leave it run for more extended time. This will make your VolumeCopy creation take longer but decrease the performance hit. As this value can be adjusted dynamically you can increase it when host processing is slower.

You can select the copy priority when you are creating a new logical drive copy, or you can change it later using the Copy Manager. The copy priority rates are lowest, low, medium, high, and highest.

Note: The lowest priority rate supports I/O activity, but the logical drive copy takes longer. The highest priority rate supports the logical drive copy, but I/O activity can be affected.

FlashCopy function

If you no longer need a FlashCopy logical drive, you should disable it. As long as a FlashCopy logical drive is enabled, your storage subsystem performance is impacted by the copy-on-write activity to the associated FlashCopy repository logical drive. When you disable a FlashCopy logical drive, the copy-on-write activity stops.

If you disable the FlashCopy logical drive instead of deleting it, you can retain it and its associated repository. Then, when you need to create a different FlashCopy of the same base logical drive, you can use the re-create option to reuse a disabled FlashCopy. This takes less time.

4.5.3 Storage server modification functions

The DS4000 Storage Servers have many modification functions that can be used to change, tuning, clean, or redefine the storage dynamically. Some of these functions are useful to help improve the performance as well. However, all of these will have an impact on the performance of the storage server and its host I/O processing. All of these functions use the *modification priority* rates to determine their process priority. Values to choose from are lowest, low, medium, high, and highest.

Note: The lowest priority rate favors system performance, but the modification operation takes longer. The highest priority rate favors the modification operation, but system performance can be compromised.

In the following sections we describe the use of each of these functions and their impact on performance.

Media Scan

Media Scan is a background check performed on all logical drives in the DS4000 storage server when selected to ensure that the blocks of data are good. This is accomplished by reading the logical drives one data stripe at a time into cache, and if successful it moves on to the next stripe. If a bad block is encountered, it will retry three times to read the block, and then go into its recovery process to rebuild the data block.

Media scan is configured to run on selected logical drives; and has a parameter for defining the maximum amount of time allowed to complete its run through all the logical drives selected. If the media scan process sees it is reaching its maximum run time and calculates that it is not going to complete in the time remaining, it will increase its priority and can impact host processing. Generally, it has been found that media scan scheduled with a “30 day” completion schedule, is able to complete if controller utilization does not exceed 95%. Shorter schedules would require lower utilization rates to avoid impact.

Best practice: Setting media scan to 30 days has been found to be a good general all around value to aid in keeping media clear and server background process load at an acceptable level.

Defragmenting an array

A fragmented array can result from logical drive deletion resulting in free space nodes or not using all available free capacity in a free capacity node during a logical drive creation.

Because creation of new logical drives cannot spread across several free space nodes, the logical drive size is limited to the greatest amount of a free space node available, even if there is more free space in the array. The array needs to be defragmented first to consolidate all

free space nodes to one free capacity node for the array. Then, a new logical drive can use the whole available free space.

Use the defragment option to consolidate all free capacity on a selected array. The defragmentation runs concurrently with normal I/O; it impacts performance, because the data of the logical drives must be moved within the array. Depending on the array configuration, this process continues to run for a long period of time.

Important: Once this procedure is started, it cannot be stopped; and no configuration changes can be performed on the array while it is running.

The defragmentation done on the DS4000 Storage Server only applies to the free space nodes on the array. It is not connected to a defragmentation of the file system used by the host operating systems in any way.

Copyback

Copyback refers to the copying of data from a hot-spare drive (used as a standby in case of possible drive failure) to a replacement drive. When you physically replace the failed drive, a copyback operation automatically occurs from the hot-spare drive to the replacement drive.

Initialization

This is the deletion of all data on a drive, logical drive, or array. In previous versions of the storage management software, this was called format.

Dynamic Segment Sizing (DSS)

Dynamic Segment Sizing describes a modification operation where the segment size for a select logical drive is changed to increase or decrease the number of data blocks that the segment size contains. A segment is the amount of data that the controller writes on a single drive in a logical drive before writing data on the next drive.

Dynamic Reconstruction Rate (DRR)

Dynamic Reconstruction Rate is a modification operation where data and parity within an array are used to regenerate the data to a replacement drive or a hot spare drive. Only data on a RAID-1, -3, or -5 logical drive can be reconstructed.

Dynamic RAID Level Migration (DRM)

Dynamic RAID Level Migration describes a modification operation used to change the RAID level on a selected array. The RAID level selected determines the level of performance and parity of an array.

Dynamic Capacity Expansion (DCE)

Dynamic Capacity Expansion describes a modification operation used to increase the available free capacity on an array. The increase in capacity is achieved by selecting unassigned drives to be added to the array. After the capacity expansion is completed, additional free capacity is available on the array for the creation of other logical drives. The additional free capacity can then be used to perform a Dynamic logical drive Expansion (DVE) on a standard or FlashCopy repository logical drive.

Dynamic logical drive Expansion (DVE)

Dynamic logical drive Expansion is a modification operation used to increase the capacity of a standard logical drive or a FlashCopy repository logical drive. The increase in capacity is achieved by using the free capacity available on the array of the standard or FlashCopy repository logical drive.

4.5.4 Storage server parameters

Settings on the DS4000 Storage Server are divided into two groups. The storage server wide parameters which affect all workloads that reside on the storage. And settings which are specific to the array or logical drive where the data resides.

Cache blocksize selection

On the DS4000 Storage Server the cache blocksize is a variable value that can be set to 4K or 16K. The main goal with setting this value is to not waste space. This is a storage server wide parameter, and when set, it is the value to be used by all cache operations.

For example, if the I/O of greatest interest is that from your database operations during the day rather than your weekly backups, you would want to tune this value to handle the high transactions best. Knowing that the higher transactions will have smaller I/O size, using the 4K setting is generally best for transaction intense environments.

Best practice: Set the cache blocksize to 4K for the DS4000 system normally for transaction intense environments.

In a throughput intense environment as we discussed earlier, you would want to get as much data into cache as possible. In this environment it is generally best to use the 16K blocksize for the cache.

Best practice: Set the cache blocksize to 16K for the DS4000 system normally for throughput intense environments

In mixed workload environments, you must decide which workload type is most critical and set the system wide settings to best handle your business needs.

Tip: Throughput operations though impacted by smaller cache blocksize can still perform reasonable if all other efforts have been accounted for. Transaction based operations are normally the higher concern, and therefore should be the focus for setting the server wide values if applicable.

Cache flush control settings

In addition to the cache blocksize the DS4000 Storage Server also has a cache control which determines the amount of data that can be held in write cache. With the *cache flush* settings you can determine what level of write cache usage can be reached before the server will start to flush the data to disks, and at what level the flushing will stop.

By default, these parameters are set to the value of “80” for each. This means that the server will wait until 80% of the write cache is used before it will flush the data to disk. In a fairly active write environment this value could be far too high. You can adjust these settings up and down until you find the value that best suites your environment. If the values are different from each other then back-end drive inactive time increases, and you have surging with peaks and valleys occurring instead of a steady usage of back-end disks.

You can also vary the maximum amount of time the write data can remain in cache prior to being forced out, and written to disks. This value by default is set to ten seconds but can be changed by using the Storage Manager command line interface command below:

```
'set logical Drive [LUN] cacheflushModifier=[new_value];'
```

Best practice: Start with “Start/Stop flush settings of 50/50, and adjust from there. Always keep them equal to each other.

4.5.5 Disk drive types

With the DS4000 Storage server there are many different types of disk drives that are available for you to choose from. There are both the Fibre Channel, as well as the Serial ATA (SATA) drives. The following is a table of the drives types that are available. We recommend using the 15K RPM models for the highest performance. The 10K RPM drives run a close second; while the 7200 RPM drives are the slowest. SATA drives can be used in lower transaction intense environments where maximum performance needs are less important, and high storage capacity, or price are main concerns. SATA drives do provide good throughput performance and can be a very good choice for these environments.

Drive sizes available for selecting from are 36 GB, 73 GB, 146 GB, and a 300 GB/10K RPM on the Fibre Channel side; and a 250 GB, 400 GB, or 500 GB SATA drive. When selecting a drive size for a performant DS4000 environment you must consider how much data will reside on each disk. If large drives are used to store a heavy transaction environment on fewer drives, then the performance will be impacted. If using large size drives and high numbers of them, then how the drive will be used becomes another variable. Some cases, where you prefer RAID 1 to RAID 5 a larger drive may be a reasonable cost compromise; but only testing with your real data and environment can show for sure.

Best practice: For transaction intense workload, we recommend using 36 GB or 73GB drives for the best performance.

The current DS4000 Storage Servers support a drive side I/O queue depth of “16” for the Fibre Channel disks. The SATA drives support only single I/Os. Refer also to 2.2.8, “Selecting drives” on page 35, and Table 2-3 on page 36 for additional information.

Arrays and logical drives

When setting up the DS4000 Storage Server the configuration of the arrays and logical drives is most likely the single most critical piece in your planning. Understanding how your workload will use the storage is crucial to the success of your performance, and your planning of the arrays and logical drives to be used.

RAID array types

When configuring a DS4000 Storage Server for the transaction intense environment you need to consider also whether it will be a read or write intensive workload. As mentioned earlier, in a database environment we actually have two separate environments with the table space, and the journals and logs. As the table space is normally high reads and low writes, and the journals and logs are high writes with low reads. This environment is best served by two different RAID types.

RAID 0, which is striping, without mirroring or parity protection, is generally the best choice for almost all environments for max performance; however, there is no protection built into RAID 0 at all, and a drive failure requires a complete restore. For protection it is necessary to look toward one of the other RAID types.

On the DS4000 Storage server RAID 1 is disk mirroring for the first pair of disks, and for larger arrays of four or more, disks mirroring and striping (RAID10 model) is used. This RAID type is a very good performer for high random write environments. It outperforms RAID 5 due to additional reads that RAID 5 requires in performing its parity check when doing the write

operations. With RAID1 there are two writes performed per operation; where as with RAID 5 there are two reads and two writes required for the same operation, totaling four I/Os.

A common use for RAID 1 is for the mailserver environment where random writes can frequently out-weigh the reads.

Some people feel strongly about database journals and logs also belonging on RAID 1 as well. This is something that should be reviewed for your specific environment. As these processes are generally sequential write intensive I/Os, you may find that RAID 5 with proper layout for the host I/O size can give you the same if not better performance with “full stripe write” planning.

RAID 5, however, is better than RAID 1 in high random read environments. This is due to there being a greater number of disks and the ability to have less seeking. This can make RAID 5 superior to RAID 1 when handling the OLTP type workload with the higher read and lower writes, even with increased table sizes.

Tip: There are no guaranteed choices as to which type of RAID to use; as this is very much dependent on the workload read and write activity. A good general guide may be to consider using RAID 1 if random writes exceed about 25%, with a peak sustained I/O rate exceeds 50% of the storage server’s capacity.

In the sequential high throughput environment RAID 5 can exactly perform excellent, as it can be configured to perform just one additional parity write when using “full stripe writes” (or “full stride writes”) to perform a large write I/O, as compared to the two writes per data drive (self, and its mirror) that are needed by RAID 1. This model is a definite advantage for RAID 5.

So with these differences, you can either place the database journals and logs on a RAID 1 or a RAID 5, depending on how sequential and well fit your data is. Testing of both types of arrays for your best performance is recommended.

Best practice: The differences described above outline the major reasons for our recommendation to keep the journals and logs on different arrays than the table spaces for database applications.

With the amount of variance that can exist with each customer environment, it is strongly recommended that you test your specific case for best performance; and decide which layout to use. With the write cache capability, in many cases RAID 5 write penalties are not noticed, as long as the back-end disk configuration is capable of keeping up with the front-end I/O load so processing is not being slowed. This again points to ensuring that proper spreading is done for best performance.

Number of disks per array

In the transaction intense environment it is more important to ensure that there are enough disk drives configured to perform the I/Os demanded by the host application, than to focus on the amount of possible storage space on the storage server.

With the DS4000 you can purchase 36 GB, 73 GB, 146 GB, or 300 GB Fibre Channel disk drives. Obviously, with the larger drives you can store more data using fewer drives.

Transaction intensive workload

In a transaction intense environment, you want to have higher drive numbers involved. This can be done by creating larger arrays with more disks of a smaller size. The DS4000 can

have a maximum of 30 drives per array/logical drive (although operating system limitations on a maximum logical drive capacity may restrict the usefulness of this capability).

Best practice: For high transactions environment, logical drives should be built on arrays with highest even number of data drives supported (+ 1parity) when using RAID 5; and highest number of supported and available drives / 2 when using RAID 10.

For large size databases consider using the host volume management software to build the database volume to be used for the application. Build the volume across sets of logical drives laid out per the RAID type discussion above. In using multiple arrays you will also be able to increase the controllers which are involved in handling the load therefore getting full use of the storage servers resources.

For example: If needing to build a database that is 1 TB in size, you can use five 300 GB drives in a 4 + 1parity RAID 5 single array/logical drive; or you could create two RAID 5 arrays of 8 + 1parity using 73 GB drives, giving two 584 GB logical drives on which to build the 1 TB database. In most cases the second method for large databases will work best, as it brings twelve more disks into play for handling the host side high transaction workload.

Large throughput workload

In the large throughput environment, it typically does not take high numbers of disks to reach the maximum sustained throughput. Considering that this type of workload is usually made of sequential I/O, which reduces disk latency, in most cases about 20 to 28 drives are enough to reach the maximum throughput.

This does, however, require that the drives be spread evenly across the DS4000 storage server to best utilize the server bandwidth. The DS4000 is optimized in its firmware to give increased throughput when the load is spread across all parts. Here, bringing all the DS4000 storage server resources into play is extremely important. Keeping the drives, channels, and bus busy with high data throughput is the winning answer. This is also a perfect model for using the high capacity drives, as we are looking to push a large volume of data and it will likely be large blocks of sequential reads and writes.

Consider building smaller arrays with single logical drives for higher combined throughput.

Best practice: For high throughput, logical drives should be built on arrays with 4+1, or 8+1 drives in them when using RAID 5. Data drive number and *segment size* should equal host I/O blocksize for full stripe write. Use multiple logical drives on separate arrays for maximum throughput.

An example configuration for this environment would be to have a single logical drive /array with 16+1 parity 300 GB disks doing all the transfers through one single path and controller; An alternative would be two 8 + 1parity defined to the two controllers using separate paths, doing two separate streams of heavy throughput in parallel and filling all the channels and resources at the same time. This keeps the whole server busy with a cost of one additional drive.

Further improvements may be gained by splitting the two 8 + 1 parity into four 4 + 1 parity arrays giving four streams, but addition of three drives would be needed. A main consideration here is to plan for the array data drive count to be a number such that the host I/O blocksize can be evenly spread using one of the DS4000 segment size selections. This will enable the full stripe write capability discussed in the next section.

Array and logical drive creation

An array is the grouping of drives together in a specific RAID type format on which the logical unit (logical drive) will be built for presentation to the hosts. As described in 3.3.2, “Creating arrays and logical drives” on page 106, there are a number of ways to create an array on the DS4000 Storage Server using the Storage Manager Client.

For best layout and optimum performance, we recommend that you manually select the drives when defining arrays. Drives should be across expansions for protection, and is best for them to stagger the across the odd and even slots for balance across the drive channel loops. Frequently this is accomplished by selecting them in an orthogonal manner. However you plan them, try to focus on keeping the load spread evenly.

With the DS4800 Storage Server the layout recommendation is slightly different. With the small configuration, we recommend that you try to bring as many channels into play as you can. When large disk expansion configurations are installed, we recommend that arrays and logical drives be built across expansions that are on a redundant channel drive loop pair as described in “Enclosure loss protection planning” on page 43 and 3.3.2, “Creating arrays and logical drives” on page 106.

A logical drive is the portion of the array that is presented to the host from the storage server. A logical drive can be equal to the entire size of the array, or just a portion of the array. A logical drive will be striped across all the data disks in the array. Generally, we recommend that you try to keep the number of logical drives on a array to a low number. However, in some cases this is not possible, and then planning of how the logical drives are used becomes very important. You must remember that each logical drive will have its own I/Os from its hosts queuing up against the drives that are in the same array. Multiple heavy transaction applications using the same array of disks can result in that array having poor performance for all its logical drives.

Best practice: We recommend that you create a single logical drive for each array when possible.

When configuring multiple logical drives on an array, try to spread out their usage evenly to have a balanced workload on the disks making up the arrays. A major point to remember here is to try to *keep all the disks busy*. Also, you will want to tune the logical drive separately for their specific workload environments.

Logical drive segments

The segment size is the maximum amount of data that is written or read from a disk per operation before the next disk in the array is used. As mentioned earlier, we recommend that for small host I/Os, the segment size be as large or larger than the host I/O size. This is to prevent the need to access a second drive for a single small host I/O. In some storage servers having the segment size equal to the host I/O size is recommended. This is not the case with the DS4000 servers.

There is no advantage in using smaller sizes with RAID 1; only in a few instances does this help with RAID 5 (which we discuss later). As the only amount of data written to cache is that which is to be written, for the I/O, there is no cache penalty either. As mentioned earlier in the host sections aligning data on segment boundaries is very important to performance. With larger segment sizes, there are less occasions of having misaligned boundaries impacting your performance as more small I/O boundaries reside within a single segment decreasing the chance of a host I/O spanning multiple drives. This can be used to help eliminate the effect of poor layout of the host data on the disks due to boundary differences.

Best practice: With the DS4000 Storage Server, we recommend that the segment size be 64 KB to 128 KB for most high transaction workloads.

With high throughput workload, the focus is on moving high throughput in fewer I/Os. This workload is generally sequential in nature.

Best practice: In the throughput environment, you want the *stripe size* to be equal to, or an even multiple of, the host I/O size.

These environments best work when we define the I/O layout to be equal to or an even multiple of the logical drive *stripe width*. The total of all the segments for one pass of all the back-end data disks is a *stripe*. So, large segment sizes that can equal the I/O size may be desired to accomplish the higher throughput you are looking for. For high read throughput, you want to have large segments (128K or higher) to get the most from each stripe. If the host I/O is 512 KB or larger, you would want to use at least a 256 KB segment size.

When the workload is high writes, and we are using RAID 5, we can use a method known as *full stripe (stride) write* which may work well to improve your performance. With RAID 5 the parity is based on the value calculated for a stripe. So, when the I/O being written is spread across the entire stripe width, no reads are required to calculate the parity; and the I/O completes with fewer back-end I/Os being required. This design may use a smaller segment size to align the host I/O size with the size of the stripe width. This type of management requires that very few host I/Os not equal a full stripe width.

The decrease in the overhead read operations is the advantage you are looking for. You must be very careful when implementing this type of layout to ensure that your data pattern does not change, and decrease its effectiveness. However, this layout may work well for you in a write intense environment. Due to the small size of segments, reads may suffer, so mixed I/O environments may not fair well. This would be worth testing if your writes are high.

Logical drive cache settings

Now, to help enhance the use of the system data cache, the DS4000 Storage Server has some very useful tuning parameters which help the specific logical drive in its defined environment. One such parameter is the *cache read-ahead multiplier* (see also 2.3.7, “Cache parameters” on page 53). This parameter is used to increase the number of segments that are read into cache to increase the amount of data that is readily available to present to the host for sequential I/O requests. To avoid excess read I/Os in the random small transaction intense environments you should disable the cache read-ahead multiplier for the logical drive by setting it to 0.

In the throughput environment where you want more throughput faster, you should generally enable this parameter to deliver more than one segment to the cache at a time.

- ▶ With pre-6.1x.xx firmware code this value could be set by the user to what was believed to be the value needed. For this code base we recommend the start value to be set to “4”, and adjusted as needed from there. When enabled the controller will read ahead into cache the next 4 sequential segments for immediate use from cache.
- ▶ With 6.1x and later code this value can be set to either “0” to disable the read-ahead feature, or “any value other than 0” to enable it. The 6.1x or later code will analyze the data pattern and determine the best value of read-ahead for the specific logical drive; and then use that value to pull additional cached segments for follow-on I/O requests to improve the performance. The number of segments read in advance are determined by the storage server using an algorithm of past I/O patterns to the logical drive. With the newer code you cannot set a value for the logical drive to stay with.

Best practice: For *high throughput with sequential I/O*, enable the cache read-ahead multiplier. For high transactions with random I/O, disable it.

For write I/O in a transaction based environment you can *enable write cache*, and *write cache mirroring* for cache protection. This allows the write I/Os to be acknowledged even before they are written to disks as the data is in cache, and backed up by the second mirror in the other controller's cache. This improves write performance dramatically; this sequence is actually a set of two options doing both write caching, and mirroring to the second cache. If you are performing a process that can handle loss of data, and can be restarted, you may chose to disable the mirroring, and see very high write performance.

In most transaction intense environments where the data is a live OLTP update environment, this is not an option that can be chosen; however, in cases like table updates, where you are loading in new values, and can restart the load over, this can be a way of greatly reducing the load time; and therefore shortening your downtime window. As much as three times the performance can be gained with this action.

For write I/O in the throughput sequential environment, these two parameters again come into play and can give you the same basic values. It should be noted that many sequential processes are more likely to be able to withstand the possible interrupt of data loss with the no cache mirroring selection, and therefore are better candidates for having the mirroring disabled.

Tip: The setting of these values is dynamic and can be varied as needed online. Starting and stopping of the mirroring may be implemented as a part of an update process.

In addition to usage of write cache to improve the host I/O response, you also have a control setting that can be varied on a per logical drive basis that defines the amount of time the write can remain in cache. This value by default is set to ten seconds, which generally has been found to be a very acceptable time; in cases where less time is needed, it can be changed by using the following Storage Manager command line interface command:

```
set logical Drive [LUN] cacheflushModifier=[new_value];
```

4.5.6 Additional NVSRAM parameters of concern

There are a number of DS4000 parameters that are defined specifically for the host type that is planned to be connected to the storage. These parameters are stored in the NVSRAM values that are defined for each host type. Two of these parameters can impact performance if not properly set for the environment. NVSRAM settings requiring change must be made using the Storage Manager Enterprise Management window and selecting the script execution tool.

The *Forced Unit Access* setting is used to instruct the DS4000 Storage Server to not use cache for I/O, but rather go direct to the disks. This parameter should be configured to ignore.

The *Synchronize Cache* setting is used to instruct DS4000 Storage Server to honor the SCSI cache flush to permanent storage command when received from the host servers. This parameter should be configured to ignore.

4.6 Fabric considerations

When connecting the DS4000 Storage Server to your SAN fabric it is best to consider what all the other devices, and servers are that will share the fabric network. This will be related to how you configure your zoning. See 2.1, “Planning your SAN and storage server” on page 14, for recommendations and details on how to establish the SAN infrastructure for your environment. Remember that a noisy fabric is a slow fabric. Unnecessary traffic makes for poor performance.

Specific SAN switch settings that are of particular interest to the DS4000 storage Server environment, and can impact performance are those that help to ensure *in-order-delivery* (IOD) of frames to the endpoints. The DS4000 cannot manage out of order frames, and retransmissions will be required for all frames of the transmitted packet. See your specific switch documentation for details on configuring parameters.



DS4000 tuning with typical applications

In this chapter we provide some general suggestions and tips to consider when implementing certain popular applications with the DS4000 Storage Servers.

Our intent is not to present a single way to set up your solution. Every situation and implementation will have their own specific or special needs.

This chapter provides general guidance as well as several tips for the following software products:

- ▶ IBM DB2
- ▶ Oracle Database
- ▶ Microsoft SQLserver
- ▶ IBM Tivoli Storage Manager
- ▶ Microsoft Exchange

5.1 DB2 database

In this section we discuss the usage of the DS4000 Storage Server with a DB2 database. We discuss the following topics:

- ▶ Data location
- ▶ Database structure
- ▶ Database RAID type
- ▶ Redo logs RAID type
- ▶ Volume management

5.1.1 Data location

With the DB2 applications, there are generally two types of data:

- ▶ Data consisting of the application programs, indexes, and tables, and stored in *table spaces*.
- ▶ Recovery data, made up of the database logs, archives, and backup management.

Generally, in an OLTP environment it is recommended to store these two data types separately: that is, on separate logical drives, on separate arrays. Under certain circumstances it can be advantageous to have both logs and data co-located on the same logical drives, but these are special cases and require testing to ensure that the benefit will be there for you.

5.1.2 Database structure

Table spaces can be configured in three possible environments:

- ▶ Database Managed Storage (DMS) table space
- ▶ System Managed Storage (SMS) table space
- ▶ Automatic Storage (AS) table space — new with V8.2.2

In a DMS environment, all the DB2 objects (data, indexes, large object data (LOB) and long field (LF)) for the same table space are stored in the same files. DB2 also stores metadata with these files as well for object management.

In an SMS environment, all the DB2 objects (data, indexes, LOB and LF) for the same table space are stored in separate files in the directory.

Restriction: When using concurrent or direct I/O (CIO or DIO) with earlier than AIX 5.2B, you must separate the LOB and LF files on separate table spaces due to I/O alignment issues.

In both DMS and SMS tablespace environments, you must define the container type to be used; either file system, or raw device.

In the AS tablespace environment, there are no containers defined. This model has a single management method for all the table spaces on the server that manages where the data is located for them on the storage.

In all cases, striping of the data is done on an *extent* basis. An extent can only belong to one object.

DB2 performs data retrieval by using three type of I/O *prefetch*:

- ▶ **RANGE:** Sequential access either in the query plan or through sequential detection at run time. Range request can be affected most by poor configuration settings.
- ▶ **LIST:** Prefetches a list of pages that are not necessarily in sequential order.

Note: DB2 will convert a LIST request to RANGE if it detects that sequential ranges exist.

- ▶ **LEAF:** Prefetches an index leaf page and the data pages pointed to by the leaf.
 - LEAF page is done as a single I/O.
 - Data pages on a leaf are submitted as a LISTrequest.

Prefetch is defined by configuring the application for the following parameter settings:

- ▶ **PREFETCHSIZE (PS):** A block of contiguous pages requested. The block is broken up into prefetch I/Os and placed on the prefetch queue based on the level of I/O parallelism that can be performed.
 - PS should be equal to the size of all the DS4000 logical drives stripe sizes so that all drives that make up the container are accessed for the prefetch request. For example if a container resides across two logical drives that were created on two separate RAID arrays of 8+1p, then when the prefetch is done, all 18 drives would be accessed in parallel.
- ▶ Prefetch is done on one extent at a time; but can be paralleled if possible with layout.
- ▶ **EXTENTSIZE (ES):** This is both the *unit of striping granularity*, and the *unit of prefetch I/O size*. Good performance of prefetch is dependent on a well configured ES:
 - Chose an extent size that is equal to or a multiple of the DS4000 logical drives segment size.

Best practice: The recommended ES should be a multiple of the segment size, and be evenly divisible into the stripe size.

- In general, you should configure the extent size to be between 128 KB and 1 MB, but at least should be equal to 16 pages. DB2 supports page sizes equal to 4 KB, 8 KB, 16 KB, or 32 KB in size. This means that an ES should not be less than 64 KB (16 X 4 KB (DB2's smallest page size)).
- ▶ Prefetch I/O parallelism for DS4000 performance requires DB2_PARALLEL_IO to be enabled.

This allows you to configure for all or one table space to be enabled for it.

- ▶ **NUM_IOSERVERS:** The number of parallel I/O requests that you will be able to perform on a single table space.
- ▶ With V8.2 of DB2, a new feature AUTOMATIC_PREFETCHSIZE was introduced.

A new database table space will have DFT_PREFETCH_SZ= AUTOMATIC.

The AUTOMATIC setting assumes a RAID 5 array of 6+1p, and will not work properly with the recommended 8+1p size array. See DB2 documentation for details on proper settings to configure this new feature.

Figure 5-1 provides a diagram showing how all these pieces fit together.

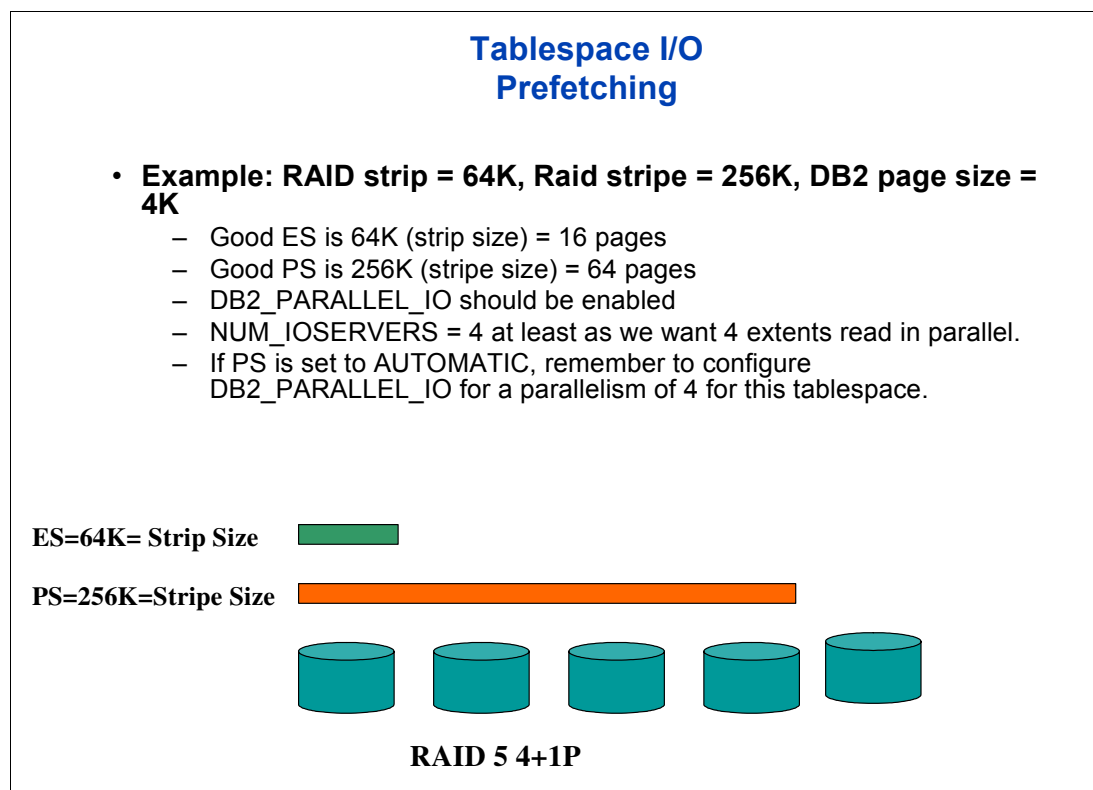


Figure 5-1 Diagram of table space I/O prefetch structure

5.1.3 Database RAID type

In many cases, OLTP environments contain a fairly high level of read workload. This is an area where your application may vary and behavior is very unpredictable. So you should try to test performance with your actual application and data.

In most cases, it has been found that laying out the datafile tables across a number of logical drives that were created across several RAID 5 arrays of 8+1 parity disks, and configured with a segment size of 64 KB or 128 KB, is a good starting point to begin testing with. This, coupled with host recommendations to help avoid offset and striping conflicts, would seem to provide a good performance start point to build from. A point to remember is that high write percentages may result in a need to use RAID 10 arrays rather than the RAID 5. This is environment specific and will require testing to determine. A rule of thumb is, if there are greater than 25–30% writes, then you may want to look at RAID 10 over RAID 5.

Best practice: Spread the containers across as many drives as possible, and ensure that the logical drive spread is evenly shared across the DS4000 Storage server's resources. Use multiple arrays where larger containers are needed.

DB2 uses a block based buffer pool to load the prefetched RANGE of I/O into. Though the RANGE is a sequential prefetch of data; in many cases the available blocks in the buffer may not be sequential. This can result in a performance impact. To assist with this management, some operating system primitives can help. These are VECTOR or SCATTER/GATHER I/O primitives. For some operating systems, you may need to enable the DB2_SCATTERED_IO to accomplish this function. There are also page cleaning parameters that can be configured

to help clear out the old or cold data from the memory buffers. See your DB2 documentation for details and recommendations.

5.1.4 DB2 logs and archives

The DB2 logs and archive files generally are high write workloads, and sequential in nature. We recommend that they be placed on RAID 10 logical drives.

As these are critical files to protect in case of failures, we recommend that you keep two full copies of them on separate disk arrays in the storage server. This is to protect you from the extremely unlikely occurrence of a double disk failure, which could result in data loss.

Also, as these are generally smaller files and require less space, we suggest that two separate arrays of 1+1 or 2+2 RAID1 be used to hold the logs and the mirror pair separately.

Logs in DB2 use the operating system's default blocksize for I/O (generally 4K) of sequential data at this time. As the small write size is has no greater penalty on the DS4000 Storage Server with higher segment size, our recommendation is that you configure the logical drive with a 64 KB or 128 KB segment size.

We also recommend that redo logs be placed on raw devices or volumes on the host system verses file system.

5.2 Oracle databases

In this section we discuss the usage of the DS4000 Storage Server with an Oracle database application environment. We discuss the following topics:

- ▶ Data types
- ▶ Data location
- ▶ Database RAID and disk type
- ▶ Redo logs RAID type
- ▶ Volume management

5.2.1 Data types

Oracle stores data at three levels:

- ▶ The first or lowest level consists of data blocks. These are also called logical blocks or pages.

One data block corresponds to a specific number of bytes corresponding to physical database space on the disk. A data block is the smallest unit of data used by a database. In contrast, at the physical, operating system level, all data is stored in bytes. Each operating system has a block size. Oracle requests data in multiples of Oracle data blocks, not operating system blocks.

- ▶ The second level consist of extents.

An extent is a specific number of contiguous data blocks. These are allocated for storing a specific type of information.

A table space that manages its extents locally can have either uniform extent sizes or variable extent sizes that are determined automatically by the system.

- For uniform extents, the default size of an extent is 1 MB.
- For system-managed extents, Oracle determines the optimal size of additional extents, with a minimum extent size of 64 KB. If the table spaces are created with a *segment*

space management auto, and if the database block size is 16 KB or higher, then Oracle manages segment size by creating extents with a minimum size of 1 MB. This is the default for permanent table spaces.

- ▶ The third level of database storage greater than an extent is called a segment.

A segment is a set of extents, each of which has been allocated for a specific data structure and all of which are stored in the same table space. Oracle allocates space for segments in units of one extent, as extents are allocated as needed. The extents of a segment may or may not be contiguous on disk. A segment and all its extents are stored in one table space. Within a table space, a segment can include extents from more than one file. The segment can span datafiles. However, each extent can contain data from only one datafile.

5.2.2 Data location

With Oracle applications, there are generally two types of data:

- ▶ Primary data, consisting of application programs, indexes, and tables
- ▶ Recovery data, consisting of database backups, archive logs, and redo logs

For data recovery reasons, in the past there has always been a recommendation that several categories of the RDBMS files be isolated from each other and placed in separate physical disk locations. This required that redo logs be separated from your data, indexes be separated from the tables, and rollback segments as well. Today, the recommendation is to keep user datafiles separated from any files needed to recover from any datafile failure.

This strategy ensures that the failure of a disk that contains a datafile does not also cause the loss of the backups or the redo logs needed to recover the datafile.

Since indexes can be rebuilt from the table data, it is not critical that they be physically isolated from the recovery-related files.

Since the Oracle control files, online redo logs, and archived redo logs are crucial for most backup and recovery operations, we recommend that at least two copies of these files be stored on different RAID arrays, and that both sets of these files should be isolated from your base user data as well.

In most cases with Oracle, the user data application workload is transaction based with high random I/O activity. With an OLTP application environment, you may see that an 8 KB database block size has been used, while with Data Warehousing applications, a 16 KB database block size is typical. Knowing what these values are set to, and how the devices on which the datafiles reside was formatted, can help in prevention of added disk I/O due to layout conflicts. For additional information see the previous discussion on host parameters in 4.3, “Host considerations” on page 136.

5.2.3 Database RAID and disk types

In many cases, OLTP environments contain a fairly high level of read workload. This is an area where your application may vary and behavior is very unpredictable. You should try to test performance with your actual application and data.

With the default extent size of 1 MB, the segment size should be selected to allow a full stripe write across all disks. In most cases, it has been found that laying out the datafile tables across a number of logical drives that were created across several RAID 5 arrays of 8+1 disks, and configured with a segment size of 64 KB or 128 KB, is a good starting point. This,

coupled with host recommendations to help avoid offset and striping conflicts, seems to provide a good performance start point to build from.

Keep in mind also that high write percentages may result in a need to use RAID 10 arrays rather than RAID 5. This is environment specific and will require testing to determine. A rule of thumb is if there are greater than 25–30% writes, then you may want to look at RAID 10 over RAID 5.

Best practice: Spread the datafiles across as many drives as possible, and ensure that the logical drive spread is evenly shared across the DS4000 Storage servers resources.

Enclosure loss protection should always be employed. This will ensure that in the unlikely event of an enclosure failure, then all the arrays and databases will continue to operate, although in a degraded state. Without enclosure loss protection, any failure of an enclosure will have major impact on those LUNs residing within that enclosure.

Use 15K rpm drives to ensure maximum throughput. This gives a 20–30% performance increase over 10K rpm disks. Using more lower-capacity, high-speed drives spreads the load across many spindles. This gives definite throughput advantages, but does use more disks.

Ensure that enough spare drives are on hand to ensure that any disk failure can fail over onto the spare disk. Also ensure that for each type and speed of disk utilized that enough spares are employed to cater. It is possibly good practice to use only high-speed versions of the drives used, as this will ensure that when a spare is in use that it does not degrade the array by having a slower speed drive introduced on a disk failure. A slower disk introduced to any array will cause the array to run at the slower speed.

5.2.4 Redo logs RAID type

The redo logs and control files of Oracle generally are both high write workloads, and sequential in nature. As these are critical files, we recommend keeping two full copies on separate disk arrays in the storage server. This is to protect against the (extremely) unlikely occurrence of a double disk failure, which could result in data lost.

We recommend that they be placed on RAID 10 logical drives. Avoid RAID 5 for these logs, dedicate a set of disks for the redo logs.

As these are generally smaller files and require less space, we suggest that two separate arrays of 1+1 or 2+2 RAID1 be used to hold the logs and the mirror pair separately.

Redo logs in Oracle use a 512 byte I/O blocksize of sequential data at this time. As the small write size has no greater penalty on the DS4000 Storage Server with higher segment size, our recommendation is to configure the logical drive with a 64 KB or 128 KB segment size.

We also recommend that you place redo logs on raw devices or volumes on the host system rather than the file system.

Also, we recommend keeping the archive logs on separate disks, as disk performance may degrade when the redo logs are being archived.

5.2.5 TEMP table space

If the files with high I/O are datafiles that belong to the TEMP table space, then investigate whether to tune the SQL statements performing disk sorts to avoid this activity, or to tune the sorting process. After the application has been tuned and the sort process checked to avoid

unnecessary I/O, if the layout is still not able to sustain the required throughput, then consider separating the TEMP table space onto its own disks.

5.2.6 Cache memory settings

The default cache memory on the DS4000 Storage Server is 4 KB, as 8 KB is an optimal database block size. Setting the cache block size to 16 KB will allow for better use of the cache memory.

Setting the Start/Stop Flushing size to 50/50 is a good place to start. Monitor the cache settings for best performance and tune the settings to requirements. Refer to Figure 5-2.

If Oracle Automatic Storage Management (ASM) is utilized to do the mirroring, then you can disable write cache mirroring on the storage subsystem. This will improve performance, but carries a risk in case of failure at the storage level since you will be relying on ASM to manage the disk writes cache.

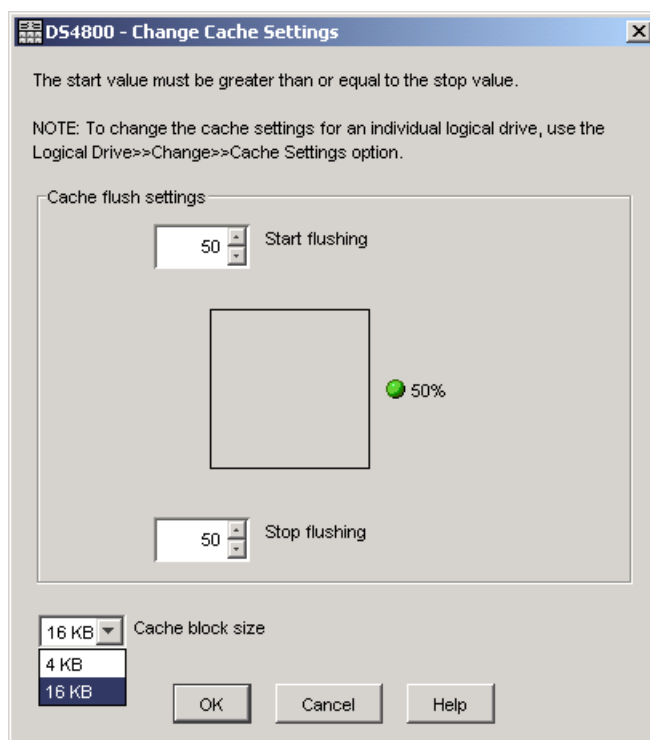


Figure 5-2 Start/stop flush settings and cache block size

5.2.7 Load balancing between controllers

Balance the load between the two controllers to ensure that the workload is evenly distributed. Load balancing is achieved by creating the LUNs/arrays evenly on each controller to balance I/Os and bandwidth requirements.

5.2.8 Volume management

Generally, the fewer volume groups, the better. However, if multiple volume groups are needed by the host volume manager due to the size of the databases, try to spread the logical drives from each array across all of the groups evenly. You should keep the two sets

of recovery logs and files in two separate volume groups. Therefore, as a rule, you would want to start with a minimum of three volume groups for your databases.

5.2.9 Performance Monitoring

With Oracle it is important to optimize the system to keep it running smoothly. In order to know if the system performs, a baseline is required so that you can compare the current state with a previously known state. Without the baseline, performance is based on perception rather than a real, documented measurement.

A common pitfall is to mistake the symptoms of a problem for the actual problem itself. It is important to recognize that many performance statistics indicate the symptoms, and that identifying the symptom is not sufficient data to implement a remedy.

For example, consider the situation of a slow physical I/O. Generally, this is caused by poorly configured disks. However, it could also be caused by a significant amount of unnecessary physical I/O on those disks issued by poorly tuned SQL statements or queries.

Ideally, baseline data gathered should include the following:

- ▶ Application statistics
- ▶ Database statistics
- ▶ Operating system statistics

Application statistics consist of figures such as active session histories, transaction volumes, and response times.

Database statistics provide information about the type of load on the database, as well as the internal and external resources used by the database.

Operating system statistics are CPU, memory, disk, and network statistics. For Windows environments, Windows Performance Monitor can be used to gather performance data. For UNIX and Linux environments, a range of tools can be employed to gather the performance data. Refer to Table 5-1.

Table 5-1 Linux tools commonly used to gather performance data

Component	Linux/UNIX tool to collect statistics
CPU	sar, vmstat, mpstat, or iostat
Memory	sar, vmstat
Disk	sar, iostat
Network	netstat

CPU statistics

CPU utilization is the most important operating system statistic in the tuning process. Make sure to gather CPU utilization for the entire system and for each individual CPU in a multiprocessor environments. The utilization figure for each CPU can help detect single threading and scalability issues.

Most operating systems report CPU usage as time spent in user mode and time spent in kernel mode. These additional statistics allow better analysis of what is actually being executed on the CPU.

On an Oracle data server, where there is generally only one application running, the server runs database activity in user mode. Activities required to service database requests, such as scheduling, synchronization, and memory management, run in kernel mode.

Virtual memory statistics

Virtual memory statistics should mainly be used as a check to validate that there is very little paging on the system. System performance degrades rapidly and unpredictably when excessive paging activity occurs.

Individual process memory statistics can detect memory leaks due to a programming fault to deallocate memory. These statistics should be used to validate that memory usage does not rapidly increase after the system has reached a steady state after startup.

Disk statistics

Because the database resides on a set of disks, the performance of the I/O subsystem is very important to the performance of the database. Most operating systems provide extensive statistics about disk performance. The most important disk statistics are the throughput, current response times and the length of the disk queues. These statistics show if the disk is performing optimally or if the disk is being overworked.

Measure the normal performance of the I/O system. If the response times are much higher than the normal performance value, then it is performing badly or is overworked. If disk queues start to exceed two, then the disk subsystem is a potential bottleneck of the system.

Network statistics

Network statistics can be used in much the same way as disk statistics to determine whether a network or network interface is overloaded or not performing optimally. For today's range of networked applications, network latency can be a large portion of the actual user response time. For this reason, these statistics are a crucial debugging tool.

5.3 Microsoft SQL Server

This section describes some of the considerations for Microsoft SQL server and the DS4000 Storage Server environment. If you have not done so, review "Windows operating system settings" on page 140. As with all recommendations, these settings should be checked to ensure that they suit your specific environment. Testing your own applications with your own data is the only true measurement.

This section includes the following topics:

- ▶ Allocation unit size and SQL Server
- ▶ RAID levels
- ▶ Disk drives
- ▶ File locations
- ▶ Transaction logs
- ▶ Databases
- ▶ Maintenance plans

5.3.1 Allocation unit size

When running on Windows 2000 and Windows 2003, SQL Server should be installed on disks formatted using NTFS. NTFS gives better performance and security to the file system. In Windows 2000 and Windows 2003, setting the file system allocation unit size to 64Kb will improve performance. Allocation unit size is set when a disk is formatted.

Adjusting the allocation unit other than the default does affect features, for example, file compression. Use this setting first in a test environment to ensure that it gives the desired performance level and that the required features are enabled.

For more information about formatting an NTFS disk, see the Windows 2000 and Windows 2003 documentation.

5.3.2 RAID levels

Redundancy and performance are required for the SQL environment.

- ▶ RAID 1 or RAID 10 should be used for the databases, tempdb, and transaction logs.
- ▶ RAID 1, RAID 5, or RAID 10 can be used for the maintenance plans.

5.3.3 File locations

As with all database applications, we recommend that the database files and the transaction logs be kept on separate logical drives, and separate arrays, for best protection. Also, the tempdb and the backup area for any maintenance plans should be separated as well. Limit other uses for these arrays to minimize contention.

It is not a good idea to place any of the database, transaction logs, maintenance plans, or tempdb files in the same location as the operating system page file.

5.3.4 User database files

General recommendations for user database files are as follows:

- ▶ Create the databases on a physically separate RAID array. The databases are being constantly being read from and written to; therefore, using separate, dedicated arrays does not interfere with other operations such as the transaction logs, or maintenance plans. Depending upon the current size of the databases and expected growth, either a RAID 1 or RAID 10 array could give best performance and redundancy. RAID 5 could also be used, but with a slightly lower performance. Data redundancy is critical in the operation of the databases.
- ▶ The speed of the disk will also affect performance: Use the 15K RPM disks rather than 10K RPM disks. Avoid using SATA drives for the databases.
- ▶ Spread the array over many drives. The more drives that the I/O operations are being sent to, the better the performance. Keep in mind that best performance is between 5 to 12 drives.

DS4000 array settings

Here are the array settings:

- ▶ Segment size 64K or 128K (dependent on I/O profile)
- ▶ Read cache on
- ▶ Write cache on
- ▶ Write cache with mirroring on
- ▶ Read ahead multiplier enabled (1)

5.3.5 Tempdb database files

Tempdb is a default database created by SQL Server. It is used as a shared working area for a variety of activities, including temporary tables, sorting, subqueries, and aggregates with

GROUP BY or ORDER BY queries using DISTINCT (temporary worktables have to be created to remove duplicate rows), cursors, and hash joins.

It is good to enable tempdb I/O operations to occur in parallel to the I/O operations of related transactions. As tempdb is a scratch area and very update intensive, use RAID 1 or RAID 10 to achieve optimal performance benefits. RAID 5 is not recommended. The tempdb is reconstructed with each server restart.

The ALTER DATABASE command can be used to change the physical file location of the SQL Server logical file name associated with tempdb; hence the actual tempdb database.

Here are some general recommendations for the physical placement and database options set for the tempdb database:

- ▶ Allow the tempdb database to expand automatically as needed. This ensures that queries generating larger than expected intermediate result sets stored in the tempdb database are not terminated before execution is complete.
- ▶ Set the original size of the tempdb database files to a reasonable size to prevent the files from automatically expanding as more space is needed. If the tempdb database expands too frequently, performance can be affected.
- ▶ Set the file growth increment percentage to a reasonable size to avoid the tempdb database files from growing by too small a value. If the file growth is too small compared to the amount of data being written to the tempdb database, then tempdb may need to constantly expand, thereby affecting performance.
- ▶ If possible, place the tempdb database on its own separate logical drive to ensure good performance. Stripe the tempdb database across multiple disks for better performance.

DS4000 array settings

Here are the array settings:

- ▶ Segment size 64K or 128K (dependent on I/O profile)
- ▶ Read cache on
- ▶ Write cache on
- ▶ Write cache with mirroring on
- ▶ Read ahead multiplier disabled (0)

5.3.6 Transaction logs

General recommendations for creating transaction log files are as follows:

- ▶ Transaction logging is primarily sequential write I/O, favoring RAID 1 or RAID 10. Note that RAID 5 is not recommended. Given the criticality of the log files, RAID 0 is not recommended either, despite its improved performance.

There are considerable I/O performance benefits to be gained from separating transaction logging activity from other random disk I/O activity. Doing so allows the hard drives containing the log files to concentrate on sequential I/O. Note that there are times when the transaction log will need to be read as part of SQL Server operations such as replication, rollbacks, and deferred updates. SQL Servers that participate in replication should pay particular attention to making sure that all transaction log files have sufficient disk I/O processing power because of the read operations that frequently occur.

- ▶ The speed of the disk will also affect performance. Whenever possible, use the 15K rpm disks rather than 10K rpm disks. Avoid using SATA drives for the transaction logs.
- ▶ Set the original size of the transaction log file to a reasonable size to prevent the file from automatically expanding as more transaction log space is needed. As the transaction log

expands, a new virtual log file is created, and write operations to the transaction log wait while the transaction log is expanded. If the transaction log expands too frequently, performance can be affected.

- ▶ Set the file growth increment percentage to a reasonable size to prevent the file from growing by too small a value. If the file growth is too small compared to the number of log records being written to the transaction log, then the transaction log may need to expand constantly, affecting performance.
- ▶ Manually shrink the transaction log files rather than allowing Microsoft SQL Server to shrink the files automatically. Shrinking the transaction log can affect performance on a busy system due to the movement and locking of data pages.

DS4000 array settings

Here are the array settings:

- ▶ Segment size 64K or 128K (dependent on I/O profile)
- ▶ Read cache on
- ▶ Write cache on
- ▶ Write cache with mirroring on
- ▶ Read ahead multiplier disabled (0)

5.3.7 Maintenance plans

Maintenance plans are used to perform backup operations with the database still running. For best performance, it would be advisable to place the backup files in a location that is separate from the database files. Here are some general recommendations for maintenance plans:

- ▶ Maintenance plans allow you to back up the database while it is still running. The location for the database backups should be in a dedicated array that is separate from both the databases and transaction logs. For the most part, these are large sequential files.
- ▶ This array needs to be much larger than the database array, as you will keep multiple copies of the database backups and transaction log backups. A RAID 5 array will give good performance and redundancy.
- ▶ The speed of the disk will also affect performance, but will not be as critical as the database or transaction log arrays. The preference is to use 15K disks for maximum performance, but 10K or even SATA drives could be used for the maintenance plans; this depends on your environment's performance needs.
- ▶ Spread the array over many drives. The more drives that the I/O operations are being sent to, the better the performance. Keep in mind that best performance is between 5 to 12 drives. For more details on array configurations see "Arrays and logical drives" on page 155.
- ▶ Verify the integrity of the backup upon completion. Doing this performs an internal consistency check of the data and data pages within the database to ensure that a system or software problem has not damaged data.

DS4000 array settings

Here are the array settings:

- ▶ Segment size 128K or higher (dependent on I/O profile)
- ▶ Read cache on
- ▶ Write cache on
- ▶ Write cache with mirroring off
- ▶ Read ahead multiplier enabled (1)

5.4 IBM Tivoli Storage Manager backup server

With a TSM backup server environment, the major workload to be considered is the backup and restore functions which are generally throughput intensive environments. Therefore, try to ensure that the DS4000 Storage Server's server-wide settings are set for the high throughput recommended settings for cache blocksize. For a DS4000 Storage Server dedicated to the TSM environment, the cache blocksize should be set to 16 KB.

Best practice: For a DS4000 Storage Server dedicated to the TSM environment, the cache blocksize should be set to 16 KB.

The TSM application has two different sets of data storage needs. TSM uses an instance database to manage its storage operations and storage pools for storing the backup data.

In the following section, we use an example of a site with three TSM host servers sharing a DS4000 Storage Server, and managing twelve TSM instances across them. These servers manage the data stored in a 16 TB storage pool that is spread across them in fairly even portions. It is estimated that the database needs for this will be about 1.7 TB in size, giving us about 100-150 GB per instance.

The customer has chosen to use 146 GB FC drives for the databases, and 250 GB SATA drives for the storage pools.

Here are some general guidelines we followed for creating the TSM databases:

- ▶ For a DS4000 Storage Server being used for the TSM databases, you have a fairly high random write workload. We used the following general guideline recommendations for our TSM site with three TSM host servers (TSM1,2 and 3), and database needs to handle twelve TSM instances per server:
 - Use a RAID 10 array, with four or more drives (remember, the higher the drive count, the better the performance with high transaction workloads). If you have a number of TSM host servers, create a large RAID 10 array out of which you can create the logical drives that will handle the database needs for all the hosts applications. In our scenario we created a single RAID 10 array of 13 x 13 drives.
 - With TSM databases of 1-100 and 11-150 GB size being requested, we created logical drives of 50 GB striped across the above RAID 10 array; giving us 35 logical drives.
 - For TSM databases, we have found that the logical drives should have a large segment size defined. The recommendation of 256K has been found to work well.
 - Ensure that cache read-ahead is disabled by setting it to “0”.
- ▶ Use partition mapping to assign each TSM server the logical drives it will use for the databases it will have. Ensure that the correct host type setting is selected.
- ▶ Since the fabric connections will support both high transaction, and high throughput, you will need to set the host and any HBA settings available for high I/O support and high throughput both.
 - HBA setting for high IOPS to be to be queued.
 - Large blocksize support, in both memory and I/O transmission drivers.
- ▶ Set the host device and any logical drive specific settings for high transaction support, especially ensure you have a good queue depth level set for the logical drives being used. We recommend starting with a queue depth of “32” per logical drive, and 256 per host adapter.
- ▶ Using the host volume manager, we created a large volume group in which we placed all of the logical drives we have assigned to handle each of the instances managed by the

TSM server. As an example, suppose TSM1 has four instances to handle, each requiring 150 GB databases. In this case we have twelve logical drives for which we will build the volume group. Create the volume group using the following parameters:

- Logical drives need to be divided into small partitions to be used for volume creation across them.

Tip: We recommend using a partition size that is one size larger than the minimum allowed.

- For each TSM instance, create a volume of 150 GB in size spread across three of the logical drives using *minimum interpolicy*.
- Configure the TSM database volume to have a file system on it.
- Each TSM instance to be on its own separate file system built as defined in the steps above.
- In the TSM application create ten files for each instance; and define in a round-robin fashion to spread workload out across them.

Some general guidelines for TSM storage pools are as follows:

- ▶ Create as many RAID 5 arrays using a 4+1 parity scheme, as you can using the drives you have allocated for the storage pools. In the example above we have enough drives to create sixteen arrays.
- ▶ Create a logical drive of equal size on each of the arrays for each of the TSM host servers (in our example this is three).
 - Make each of the logical drives of equal size. For example, a 4+1p RAID 5 of 250 GB SATA drives could give us about 330 GB if divided by three. A good plan would be to have an even number of arrays to spread, if dividing arrays into an odd number of logical drives.

Best practice: The objective is to spread the logical drives evenly across all resources. Therefore, if configuring an odd number of logical drives per array, it is a good practice to have an even number of arrays.

- Use a segment size of 512 KB for very large blocksize and high sequential data.
- Define cache read-ahead to “1” (enabled) to ensure we get best throughput.
- ▶ Define one logical drive from each array to each TSM host server. Use partition mapping to assign each host its specific logical drives. Ensure the correct host type setting is selected.
- ▶ Using the host volume manager, create a large volume group containing one logical drive from each array defined for the specific storage pool’s use on the DS4000 Storage Server as outlined above.
 - These logical drives need to be divided into small partitions to be used for volume creation across them.
 - Create two raw volumes of even spread and size across all of the logical drives.
 - With high throughput workloads we recommend that you set the logical drive queue depth settings to a lower value like 16.

- ▶ The TSM application has a storage pool parameter setting that can be varied for use with tape drives.
txngroupmax=256 (change to 2048 for tape configuration support).

5.5 Microsoft Exchange

This section builds on Microsoft best practices, making recommendations based around storage design for deploying Microsoft Exchange 2003 messaging server on the family of DS4000 storage systems.

The configurations described here are based on Exchange 2003 storage best practice guidelines and a series of lengthy performance and functionality tests. The guidelines can be found at:

<http://www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/fa839f7d-f876-42c4-a335-338a1eb04d89.mspx>

This section is primarily concerned with the storage configuration and does not go into the decisions behind the Exchange 2003 configurations referenced. For more information about Exchange design, please use the URL:

<http://www.microsoft.com/technet/prodtechnol/exchange/2003/library/default.mspx>

We assume that:

- ▶ Exchange 2003 Enterprise Edition is running in a standalone configuration (non-clustered).
- ▶ Windows 2003 operating system, page file, and all application binaries are located on locally attached disks.
- ▶ All additional data, including Exchange logs, storage groups (SG), SMTP queues, and RSG (Recovery Storage Groups) are located on a DS4000 Fibre Channel Storage System.

5.5.1 Exchange configuration

All Exchange data is located in the Exchange store, consisting of three major components:

- ▶ Jet database (.edb file)
- ▶ Streaming database (.stm file)
- ▶ Transaction log files (.log files)

Each Exchange store component is written to differently. Performance will be greatly enhanced if the .edb files and corresponding .stm files are located on the same storage group, on one array, and the transaction log files are placed on a separate array.

The following list shows how the disk read/writes are performed for each Exchange store component.

- ▶ Jet database (.edb file):
 - Reads and writes are random
 - 4 KB page size
- ▶ Streaming database (.stm file):
 - Reads and writes are sequential
 - Variable page size that averages 8 KB in production

- ▶ Transaction log files (.log files):
 - 100% sequential writes during normal operations
 - 100% sequential reads during recovery operations
 - Writes vary in size from 512 bytes to the log buffer size, which is 5 MB

Additional activities that affect I/O

Here is a list of such activities:

- ▶ Zero out deleted database pages
- ▶ Content indexing
- ▶ SMTP mail transmission
- ▶ Paging
- ▶ MTA message handling
- ▶ Maintenance mode
- ▶ Virus scanning

User profiles

Table 5-2 lists mailbox profiles that can be used as a guideline for capacity planning of Exchange mailbox servers. These profiles represent mailbox access for the peak of an average user Outlook® (or Messaging Application Programming Interface,MAPI, based) client within an organization.

Table 5-2 User profiles and corresponding usage patterns

User type	Database volume IOPS	Send/receive per day	Mailbox size
Light	0.18	10 sent/50 received	< 50 MB
Average	0.4	20 sent/100 received	50 MB
Heavy	0.75	30 sent/100 received	100 MB

5.5.2 Calculating theoretical Exchange I/O usage

To estimate the number of IOPS an Exchange configuration may need to support, you can use the following formula:

Number of users (mailboxes) x I/O profile of user = required IOPS for database drives

Consider the following example:

1500 (users/mailboxes) x 0.75 (heavy user) = 1125 IOPS

Using a ratio of two reads for every write, which is 66% read and 33% writes, you would plan for 742.5 IOPS for read and 371.5 IOPS for writes.

All writes are committed to the log drive first and then written to the database drive. Approximately 10% of the total IOPS seen on the database drive will be seen on the log drive. The reason for a difference between the log entries and the database is that the log entries are combined to provide for better streaming of data.

Therefore, 10% of the total 1125 IOPS seen on the database drive will be seen on the log drive:

$1125/100 \times 10 = 112.5$

In this example, the drives would have to support the following IOPS:

- ▶ Logs = 112.5 IOPS
- ▶ Database = 1125 IOPS
- ▶ Total = 1237.5 IOPS

Note: It is assumed that Exchange is the only application running on the server. If other services are running, then the I/O profile would have to be amended to take into account the additional tasks running.

5.5.3 Calculating Exchange I/O usage from historical data

If an Exchange environment is already deployed, then historical performance data can be used to size the new environment.

This data can be captured with the Windows Performance Monitor using the following counters:

- ▶ Logical disk
- ▶ Physical disk
- ▶ Processor
- ▶ MS Exchange IS

To get the initial IOPS of the Exchange database, monitor the Logical Disk → Disk Transfers/sec → Instance=Drive letter that houses the Exchange Store database. (Add all drive letters that contain Exchange Database files).

This will need to be monitored over time to determine times of peak load.

Below is an example of how to calculate the I/O requirements for an Exchange deployment based on a DS4000 Storage System using RAID 10 arrays and having all Exchange transaction log and storage groups on their own individual arrays.

Assume the following values:

- ▶ Users/mailboxes = 1500
- ▶ Mailbox size = 50 MB
- ▶ Database IOPS = 925

To calculate the individual IOPS of a user divide database IOPS by the number of users.

$$925/1500 = 0.6166$$

To calculate the I/O overhead of a given RAID level, you need to have an understanding of the different RAID types. RAID 0 has no RAID penalty, as it is just a simple stripe over a number of disks — so, 1 write will equal 1 I/O. RAID 1 or RAID 10 is a mirrored pair of drives or multiple mirrored drives striped together, because of the mirror, it means that for every write committed 2 I/Os will be generated. RAID 5 uses disk striping with parity so, for every write committed, this will often translate to 4 I/Os, due to the need to read the original data and parity before writing the new data and new parity. For further information about RAID levels, please refer to 2.3.1, “Arrays and RAID levels” on page 37.

To calculate the RAID I/O penalty in the formulas below, use the following substitutions:

- ▶ RAID 0 = 1
- ▶ RAID 1 or RAID 10 = 2
- ▶ RAID 5 = 4

Using the formula:

$$[(\text{IOPS/mailbox} \times \text{READ RATIO}\%)] + [(\text{IOPS/mailbox} \times \text{WRITE RATIO}\%) \times \text{RAID penalty}]$$

Gives us:

$$[(925/1500 \times 66/100) = 0.4069] + [(925/1500 \times 33/100) = 0.2034 \times 2] = 0.4069$$

That is a total of 0.8139 IOPS per user, including the penalty for RAID 10.

Exchange 2003 EE supports up to four storage groups (SGs) with five databases within each storage group. Therefore, in this example, the Exchange server will have the 1500 users spread over three storage groups with 500 users in each storage group. The fourth storage group will be used as a recovery storage group (RSG). More information about RSGs can be found at the following URL:

<http://www.microsoft.com/technet/prodtechnol/exchange/guides/UseE2k3RecStorGrps/d42ef860-170b-44fe-94c3-ec68e3b0e0ff.mspx>

- To calculate the IOPS required for a storage group to support 500 users, apply the following formula:

$$\begin{aligned} \text{Users per storage group} \times \text{IOPS per user} &= \text{required IOPS per storage group} \\ 500 \times 0.8139 &= 406.95 \text{ IOPS per storage group} \end{aligned}$$

A percentage should be added for non-user tasks such as Exchange Online Maintenance, anti-virus, mass deletions, and various additional I/O intensive tasks. Best practice is to add 20% to the per user I/O figure.

- To calculate the additional IOPS per storage group, use the following formula:

$$\begin{aligned} \text{Users per storage group} \times \text{IOPS per user} \times 20\% &= \text{overhead in IOPS per storage group} \\ 500 \times 0.8139/100 \times 20 &= 81.39 \text{ IOPS overhead per SG} \end{aligned}$$

- The final IOPS required per storage group is determined by adding the user IOPS per storage group with the overhead IOPS per storage group.

$$\begin{aligned} \text{User IOPS per SG} + \text{overhead in IOPS per SG} &= \text{total IOPS per SG} \\ 406.95 + 81.39 &= 488.34 \end{aligned}$$

- The total required IOPS for the Exchange server in general would be as follows:

$$\begin{aligned} \text{Total IOPS per SG} \times \text{total number of SGs} \\ 488.34 \times 3 &= 1465.02 \end{aligned}$$

- The new IOPS user profile is obtained by dividing the total IOPS by the total number of users.

$$1465.02/1500 = 0.9766 \text{ IOPS per user}$$

- Taking this last figure and rounding it up gives us a 1.0 IOPS per user. This figure allows for times of extraordinary peak load on the server.

- Multiplying by the 1500 users supported by the server gives a figure of 1500 IOPS across all three storage groups, divided by the three storage groups on the server, means that each storage group will have to be able to sustain 500 IOPS.

- Microsoft best practice recommends that log drives be designed to take loads equal to 10% of those being handled by the storage group logical drive.

$$500/100 \times 10 = 50 \text{ IOPS}$$

- Microsoft best practice recommends that log files be kept on separate spindles (physical disks) from each other and the storage groups.

After extensive testing, it has been determined that a RAID 1 (mirrored pair) provides the best performance for Exchange transaction logs on the DS4000 series, which is consistent with Microsoft best practices. In addition, Microsoft recommends that the storage groups be

placed on RAID 10. Again, this has proved to provide the best performance however, RAID 5 will also provide the required IOPS performance in environments where the user I/O profile is less demanding.

Taking the same data used in the example above for RAID 10, for the Exchange storage groups only and substituting the RAID 10 penalty, which is 2 I/Os for the RAID 5 penalty, which could be up to 4 I/Os, the new RAID 5 storage groups would have to each deliver an additional 500 IOPS than for the RAID 10 configuration.

5.5.4 Path LUN assignment (RDAC/MPP)

RDAC/MPP (multi-path proxy) driver is an added layer in the OS driver stack. Its function is to provide multiple data paths to a storage system's logical drive transparent to the software above it, especially applications. (Refer also to 2.2.5, "Multipath driver selection" on page 27).

The features of the MPP driver include:

- ▶ Auto-discovery of multiple physical paths to the media
- ▶ Mapping of multiple data paths to a logical drive into a single, highly-reliable virtual path which is presented to the OS
- ▶ Transparent failover on path-related errors
- ▶ Automatic fail back when a failed path is restored to service
- ▶ Logging of important events

To optimize read/write performance, the various logical drives are assigned to specific paths as detailed in Table 5-3. This is to keep logical drives with similar read/write characteristics grouped down the same paths.

Table 5-3 Logical drives and path assignment

Path A	Path B
F: Logs 1	K: Storage Group 1
G: Logs 2	L: Storage Group 2
H: Logs 3	M: Storage Group 3
I: RSG Log area if required	N: Scratch disk area (and RSG if needed)
J: SMTP queues	

Note: In this table, the preferred paths have been set for optimal performance. In the event of a path failure, RDAC/MPP will act as designed and fail the logical drives between paths as appropriate.

5.5.5 Storage sizing for capacity and performance

Mailbox quota, database size, and the number of users are all factors that you have to consider during capacity planning. Considerations for additional capacity should include NTFS fragmentation, growth, and dynamic mailbox movement. Experience has determined that it is always a best practice to double capacity requirements wherever possible to allow for unplanned needs, for example:

- ▶ A 200 MB maximum mailbox quota
- ▶ About 25 GB maximum database size*
- ▶ Total mailbox capacity to be 1500 users
- ▶ Mailboxes to be located on one Exchange server

Maximum database size for Exchange 2003 SE (Standard Edition) is 16 GB increasing to 75 GB with sp2 and the theoretical limit for Exchange 2003 EE (Enterprise Edition) is 16 TB.

Exchange SE supports four storage groups with one mailbox store and one public folder database store. Exchange EE supports four storage groups with five mailbox or public folder database stores, located within each storage group to a maximum of 20 mailbox or public store databases across the four storage groups. For more information, refer to the URL:

<http://support.microsoft.com/default.aspx?scid=kb;en-us;822440>

Using the above data, the capacity planning was done as follows:

Database size / maximum mailbox size = number of mailboxes per database
25GB / 200MB = 125 mailboxes per database

There is a maximum of five databases per storage group:

Maximum mailboxes per database x database instances per SG = maximum mailboxes per SG
125 x 5 = 625 mailboxes per SG

There are three active storage groups on the Exchange server:

Storage groups per server x maximum mailboxes per SG = maximum mailboxes per server
3 x 625 = 1875 mailboxes per server

In addition to the database storage requirements listed above, logical drives/capacity will also need to be provided for the following:

- ▶ Log files
- ▶ Extra space added to each storage group for database maintenance and emergency database expansion
- ▶ A 50 GB logical drive for the SMTP and MTA working directories
- ▶ Additional logical drive capacity for one additional storage group, for spare capacity or as a recovery group to recover from a database corruption
- ▶ An additional logical drive for use with either the additional storage group or recovery storage group

Logical view of the storage design with capacities

Figure 5-3 shows a logical view of the storage design.

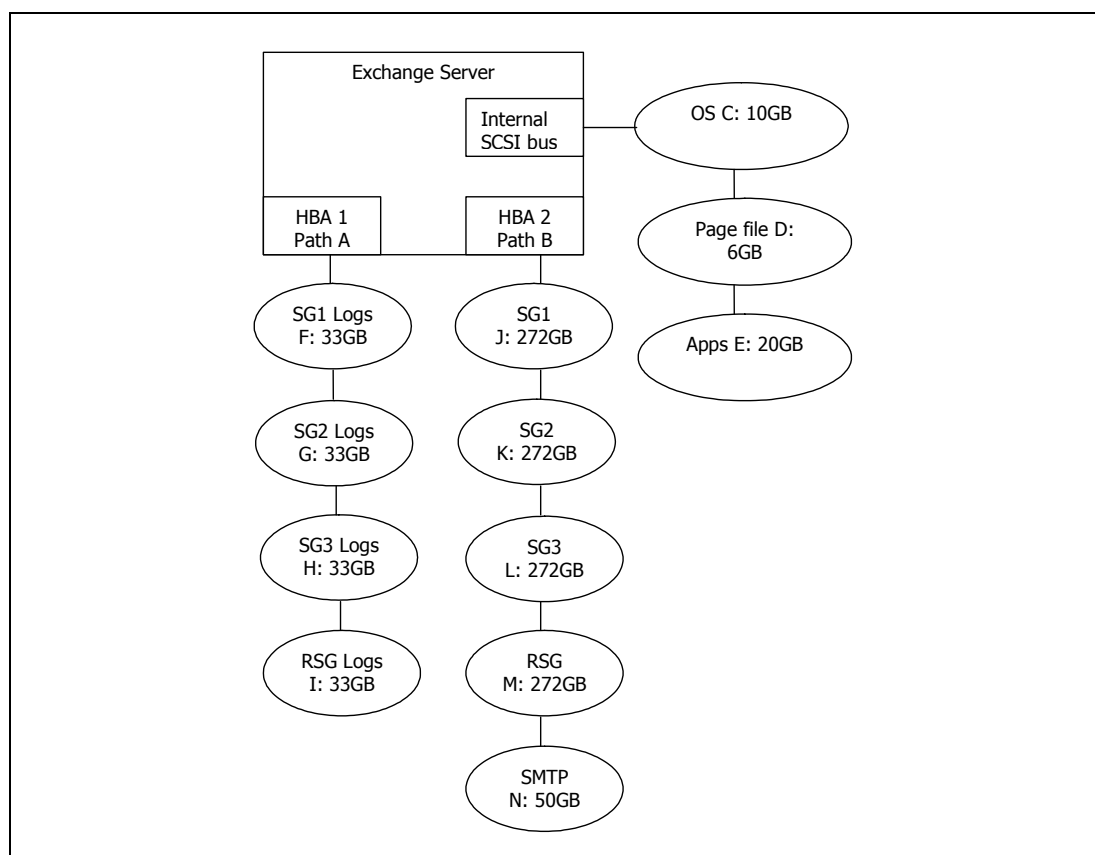


Figure 5-3 Logical view of the storage design with capacities

Table 5-4 details the drive letters, RAID levels, disks used, capacity and role of the logical drives that will be presented to Windows.

Table 5-4 Logical drive characteristics

Drive Letter	Size (GB)	Role	Location	RAID level	Array	Disks used
C	10 GB	Operating system	Local	RAID1	N/A	N/A
D	6 GB	Windows page file	Local	RAID1	N/A	N/A
E	18 GB	Applications	Local	RAID1	N/A	N/A
F	33 GB	SG1 Logs	SAN	RAID1	1	2 x 36 GB 15k
G	33 GB	SG2 Logs	SAN	RAID1	2	2 x 36 GB 15k
H	33 GB	SG3 Logs	SAN	RAID1	3	2 x 36 GB 15k
I	33 GB	RSG Logs	SAN	RAID1	4	2 x 36 GB 15k
J	272 GB	SG1 + maintenance	SAN	RAID10	5	8 x 73 GB 15k

Drive Letter	Size (GB)	Role	Location	RAID level	Array	Disks used
K	272 GB	SG2 + maintenance	SAN	RAID10	6	8 x 73 GB 15k
L	272 GB	SG3 + maintenance	SAN	RAID10	7	8 x 73 GB 15k
M	272 GB	Recovery Storage Group + maintenance	SAN	RAID10	8	8 x 73 GB 15k
N	50 Gb	SMTP Queues & MTA data	SAN	RAID10	9	4 x 73 GB 15k

5.5.6 Storage system settings

Use the following settings:

- ▶ Global cache settings
 - 4k
 - Start flushing 50%
 - Stop flushing 50%
- ▶ Array settings
 - Log drives - RAID 1
 - Segment size 64 KB
 - Read ahead 0
 - Write cache on
 - Write cache with mirroring on
 - Read cache on
 - Storage group settings - RAID 10 or 5 depending on user I/O profile
 - Segment size 64 KB
 - Read ahead 0
 - Write cache on
 - Write cache with mirroring on
 - Read cache on

5.5.7 Aligning Exchange I/O with storage track boundaries

With a physical disk that maintains 64 sectors per track, Windows always creates the partition starting at the sixty-fourth sector, therefore misaligning it with the underlying physical disk. To be certain of disk alignment, use diskpart.exe, a disk partition tool. The diskpart.exe utility is contained within Windows Server 2003 and Windows 2000 Server and can explicitly set the starting offset in the master boot record (MBR). By setting the starting offset, you can track alignment and improve disk performance. Exchange Server 2003 writes data in multiples of 4 KB I/O operations (4 KB for the databases and up to 32 KB for streaming files). Therefore, make sure that the starting offset is a multiple of 4 KB. Failure to do so may cause a single I/O operation to span two tracks, causing performance degradation.

Important: Using the diskpart utility to align storage track boundaries will be data destructive. When used against a disk, all data on the disk will be wiped out during the storage track boundary alignment process. Therefore, if the disk on which you will run diskpart contains data, back up the disk before performing the following procedure.

For more information, please refer to:

<http://www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/0e24eb22-fbd5-4536-9cb4-2bd8e98806e7.msp>

The diskpart supersedes the functionality previously found in diskpar.exe. Both diskpar and diskpart should only be used if the drive is translated as 64 sectors per track.

Additional considerations

It is important in planning an Exchange configuration to recognize that disks not only provide capacity but determine performance.

After extensive testing, it has been proven that the log drives perform best on a RAID 1 mirrored pair.

In line with Microsoft Exchange best practices, it has also been proven that RAID 10 for storage groups outperforms other RAID configurations. That said, depending on the user profile, RAID 5 will provide the required IOPS performance in some instances.

Laying out the file system with diskpart provides additional performance improvements and also reduces the risk of performance degradation over time as the file system fills.

Dedicating HBAs for specific data transfer has a significant impact on performance. For example, HBA1 to controller A for log drives (sequential writes), and HBA 2 to controller B for storage groups (random read and writes).

Disk latency can be further reduced by adding additional HBAs to the server. Four HBAs with logs and storage groups assigned to pairs of HBAs provided a significant reduction in disk latency and also reduces performance problems in the event of a path failure.

The Windows Performance Monitor can be used effectively to monitor an active Exchange server. Here are the counters of interest:

- ▶ Average disk sec/write
- ▶ Average disk sec/read
- ▶ Current disk queue length
- ▶ Disk transfers/sec

The average for the average disk sec/write and average disk sec/read on all database and log drives must be less than 20 ms.

The maximum of the average disk sec/write and average disk sec/read on all database and log drives must be less than 40 ms.



Analyzing and measuring performance

When implementing a storage solution, whether it is directly attached to a server, connected to the enterprise network (NAS), or on its own network (SAN — Fibre Channel or iSCSI), it is important to know just how well the storage performs. If you do not have this information, growth is difficult since you cannot properly manage it.

There are many different utilities and products that can help you measure and analyze performance. We introduce and review a few of them in this chapter:

- ▶ IOmeter
- ▶ Xdd
- ▶ Storage Manager Performance Monitor
- ▶ Various AIX utilities
- ▶ QLogic SANSurfer
- ▶ MPPUTIL in Windows 2000/2003
- ▶ Microsoft Windows Performance Monitor

Additional tools for performance analysis and measurement are presented in Chapter 7, “IBM TotalStorage Productivity Center for Disk” on page 231, and Chapter 8, “Disk Magic” on page 265.

6.1 Analyzing performance

To determine where a performance problem exists, it is important to gather data from all the components of the storage solution. It is not uncommon to be misled by a single piece of information and lulled into a false sense of knowing the cause of a poor system performance, only to realize that another component of the system is truly the cause.

In this section we look at what utilities, tools and monitors are available to help you analyze what is actually happening within your environment.

As we have seen in Chapter 4, “DS4000 performance tuning” on page 133, storage applications can be categorized according to two types of workloads: transaction based or throughput based.

- ▶ Transaction performance is generally perceived to be poor when these conditions occur:
 - Random reads/writes are exceeding 20ms (without write cache)
 - Random writes are exceeding 2ms with cache enabled
 - I/Os are queuing up in the operating system I/O stack (due to bottleneck)
- ▶ Throughput performance is generally perceived to be poor when the disk capability is not being reached. Causes of this can stem from the following situations:
 - With reads, read-ahead is being limited, preventing higher amounts of immediate data available.
 - I/Os are queuing up in the operating system I/O stack (due to bottleneck)

We discuss the following areas to consider:

- ▶ Gathering host server data
- ▶ Gathering fabric network data
- ▶ Gathering DS4000 storage server data

6.1.1 Gathering host server data

When gathering data from the host systems to analyze performance, it is important to gather the data from all host attached even though some may not be seeing any slowness. Indeed, a performant host may be impacting the others with its processing.

Gather all the statistics you can from the operating system tools and utilities. Data from these will help you when comparing them to what is seen with other measurement products. Utilities vary from operating system to operating system, so check with your administrators, or the operating system vendors.

Many UNIX type systems offer utilities that report disk I/O statistics and system statistics like: iostat, sar, vmstat, filemon, nmon, and iotop to mention a few. All are very helpful with determining where the poor performance originates. With each of these commands you want to gather a sample period of statistics. Gathering one minute to 15 minutes worth of samples during the slow period will give you a fair sampling of data to review.

In the example shown in Figure 6-1, we see the following information:

- ▶ Interval time = (894 + 4 KB)/15 KBps = 60 sec
- ▶ Average I/O size = 227.8 KBps/26.9 tps = 8.5 KB
- ▶ Estimated average I/O service time = 0.054/26.9 tps = 2 ms
- ▶ tps < 75 No I/O bottleneck!
- ▶ Disk service times good: No I/O bottleneck
- ▶ Disks not well balanced: hdisk0, hdisk1, hdisk5?

tty:	tin	tout	avg-cpu:	% user	% sys	% idle	% iowait
	24.7	71.3		8.3	2.4	85.6	3.6
Disks:	% tm_act	Kbps	tps	Kb_read	Kb_wrtn		
hdisk0	2.2	19.4	2.6	268	894		
hdisk1	1.0	15.0	1.7	4	894		
hdisk2	5.0	231.8	28.1	1944	11964		
hdisk4	5.4	227.8	26.9	2144	11524		
hdisk3	4.0	215.9	24.8	2040	10916		
hdisk5	0.0	0.0	0.0	0	0		

Figure 6-1 Example AIX iostat report

The information shown here does not necessarily indicate a problem. Hdisk5 may be a new drive that is not yet being used. All other indications appear to be within expected levels.

Windows operating systems offer device manager tools which may have performance gathering capabilities in them. Also, many third party products are available and frequently provide greater detail and graphical presentations of the data gathered.

6.1.2 Gathering fabric network data

To ensure the host path is clean and operating as desired, it is advised to gather any statistical information available from the switches or fabric analyzers for review of the switch configuration settings, and any logs or error data that may be gathered. This can be critical in determining problems that may cross multiple switch fabrics or other extended network environments.

IBM 2109 and Brocade switches offer a supportshow tool that enables you to run a single command and gather all support data at one time. McData type switches also have a similar capability in their ECM function. As well the Cisco switches offer this in their **show tech** command.

Additionally, you want to gather the host bus adapter parameter settings to review and ensure they are configured for the best performance for your environment. Many of these adapters have BIOS type utilities which will provide this information. Some operating systems (like AIX) can also provide much of this information through system attribute and configuration setting commands.

6.1.3 Gathering DS4000 storage server data

When gathering data from the DS4000 for analysis, there are two major functions you can use to document how the system is configured and how it performs. These two items functions are the Performance Monitor and Collect All Support Data.

- Performance Monitor

Using the performance Monitor, the DS4000 Storage Server can provide a point in time presentation of its performance at a specified time interval for a specified number of occurrences. This is useful to compare with the host data collected at the same time and can be very helpful in determining hot spots or other tuning issues to be addressed.

We provide details on the Performance Monitor in 6.4, “Storage Manager Performance Monitor” on page 204.

- Collect All Support Data

This function can be run from the “IBM TotalStorage DS4000 Storage Manager Subsystem Management” window by selecting **Advanced** → **TroubleShooting** → **Collect All Support Data**.

This will create a zip file of all the internal information of the DS4000 Storage Server for review by the support organization. This includes the storage Subsystems Profile, majorEventLog, driveDiagnosticData, NVSRAM data, readLinkStatus, performanceStatistics, and many others. This information, when combined and compared to the Performance Monitor data, will give a good picture of how the DS4000 Storage Server sees its workload being handled, and any areas it sees that are having trouble.

The performanceStatistics file provides you with a far greater amount of time coverage of data gathering, and a further breakout of the I/O details of what the storage server’s workload has been. In this spreadsheet based log, you can see what your read and write ratio are for all the logical drives, controllers, and the storage server’s workload. Also, you can view the cache hits and misses for each type (read and write).

6.2 Iometer

Iometer is a tool to generate workload and collect measurements for storage servers that are either directly attached or SAN attached to a host application server. Iometer was originally developed by Intel® Corporation, and is now maintained and distributed under an Intel Open Source License; the tool is available for download at:

<http://www.iometer.org>

6.2.1 Iometer components

Iometer consists of two programs, Iometer and Dynamo:

- *Iometer* is the controlling program. It offers a graphical interface to define the workload, set operating parameters, and start and stop tests. It can be configured in a number of ways that allow very granular testing to take place so that it can test multiple scenarios. After completion of tests, it summarizes the results in output files. Only one instance of Iometer should be running at a time, typically on the server machine.
- *Dynamo* is the workload generator. It has no user interface. Dynamo performs I/O operations and records performance information as specified by Iometer, then returns the data to Iometer. There can be more than one copy of Dynamo running at a time and it must be installed on each system for which you would like to gather performance results.

Dynamo is multi-threaded; each copy can simulate the workload of multiple client programs. Each running copy of Dynamo is called a *manager*; each thread within a copy of Dynamo is called a *worker*.

For a list of supported platforms, refer to:

<http://www.iometer.org/doc/matrix.html>

Iometer is extraordinarily versatile; refer to the Iometer user guide for more information.

In the sections that follow, we go over some basic configurations and explain how to run tests locally.

6.2.2 Configuring Iometer

Iometer is a completely simulated workload test.

Iometer (Dynamo) provides *workers* for each processor in the system, and it is recommended to have one worker for each processor in the system to keep all processors busy. This allows you to have multiple parallel I/Os issued and thus keep the disks in the storage server busy, as would be the case with intensive, high performance server applications.

Note that Iometer allows you to configure a specific access pattern for each worker (a set of parameters that defines how the worker generates workload and accesses the target).

Figure 6-2 (topology pane) shows an example where Iometer is running on a server (RADON) where one worker is also running. Note that ideally a test environment should include at least two or more machines, one running the main Iometer interface (IOMETER.EXE), the other ones each running the workload generator module (DYNAMO.EXE).

The right pane in Figure 6-2 consists of several tabs; the default tab is the Disk Targets tab, which lists the drives that are available to use for tests.

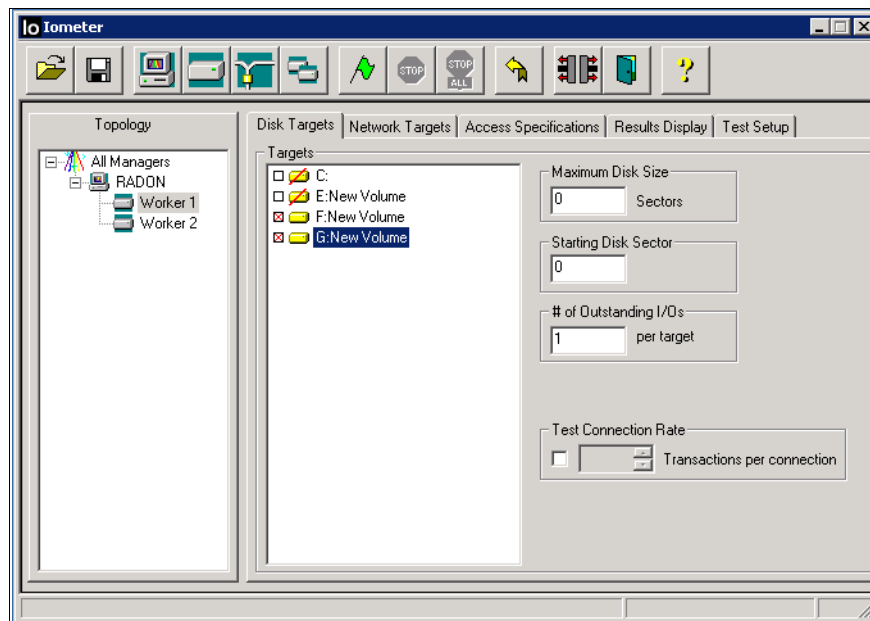


Figure 6-2 Disk targets

Define more workers if you need to simulate the activity that would be generated by more applications.

Before running any Iometer test, you must select access specifications from the Access Specifications tab as shown in Figure 6-3. The information on this tab specifies the type of I/O operations that will be simulated and performed by Iometer, allowing you to customize tests that match the characteristics of your particular application. The example illustrated in Figure 6-3 shows I/O tests with 512-byte and 32-KB chunks of data and a read frequency of 50% and 25%, respectively.

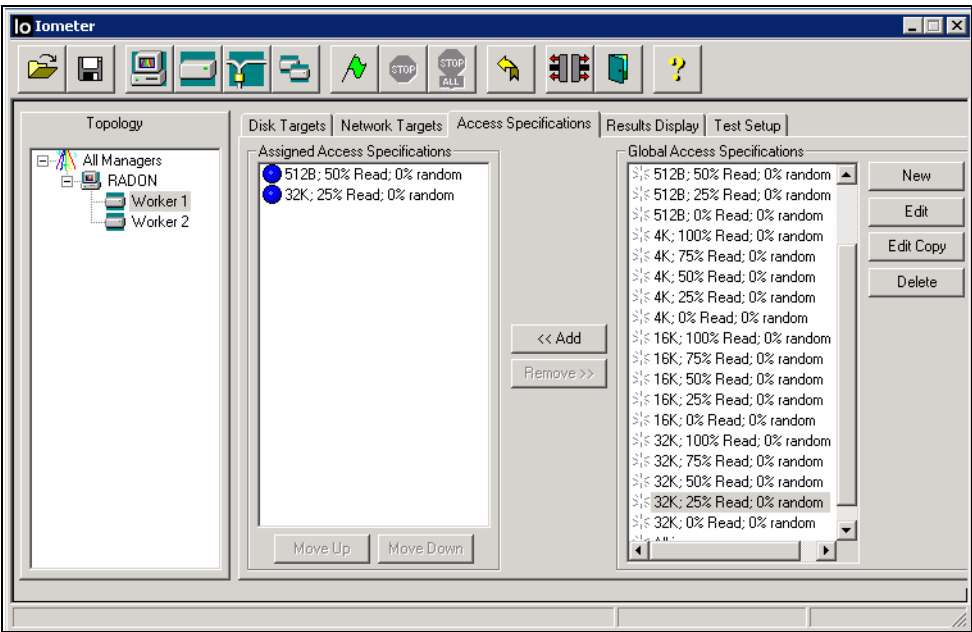


Figure 6-3 Access Specifications

To control precisely the access specifications, you can create new ones or edit existing ones by clicking the **New** or **Edit** buttons. See Figure 6-4.

Figure 6-4 Creating your own test

An access pattern contains the following variables:

- ▶ Transfer Request Size: A minimal data unit to which the test can apply.
- ▶ Percent Random/Sequential Distribution: Percentage of random requests. The other are, therefore, sequential.
- ▶ Percent Read/Write Distribution: Percentage of requests for reading: Another important variable which is not directly included in the access pattern (# of Outstanding I/Os) defines a number of simultaneous I/O requests for the given worker and, correspondingly, the disk load.

This gives you enormous flexibility and allows very granular analysis of the performance by changing one parameter at a time from test to test.

The various access specifications you define can be saved (and reloaded) in a file (this is a text file with the extension .icf. You can copy/paste the sample shown, into your environment.

To load the file in Iometer, click the **Open Test Configuration File** icon. See Example 6-1.

Example 6-1 Access specification file for Iometer (workload.icf)

```
'Access specifications
'Access specification name,default assignment
2K OLTP,1
'size,% of size,% reads,% random,delay,burst,align,reply
2048,100,67,100,0,1,0,0
'Access specification name,default assignment
4K OLTP,1
```

```

'size,% of size,% reads,% random,delay,burst,align,reply
4096,100,67,100,0,1,0,0
'Access specification name,default assignment
8K OLTP,1
'size,% of size,% reads,% random,delay,burst,align,reply
8192,100,67,100,0,1,0,0
'Access specification name,default assignment
32 Byte Data Streaming Read,3
'size,% of size,% reads,% random,delay,burst,align,reply
32,100,100,0,0,1,0,0
'Access specification name,default assignment
32 Byte Data Streaming Write,3
'size,% of size,% reads,% random,delay,burst,align,reply
32,100,0,0,0,1,0,0
'Access specification name,default assignment
512 Byte Data Streaming Read,2
'size,% of size,% reads,% random,delay,burst,align,reply
512,100,100,0,0,1,0,0
'Access specification name,default assignment
512 Byte Data Streaming Write,2
'size,% of size,% reads,% random,delay,burst,align,reply
512,100,0,0,0,1,0,0
'Access specification name,default assignment
8K Data Streaming Read,3
'size,% of size,% reads,% random,delay,burst,align,reply
8192,100,100,0,0,1,0,0
'Access specification name,default assignment
8K Data Streaming Write,3
'size,% of size,% reads,% random,delay,burst,align,reply
8192,100,0,0,0,1,0,0
'Access specification name,default assignment
64K Data Streaming Read,1
'size,% of size,% reads,% random,delay,burst,align,reply
65536,100,100,0,0,1,0,0
'Access specification name,default assignment
64K Data Streaming Write,1
'size,% of size,% reads,% random,delay,burst,align,reply
65536,100,0,0,0,1,0,0
'Access specification name,default assignment
TCP/IP Proxy Transfer,3
'size,% of size,% reads,% random,delay,burst,align,reply
350,100,100,100,0,1,0,11264
'Access specification name,default assignment
File Server,2
'size,% of size,% reads,% random,delay,burst,align,reply
512,10,80,100,0,1,0,0
1024,5,80,100,0,1,0,0
2048,5,80,100,0,1,0,0
4096,60,80,100,0,1,0,0
8192,2,80,100,0,1,0,0
16384,4,80,100,0,1,0,0
32768,4,80,100,0,1,0,0
65536,10,80,100,0,1,0,0
'Access specification name,default assignment
Web Server,2
'size,% of size,% reads,% random,delay,burst,align,reply
512,22,100,100,0,1,0,0
1024,15,100,100,0,1,0,0
2048,8,100,100,0,1,0,0
4096,23,100,100,0,1,0,0

```

```
8192,15,100,100,0,1,0,0
16384,2,100,100,0,1,0,0
32768,6,100,100,0,1,0,0
65536,7,100,100,0,1,0,0
131072,1,100,100,0,1,0,0
524288,1,100,100,0,1,0,0
'End access specifications
```

The next step is to configure the Test Setup as shown in Figure 6-5. This allows you to set parameters, such as the length of the test. For example, if you want to get a true average reading of your disk system performance, you might want to run the test for hours instead of seconds to minimize the impact of possible events, such as a user pulling a large file from the system. Figure 6-5 shows what parameters can be adjusted.

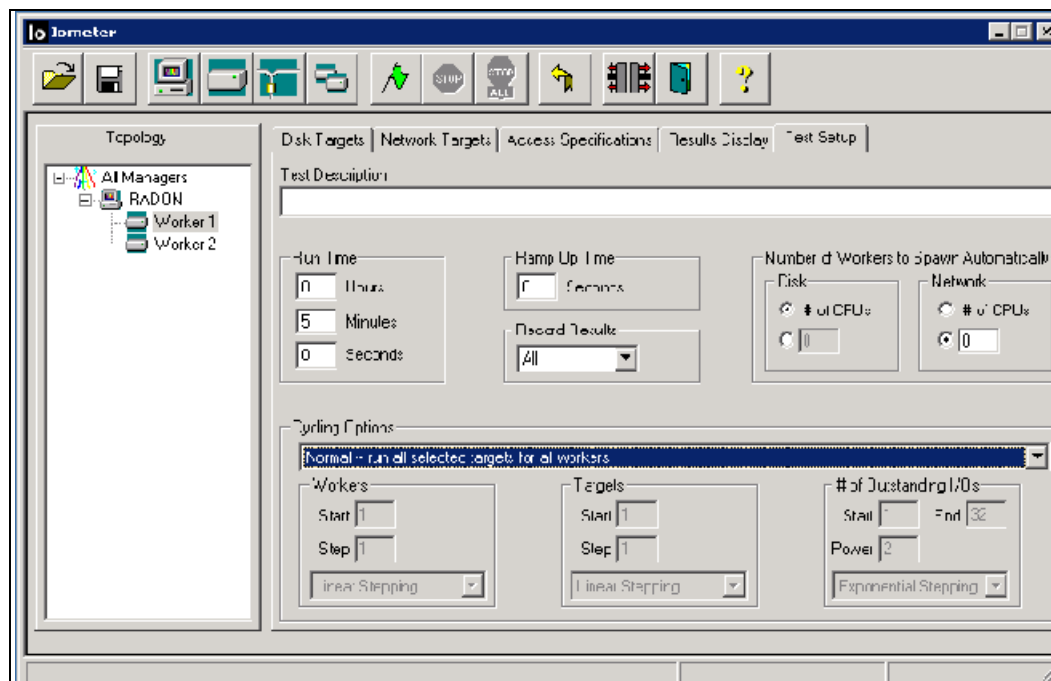


Figure 6-5 Test setup

Of importance here is the # of Outstanding I/Os parameter. It specifies the maximum number of outstanding asynchronous I/O operations per disk (also known as the queue depth) that each selected worker will maintain at any one time for each of the selected disks. The default for that parameter is 1. Real applications, on average, have a number of outstanding I/Os of around 60, and more than 200 for highly intensive I/Os applications.

Once you have configured your test, click the green flag on the menu bar to start running the test. Click the **Results Display** tab to get a real-time view of your running test.

6.2.3 Results Display

The Results Display tab shows performance statistics while a test is running. A sample results display is shown in Figure 6-6. You can choose which statistics are displayed, which managers or workers are included, and how often the display is updated. You can drill down on a particular result and display its own window by pressing the right-arrow at the right of each statistic displayed; this displays a window similar to the one shown in Figure 6-7.

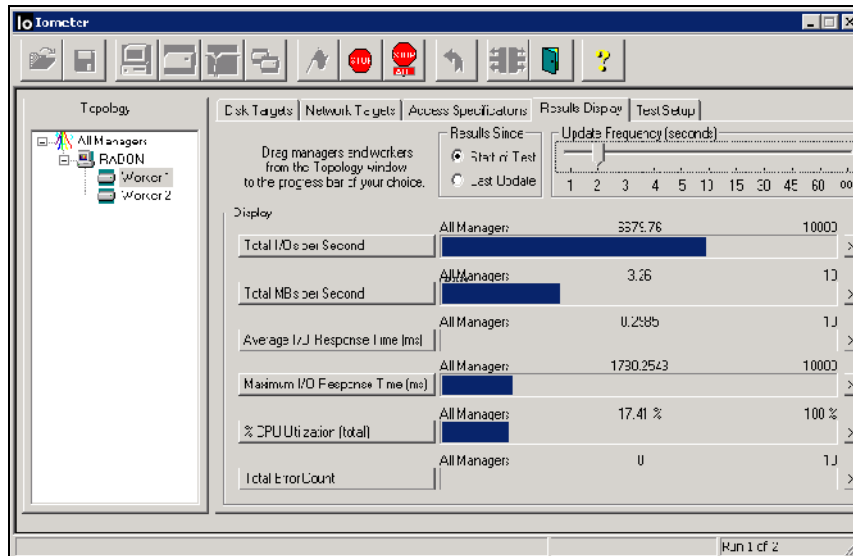


Figure 6-6 Results display



Figure 6-7 Drill down test display

The following results are included:

- ▶ **Total I/Os Per Second:** An average number of requests implemented per second. A request consists of positioning and read/write of the unit of the corresponding size.
- ▶ **Total MBs Per Second:** This is the same, but in other words, if the patterns are working with the units of the same size (Workstation and Database), this is just multiplication of Total I/Os Per Second by the unit's size.
- ▶ **Average I/O Response Time:** For linear loading (1 outstanding I/O) - this is again the same as Total I/Os Per Second (Total I/Os Per Second = 1000 milliseconds / Average I/O)

Response Time). With load increase, the value rises, but not arc-wise. The result depends on optimization of drive firmware, bus, and OS.

- ▶ CPU Effectiveness, or I/Os per % CPU Utilization.

It is important to note that the default view for each value on the Results Display panel is a sum total of ALL MANAGERS. To choose a specific manager or worker whose statistics are displayed by a particular chart, drag the desired worker or manager from the topology pane to the corresponding button.

The Result Display tab provides quick access to a lot of information about how the test is performing.

6.3 Xdd

Xdd is a tool for measuring and analyzing disk performance characteristics on single systems or clusters of systems. It was designed by Thomas M. Ruwart from I/O Performance, Inc. to provide consistent and reproducible performance of a sustained transfer rate of an I/O subsystem. It is a command-line based tool that grew out of the UNIX world and has been ported to run in Window's environments as well.

Xdd is a free software program distributed under a GNU General Public License. Xdd is available for download at:

<http://www.ioperformance.com/products.htm>

The Xdd distribution comes with all the source code necessary to install Xdd and the companion programs for the timeserver and the gettime utility programs.

6.3.1 Xdd components and mode of operation

There are three basic components to Xdd:

- ▶ The actual xdd program
- ▶ The timeserver program
- ▶ The gettime program

The timeserver and gettime programs are used to synchronize the clocks of all the servers that run the xdd program simultaneously, providing a consistent time stamping to accurately correlate xdd events in multiple server systems (Figure 6-8).

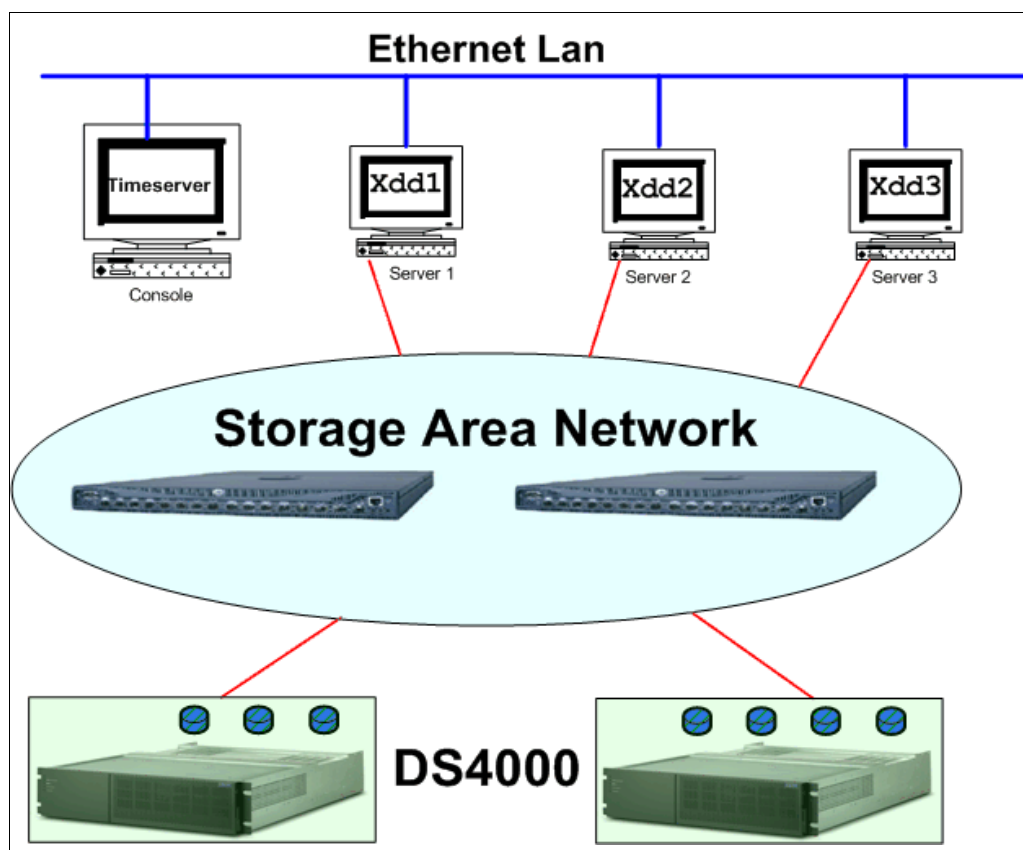


Figure 6-8 Xdd components in a multiple server system

Mode of operation

Xdd performs data transfer operations between memory and a disk device and collects performance information. Xdd creates one thread for every device or file under test. Each I/O operation is either a read or write operation of a fixed size known as the *request size*.

Multiple passes feature

For reproducibility of the results, an Xdd run must include several *passes*. Each pass executes some number of I/O requests on the specified target at the given request size. In general, each pass is identical to the previous passes in a run with respect to the request size, the number of requests to issue, and the access pattern. Passes are run one after another with no delay between passes unless a pass delay is specified.

Basic operations

Xdd is invoked through a command line interface. All parameters must be specified upon invocation, either directly on the command line or through a setup file.

- ▶ The operation to perform is specified by the `-op` option and can be either read or write.
- ▶ You can mix read and write operations using the `-rwratio` parameter.
- ▶ The request size is specified by `-reqsize` in units of blocks (1024 bytes); the blocksize can be overridden by the `-blocksize` option.

- ▶ All requests are sent to a target (disk device or file), and an Xdd process can operate on single or multiple targets, specified through the `-targets` option.
- ▶ Each Xdd thread runs independently until it has either completed all its I/O operations (number of transfers `-numreqs` or number of Megabytes `-mbytes`) or reached a time limit (`-timelimit` option)
- ▶ Several options are also available to specify the type of access pattern desired (sequential, staggered sequential, or random).

The de-skew feature

When testing a large number of devices (targets), there can be a significant delay between the time Xdd is started on the first target and the time it starts on the last one. Likewise, there will be a delay between when Xdd finishes on the first and last target. This causes the overall results to be skewed.

Xdd has a *de-skew* option that reports the bandwidth when all targets are active and transferring data (the amount of data transferred by any given target during the de-skew window is simply the total amount of data it actually transferred minus the data it transferred during the front-end skew period, and minus the data it transferred during the back-end skew period). The de-skewed data rate is the total amount of data transferred by all targets during the de-skew window, divided by the de-skew window time.

Read-behind-write feature

Xdd also has a read-behind-write feature. For the same target, Xdd can launch two threads: a writer thread, and a reader thread. After each record is written by the writer thread, it will block until the reader thread reads the record before it continues.

6.3.2 Compiling and installing Xdd

As part of our experiments in writing this book, we compiled and installed Xdd in an AIX environment.

To make Xdd, timeserver, and gettime in an AIX environment, first download the file `xdd63-1.030305.tar.gz`, then uncompress and extract the contents of the file with the **gzip** and **tar** commands, respectively:

```
# gzip -d xdd63-1.030305.tar.gz
# tar -xvf xdd63-1.030305.tar
```

Important: Verify the file `aix.makefile` before execution. It may contain undesirable control characters.

Open the file with an editor or browser and look at the end of each line for a ^M special character. If it is found, it must be removed. See Figure 6-9.

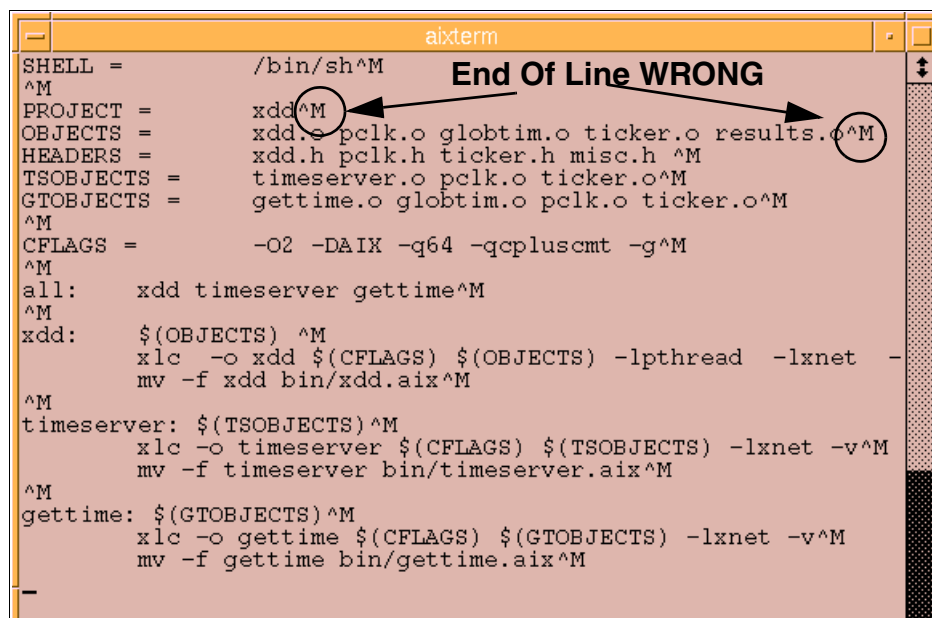


Figure 6-9 Example of an unwanted end-of-line control character

The best solution is to transfer all files with ftp in binary mode to a Windows system and transfer the file again with ftp in ascii mode.

Once the file is corrected (no ^M end of line), you can compile by issuing the following command:

```
# make -f aix.makefile all
```

This uses the **xlc** compiler and associated libraries. Ensure that the required libraries are installed for correct compilation and linking.

Verify the required file sets with:

```
# ls1pp -l | grep vac
vac.C                6.0.0.0  COMMITTED  C for AIX Compiler
vac.C.readme.ibm     6.0.0.0  COMMITTED  C for AIX iFOR/LS Information
vac.lic              6.0.0.0  COMMITTED  C for AIX Licence Files
vac.msg.en_US.C      6.0.0.0  COMMITTED  C for AIX Compiler Messages -
vac.C                6.0.0.0  COMMITTED  C for AIX Compiler
```

If you encounter a problem with your compiler, try the following steps:

```
cd /usr/opt/ifor/bin
i4cfg -stop
cd /var/ifor
rm *.dat
rm *.err
rm *.out
rm *.idx
rm i4ls.ini
```

To clear the data about the nodelock server, recreate i4ls.ini. Enter:

```
cd /usr/opt/ifor/bin
i4cnvini
```

Start the concurrent nodelock server and set it to restart on a system reboot. Enter:

```
i4cfg -a n -n n -S a -b null
i4cfg -start
```

Verify that the nodelock license server daemon, i4llmd, is running (others may start). Enter:

```
i4cfg -list
```

Register the concurrent nodelock license by entering the following command

```
i4blt -a -f /usr/vac/cforaix_cn.lic -T 10 -R "root"
```

Try to compile again with:

```
make -f aix.makefile
```

6.3.3 Running the xdd program

Xdd has a command-line interface that requires all the run-time parameters to be specified either on the **xdd** invocation command line or in a setup file. The format of the setup file is similar to the **xdd** command line in that the options can simply be entered into the setup file the same as they would be seen on the command line. The following example shows an **xdd** invocation using just the command line and the same invocation using the setup file along with the contents of the setup file.

Using the command line:

```
xdd -op read -targets 1 /dev/scsi/disk1 -reqsize 8 -numreqs 128 -verbose
```

Using a setup file:

```
xdd -setup xddrun.txt
```

Where the setup file xddrun.txt is an ASCII text file that contains the following Example 6-2

Example 6-2

```
-op read -targets 1 /dev/scsi/disk1
-reqsize 8
-numreqs 128
-verbose
```

Under Windows, you must replace “/dev/scsi/disk1” with “\\.\physicaldrive1” where physicaldrive1 is the disk as you can see it in the Windows disk storage manager.

Attention: Pay particular attention to which disk you use for the writing test, since all data will be lost on the target write disk.

Xdd examples in Windows

Enter the following command:

```
xdd -op read -targets 1 \\.\physicaldrive3 -reqsize 128 -mbytes 64 -passes 3
-verbose
```

This is a very basic test that will read sequentially from target device disk 3 starting at block 0 using a fixed request size of 128 blocks until it has read 64 megabytes.

It will do this 3 times and display performance information for each pass. The default block size is 1024 bytes per block, so the request size in bytes is 128 KB (128 * 1024 bytes).

Please note that all these options need to be on a single command line unless they are in the setup file, where they can be on separate lines.

Xdd examples in AIX

After compilation, the xdd executable file for AIX resides in the bin directory. In our case, it is called xdd.aix and is under /IBM/xdd/xdd62c/bin.

```
./xdd.aix -op read -targets 1 /dev/hdisk3 -reqsize 128 -mbytes 4 -passes 5
-verbose
```

This command will read sequentially from target disk /dev/hdisk3 starting at block0 using a fixed request size of 128 blocks, until it has read 4 Mb (4*1024*1024 bytes). The command runs for five times (passes) and displays performance information for each pass.

Example 6-3 Xdd example in AIX

Seconds before starting, 0										
		T	Q	Bytes	Ops	Time	Rate	IOPS	Latency	%CPU
TARGET	PASS0001	0	1	4194304	32	0.087	48.122	367.15	0.0027	22.22
TARGET	PASS0002	0	1	4194304	32	0.087	48.201	367.74	0.0027	37.50
TARGET	PASS0003	0	1	4194304	32	0.086	48.494	369.98	0.0027	33.33
TARGET	PASS0004	0	1	4194304	32	0.087	48.397	369.24	0.0027	22.22
TARGET	PASS0005	0	1	4194304	32	0.087	48.393	369.21	0.0027	75.00
TARGET	Average	0	1	20971520	16	0.434	48.321	368.66	0.0027	37.21
	Combined	1	1	20971520	16	0.434	48.321	368.66	0.0027	37.21
Ending time for this run, Wed Nov 9 07:43:58 2005										

In Example 6-4 we compare the latency value between an internal SCSI disk (hdisk2) and an FC disk (hdisk4) in a DS4000 storage server. For this test, we use hdisk2 for internal SCSI and hdisk4 for DS4000 fibre disk.

Example 6-4 Xdd example of SCSI disk in AIX

```
./xdd.aix -op read -targets 1 /dev/hdisk2 -reqsize 128 -mbytes 256 -passes 5 -verbose
```

Seconds before starting, 0										
		T	Q	Bytes	Ops	Time	Rate	IOPS	Latency	%CPU
TARGET	PASS0001	0	1	268435456	2048	9.486	28.298	215.89	0.0046	26.87
TARGET	PASS0002	0	1	268435456	2048	9.476	28.327	216.12	0.0046	25.95
TARGET	PASS0003	0	1	268435456	2048	9.476	28.327	216.12	0.0046	26.50
TARGET	PASS0004	0	1	268435456	2048	9.476	28.327	216.12	0.0046	27.22
TARGET	PASS0005	0	1	268435456	2048	9.483	28.306	215.96	0.0046	27.11
TARGET	Average	0	1	1342177280	10240	47.399	28.317	216.04	0.0046	26.73
	Combined	1	1	1342177280	10240	47.399	28.317	216.04	0.0046	26.73
Ending time for this run, Wed Nov 9 08:06:30 2005										

Example 6-5 Xdd example of FC disk in AIX

```
./xdd.aix -op read -targets 1 /dev/hdisk4 -reqsize 128 -mbytes 256 -passes 5 -verbose
```

```
Seconds before starting, 0
      T Q Bytes      Ops Time      Rate IOPS      Latency %CPU
TARGET PASS0001    0 1 268435456 2048 5.718 46.946 358.17 0.0028 36.19
TARGET PASS0002    0 1 268435456 2048 5.604 47.897 365.42 0.0027 37.68
TARGET PASS0003    0 1 268435456 2048 5.631 47.669 363.69 0.0027 34.99
TARGET PASS0004    0 1 268435456 2048 5.598 47.955 365.87 0.0027 38.57
TARGET PASS0005    0 1 268435456 2048 5.574 48.162 367.45 0.0027 37.46
TARGET Average    0 1 1342177280 10240 28.12 47.722 364.09 0.0027 36.97
      Combined    1 1 1342177280 10240 28.12 47.722 364.09 0.0027 36.97
Ending time for this run, Wed Nov 9 08:07:26 2005
```

The latency statistics are very different between the SCSI disk (0.0046) and the DS4000 disk (0.0027).

When creating new arrays for a DS4000, you can use the Xdd tool as illustrated in Example 6-5 to experiment and evaluate different RAID array topologies and identify the one that best suits your application.

You can also use Xdd to gather write statistics. In this case remember that all data on the target device will be lost. A write test is illustrated in Example 6-6. Options for the test are specified through the xdd.set option file.

```
# cat xdd.set
-blocksize 1024
-reqsize 128
-mbytes 4096
-verbose
-passes 5 -timelimit 10
#
```

Example 6-6 Xdd write test

```
./xdd.aix -op read -targets 1 /dev/hdisk4 -setup xdd.set
```

```
Seconds before starting, 0
      T Q Bytes      Ops Time      Rate      IOPS      Latency %CPU
TARGET PASS0001    0 1 478019584 3647 10.003 47.790 364.61 0.0027 33.90
TARGET PASS0002    0 1 480641024 3667 10.002 48.053 366.62 0.0027 35.86
TARGET PASS0003    0 1 480116736 3663 10.001 48.007 366.26 0.0027 39.00
TARGET PASS0004    0 1 478412800 3650 10.002 47.834 364.94 0.0027 38.60
TARGET PASS0005    0 1 472252416 3603 10.002 47.215 360.22 0.0028 37.50
TARGET Average    0 1 238944256 18230 50.010 47.780 364.53 0.0027 36.97
      Combined    1 1 238944256 18230 50.010 47.780 364.53 0.0027 36.97
Ending time for this run, Wed Nov 9 08:44:09 2005
```

Another interesting option allowed with Xdd is to write on file systems. As shown in Example 6-7 and Example 6-8, we can test our jfs or jfs2 file system to decide which one to use with our application. Obviously, it is necessary to know how the application writes: sequential or random, blocksize, and so on.

Example 6-7 Xdd write on jfs file system

```
./xdd.aix -op write -targets 1 /fsds3J1/file3 -blocksize 512 -reqsize 128 -mbytes 64 \
-verbose -passes 3
```

```
Seconds before starting, 0
T Q Bytes Ops Time Rate IOPS Latency %CPU
TARGET PASS0001 0 1 67108864 1024 0.188 356.918 5446.14 0.0002 100.00
TARGET PASS0002 0 1 67108864 1024 0.582 115.391 1760.72 0.0006 27.59
TARGET PASS0003 0 1 67108864 1024 0.557 120.576 1839.85 0.0005 33.93
TARGET Average 0 1 20132659 3072 1.326 151.810 2316.44 0.0004 40.60
Combined 1 1 20132659 3072 1.326 151.810 2316.44 0.0004 40.60
Ending time for this run, Wed Nov 9 09:27:12 2005
```

Example 6-8 Xdd write on jfs2 file system

```
./xdd.aix -op write -targets 1 /fsds1J2/file1 -blocksize 512 -reqsize 128 -mbytes\ 64
-verbose -passes 3
```

```
Seconds before starting, 0
T Q Bytes Ops Time Rate IOPS Latency %CPU
TARGET PASS0001 0 1 67108864 1024 0.463 144.970 2212.07 0.0005 82.61
TARGET PASS0002 0 1 67108864 1024 0.203 330.603 5044.61 0.0002 100.00
TARGET PASS0003 0 1 67108864 1024 0.314 213.443 3256.88 0.0003 78.12
TARGET Average 0 1 201326592 3072 0.980 205.369 3133.69 0.0003 84.69
Combined 1 1 201326592 3072 0.980 205.369 3133.69 0.0003 84.69
Ending time for this run, Wed Nov 9 09:31:46 2005
```

6.4 Storage Manager Performance Monitor

The Storage Performance Monitor is a tool built into the DS4000 Storage Manager client. It monitors performance on each logical drive, and collects information such as:

- ▶ Total I/Os
- ▶ Read percentage
- ▶ Cache hit percentage
- ▶ Current KB/sec and maximum KB/sec
- ▶ Current I/O per sec and maximum I/O per sec

This section describes how to use data from the Performance Monitor and what tuning options are available in the Storage Manager for optimizing the storage server performance.

6.4.1 Starting the Performance Monitor

You launch the Performance Monitor from the SMclient Subsystem Management window by either:

- ▶ Selecting the **Monitor Performance** icon
- or
- ▶ Selecting the **Storage Subsystem** → **Monitor Performance** pull-down menu option

- Selecting the storage subsystem node in the Logical View or Mappings View, then choosing **Monitor Performance** from the right-mouse pop-up menu

The Performance Monitor window opens up with all logical drives displayed as shown in Figure 6-10.

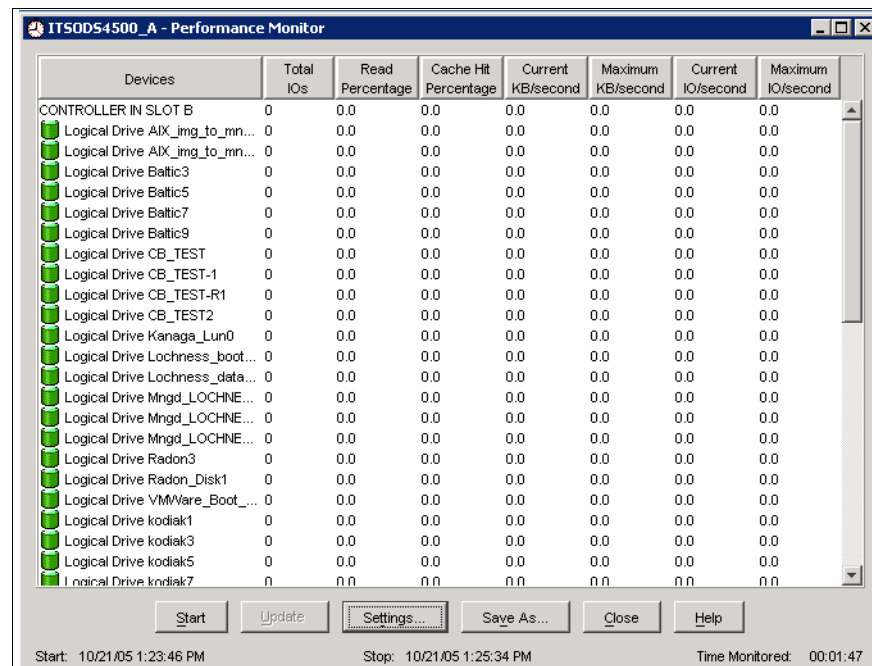


Figure 6-10 Performance Monitor

Table 6-1 describes the information collected by the Performance Monitor.

Table 6-1 Information collected by Performance Monitor

Data field	Description
Total I/Os	Total I/Os performed by this device since the beginning of the polling session.
Read percentage	The percentage of total I/Os that are read operations for this device. Write percentage can be calculated as 100 minus this value.
Cache hit percentage	The percentage of reads that are processed with data from the cache rather than requiring a read from disk.
Current KBps	Average <i>transfer rate</i> during the polling session. The transfer rate is the amount of data in Kilobytes that can be moved through the I/O Data connection in a second (also called <i>throughput</i>).
Maximum KBps	The maximum transfer rate that was achieved during the Performance Monitor polling session.
Current I/O per second	The average number of I/O requests serviced per second during the current polling interval (also called an <i>I/O request rate</i>).
Maximum I/O per second	The maximum number of I/O requests serviced during a one-second interval over the entire polling session.

The fields are:

- Total I/Os

This data is useful for monitoring the I/O activity of a specific controller and a specific logical drive, which can help identify possible high-traffic I/O areas.

If I/O rate is slow on a logical drive, try increasing the array size.

You might notice a disparity in the Total I/Os (workload) of controllers, for example, the workload of one controller is heavy or is increasing over time, while that of the other controller is lighter or more stable. In this case, consider changing the controller ownership of one or more logical drives to the controller with the lighter workload. Use the logical drive Total I/O statistics to determine which logical drives to move.

If you notice that the workload across the storage subsystem (Storage Subsystem Totals Total I/O statistic) continues to increase over time, while application performance decreases, this might indicate the need to add additional storage subsystems to your installation so that you can continue to meet application needs at an acceptable performance level.

- Read percentage

Use the read percentage for a logical drive to determine actual application behavior. If there is a low percentage of read activity relative to write activity, consider changing the RAID level of an array from RAID-5 to RAID-1 for faster performance.

- Cache hit percentage

A higher percentage is desirable for optimal application performance. There is a positive correlation between the cache hit percentage and I/O rates.

The cache hit percentage of all of the logical drives might be low or trending downward. This might indicate inherent randomness in access patterns, or at the storage subsystem or controller level, this can indicate the need to install more controller cache memory if you do not have the maximum amount of memory installed.

If an individual logical drive is experiencing a low cache hit percentage, consider enabling cache read ahead for that logical drive. Cache read ahead can increase the cache hit percentage for a sequential I/O workload.

Determining the effectiveness of a logical drive cache read-ahead multiplier

To determine if your I/O has sequential characteristics, try enabling a conservative cache read-ahead multiplier (four, for example). Then, examine the logical drive cache hit percentage to see if it has improved. If it has, indicating that your I/O has a sequential pattern, enable a more aggressive cache read-ahead multiplier (eight, for example). Continue to customize logical drive cache read-ahead to arrive at the optimal multiplier (in the case of a random I/O pattern, the optimal multiplier is zero).

- Current KB/sec and maximum KB/sec

The “Current KB/sec” value is the average size of the amount of data that was transfer over one second during a particular *interval period* that was monitored. The “Maximum KB/sec” value is the highest amount that was transferred over any one second period, during all of the interval periods in the *number of iterations* that were ran for a specific command. This value can show you when peak transfer rate period was detected during the command runtime.

Tip: If Maximum KB/sec is the same as last interval’s Current KB/sec, we recommend extending the number of iterations to see when the peak rate is actually reached, as this may be on the rise.

The transfer rates of the controller are determined by the application I/O size and the I/O rate. Generally, small application I/O requests result in a lower transfer rate, but provide a faster I/O rate and shorter response time. With larger application I/O requests, higher throughput rates are possible. Understanding your typical application I/O patterns can help you determine the maximum I/O transfer rates for a given storage subsystem.

Consider a storage subsystem, equipped with Fibre Channel controllers, that supports a maximum transfer rate of 100 Mbps (100,000 KB per second). Your storage subsystem typically achieves an average transfer rate of 20,000 KB/sec. (The typical I/O size for your applications is 4 KB, with 5,000 I/Os transferred per second for an average rate of 20,000 KB/sec.) In this case, I/O size is small. Because there is system overhead associated with each I/O, the transfer rates will not approach 100,000 KB/sec. However, if your typical I/O size is large, a transfer rate within a range of 80,000 to 90,000 KB/sec might be achieved.

- Current I/O per second and maximum I/O per second

The *Current IO/sec* value is the average number of I/Os serviced in one second during a particular *interval period* that was monitored. The *Maximum IO/sec* value is the highest number of I/Os serviced in any one second period, during all of the interval periods in the *number of iterations* that were ran for a specific command. This value can show you when the peak I/O period was detected during the command runtime.

Tip: If Maximum I/Ops is the same as the last interval's Current I/Ops, we recommend extending the number of iterations to see when the peak is actually reached, as this may be on the rise.

Factors that affect I/Os per second include access pattern (random or sequential), I/O size, RAID level, segment size, and number of drives in the arrays or storage subsystem. The higher the cache hit rate, the higher the I/O rates.

Performance improvements caused by changing the segment size can be seen in the I/Os per second statistics for a logical drive. Experiment to determine the optimal segment size, or use the file system or database block size.

Higher write I/O rates are experienced with write caching enabled compared to disabled. In deciding whether to enable write caching for an individual logical drive, consider the current and maximum I/Os per second. You should expect to see higher rates for sequential I/O patterns than for random I/O patterns. Regardless of your I/O pattern, we recommend that write caching be enabled to maximize I/O rate and shorten application response time.

6.4.2 Using the Performance Monitor

The Performance Monitor queries the Storage subsystem at regular intervals. To change the polling interval and to select only the logical drives and the controllers you wish to monitor, click the **Settings** button.

To change the polling interval, choose a number of seconds in the spin box. Each time the polling interval elapses, the Performance Monitor re-queries the storage subsystem and updates the statistics in the table. If you are monitoring the storage subsystem in real time, update the statistics frequently by selecting a short polling interval, for example, five seconds. If you are saving results to a file to look at later, choose a slightly longer interval, for example, 30 to 60 seconds, to decrease the system overhead and the performance impact.

The Performance Monitor will not dynamically update its display if any configuration changes occur while the monitor window is open (for example, creation of new logical drives, change

in logical drive ownership, and so on). The Performance Monitor window must be closed and then reopened for the changes to appear.

Note: Using the Performance Monitor to retrieve performance data can affect the normal storage subsystem performance, depending on how many items you want to monitor and the refresh interval.

If the storage subsystem you are monitoring begins in or transitions to an unresponsive state, an informational dialog box opens, stating that the Performance Monitor cannot poll the storage subsystem for performance data.

The Performance Monitor is a real time tool; it is not possible to collect performance data over time with the Storage Manager GUI. However, you can use a simple script to collect performance data over some period of time and analyze it later.

The script can be run from the command line with SMcli or from the Storage Manager Script Editor GUI. From the Storage Manager Enterprise management window, select **Tools** → **Execute Script**. Data collected while executing the script is saved in the file and directory specified in the storage Subsystem file parameter.

Using the script editor GUI

A sample of the script editor GUI is shown in Figure 6-11.

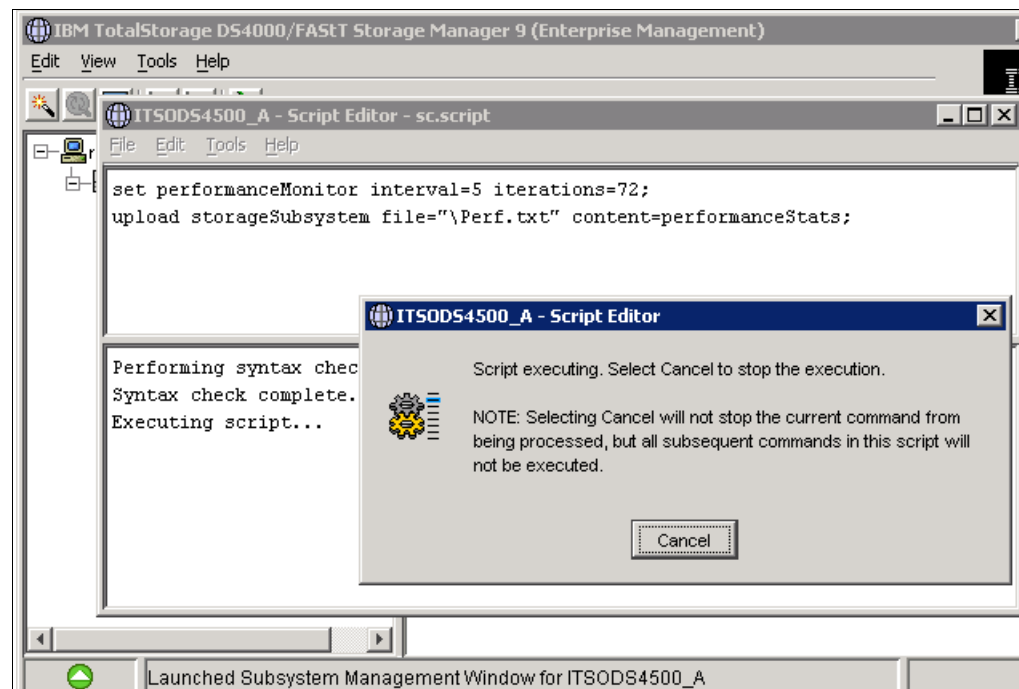


Figure 6-11 Script to collect Performance Monitor data overtime

A sample of the output file is shown in Figure 6-12.

```
Performance Monitor Statistics for Storage Subsystem: ITSODS4500_A
Date/Time: 10/26/05 5:04:08 PM
Polling interval in seconds: 5

Storage Subsystems,Total,Read,Cache Hit,Current,Maximum,Current,Maximum
,Ios,Percentage,Percentage,KB/second,KB/second,IO/second,IO/second
Capture Iteration: 1
Date/Time: 10/26/05 5:04:08 PM
CONTROLLER IN SLOT B,416.0,100.0,100.0,85196.8,85196.8,83.2,83.2,
Logical Drive Kanaga_Lun0,416.0,100.0,100.0,85196.8,85196.8,83.2,83.2,
CONTROLLER IN SLOT A,416.0,0.0,0.0,85196.8,85196.8,83.2,83.2,
Logical Drive Kanaga_Lun1,416.0,0.0,0.0,85196.8,85196.8,83.2,83.2,
STORAGE SUBSYSTEM TOTALS,832.0,50.0,100.0,170393.6,170393.6,166.4,166.4,

Capture Iteration: 2
Date/Time: 10/26/05 5:04:13 PM
CONTROLLER IN SLOT B,848.0,100.0,100.0,88473.6,88473.6,86.4,86.4,
Logical Drive Kanaga_Lun0,848.0,100.0,100.0,88473.6,88473.6,86.4,86.4,
CONTROLLER IN SLOT A,848.0,0.0,0.0,88473.6,88473.6,86.4,86.4,
Logical Drive Kanaga_Lun1,848.0,0.0,0.0,88473.6,88473.6,86.4,86.4,
STORAGE SUBSYSTEM TOTALS,1696.0,50.0,100.0,176947.2,176947.2,172.8,172.8,

Capture Iteration: 3
Date/Time: 10/26/05 5:04:18 PM
CONTROLLER IN SLOT B,1272.0,100.0,100.0,86835.2,88473.6,84.8,86.4,
Logical Drive Kanaga_Lun0,1272.0,100.0,100.0,86835.2,88473.6,84.8,86.4,
CONTROLLER IN SLOT A,1272.0,0.0,0.0,86835.2,88473.6,84.8,86.4,
Logical Drive Kanaga_Lun1,1272.0,0.0,0.0,86835.2,88473.6,84.8,86.4,
STORAGE SUBSYSTEM TOTALS,2544.0,50.0,100.0,173670.4,176947.2,169.6,172.8,
```

Figure 6-12 Script output file

Using the Command Line Interface CLI

You can also use the SMcli to gather Performance Monitor data over a period of time.

Example 6-9 shows a test_script file for execution under AIX, using the ksh test_script:

```
# cat test_script
```

Example 6-9 Test script

```
#!/bin/ksh
#The information is captured from a single Linux/Aix server by running the
# following "Storage Manager Command Line Interface Utility" Linux/Aix command
CMD='set session performanceMonitorInterval=60 performanceMonitorIterations=2; \
show allLogicalDrives performanceStats;'
/usr/SMclient/SMcli -e -S 9.1.39.26 9.1.39.27 -c "$CMD"
#(Note; this will get you a run every minute for 2 time; if run every 10 minutes for 10
# times set the
# "performanceMonitorinterval=600"
# "performanceMonitorIterations=10"
```

The first executable line sets the CMD variable; the second executable line invokes the SMcli command. Note that for the -S parameter, it is necessary to specify the IP address of both DS4000 controllers (A and B).

The output resulting from the script execution can be redirected to a file by typing the command:

```
ksh test_script > test_output
```

This test_script collects information for all logical drives, but it is possible to select specific logical drives. Example 6-10 shows how to select only two drives, Kanaga_lun1 and Kanaga_lun2.

Example 6-10 Test script for two logical drives

```
#!/bin/ksh
CMD='set session performanceMonitorInterval=60 performanceMonitorIterations=2;\
show logicaldrives [Kanaga_lun1 Kanaga_lun2] performanceStats;'
/usr/SMclient/SMcli -e -S 9.1.39.26 9.1.39.27 -c "$CMD"
```

We created a file called test_output_twoluns:

```
ksh test_output_twoluns > output_twoluns
```

The output file (called output_twoluns) is shown in Example 6-11.

Example 6-11 Output file: output_twoluns

```
Performance Monitor Statistics for Storage Subsystem: ITS0DS4500_A
Date/Time: 11/10/05 2:06:21 PM
Polling interval in seconds: 60
Storage Subsystems,Total,Read,Cache Hit,Current,Maximum,Current,Maximum
,I/Os,Percentage,Percentage,KB/second,KB/second,I/O/second,I/O/second
Capture Iteration: 1
Date/Time: 11/10/05 2:06:22 PM
CONTROLLER IN SLOT A,689675.0,100.0,100.0,45224.6,45224.6,11306.1,11306.1,
Logical Drive Kanaga_lun1,689675.0,100.0,100.0,45224.6,45224.6,11306.1,11306.1,
CONTROLLER IN SLOT B,518145.0,100.0,100.0,33976.7,33976.7,8494.2,8494.2,
Logical Drive Kanaga_lun2,518145.0,100.0,100.0,33976.7,33976.7,8494.2,8494.2,
STORAGE SUBSYSTEM TOTALS,1207820.0,100.0,100.0,79201.3,79201.3,19800.3,19800.3,
Capture Iteration: 2
Date/Time: 11/10/05 2:07:23 PM
CONTROLLER IN SLOT A,1393595.0,100.0,100.0,46158.7,46158.7,11539.7,11539.7,
Logical Drive Kanaga_lun1,1393595.0,100.0,100.0,46158.7,46158.7,11539.7,11539.7,
CONTROLLER IN SLOT B,518145.0,100.0,100.0,0.0,33976.7,0.0,8494.2,
Logical Drive Kanaga_lun2,518145.0,100.0,100.0,0.0,33976.7,0.0,8494.2,
STORAGE SUBSYSTEM TOTALS,1911740.0,100.0,100.0,46158.7,79201.3,11539.7,19800.3,
```

All the data saved in the file is comma delimited so that the file can be easily imported into a spreadsheet for easier analysis and review (Figure 6-13).

The screenshot shows a Microsoft Excel window with the file name 'out.mo2'. The spreadsheet has a single sheet named 'Storage Subsystems'. The data is organized into two main sections, one for 'Capture Iteration: 1' and another for 'Capture Iteration: 2'. Each section includes a header row for 'Storage Subsystems' with columns for 'Total IOs', 'Read Percentage', 'Cache Hit Percentage', 'Current KB/second', 'Maximum KB/second', 'Current IO/second', and 'Maximum IO/second'. The data rows list performance metrics for 'CONTROLLER IN SLOT A', 'Logical Drive Kanaga_lun1', 'CONTROLLER IN SLOT B', and 'Logical Drive Kanaga_lun2', followed by a 'STORAGE SUBSYSTEM TOTALS' row. The status bar at the bottom indicates 'Ready' and 'NUM'.

	A	B	C	D	E	F	G	H
1	Storage Subsystems	Total	Read	Cache Hit	Current	Maximum	Current	Maximum
2		IOs	Percentage	Percentage	KB/second	KB/second	IO/second	IO/second
3	Capture Iteration: 1							
4	Date/Time: 11/10/05 2:06:22 PM							
5	CONTROLLER IN SLOT A	689675	100	100	45224.6	45224.6	11306.1	11306.1
6	Logical Drive Kanaga_lun1	689675	100	100	45224.6	45224.6	11306.1	11306.1
7	CONTROLLER IN SLOT B	518145	100	100	33976.7	33976.7	8494.2	8494.2
8	Logical Drive Kanaga_lun2	518145	100	100	33976.7	33976.7	8494.2	8494.2
9	STORAGE SUBSYSTEM TOTALS	1207820	100	100	79201.3	79201.3	19800.3	19800.3
10								
11	Capture Iteration: 2							
12	Date/Time: 11/10/05 2:07:23 PM							
13	CONTROLLER IN SLOT A	1393595	100	100	46158.7	46158.7	11539.7	11539.7
14	Logical Drive Kanaga_lun1	1393595	100	100	46158.7	46158.7	11539.7	11539.7
15	CONTROLLER IN SLOT B	518145	100	100	0	33976.7	0	8494.2
16	Logical Drive Kanaga_lun2	518145	100	100	0	33976.7	0	8494.2
17	STORAGE SUBSYSTEM TOTALS	1911740	100	100	46158.7	79201.3	11539.7	19800.3
18								

Figure 6-13 Importing results into a spreadsheet

The following is a description of all the values saved.

6.4.3 Using the Performance Monitor: Illustration

To illustrate the use of the Performance Monitor, we suppose that we want to find the optimal value for the *max_xref_size* parameter of an HBA.

We ran several tests with different values of the parameter. The test environment consists of the following components: DS4500, AIX5.2, Brocade switch.

Two LUNs, Kanaga_lun0 and Kanaga_lun1 of 20Gb each, are defined on DS45000 and reside in separate arrays. Kanaga_lun0 is in array 10, defined as RAID 0 and consists of three physical drives; Kanaga_lun1 is in Array 11 (RAID 0) with four physical drives. Kanaga_lun0 is on the preferred controller B, and kanaga_lun1 is on controller A (Figure 6-14).

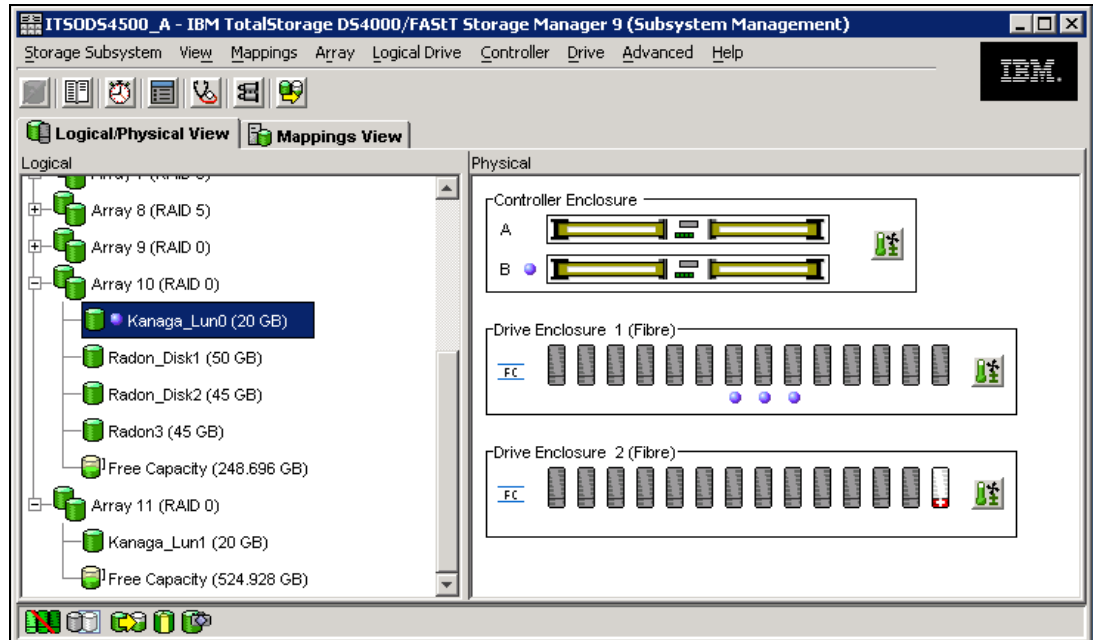
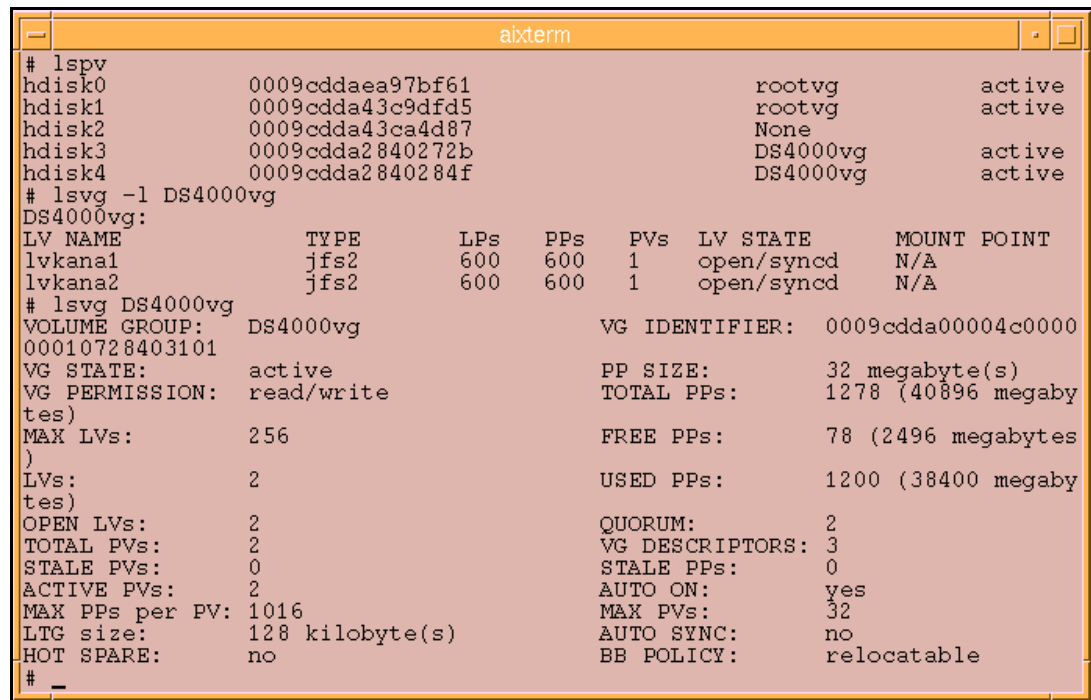


Figure 6-14 Logical/Physical view of test LUNs

We assign to LUNs to AIX, create one volume group, DS4000vg, with 32 PPSize and create two different logical volumes: lvkana1 with 600 Physical Partitions on hdisk 3 “lun Kanaga_lun1” and lvkana2 with 600 Physical Partitions on hdisk4 “lun Kanaga_lun0” (Figure 6-15).



```
# lspv
hdisk0          0009cddaea97bf61          rootvg          active
hdisk1          0009cdda43c9dfd5          rootvg          active
hdisk2          0009cdda43ca4d87          None
hdisk3          0009cdda2840272b          DS4000vg        active
hdisk4          0009cdda2840284f          DS4000vg        active
# lsvg -l DS4000vg
DS4000vg:
LV NAME          TYPE          LPs          PPs          PVs          LV STATE        MOUNT POINT
lvkana1          jfs2          600          600          1            open/syncd      N/A
lvkana2          jfs2          600          600          1            open/syncd      N/A
# lsvg DS4000vg
VOLUME GROUP:    DS4000vg                VG IDENTIFIER:   0009cdda00004c0000
00010728403101
VG STATE:         active                    PP SIZE:         32 megabyte(s)
VG PERMISSION:    read/write              TOTAL PPs:       1278 (40896 megaby
tes)
MAX LVs:          256                      FREE PPs:        78 (2496 megabytes
)
LVs:              2                      USED PPs:        1200 (38400 megaby
tes)
OPEN LVs:         2                      QUORUM:          2
TOTAL PVs:        2                      VG DESCRIPTORS:  3
STALE PVs:        0                      STALE PPs:       0
ACTIVE PVs:       2                      AUTO ON:         yes
MAX PPs per PV:   1016                   MAX PVs:         32
LTG size:         128 kilobyte(s)         AUTO SYNC:       no
HOT SPARE:        no                     BB POLICY:       relocatable
#
```

Figure 6-15 AIX volumes and volume group

With the **dd** command, we simulate sequential read/write from the source hdisk3 to the target hdisk4:

```
dd if=/dev/lvkana1 of=/dev/lvkana2 bs=8192k
```

Attention: This test deletes all data on lvkana2. If you are reproducing this test, be sure you can delete data on the target logical volume output file.

Before running the **dd** command, check the default value for **max_xfer_size** (the default value is 0x100000). Use the following commands:

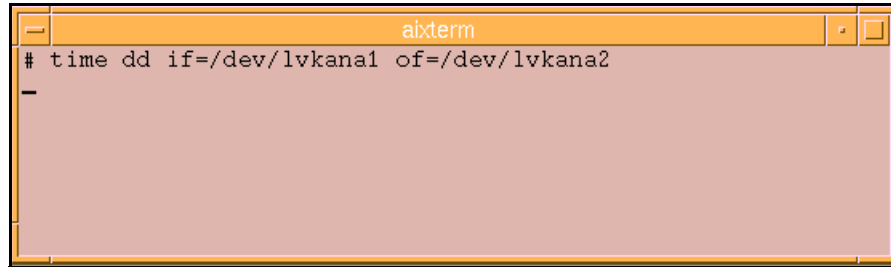
```
lsattr -El fcs0 | grep max_xref_size
```

```
max_xfer_size is equal to 0x100000
```

```
lsattr -El fcs1 | grep max_xref_size
```

```
max_xfer_size is equal to 0x100000
```

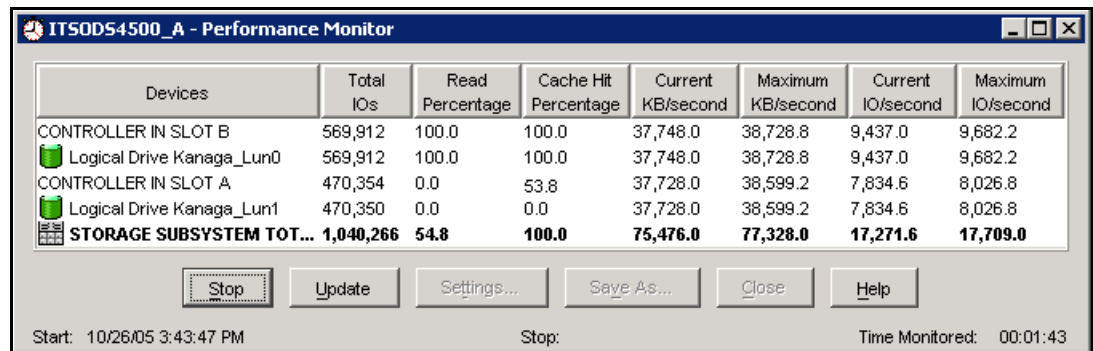
Now, you can start the **dd** command (Figure 6-16).



```
# time dd if=/dev/lvkana1 of=/dev/lvkana2
-
```

Figure 6-16 Running the **dd** command

During the execution, use the SM Performance Monitor to observe the LUNs on the DS4000 storage server (Figure 6-17).

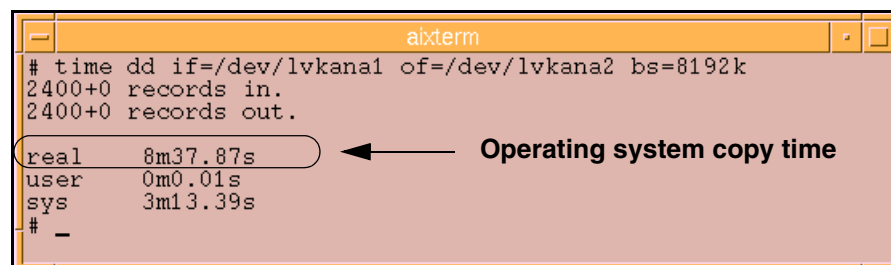


Devices	Total IOs	Read Percentage	Cache Hit Percentage	Current KB/second	Maximum KB/second	Current IO/second	Maximum IO/second
CONTROLLER IN SLOT B	569,912	100.0	100.0	37,748.0	38,728.8	9,437.0	9,682.2
Logical Drive Kanaga_Lun0	569,912	100.0	100.0	37,748.0	38,728.8	9,437.0	9,682.2
CONTROLLER IN SLOT A	470,354	0.0	53.8	37,728.0	38,599.2	7,834.6	8,026.8
Logical Drive Kanaga_Lun1	470,350	0.0	0.0	37,728.0	38,599.2	7,834.6	8,026.8
STORAGE SUBSYSTEM TOT...	1,040,266	54.8	100.0	75,476.0	77,328.0	17,271.6	17,709.0

Start: 10/26/05 3:43:47 PM Stop: Time Monitored: 00:01:43

Figure 6-17 Observe with Performance Monitor

Upon termination, the **dd** command indicates how long it took to make the copy (Figure 6-18).



```
# time dd if=/dev/lvkana1 of=/dev/lvkana2 bs=8192k
2400+0 records in.
2400+0 records out.
real    8m37.87s
user    0m0.01s
sys     3m13.39s
# -
```

← Operating system copy time

Figure 6-18 **dd** output

Now, using a simple script called **chxfer** and shown in Example 6-12, we change the value of **max_xfer_size** for the two HBAs (Example 6-12).

Example 6-12 Script chxfer

```

if [[ $1 -eq 100 || $1 -eq 200 || $1 -eq 400 || $1 -eq 1000 ]]
then
    varyoffvg DS4000vg
    rmdev -dl dar0 -R
    rmdev -dl fscsi0 -R
    rmdev -dl fscsi1 -R
    chdev -l fcs0 -a max_xfer_size=0x"$1"000
    chdev -l fcs1 -a max_xfer_size=0x"$1"000
    cfgmgr
    varyonvg DS4000vg
else
    echo "paramitter value not valid. use 100 200 400 1000"
fi

```

The script does a **varyoffvg** for the DS4000vg volume group, removes all drivers under **fcs0** and **fcs1**, changes the **max_xfer_size** value on both **fcs** adapters, and reconfigures all devices. These steps are necessary to set a new **max_xfer_size** value on the adapters.

Run the script by entering (at the AX command line):

```
# ksh chxfer 200
```

Restart the **dd** test and look at the results using the Performance Monitor (Figure 6-19).

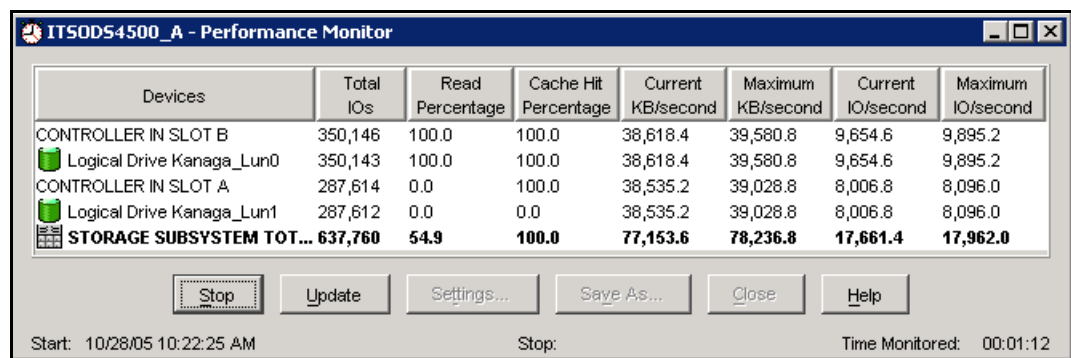


Figure 6-19 New observation in Performance Monitor

Again, note the time upon completion of the **dd** command (Figure 6-20).

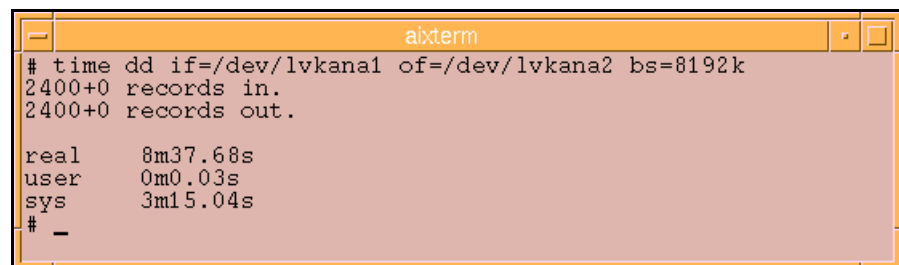


Figure 6-20 dd command with max_xfer_size set to 0x200000

Repeat the steps for other values of the `max_xfer_size` parameter and collect the results as shown in Table 6-2. Now we are able to analyze the copy time.

Table 6-2 Results for dd command

max_xfer_size	time dd command	SM perform max I/O	SM current kbps
0x100000	8min 37.87 sec	8,026	38,599
0x200000	8min 37.32 sec	17,962	78,236
0x400000	8min 36.60 sec	17,954	83,761
0x1000000	8min 36.76 sec	17,772	77,363

In our example, the best value for `max_xfer_size` is 0x200000. Although this does not get the lowest time for executing the `dd` command, it performs the I/O operations more efficiently.

Now we can run another test for a big sequential I/O. In the previous test we used the `/dev/rlvkana1` and `/dev/rlvkana2` “block access device” in which case AIX generates sequential block of 4Kb for read and write. In this new test we use `/dev/rlvkana1` and `/dev/rlvkana2` “raw access device” for which AIX generates sequential blocks of 8192Kb.

Results of the two tests for the same value of `max_xfer_size` are shown in Table 6-3.

Table 6-3 Test results - max_xfer_size

max_xfer_size	time dd command	SM perform max I/O	Sm current kbps
0x200000 Block 4K	8min 37.68 sec	3,600 - 3,800	38,000
0x200000 Row 8192K	3min 41.24 sec	88,000 - 90,000	901,000

You can see a significant time reduction for the copy operation, using the same value recommended before of 0x200000, but changing the block size from 4K to 8 MB. The 8 MB block allows for much greater throughput than the 4K block.

With the SM Performance Monitor you can view which controller works, and can then decide to move some volumes from one controller to the other to balance the performance among controllers.

You can also see if some LUNs work more than others (on the same host), and can then decide if possible to move data from on one logical drive to another to balance the throughput among the drives. Note, however, that the Performance Monitor cannot show the throughput for a physical drive.

6.5 AIX utilities

This section reviews several command and utilities that are part of the AIX operating system and can be used to analyze storage performance. For details, please refer to the *AIX Performance Management Guide*, SC23-4876.

6.5.1 Introduction to monitoring Disk I/O

When you are monitoring disk I/O, use the following tips to determine your course of action:

- Find the most active files, file systems, and logical volumes:

Should “hot” file systems be better located on the physical drive or be spread across multiple physical drives? (**lslv**, **iostat**, **filemon**)

Are “hot” files local or remote? (**filemon**)

Does paging space dominate disk utilization? (**vmstat**, **filemon**)

Is there enough memory to cache the file pages being used by running processes? (**vmstat**, **svmon**, **vm tune**, **topas**)

Does the application perform a lot of synchronous (non-cached) file I/O?

- Determine file fragmentation:

Are “hot” files heavily fragmented? (**fileplace**)

- Find the physical volume with the highest utilization:

Is the type of drive or I/O adapter causing a bottleneck? (**iostat**, **filemon**)

Building a pre-tuning baseline

Before you make significant changes in your disk configuration or tuning parameters, it is a good idea to build a baseline of measurements that record the current configuration and performance.

Wait I/O time reporting

AIX 5.1 and later contain enhancements to the method used to compute the percentage of CPU time spent waiting on disk I/O (wio time). The method used in an AIX operating system can, under certain circumstances, give an inflated view of wio time on SMPs. The wio time is reported by the commands **sar** (%wio), **vmstat** (wa), and **iostat** (%iowait).

6.5.2 Assessing disk performance with the iostat command

Begin the assessment by running the **iostat** command with an interval parameter during your system's peak workload period or while running a critical application for which you need to minimize I/O delays.

Drive report

When you suspect a disk I/O performance problem, use the **iostat** command. To avoid the information about the TTY and CPU statistics, use the **-d** option. In addition, the disk statistics can be limited to the important disks by specifying the disk names.

Note: Remember that the first set of data represents all activity since system startup.

The following information is reported:

- disks

Shows the names of the physical volumes. These are either **hdisk** or **cd** followed by a number. If physical volume names are specified with the **iostat** command, only those names specified are displayed.

- % tm_act

Indicates the percentage of time that the physical disk was active (bandwidth utilization for the drive) or, in other words, the total time disk requests are outstanding. A drive is active

during data transfer and command processing, such as seeking to a new location. The “disk active time” percentage is directly proportional to resource contention and inversely proportional to performance. As disk use increases, performance decreases and response time increases. In general, when the utilization exceeds 70%, processes are waiting longer than necessary for I/O to complete because most UNIX processes block (or sleep) while waiting for their I/O requests to complete. Look for busy versus idle drives. Moving data from busy to idle drives can help alleviate a disk bottleneck. Paging to and from disk will contribute to the I/O load.

- Kbps

Indicates the amount of data transferred (read or written) to the drive in KB per second. This is the sum of Kb_read plus Kb_wrtn, divided by the seconds in the reporting interval.

- tps

Indicates the number of transfers per second that were issued to the physical disk. A transfer is an I/O request through the device driver level to the physical disk. Multiple logical requests can be combined into a single I/O request to the disk. A transfer is of indeterminate size.

- Kb_read

Reports the total data (in KB) read from the physical volume during the measured interval.

- Kb_wrtn

Shows the amount of data (in KB) written to the physical volume during the measured interval.

Taken alone, there is no unacceptable value for any of the above fields because statistics are too closely related to application characteristics, system configuration, and type of physical disk drives and adapters. Therefore, when you are evaluating data, look for patterns and relationships. The most common relationship is between disk utilization (%tm_act) and data transfer rate (tps).

To draw any valid conclusions from this data, you have to understand the application's disk data access patterns such as sequential, random, or combination, as well as the type of physical disk drives and adapters on the system. For example, if an application reads/writes sequentially, you should expect a high disk transfer rate (Kbps) when you have a high disk busy rate (%tm_act). Columns Kb_read and Kb_wrtn can confirm an understanding of an application's read/write behavior. However, these columns provide no information about the data access patterns.

Generally you do not need to be concerned about a high disk busy rate (%tm_act) as long as the disk transfer rate (Kbps) is also high. However, if you get a high disk busy rate and a low disk transfer rate, you may have a fragmented logical volume, file system, or individual file.

Discussions of disk, logical volume and file system performance sometimes lead to the conclusion that the more drives you have on your system, the better the disk I/O performance. This is not always true because there is a limit to the amount of data that can be handled by a disk adapter. The disk adapter can also become a bottleneck. If all your disk drives are on one disk adapter, and your hot file systems are on separate physical volumes, you might benefit from using multiple disk adapters. Performance improvement will depend on the type of access.

To see if a particular adapter is saturated, use the **iostat** command and add up all the Kbps amounts for the disks attached to a particular disk adapter. For maximum aggregate performance, the total of the transfer rates (Kbps) must be below the disk adapter throughput rating. In most cases, use 70% of the throughput rate, the -a or -A option will display this information.

6.5.3 Assessing disk performance with the vmstat command

The **vmstat** command can give the following statistics:

- The **wa** column

The **wa** column details the percentage of time the CPU was idle with pending disk I/O.

- The **disk xfer** part

To display a statistic about the logical disks (a maximum of four disks is allowed), use the following command (Example 6-13).

Example 6-13 vmstat command

# vmstat hdisk3 hdisk4 1 8																						
kthr		memory		page						faults				cpu				disk xfer				
r	b	avm	fre	re	pi	po	fr	sr	cy	in	sy	cs	us	sy	id	wa	1	2	3	4		
0	0	3456	27743	0	0	0	0	0	0	131	149	28	0	1	99	0	0	0	0			
0	0	3456	27743	0	0	0	0	0	0	131	77	30	0	1	99	0	0	0	0			
1	0	3498	27152	0	0	0	0	0	0	153	1088	35	1	10	87	2	0	11	11			
0	1	3499	26543	0	0	0	0	0	0	199	1530	38	1	19	0	80	0	59	59			
0	1	3499	25406	0	0	0	0	0	0	187	2472	38	2	26	0	72	0	53	53			
0	0	3456	24329	0	0	0	0	0	0	178	1301	37	2	12	20	66	0	42	42			
0	0	3456	24329	0	0	0	0	0	0	124	58	19	0	0	99	0	0	0	0			
0	0	3456	24329	0	0	0	0	0	0	123	58	23	0	0	99	0	0	0	0			

The **disk xfer** part provides the number of transfers per second to the specified physical volumes that occurred in the sample interval. One to four physical volume names can be specified. Transfer statistics are given for each specified drive in the order specified. This count represents requests to the physical device. It does not imply an amount of data that was read or written. Several logical requests can be combined into one physical request.

- The **in** column

This column shows the number of hardware or device interrupts (per second) observed over the measurement interval. Examples of interrupts are disk request completions and the 10 millisecond clock interrupt. Since the latter occurs 100 times per second, the **in** field is always greater than 100.

- The **vmstat -i** output

The **-i** parameter displays the number of interrupts taken by each device since system startup. But, by adding the interval and, optionally, the count parameter, the statistic since startup is only displayed in the first stanza; every trailing stanza is a statistic about the scanned interval (Example 6-14).

Example 6-14 vmstat -i command

```
# vmstat -i 1 2
```

priority	level	type	count	module(handler)
0	0	hardware	0	i_misc_pwr(a868c)
0	1	hardware	0	i_scu(a8680)
0	2	hardware	0	i_epow(954e0)
0	2	hardware	0	/etc/drivers/ascsiddpin(189acd4)
1	2	hardware	194	/etc/drivers/rsdd(1941354)
3	10	hardware	10589024	/etc/drivers/mpsdd(1977a88)
3	14	hardware	101947	/etc/drivers/ascsiddpin(189ab8c)
5	62	hardware	61336129	clock(952c4)
10	63	hardware	13769	i_softoff(9527c)
priority	level	type	count	module(handler)

0	0	hardware	0	i_misc_pwr(a868c)
0	1	hardware	0	i_scu(a8680)
0	2	hardware	0	i_epow(954e0)
0	2	hardware	0	/etc/drivers/ascsiddpin(189acd4)
1	2	hardware	0	/etc/drivers/rsdd(1941354)
3	10	hardware	25	/etc/drivers/mpsdd(1977a88)
3	14	hardware	0	/etc/drivers/ascsiddpin(189ab8c)
5	62	hardware	105	clock(952c4)
10	63	hardware	0	i_softoff(9527c)

Note: The output will differ from system to system, depending on hardware and software configurations (for example, the clock interrupts may not be displayed in the **vmstat -i** output although they will be accounted for under the in column in the normal **vmstat** output). Check for high numbers in the count column and investigate why this module has to execute so many interrupts.

6.5.4 Assessing disk performance with the sar command

The **sar** command is a standard UNIX command used to gather statistical data about the system. With its numerous options, the **sar** command provides queuing, paging, TTY, and many other statistics. With AIX, the **sar -d** option generates real-time disk I/O statistics (Example 6-15).

Example 6-15 sar command

```
# sar -d 3 3
```

AIX konark 3 4 0002506F4C00	08/26/99						
12:09:50	device	%busy	avque	r+w/s	blks/s	avwait	avserv
12:09:53	hdisk0	1	0.0	0	5	0.0	0.0
	hdisk1	0	0.0	0	1	0.0	0.0
	cd0	0	0.0	0	0	0.0	0.0
12:09:56	hdisk0	0	0.0	0	0	0.0	0.0
	hdisk1	0	0.0	0	1	0.0	0.0
	cd0	0	0.0	0	0	0.0	0.0
12:09:59	hdisk0	1	0.0	1	4	0.0	0.0
	hdisk1	0	0.0	0	1	0.0	0.0
	cd0	0	0.0	0	0	0.0	0.0
Average	hdisk0	0	0.0	0	3	0.0	0.0
	hdisk1	0	0.0	0	1	0.0	0.0
	cd0	0	0.0	0	0	0.0	0.0

The fields listed by the **sar -d** command are as follows:

► **%busy**

Portion of time device was busy servicing a transfer request. This is the same as the **%tm_act** column in the **iostat** command report.

► **avque**

Average number of requests outstanding during that time. This number is a good indicator if an I/O bottleneck exists.

► **r+w/s**

Number of read/write transfers from or to device. This is the same as **tps** in the **iostat** command report.

- ▶ **blks/s**
Number of bytes transferred in 512-byte units
- ▶ **avwait**
Average number of transactions waiting for service (queue length). Average time (in milliseconds) that transfer requests waited idly on queue for the device. This number is currently not reported and shows 0.0 by default.
- ▶ **avserv**
Number of milliseconds per average seek. Average time (in milliseconds) to service each transfer request (includes seek, rotational latency, and data transfer times) for the device. This number is currently not reported and shows 0.0 by default.

6.5.5 Assessing logical volume fragmentation with the `lslv` command

The `lslv` command shows, among other information, the logical volume fragmentation. To check logical volume fragmentation, use the command `lslv -l lvname`, as follows (Example 6-16).

Example 6-16 lslv command example

```
# lslv -l hd2
hd2:/usr
PV          COPIES      IN BAND      DISTRIBUTION
hdisk0      114:000:000    22%          000:042:026:000:046
```

The output of **COPIES** shows that the logical volume `hd2` has only one copy. The **IN BAND** column shows how well the intrapolicy, an attribute of logical volumes, is followed. The higher the percentage, the better the allocation efficiency. Each logical volume has its own intrapolicy. If the operating system cannot meet this requirement, it chooses the best way to meet the requirements. In our example, there are a total of 114 logical partitions (LP); 42 LPs are located on middle, 26 LPs on center, and 46 LPs on inner-edge. Since the logical volume intrapolicy is center, the in-band is 22% (26 / (42+26+46)). The **DISTRIBUTION** shows how the physical partitions are placed in each part of the intrapolicy; that is:

edge : middle : center : inner-middle : inner-edge

6.5.6 Assessing file placement with the `fileplace` command

The `fileplace` command displays the placement of a file's blocks within a logical volume or within one or more physical volumes.

To determine whether the `fileplace` command is installed and available, run the following command:

```
# ls1pp -lI perfagent.tools
```

Use the following command for the file big1 (Example 6-17).

Example 6-17 fileplace command

```
# fileplace -pv big1
File: big1 Size: 3554273 bytes Vol: /dev/hd10
Blk Size: 4096 Frag Size: 4096 Nfrags: 868 Compress: no
Inode: 19 Mode: -rwxr-xr-x Owner: hoetzel Group: system
Physical Addresses (mirror copy 1)                                Logical Fragment
-----
0001584-0001591 hdisk0      8 frags    32768 Bytes,   0.9%    0001040-0001047
0001624-0001671 hdisk0     48 frags   196608 Bytes,   5.5%    0001080-0001127
0001728-0002539 hdisk0    812 frags  3325952 Bytes,  93.5%    0001184-0001995

868 frags over space of 956 frags: space efficiency = 90.8%
3 fragments out of 868 possible: sequentiality = 99.8%
```

This example shows that there is very little fragmentation within the file, and those are small gaps. We can therefore infer that the disk arrangement of big1 is not significantly affecting its sequential read-time. Further, given that a (recently created) 3.5 MB file encounters so little fragmentation, it appears that the file system in general has not become particularly fragmented.

Occasionally, portions of a file may not be mapped to any blocks in the volume. These areas are implicitly filled with zeroes by the file system. These areas show as unallocated logical blocks. A file that has these holes will show the file size to be a larger number of bytes than it actually occupies (that is, the **ls -l** command will show a large size, while the **du** command will show a smaller size or the number of blocks the file really occupies on disk).

The **fileplace** command reads the file's list of blocks from the logical volume. If the file is new, the information may not be on disk yet. Use the **sync** command to flush the information. Also, the **fileplace** command will not display NFS remote files (unless the command runs on the server).

Note: If a file has been created by seeking to various locations and writing widely dispersed records, only the pages that contain records will take up space on disk and appear on a fileplace report. The file system does not fill in the intervening pages automatically when the file is created. However, if such a file is read sequentially (by the **cp** or **tar** commands, for example) the space between records is read as binary zeroes. Thus, the output of such a **cp** command can be much larger than the input file, although the data is the same.

Space efficiency and sequentiality

Higher space efficiency means files are less fragmented and probably provide better sequential file access. A higher sequentiality indicates that the files are more contiguously allocated, and this will probably be better for sequential file access.

- ▶ Space efficiency: Total number of fragments used for file storage (Largest fragment physical address - Smallest fragment physical address + 1)
- ▶ Sequentiality: Total number of fragments - Number of grouped fragments + 1 / Total number of fragments

If you find that your sequentiality or space efficiency values become low, you can use the **reorgvg** command to improve logical volume utilization and efficiency

In Example 6-17, the Largest fragment physical address - Smallest fragment physical address + 1 is: 0002539 - 0001584 + 1 = 956 fragments; total used fragments is: 8 + 48 + 812

= 868; the space efficiency is 868 / 956 (90.8%); the sequentiality is (868 - 3 + 1) / 868 = 99.8%.

Because the total number of fragments used for file storage does not include the indirect blocks location, but the physical address does, the space efficiency can never be 100% for files larger than 32 KB, even if the file is located on contiguous fragments.

6.5.7 The **topas** command

The **topas** command is a Performance Monitoring tool that is ideal for broad spectrum performance analysis. The command is capable of reporting on local system statistics such as CPU use, CPU events and queues, memory and paging use, disk performance, network performance, and NFS statistics. It can report on the top hot processes of the system as well as on Workload Manager (WLM) hot classes. The WLM class information is only displayed when WLM is active.

The **topas** command defines hot processes as those processes that use a large amount of CPU time. The **topas** command does not have an option for logging information. All information is real time.

Note: In order to obtain a meaningful output from the **topas** command, the panel or graphics window must support a minimum of 80 characters by 24 lines. If the display is smaller than this, then parts of the output become illegible.

The **topas** command requires the `perfagent.tools` fileset to be installed on the system. The **topas** command resides in `/usr/bin` and is part of the `bos.perf.tools` fileset that is obtained from the AIX base installable media.

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

```
topas [ -d number_of_monitored_hot_disks ]  
[ -h show help information ]  
[ -i monitoring_interval_in_seconds ]  
[ -n number_of_monitored_hot_network_interfaces ]  
[ -p number_of_monitored_hot_processes ]  
[ -w number_of_monitored_hot_WLM_classes ]  
[ -c number_of_monitored_hot_CPUs ]  
[ -P show full-screen process display ]  
[ -W show full-screen WLM display ]
```

Where:

► -d

Specifies the number of disks to be displayed and monitored. The default value of two is used by the command if this value is omitted from the command line. In order that no disk information is displayed, the value of zero must be used. If the number of disks selected by this flag exceeds the number of physical disks in the system, then only the physically present disks will be displayed. Because of the limited space available, only the number of disks that fit into the display window are shown. The disks by default are listed in descending order of kilobytes read and written per second KBPS. This can be changed by moving the cursor to an alternate disk heading (for example, `Busy%`).

► -h

Used to display the **topas** help.

- ▶ -i
Sets the data collection interval and is given in seconds. The default value is two.
- ▶ -n
Used to set the number of network interfaces to be monitored. The default is two. The number of interfaces that can be displayed is determined by the available display area. No network interface information will be displayed if the value is set to zero.
- ▶ -p
Used to display the top hot processes on the system. The default value of 20 is used if the flag is omitted from the command line. To omit top process information from the displayed output, the value of this flag must be set to zero. If there is no requirement to determine the top hot processes on the system, then this flag should be set to zero, as this function is the main contributor of the total overhead of the **topas** command on the system.
- ▶ -w
Specifies the number of WLM classes to be monitored. The default value of two is assumed if this value is omitted. The classes are displayed as display space permits. If this value is set to zero, then no information about WLM classes will be displayed. Setting this flag to a value greater than the number of available WLM classes results in only the available classes being displayed.
- ▶ -P
Used to display the top hot processes on the system in greater detail than is displayed with the -p flag. Any of the columns can be used to determine the order of the list of processes. To change the order, simply move the cursor to the appropriate heading.
- ▶ -W
Splits the full panel display. The top half of the display shows the top hot WLM classes in detail, and the lower half of the panel displays the top hot processes of the top hot WLM class.

Information about measurement and sampling

The **topas** command makes use of the System Performance Measurement Interface (SPMI) Application Program Interface (API) for obtaining its information. By using the SPMI API, the system overhead is kept to a minimum. The **topas** command uses the **perfstat** library call to access the **perfstat** kernel extensions. In instances where the **topas** command determines values for system calls, CPU clicks, and context switches, the appropriate counter is incremented by the kernel, and the mean value is determined over the interval period set by the -i flag. Other values such as free memory are merely snapshots at the interval time. The sample interval can be selected by the user by using the -i flag option. If this flag is omitted in the command line, then the default of two seconds is used.

6.6 Qlogic SANSurfer

In this section we discuss the Qlogic SANSurfer in the context of troubleshooting, and how to monitor Qlogic based host bus adapters.

The SANSurfer Java GUI can be used to manage the host bus adapters in servers running the SANSurfer agent. This section gives an overview on how to use the tool to make changes to the configuration of host bus adapters, perform diagnostic tests, and check the performance of the host bus adapters.

Details can be obtained from the Web site:

<http://www.ibm.com/servers/storage/support/disk>

Then search under the specific DS4000 model under TOOLS section.

For more information about QLogic SANSurfer refer to *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

6.6.1 Using the QLogic SANSurfer diagnostic tools

You must first use the Qlogic SANSurfer and connect to the server that contains the HBA that you wish to diagnose. When you click one of the host bus adapter ports in the HBA tree panel, the tab panel displays eight tabs as shown in Figure 6-21.

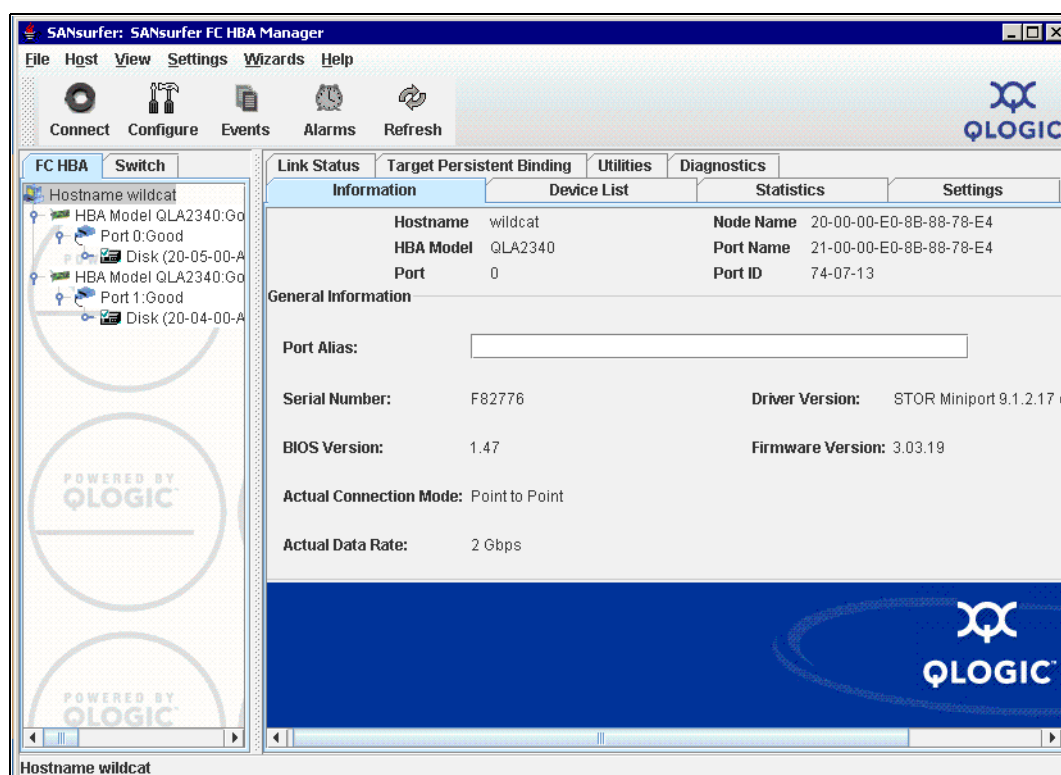


Figure 6-21 QLogic SANSurfer HBA view

The eight tabs are as follows:

- ▶ Information: Displays general information about the server and host bus adapters, such as world-wide name, BIOS version, and driver version.
- ▶ Device List: Displays the devices currently available to the host bus adapter.
- ▶ Statistics: Displays a graph of the performance and errors on the host bus adapters over a period of time.
- ▶ Settings: Displays the current settings and allows you to make remote configuration changes to the NVSRAM of the adapters. All changes require a reboot of the server.
- ▶ Link Status: Displays link information for the devices attached to an adapter connected to a host.
- ▶ Target Persistent Binding: Allows you to bind a device to a specific LUN.

- Utilities: Allows you to update the flash and NVSRAM remotely.
- Diagnostics: Allows you to run diagnostic tests remotely.

You can find detailed information about all these functions in the Users Guide that is bundled with the Qlogic SANSurfer download package.

Here we briefly introduce some of the possibilities of the Qlogic SANSurfer.

Statistics

The Statistics panel displays the following information:

- HBA Port Errors: The number of adapter errors reported by the adapter device driver (connection problem from or to switches or hubs).
- Device Errors: The number of device errors reported by the adapter device driver (I/O problems to DS4000, and so on); this usually gives the first hint about what path to the DS4000 controller has a problem.
- Reset: The number of LIP resets reported by the adapter's driver. If you get increasing numbers, there might be a communication problem between HBAs and storage.
- I/O Count: Total numbers of I/Os reported by the adapter's driver.
- IOPS (I/O per second): The current number of I/Os processed by the adapter.
- BPS (bytes per second): The current numbers of bytes processed by the adapter.

The Statistics panel is shown in Figure 6-22.

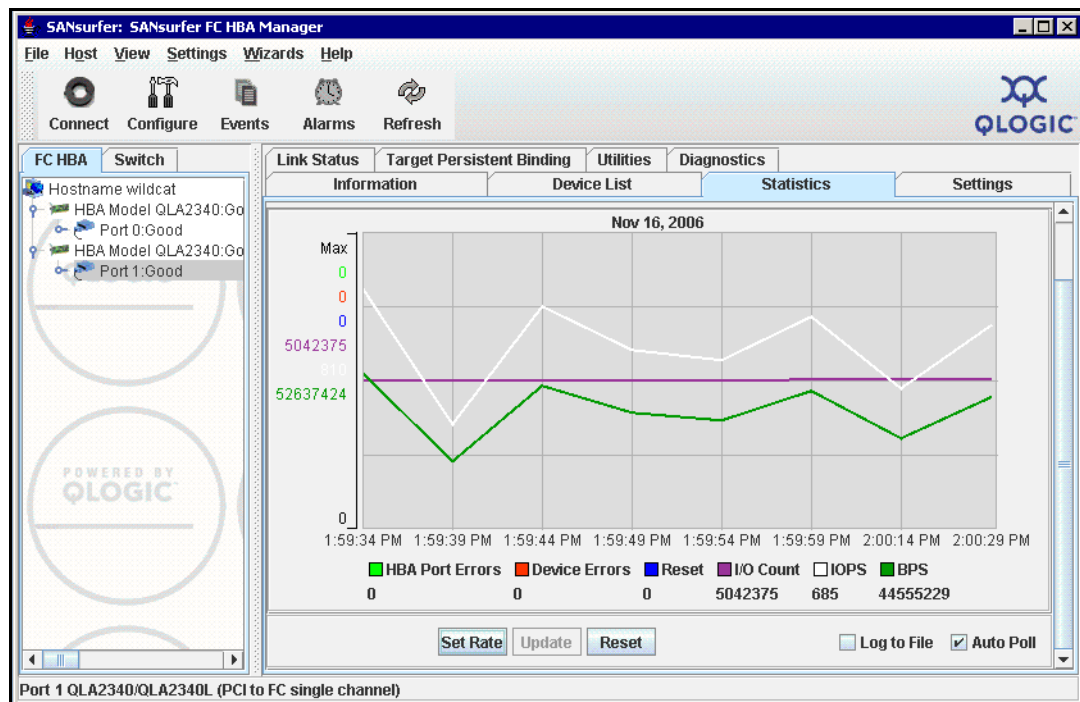


Figure 6-22 Statistics window in Qlogic SANSurfer

To get the graph working, you should select **Auto Poll**, then click **Set Rate** to set a suitable polling rate, and the graph will start to display statistics.

To keep the statistics, select **Log to File**. You will be prompted for a path to save the file. This will create a CSV file of the statistics.

Link Status

If you experience problems with connectivity or performance, or you see entries from RDAC or the HBA driver in Windows, the Link Status tab is where you should start from, to narrow down the device causing the problems (faulty cable, SFPs, and so on).

The following information can be retrieved from the Link Status window:

- ▶ **Link Failure:** The number of times the link failed. A link failure is a possible cause for a time-out (see Windows Event Log).
- ▶ **Loss of Sync:** The number of times the adapter had to re-synchronize the link.
- ▶ **Loss Signal:** The number of times the signal was lost (dropped and re-connected).
- ▶ **Invalid CRC:** The number of Cyclic Redundancy Check (CRC) errors that were detected by the device.

Diagnostics

Using the Diagnostics panel you can perform the loopback and read/write buffer tests.

- ▶ The loopback test is internal to the adapter. The test evaluates the Fibre Channel loop stability and error rate. The test transmits and receives (loops back) the specified data and checks for frame CRC, disparity, and length errors.
- ▶ The read/write buffer test sends data through the SCSI Write Buffer command to a target device, reads the data back through the SCSI Read Buffer command, and compares the data for errors. The test also compares the link status of the device before and after the read/write buffer test. If errors occur, the test indicates a broken or unreliable link between the adapter and the device.

The Diagnostics panel has three main parts:

- ▶ **Identifying Information:** Displays information about the adapter being tested
- ▶ **Test Configuration:** Contains testing options (like data patterns, number of tests, test increments)
- ▶ **Test Results:** Displays the results of a test showing whether the test passed or failed and error counters
 - For a loopback test, the test result includes the following information: Test Status, CRC Error, Disparity Error, and Frame Length Error.
 - For a read/write buffer test, the test result shows the following information: ID (Port/Loop) Status, Data Mismatch, Link Failure, Sync Loss, Signal Loss, and Invalid CRC.

6.7 MPPUTIL Windows 2000/2003

With the new RDAC driver architecture in Windows 2000 and Windows 2003 a new command line utility, called **mppUtil**, was introduced as well. Once you have installed the RDAC driver, the utility can be found in C:\Program Files\IBM_DS4000\mpp\mppUtil.exe

With mppUtil, you can display the driver internal information like controller, path and volume information for a specified DS4000. It also allows you to rescan for new LUNs as well as doing a failback initiated from the host after a volume failover has occurred. This means that you no longer need to redistribute logical drives from the Storage Manager after failover, but you can simply use the mppUtil to move the volumes back to the preferred paths.

If you want to scan for all the DS4000s that are attached to your Windows 2000/2003 host you have to use the command `mpputil -a`, as illustrated in Example 6-16 on page 221.

Example 6-18 Using mpputil in Windows

```
C:\Program Files\IBM_DS4000\mpp\mpputil -a
Hostname      = radon
Domainname    = almaden.ibm.com
Time          = GMT Fri Oct 21 21:23:49 2005
```

Info of Array Module's seen by this Host.

ID	WWN	Name
0	600a0b80001744310000000041eff538	ITS0DS4500_A

If you want to get more details about a specific DS4000, you can either use the ID or the Name as parameter of the `mppUtil` command. Use the command `mppUtil -a storage_server_name` or `mppUtil -g 0`. Both will produce the same result.

6.8 Windows Performance Monitor

Windows comes with a Performance Monitoring tool (do not confuse it with the Storage Manager Performance Monitor). It can be run from the menu under **Administrative tools** → **Performance** or by entering `perfmon` at the command line. Performance logs and alerts can be used to get statistics from the operating system about the hardware.

You can configure Performance Monitor and alerts to display current performance statistics and also to log these statistics to a log file for later examination. In Windows Server 2003, these log files can be stored in text files, binary files, or an SQL database.

There is a vast array of counters that can be monitored, but monitoring does have a slight system overhead. The more counters being monitored, the higher the overhead and impact on performance.

In the case of the DS4000 storage server, a logical drive or LUN for the DS4000 is presented as a physical drive to the host operating system. This physical disk, as seen by the OS, can be then split up into several logical drives, depending on partitioning and formatting. With Performance Monitor, you can monitor the physical disk or the logical drives it contains.

The following disk counters are available:

- ▶ % Disk Read Time
- ▶ % Disk Time
- ▶ % Disk Write Time
- ▶ % Idle Time
- ▶ Average Disk Bytes/Read
- ▶ Average Disk Bytes/Transfer
- ▶ Average Disk Bytes/Write
- ▶ Average Disk Queue Length
- ▶ Average Disk Read Queue Length
- ▶ Average Disk sec/Read
- ▶ Average Disk sec/Transfer
- ▶ Average Disk sec/Write

- ▶ Average Disk Write Queue Length
- ▶ Current Disk Queue Length
- ▶ Disk Bytes/sec
- ▶ Disk Read Bytes/sec
- ▶ Disk Reads/sec
- ▶ Disk Transfers/sec
- ▶ Disk Write Bytes/sec
- ▶ Disk Writes/sec
- ▶ Split I/O/sec

A list of all Logical Disk counters includes the list above, plus these two counters.

- ▶ % Free Space
- ▶ Free Megabytes

The counters to monitor for disk performance are:

- ▶ *Physical: Disk Reads/sec* for the number of read requests per second to the disks.
- ▶ *Physical: Disk Writes/sec* for the number of write requests per second to the disks.
- ▶ *Logical: % Free Space* for the amount of free space left on a volume. This has a big impact on performance if the disk runs out of space.
- ▶ *Physical: Average Disk Queue Length* that goes above 2 for a long period of time indicates that the disk is queuing requests continuously and that could indicate a potential bottleneck.
- ▶ *Physical: % Disk Time and % Idle Time*. Interpret the % Disk Time counter carefully. This counter may not accurately reflect utilization on multiple disk system; it is important to use the % Idle Time counter as well. These counters cannot display a value exceeding 100%.

You may have to use these counters in conjunction with other counters (memory, network traffic, and processor utilization) to fully understand your Microsoft server and get a complete picture of the system performance.

Refer to your Microsoft documentation for a full explanation of these counters and how they are used to monitor system performance.

Alerts

Microsoft Performance logs and alerts can be configured to send an alert if the counters being monitored reach a preset threshold. This is a useful tool, but if the thresholds are set too low, false alerting may occur.



IBM TotalStorage Productivity Center for Disk

This chapter discusses how to use IBM TotalStorage Productivity Center for Disk (TPC for Disk) to manage and monitor performance on the DS4000 Storage Subsystem.

The first part of this chapter is a brief overview of the overall TPC offering that includes components such as TPC for Fabric, TPC for Data, TPC for Disk, and TPC for Replication.

The second part of the chapter focuses on TPC for Disk and the steps required to manage and monitor the DS4000 from a TPC server. It also includes some examples of performance reports that TPC for Disk can generate.

For more detailed information about TPC. Refer to *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194.

7.1 IBM TotalStorage Productivity Center

IBM TotalStorage Productivity Center is an integrated set of software components that provides end-to-end storage management. This software offering provides disk and tape library configuration and management, performance management, SAN fabric management and configuration, and host-centered usage reporting and monitoring from the perspective of the database application or file system.

IBM TotalStorage Productivity Center:

- ▶ Simplifies the management of storage infrastructures
- ▶ Manages, configures, and provisions SAN-attached storage
- ▶ Monitors and tracks performance of SAN-attached devices
- ▶ Monitors, manages, and controls (through zones) SAN fabric components
- ▶ Manages the capacity utilization and availability of file systems and databases

7.1.1 TotalStorage Productivity Center structure

In this section we look at the TotalStorage Productivity Center structure from the logical and physical view.

Logical structure

The logical structure of TotalStorage Productivity Center V3.1 has three layers, as shown in Figure 7-1.

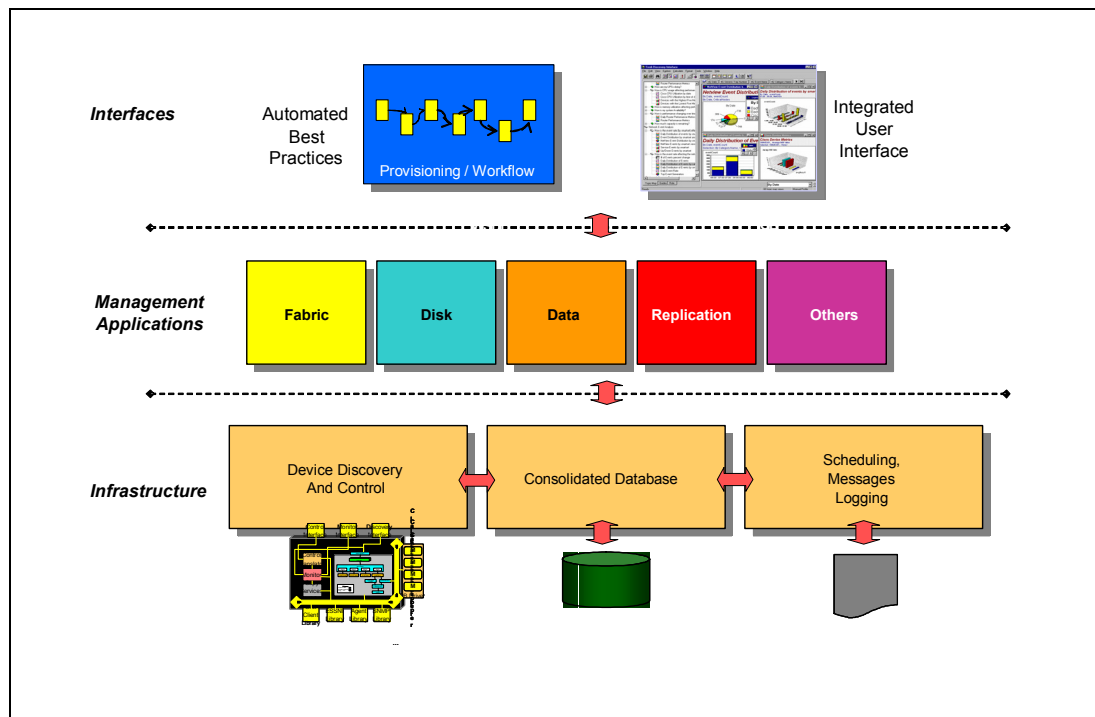


Figure 7-1 IBM TPC v3.1 logical structure

The infrastructure layer consists of basic functions such as messaging, scheduling, logging, device discovery, and a consolidated database shared by all components of TotalStorage Productivity to ensure consistent operation and performance.

The application layer consists of core TotalStorage Productivity Center management functions, that rely on the common base infrastructure to provide different disciplines of storage or data management. These application components are most often associated with the product components that make up the product suite, such as fabric management, disk management, replication management and data management.

The interface layer presents integration points for the products that make up the suite. The integrated graphical user interface (GUI) brings together product and component functions into a single representation that seamlessly interacts with the components to centralize the tasks for planning, monitoring, configuring, reporting, topology viewing, and problem resolving.

Physical Structure

IBM TotalStorage Productivity Center is comprised of the following components:

- ▶ A data component: IBM TotalStorage Productivity Center for Data (formerly IBM Tivoli Storage Resource Manager)
- ▶ A fabric component: IBM TotalStorage Productivity Center for Fabric (formerly IBM Tivoli SAN Manager)
- ▶ A disk component: IBM TotalStorage Productivity Center for Disk (formerly IBM TotalStorage Multiple Device Manager)
- ▶ A replication component (formerly IBM TotalStorage Multiple Device Manager Replication Manager)

IBM TotalStorage Productivity Center includes a centralized suite installer.

Figure 7-2 shows the TotalStorage Productivity Center V3.1 physical structure.

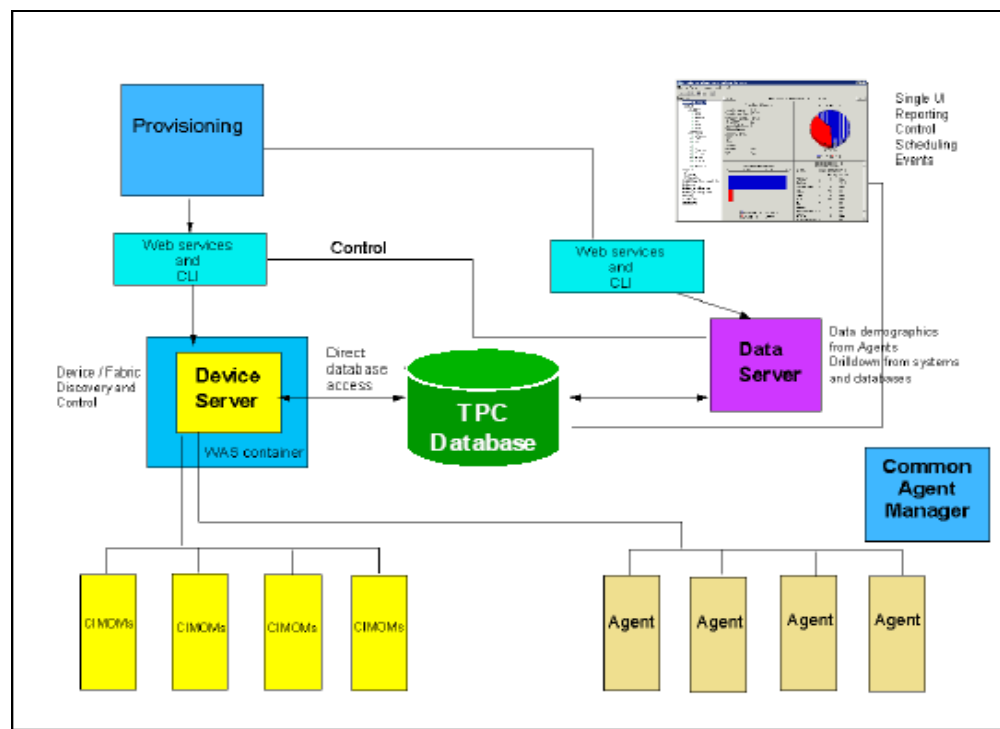


Figure 7-2 TPC V3.1 physical structure

The Data server is the control point for product scheduling functions, configuration, event information, reporting, and GUI support. It coordinates communication with agents and data collection from agents that scan file systems and databases to gather storage demographics and populate the database with results. Automated actions can be defined to perform file system extension, data deletion, and backup or archiving or event reporting when defined thresholds are encountered. The data server is the primary contact point for GUI user interface functions. It also includes functions that schedule data collection and discovery for the device server.

The device server component discovers, gathers information from, analyzes performance of, and controls storage subsystems and SAN fabrics. It coordinates communication with agents and data collection from agents that scan SAN fabrics.

The single database instance serves as the repository for all TotalStorage Productivity Center components.

The Data agents and Fabric agents gather host, application, and SAN fabric information and send this information to the data server or device server.

The GUI allows you to enter information or receive information for all TotalStorage Productivity Center components.

The command-line interface (CLI) allows you to issue commands for major TotalStorage Productivity Center functions.

7.1.2 Standards and protocols used in IBM TotalStorage Productivity Center

TPC was built upon storage industry standards. This section presents an overview the standards used within the different TPC components.

Common Information Model/Web-Based Enterprise Management

Web-Based Enterprise Management (WBEM) is a initiative of the Distributed Management Task Force (DMTF) with the objective to enable the management of complex IT environments. It defines a set of management and Internet standard technologies to unify the management of complex IT environments.

The three main conceptual elements of the WBEM initiative are:

- ▶ Common Information Model (CIM)
CIM is a formal object-oriented modeling language that is used to describe the management aspects of systems.
- ▶ xmlCIM
This is a grammar to describe CIM declarations and messages used by the CIM protocol.
- ▶ Hypertext Transfer Protocol (HTTP)
- ▶ Hypertext Transfer Protocol over Secure Socket Layer (HTTPS)

HTTP and HTTPS are used as a way to enable communication between a management application and a device that both use CIM.

The CIM Agent provides a means by which a device can be managed by common building blocks rather than proprietary software. If a device is CIM-compliant, software that is also CIM-compliant can manage the device. Using CIM, you can perform tasks in a consistent manner across devices and vendors.

The CIM/WBEM architecture defines the following elements:

- ▶ Agent code or CIM Agent

An open-systems standard that interprets CIM requests and responses as they transfer between the client application and the device. The Agent is normally embedded into a device, which can be hardware or software. When not embedded (which is the case for devices that are not CIM-ready such as the DS4000), a device provider (usually provided by the device manufacturer) is required.

- ▶ CIM Object Manager (CIMOM)

The common conceptual framework for data management that receives, validates, and authenticates the CIM requests from the client application (such as TPC for disk). It then directs the requests to the appropriate component or a device provider.

- ▶ Client application or CIM Client

A storage management program, such as TotalStorage Productivity Center, that initiates CIM requests to the CIM Agent for the device. A CIM Client can reside anywhere in the network, because it uses HTTP to talk to CIM Object Managers and Agents.

- ▶ Device or CIM Managed Object

A Managed Object is a hardware or software component that can be managed by a management application by using CIM (for example, a DS4000 storage server).

- ▶ Device provider

A device-specific handler that serves as a plug-in for the CIMOM. That is, the CIMOM uses the handler to interface with the device. There is a device provider for the DS4000.

Note: The terms *CIM Agent* and *CIMOM* are often used interchangeably. At this time, few devices come with an integrated CIM Agent. Most devices need a external CIMOM for CIM to enable management applications (CIM Clients) to talk to the device.

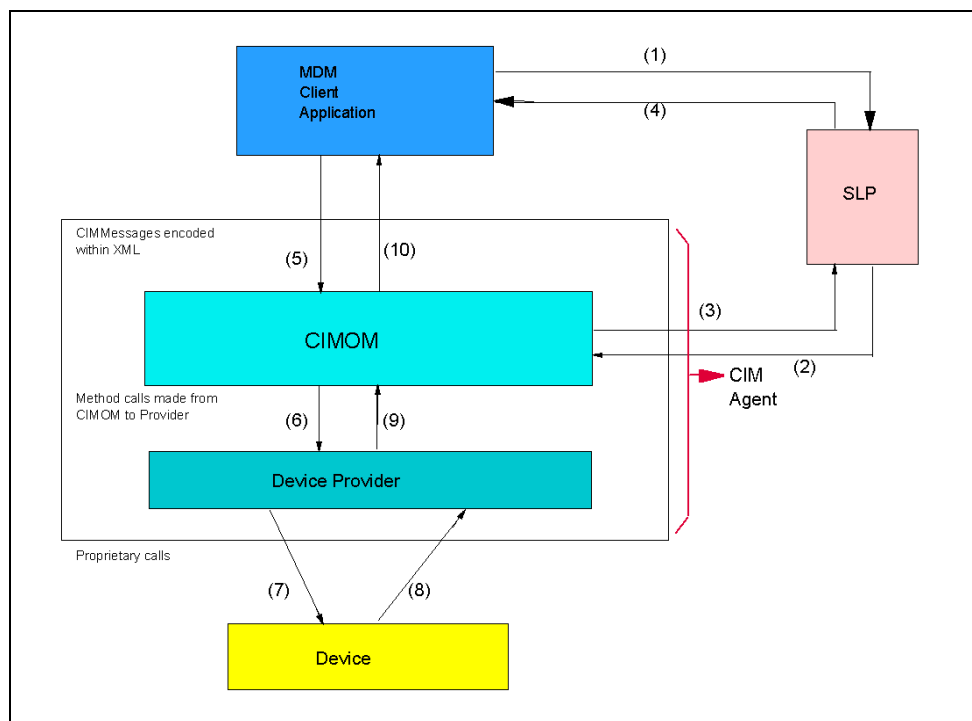


Figure 7-3 CIM architecture elements

For more information go to:

<http://www.dmtf.org/standards/wbem/>

Storage Management Initiative - Specification

The Storage Networking Industry Association (SNIA) has fully adopted and enhanced the CIM for Storage Management in its Storage Management Initiative - Specification (SMI-S). SMI-S was launched in mid-2002 to create and develop a universal open interface for managing storage devices, including storage networks.

The idea behind SMI-S is to standardize the management interfaces so that management applications can use these and provide cross-device management. This means that a newly introduced device can be immediately managed as it conforms to the standards. TPC for disk uses that standard.

Service Location Protocol

The Service Location Protocol (SLP) is an IETF standard that SLP provides as a scalable framework for the discovery and selection of network services.

SLP enables the discovery and selection of generic services, which can range in function from hardware services such as those for printers or fax machines, to software services such as those for file servers, e-mail servers, Web servers, databases, or any other possible services that are accessible through an IP network.

Traditionally, to use a particular service, an end-user or client application needs to supply the host name or network IP address of that service. With SLP, however, the user or client no longer needs to know individual host names or IP addresses (for the most part). Instead, the user or client can search the network for the desired service type and an optional set of qualifying attributes.

The SLP architecture includes three major components:

- ▶ **Service agent (SA)**

A process working on the behalf of one or more network services to broadcast the services.

- ▶ **User agent (UA)**

A process working on the behalf of the user to establish contact with some network service. The UA retrieves network service information from the service agents or directory agents.

- ▶ **Directory agent (DA)**

A process that collects network service broadcasts.

The SA and UA are required components in an SLP environment, where the SLP DA is optional.

The SMI-S specification introduces SLP as the method for the management applications (the CIM clients) to locate managed objects. In SLP, an SA is used to report to UAs that a service that has been registered with the SA is available.

7.2 Managing DS4000 using IBM TPC for Disk

TPC for Disk enables device configuration and management of SAN-attached devices from a single console. It allows you to manage network storage components based on SMI-S, such as IBM System Storage SAN Volume Controller (SVC), IBM System Storage Disk Subsystems (DS4000, DS6000, and DS8000 Series), and other storage subsystems that support the SMI-S standards.

TPC for Disk also includes performance monitoring and reporting capabilities such as:

- ▶ Collect and store performance data and provide alerts.
- ▶ Provide graphical performance reports.
- ▶ Help optimize storage allocation.
- ▶ Provide volume contention analysis.

The performance function starts with the data collection task, responsible for capturing performance statistics for the devices and storing the data in the database.

Thresholds can be set for certain performance metrics depending on the type of device. Threshold checking is performed during data collection. When performance is outside the specified boundaries, alerts can be generated.

Once performance data has been collected, you can configure TotalStorage Productivity Center for Disk to present graphical or text reports on the historical performance behavior of specified devices.

For DS4000, you can use TPC for Disk to perform storage provisioning, logical drive (LUN) creation and assignment, performance management, and reporting. TPC for Disk can monitor the disk subsystem ports, arrays, and measure logical drives throughput, IOPS, and cache rates.

Device discovery is performed by the SLP, and configuration of the discovered devices is possible in conjunction with CIM agents associated with those devices, using the standard mechanisms defined in SMI-S.

For devices that are not CIM ready, as the DS4000, the installation of a proxy application (CIM Agent) is required.

To monitor and manage a DS4000 through TPC, perform the following task sequence:

1. Install the CIM agent for DS4000.
2. Register CIMOM in TPC.
3. Probe discovered DS4000 to gather device information.
4. Create Performance Monitor job.

7.2.1 Install CIM agent for DS4000

This section discusses the implementation of the CIM agent for a DS4000, assuming that a TPC environment is already in place (we used a TPC v3.1.3 environment during the preparation of this book).

The DS4000 is not a CIM- ready device. Therefore, a DS4000 Device provider (acting as the CIM Agent) must be installed (on a host system) to bridge communications between the DS4000 and the TPC server. The device provider (agent) for the DS4000 is provided by Engenio and is called the SANtricity SMI Provider.

To install the Engenio SANtricity SMI Provider in Windows Server 2003:

1. Download the code from the following Engenio Web site:

http://www.engenio.com/products/smi_provider_archive.html

Choose the version that is intended specifically for use with IBM TotalStorage Productivity Center and download the correct operating system of the server on which the SMI-S Provider is going to be installed. At the time of writing, the SMI-S Provider version is 9.16.G0.34, which is intended specifically for use with IBM TPC 3.1.1 and later.

Download and read the readme file as well to make sure that your current storage subsystem firmware level is supported.

2. Extract the downloaded file and launch the installer to start the installation process. In our example, the installer file name is Windows_Installer_09.16.G0.34_setup.exe. Launching the executable starts the InstallShield Wizard and opens the Welcome window, as shown in Figure 7-4. Click **Next** to continue.

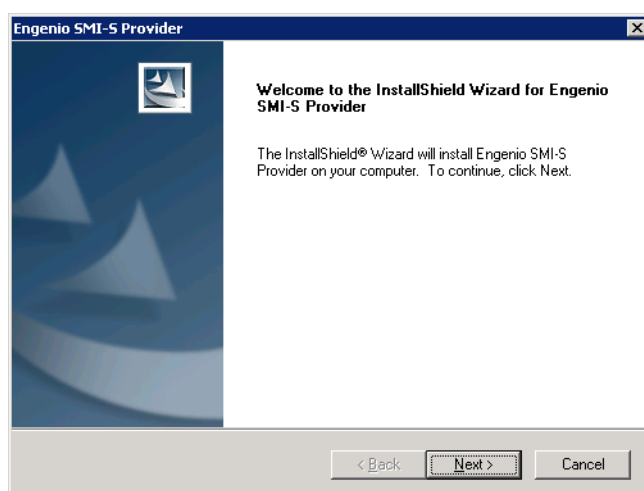


Figure 7-4 Welcome window of Engenio SMI-S Provider InstallShield Wizard

3. The License Agreement window opens. If you agree with the terms of the license agreement, click **Yes** to accept the terms and continue the installation.

4. The System Info window opens. The minimum requirements are listed along with the install system disk free space and memory attributes, as shown in Figure 7-5.

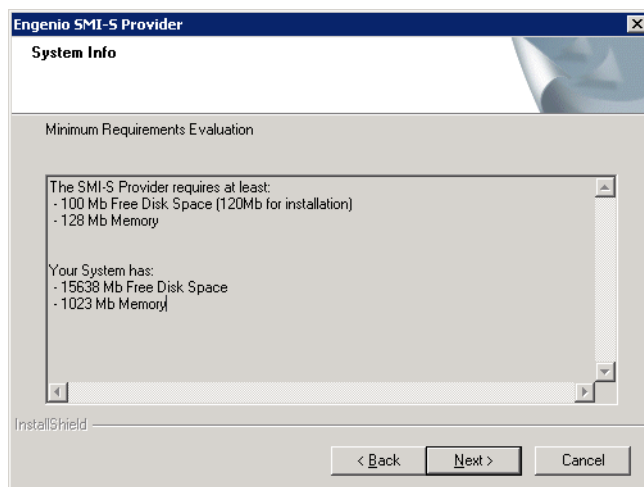


Figure 7-5 Minimum System Requirements window

5. Click **Next** if your system meets the minimum requirements.
6. The Choose Destination Location window opens. Click **Browse** to choose another location or click **Next** to begin the installation of the DS4000/FASTT CIM agent (Figure 7-6).

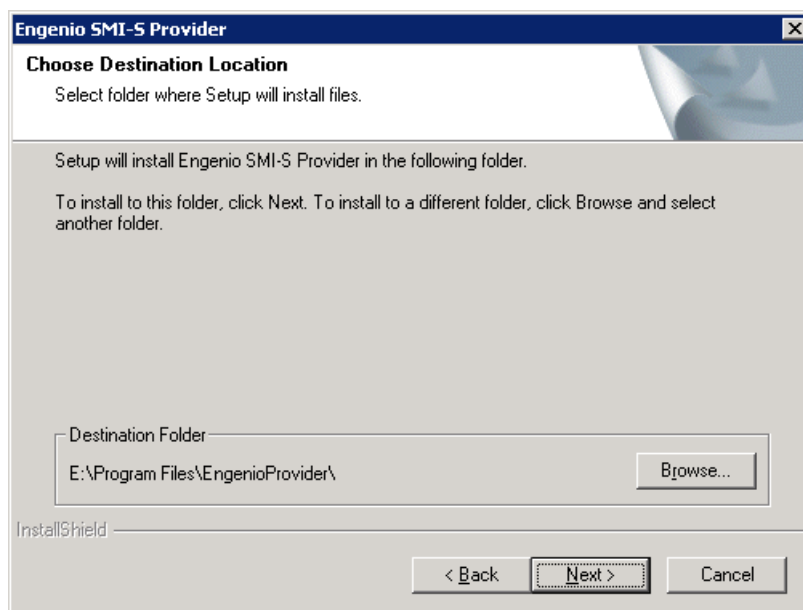


Figure 7-6 Choose Destination Location window

The InstallShield Wizard is now preparing and copying files into the destination folder.

7. In the next window, enter IP addresses or host names of devices to be discovered by the SMI-S Provider at startup, as shown in Figure 7-7. In this case, you need to enter IP addresses of your DS4000 storage subsystem.

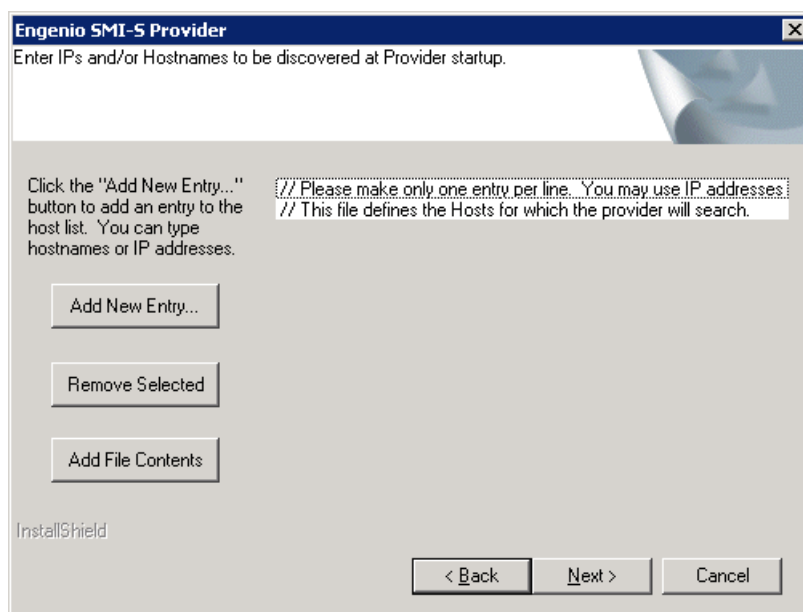


Figure 7-7 Device List window

Use the **Add New Entry** button to add the IP addresses or host names of the DS4000 devices with which this DS4000 CIM Agent will communicate. Enter one IP address or host name at a time until all of the DS4000 devices have been entered and click **Next**, as shown in Figure 7-8.

Note that you can only make one entry at a time. Click **Add New Entry** again to add the other IP address or host name of your DS4000 storage subsystem.

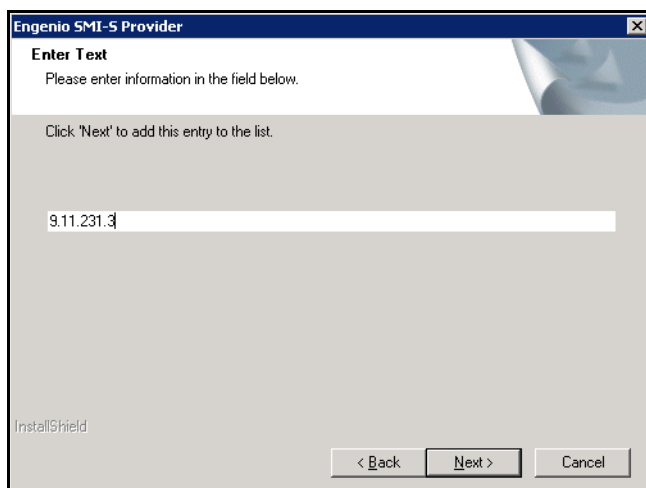


Figure 7-8 Add IP address or host name window

In our example, we enter two IP addresses, as shown in Figure 7-9.

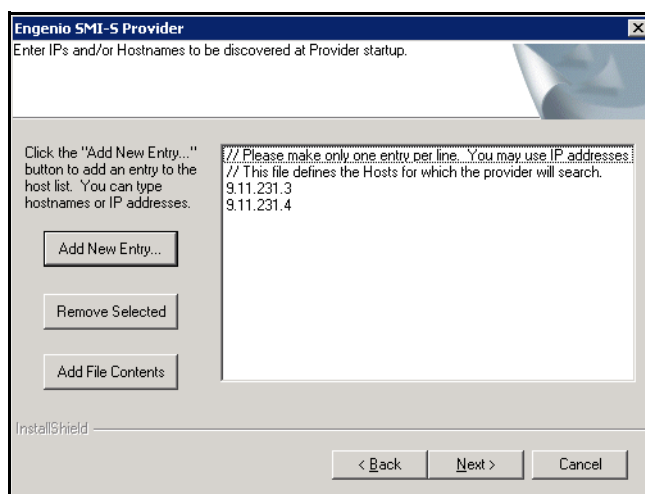


Figure 7-9 Modify device window

Important: Do not enter the IP address of a DS4000 device in multiple DS4000 CIM Agents within the same subnet. This can cause unpredictable results on the TotalStorage Productivity Center for Disk server and could cause a loss of communication with the DS4000 devices.

8. Click **Next** to start the Engenio SMI-S Provider Service. When the Service has started, the installation of the Engenio SMI-S Provider is complete. You can click Finish on the InstallShield Wizard Complete window, as shown in Figure 7-10.

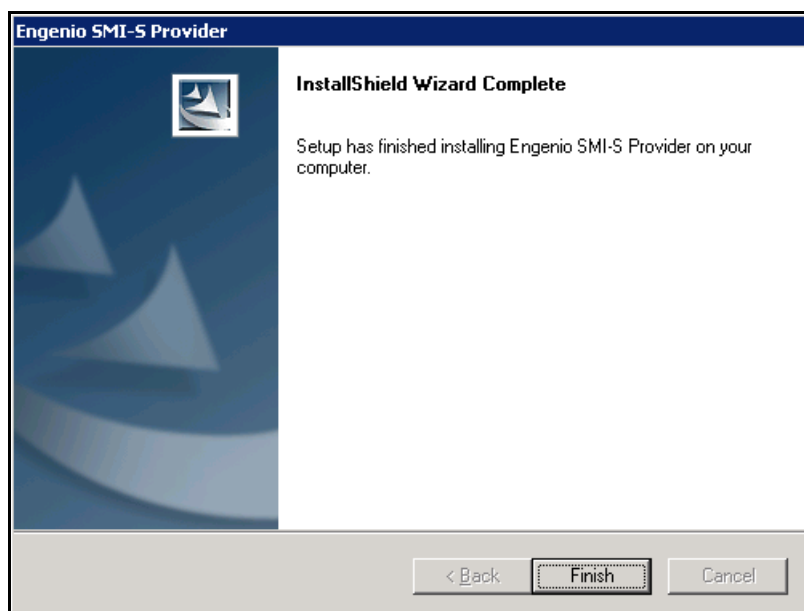


Figure 7-10 InstallShield Wizard Complete window

During the start of the service, the Engenio code processes all of the entries in the arrayhosts.txt file. The configuration is stored in an another file named:

C:\Program Files\EngenioProvider\SMI_SProvider\bin\providerStore

Every time you change anything with the registered DS4000/FaStT controllers and restart the Engenio CIMOM, and when you make a new discovery, the providerStore and arrayhosts.txt are updated with a new time stamp.

Verifying Engenio SMI-S Provider Service

To verify that the service has started, open the Windows Service window (**Start → All Programs → Administrative Tools → Services**) and checking the status of the Engenio SMI-S Provider Server service, as shown in Figure 7-11.

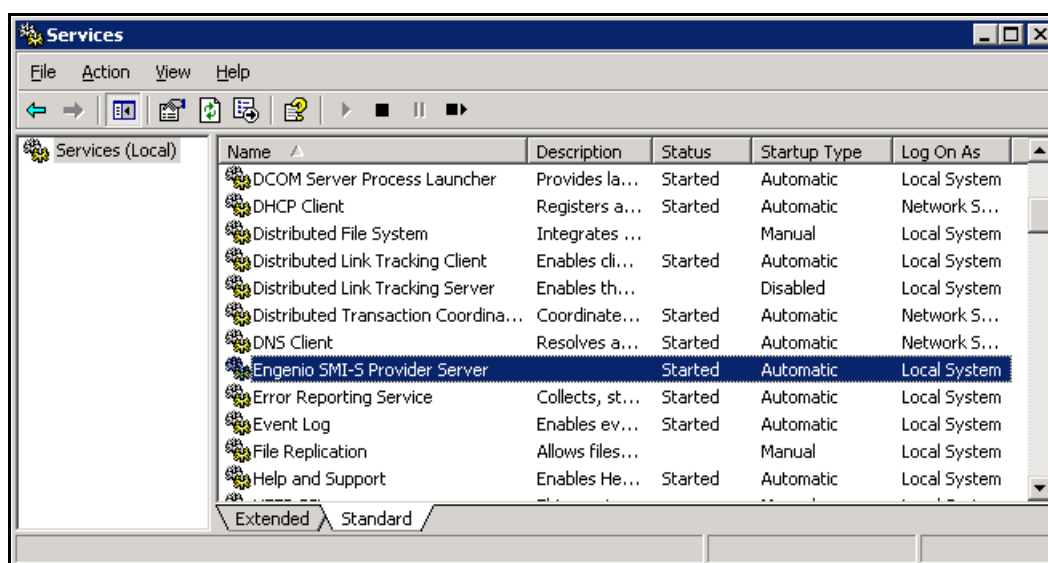


Figure 7-11 Verify Engenio SMI-S Provider Service

Use this interface to start/stop/restart the service. Note that you need to restart the service after any modification to the file arrayhosts.txt.

7.2.2 Registering the Engenio SMI-S provider in TPC

The Engenio SMI-S provider must now be registered as a CIMOM agent in TPC. This can be done by SLP-DA, if the service is installed, or manually via the CIMOM registration menu at the TPC console. In our example, we register the CIMOM manually since we did not implement SLP-DA (for more information about SLP-DA and how to register CIMOM by SLP-DA, refer to *IBM TotalStorage Productivity Center: The Next Generation*, SG24-7194).

Important: You cannot have the DS4000 management password set if you are using IBM TotalStorage Productivity Center, because the Engenio CIMOM has no capability to keep track of the user ID and password combinations that would be necessary to get into all of the managed DS4000 subsystems.

To register the CIMOM manually, open the TPC console, expand **Administrative Service** → **Agents** → **CIMOM**, then click **Add CIMOM** on the right panel, as shown in Figure 7-12.

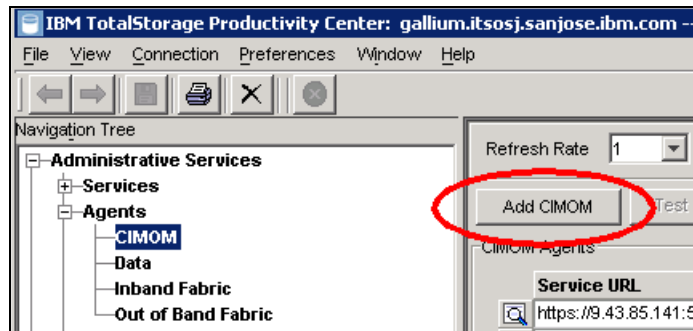


Figure 7-12 Register CIMOM manually from TPC Console

Enter the following information in the Add CIMOM window:

- ▶ Host: IP address or fully qualified domain name of the server where you installed the Engenio SMI-S Provider (*not* the IP address of your DS4000).
- ▶ Port: unsecure port 5988 (The Engenio Provider does not support secure communication at the time of writing.)
- ▶ Username: any username (The Engenio Provider does not require any specific user ID and password to authenticate.)
- ▶ Password: any password (not null)
- ▶ Password Confirm: same as above
- ▶ Interoperability: /interop
- ▶ Protocol: HTTP
- ▶ Display name: any_name to identify the CIM agent
- ▶ Description: optional

Figure 7-13 shows an example.

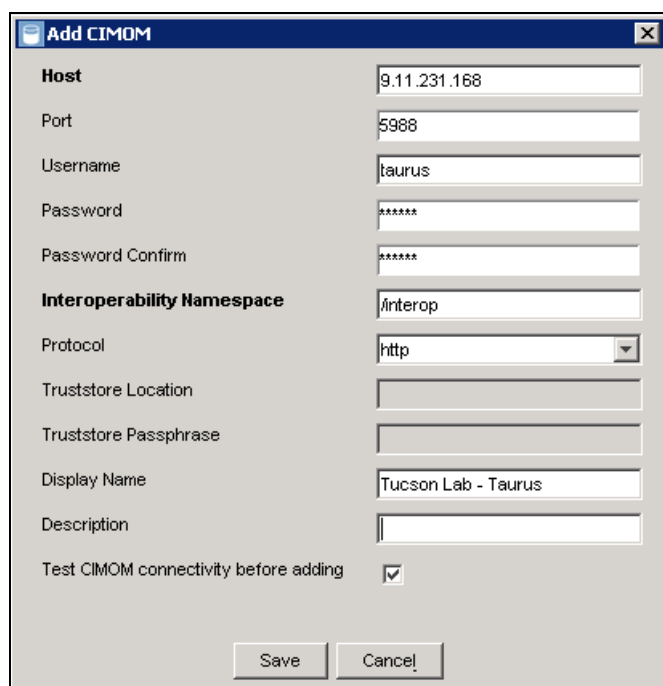
A screenshot of the 'Add CIMOM' dialog box. It contains several input fields: 'Host' with '9.11.231.168', 'Port' with '5988', 'Username' with 'taurus', 'Password' and 'Password Confirm' both with '*****', 'Interoperability Namespace' with '/interop', 'Protocol' with a dropdown menu showing 'http', 'Truststore Location' and 'Truststore Passphrase' both empty, 'Display Name' with 'Tucson Lab - Taurus', and 'Description' empty. There is a checkbox labeled 'Test CIMOM connectivity before adding' which is checked. At the bottom are 'Save' and 'Cancel' buttons.

Figure 7-13 Add CIMOM window

If the “Test CIMOM connectivity before adding” check box is checked, you will see the connectivity status after clicking **Save**, as illustrated in Figure 7-14.

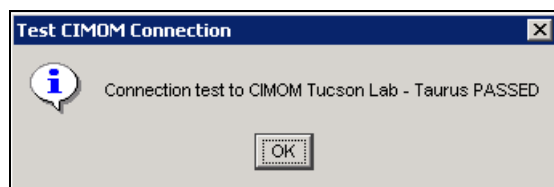
A screenshot of the 'Test CIMOM Connection' dialog box. It features an information icon (i) and the text 'Connection test to CIMOM Tucson Lab - Taurus PASSED'. At the bottom is an 'OK' button.

Figure 7-14 CIMOM connection test status

If the connection test failed, you must verify that your Engenio SMI-S Provider is properly installed in the server and that the service is running. Also, check for any connectivity or possible firewall problems.

Use DOS utilities such as ping, netstat and telnet from the TPC server to the server where Engenio SMI-S Provider is installed to check the connectivity.

Show Managed Devices

After the CIM agent was successfully registered, you can view the devices managed by the CIM agent (the Engenio SMI-S provider in our case). Highlight the **CIMOM** agent in the Navigation pane, and click **Show Managed Devices**, as shown in Figure 7-15.

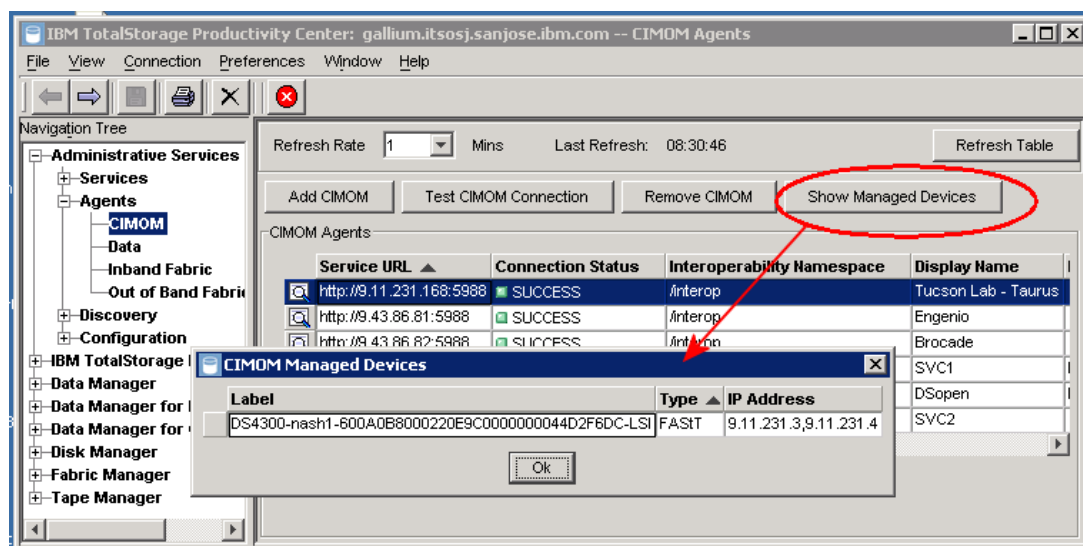


Figure 7-15 Show Managed Devices

7.2.3 Probing CIM agent

The CIMOM discovery and registration process will only detect and recognize devices bridged by the CIMOM agent. TPC needs to get more detailed information about a device to effectively manage and monitor it. This can be accomplished by running a probe job. The probe interacts with the CIM agent by sending a request for information about the configured device. The probe job must in fact be performed before you attempt to manage the device.

With the DS4000, a probe job needs to be run before volume (logical drive) management and performance monitoring can be performed on the storage subsystem.

To configure a probe job, open the TPC console, and expand **IBM TotalStorage Productivity Center** → **Monitoring** → **Probes**. Right-click **Probes** and click **Create Probe**, as shown in Figure 7-16.

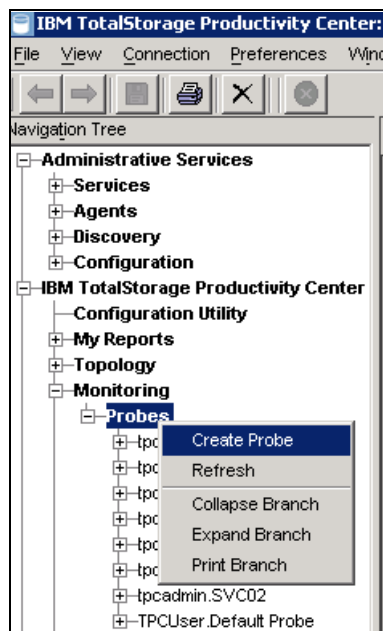


Figure 7-16 Create probe

The Create Probe panel is displayed in the right pane. Expand **Storage Subsystems**, select the DS4000 you want to probe, and click the >> button, as shown in Figure 7-17.

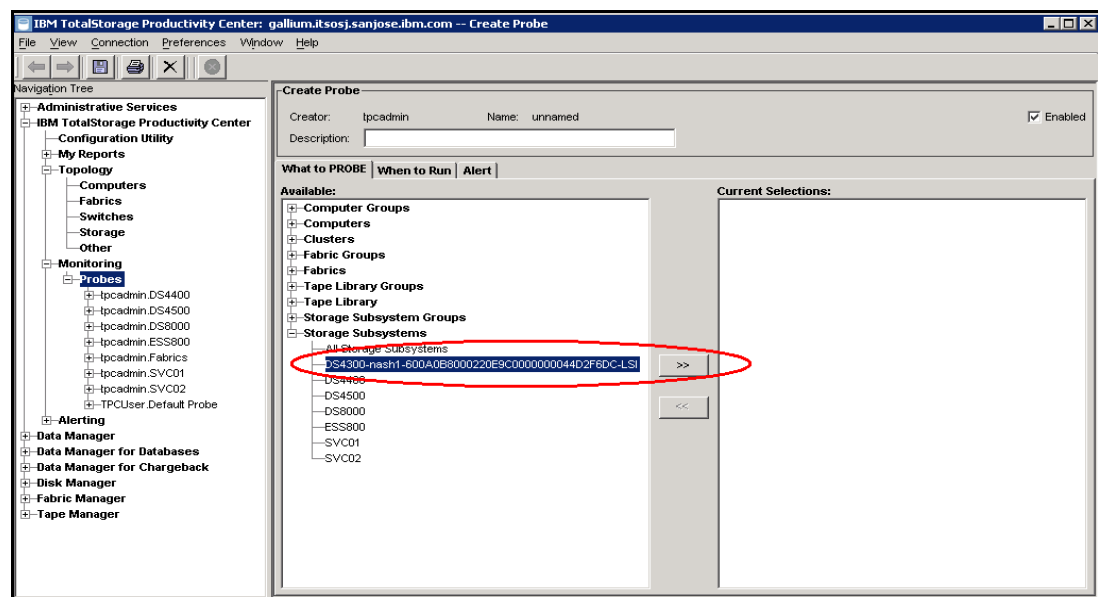


Figure 7-17 Select storage subsystem to probe

Click the **When to Run** tab, as indicated in Figure 7-18.

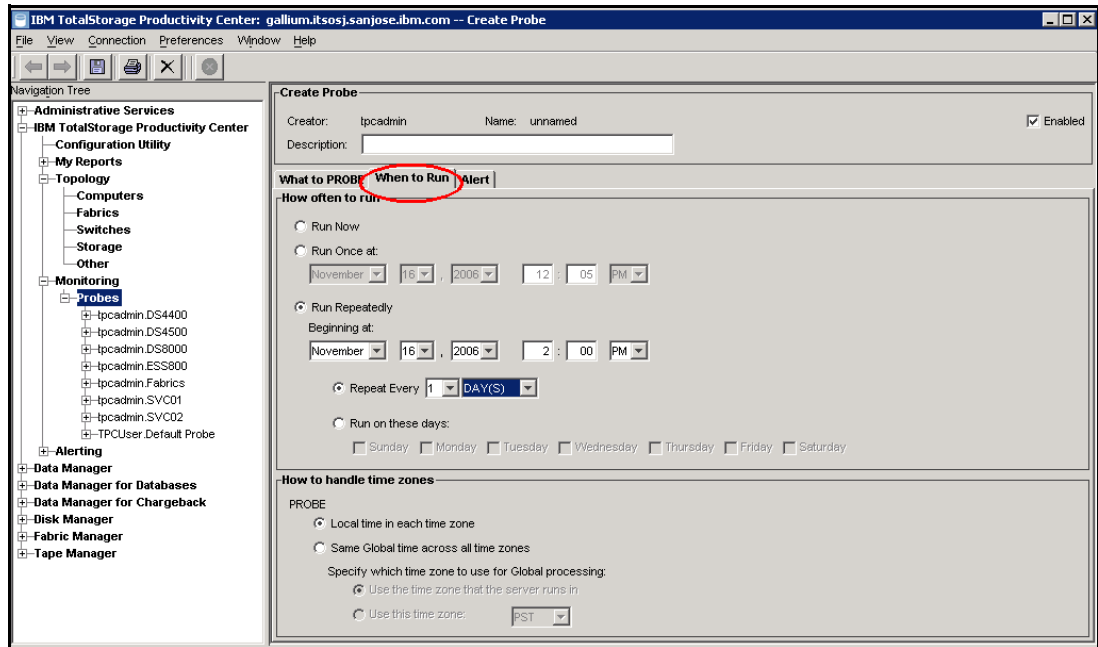


Figure 7-18 When to run the probe

On the When to Run tab, you can customize how often you want the probe job to run.

You can also specify how to handle time zones, which is especially useful when the storage subsystems to be probed are located in different time zones.

Next click the **Alert** tab, as shown in Figure 7-19.

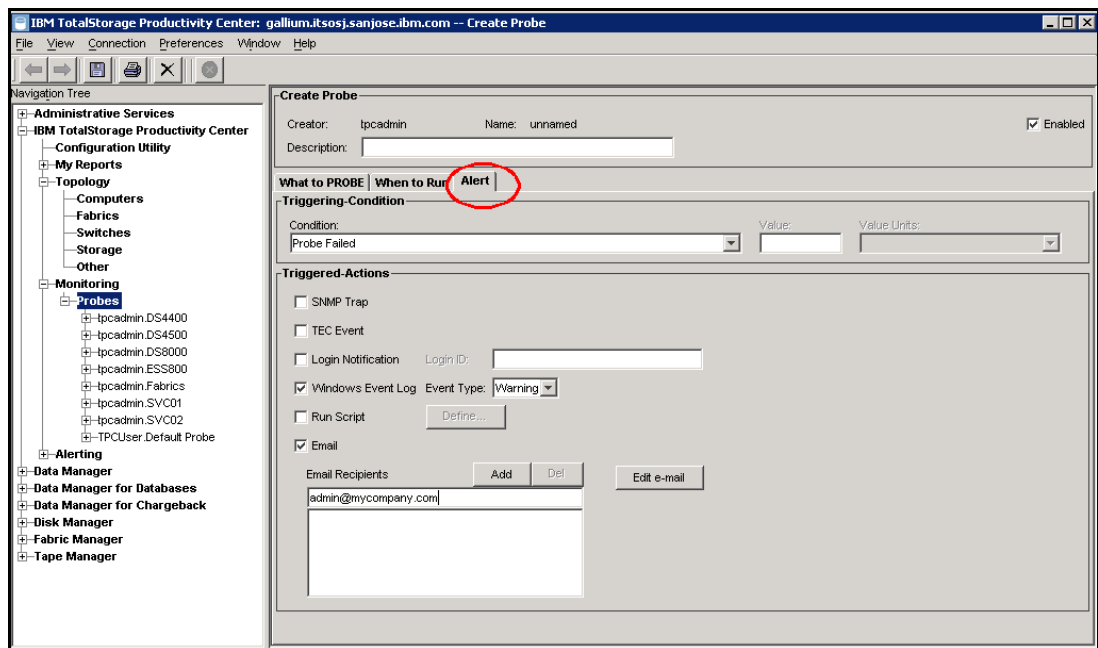


Figure 7-19 Alert when the probe failed

On the Alert tab, you can configure what actions should be automatically taken when the probe fails. There are several options available, including logging the error messages to the Windows Event Log or sending e-mail to specific recipients.

To save the probe job, click **File** → **Save**. You are prompted for a probe job name, as shown in Figure 7-20. Type the probe name of your choice and click **OK** to save and submit the job. The job will be activated according to the schedule you have configured earlier.

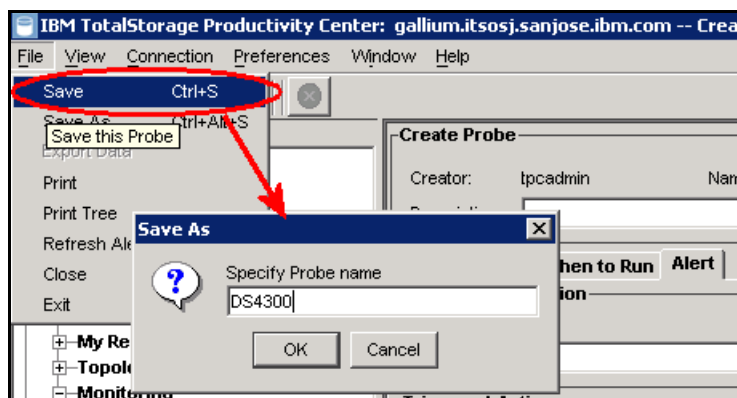


Figure 7-20 Save the probe

To check the probe job status, expand **IBM TotalStorage Productivity Center** → **Monitoring** → **Probes** and select the probe job name. In this example, the probe job name is DS4300, as shown in Figure 7-21.

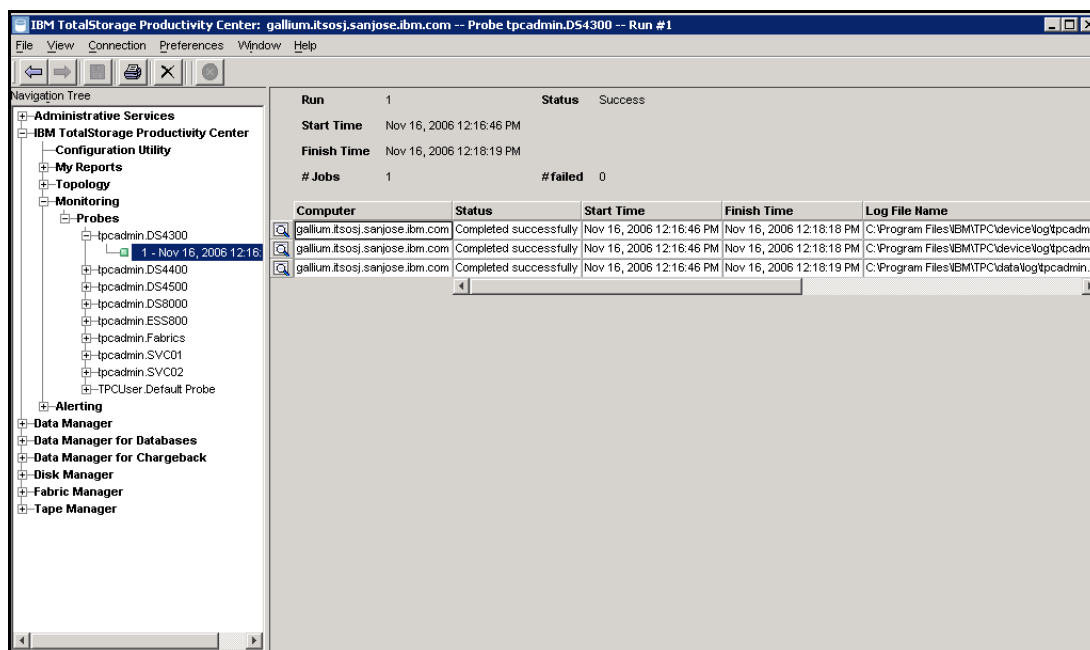


Figure 7-21 Check probe status

Viewing storage subsystem information

Once the probe job created for the storage subsystem has successfully completed, you can view more detailed information about your storage subsystem and start using TPC to manage the system.

To view detailed information about the storage subsystem, expand **Data Manager** → **Reporting** → **Asset** → **By Storage Subsystem**, as shown in Figure 7-22.

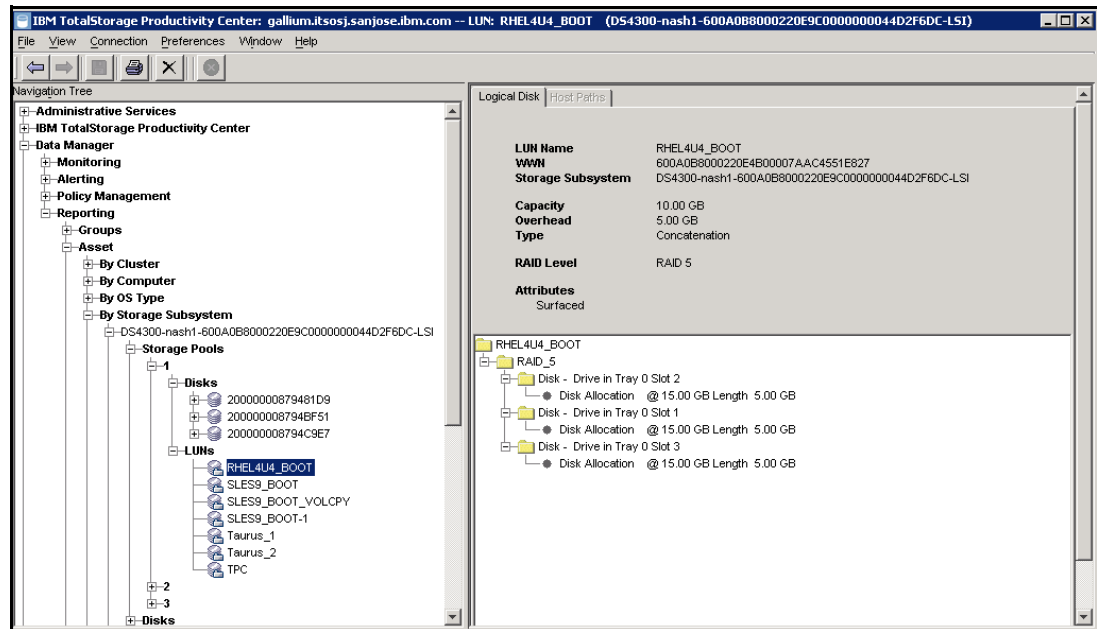


Figure 7-22 Detailed information of storage subsystem

7.2.4 Creating a Performance Monitor job

Before storage subsystem performance can be monitored, a Performance Monitor job has to be created. The job will collect performance data for the storage subsystem.

To create a Performance Monitor for a storage subsystem:

1. Expand **Disk Manager** → **Monitoring**, then right-click **Subsystem Performance Monitor** and click **Create Subsystem Performance Monitor**. In the storage subsystem tab in the right pane, select the storage subsystem to be monitored and move it to the Selected Subsystem panel, as shown in Figure 7-23.

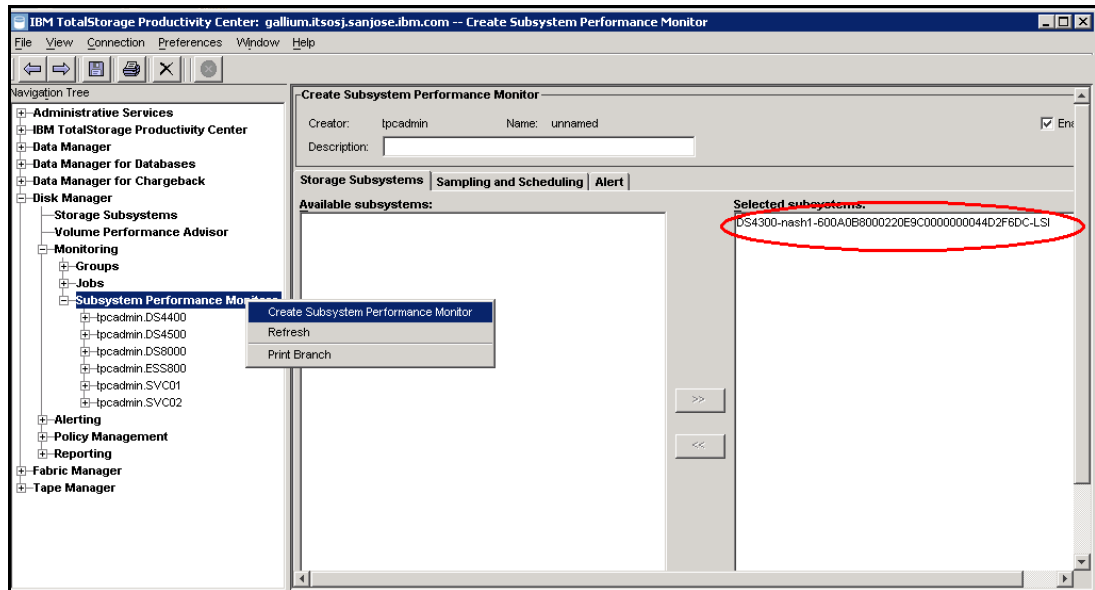


Figure 7-23 Create Performance Monitor job

2. Select the **Sampling and Scheduling** tab to configure the interval length, duration, and scheduling of the Performance Monitor job to be submitted, as shown in Figure 7-24.

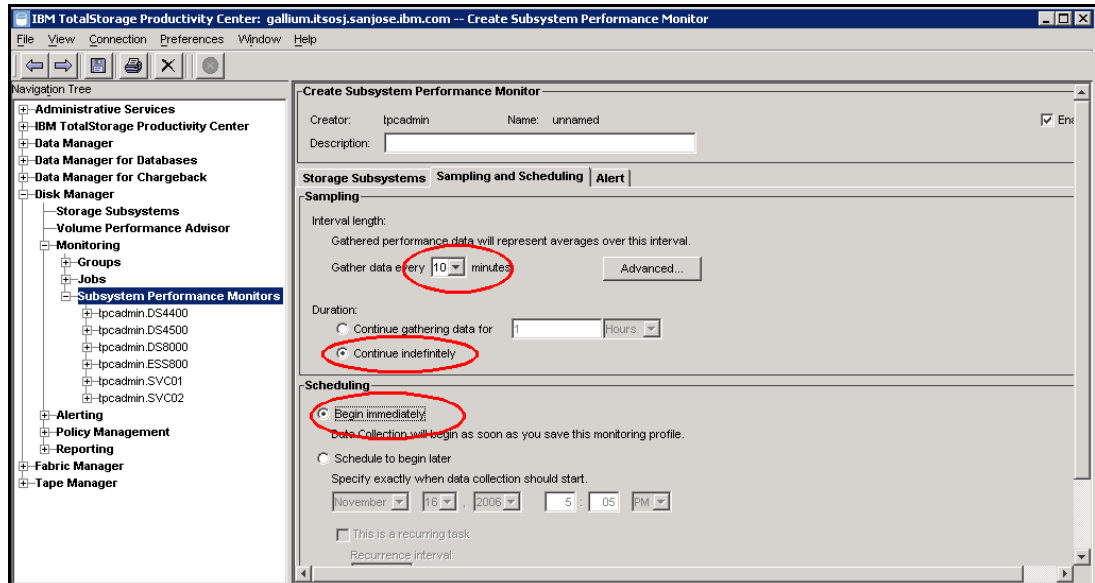


Figure 7-24 Sampling and scheduling

3. Go to the Alert tab to select what actions should be triggered should the monitor fail, as shown in Figure 7-25.

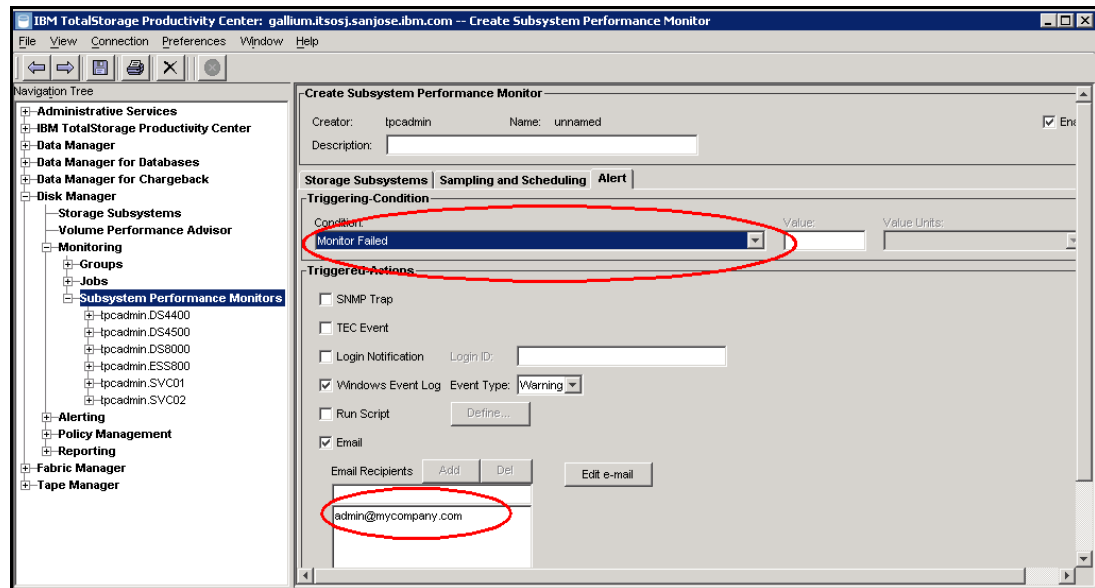


Figure 7-25 Alert tab

4. Select **File** → **Save** to save the job and specify the job name when prompted, as shown in Figure 7-26.

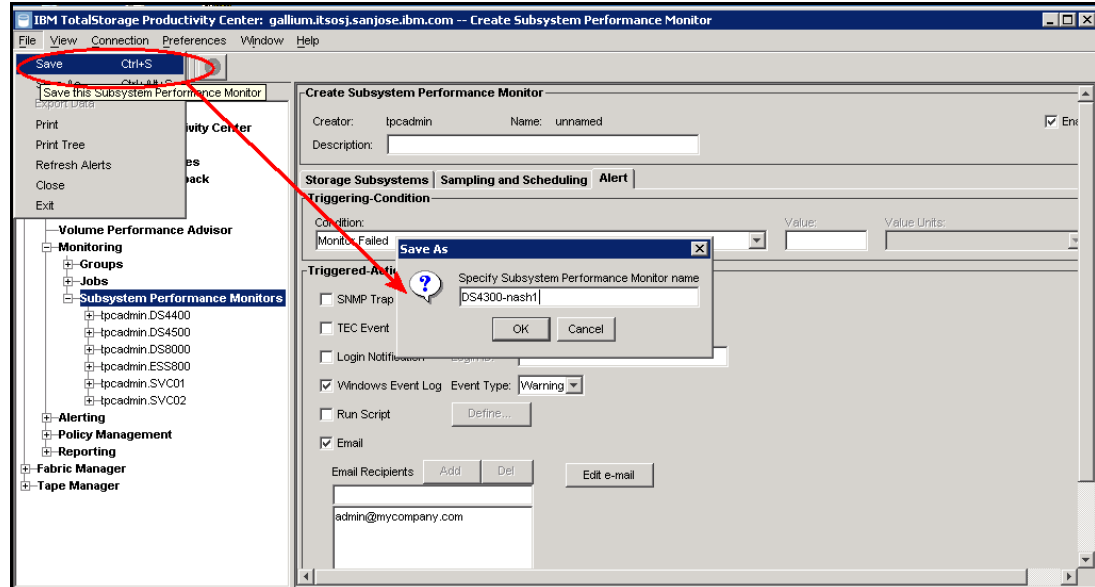


Figure 7-26 Save Performance Monitor job

Your Performance Monitor job should have been submitted now. To verify the job status, expand the job name under **Disk Manager** → **Monitoring** → **Subsystem Performance Monitor**, as shown in Figure 7-27.

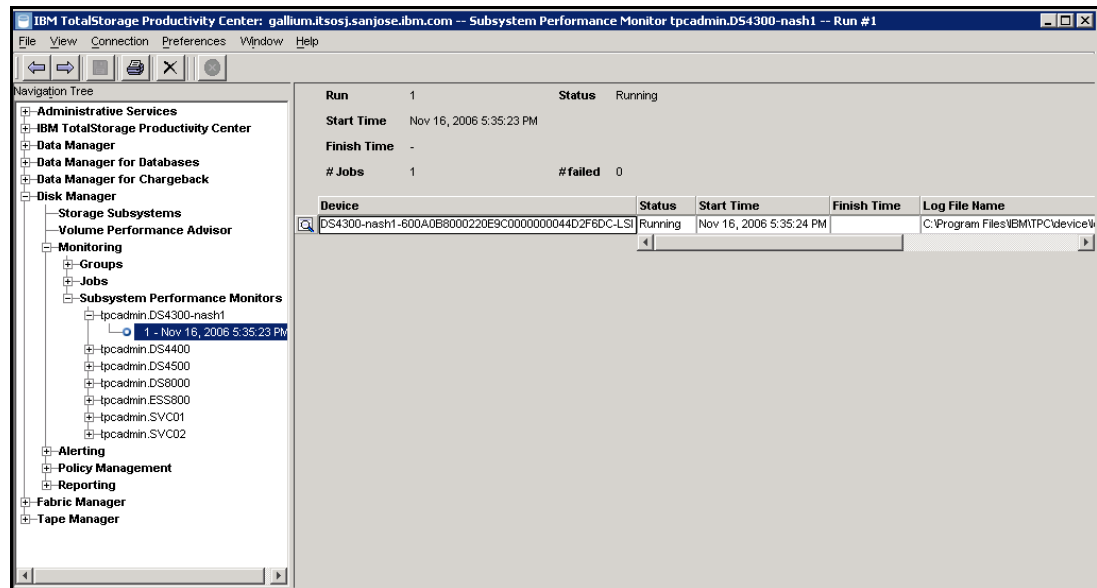


Figure 7-27 Verify Performance Monitor job

7.3 TPC reporting for DS4000

As we have discussed in Chapter 4, “DS4000 performance tuning” on page 133, the storage subsystem performance is only one piece of the performance puzzle and is impacted by many factors, including type of workload created by the application (transaction or throughput based), the system hosting the application, and performance of other SAN components.

To resolve performance issues with a storage subsystem, the storage administrator must understand the dynamics of the various performance characteristics.

A good approach is to look at current and historical data for the configuration and workloads that are not getting complaints from users, and do a trending from this performance base. In the event of performance problems, look for the changes in workloads and configuration that can cause them. TPC can help you accomplish just that by collecting performance data over time and generating performance reports.

7.3.1 DS4000 performance report

TotalStorage Productivity Center provides two types of reports: predefined reports and custom reports. All the predefined reports for storage subsystems are performance related and can be found under **IBM TotalStorage Productivity Center** → **My Reports** → **System Reports** → **Disk**.

For other non performance-related reports, you can generate custom reports under **Disk Manager** → **Reporting**, which can also be used to generate performance-related reports.

In TPC, DS4000 performance data can be displayed in a table or graphical report. It can display recent or historical performance data, which can be exported into a file for offline

analysis. Using custom reports, performance report for DS4000 can be created by storage subsystem and by volume.

For DS4000, there are several metrics/parameters can be monitored by TPC, as shown in Table 7-1.

Table 7-1 DS4000 performance metrics

Metrics	Description
Read I/O rate Write I/O rate Total I/O rate	Average number of I/O operations per second for both sequential and non-sequential read/write/total operations for a particular component over a time interval.
Read cache hits Write cache hits Total cache hits	Percentage of cache hits for non-sequential read/write/total operations for a particular component over a time interval.
Read Data Rate Write Data Rate Total Data Rate	Average number of megabytes (10 ⁶ bytes) per second that were transferred for read/write/total operations for a particular component over a time interval.
Read transfer size Write transfer size Overall transfer size	Average number of KB per I/O for read/write/total operations for a particular component over a time interval.
Total port I/O rate Total port data rate	Average number of I/O operations per second for send and receive operations for a particular port over a time interval. Average number of megabytes (10 ⁶ bytes) per second that were transferred for send and receive operations for a particular port over a time interval.
Total port transfer size	Average number of KB transferred per I/O by a particular port over a time interval.

7.3.2 Generating reports

This section provides several examples of how to generate reports using predefined and custom reports.

Example 1: DS4000 predefined performance report

In this example, we compare the overall read I/O rate of DS4300 and DS8000 storage subsystems. Remember that the results are affected by many factors including the workload on each storage subsystem. This example only illustrates the steps required to compare the performance of storage subsystems within a TPC environment.

To generate the report using TPC predefined reports:

1. Expand **IBM TotalStorage Productivity Center** → **My Reports** → **System Reports** → **Disk** to see a list of system supplied performance reports.

- Click one of the listed predefined reports (note that not all predefined performance reports support DS4000 at the time of writing). In the example, shown in Figure 7-28 we click **Subsystem Performance** report to display a list of all storage subsystems monitored by TPC with their performance metrics values.

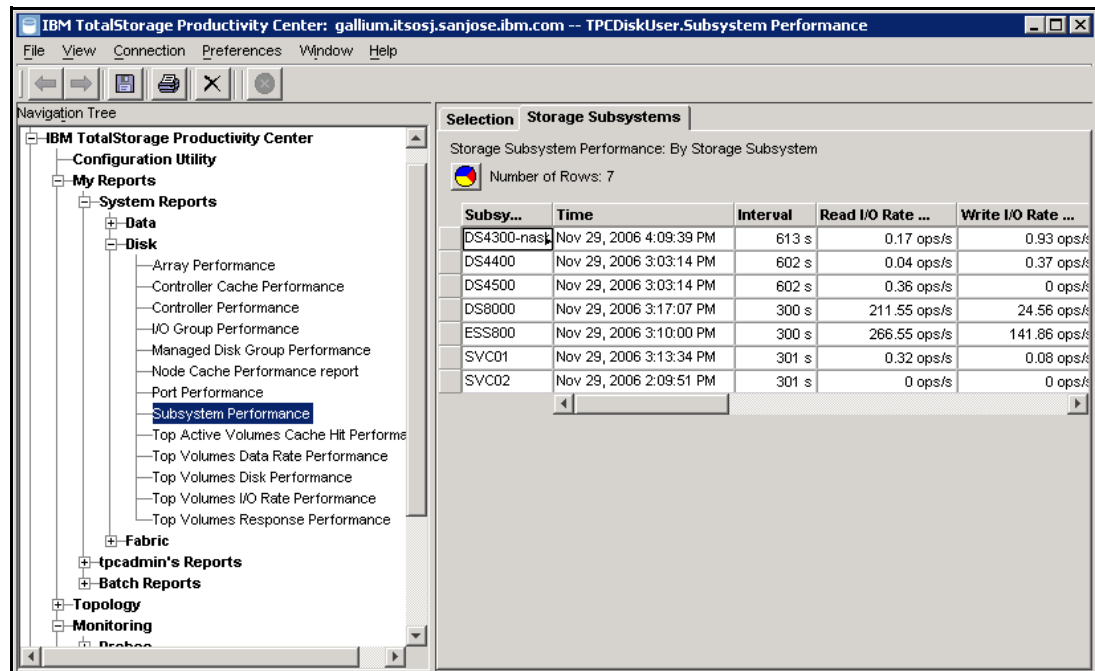


Figure 7-28 Subsystem Performance custom report

You can click the chart icon to generate a graph or click the **Selection** tab to modify this predefined report and regenerate it.

- In our example, we click the **Selection** tab because we want to generate a report that compares the overall read I/O rate of the DS4300 and the DS8000. Move to the Included column the performance metrics (Time, Interval and Overall Read I/O rate) you want to include in the report, as shown in Figure 7-29.

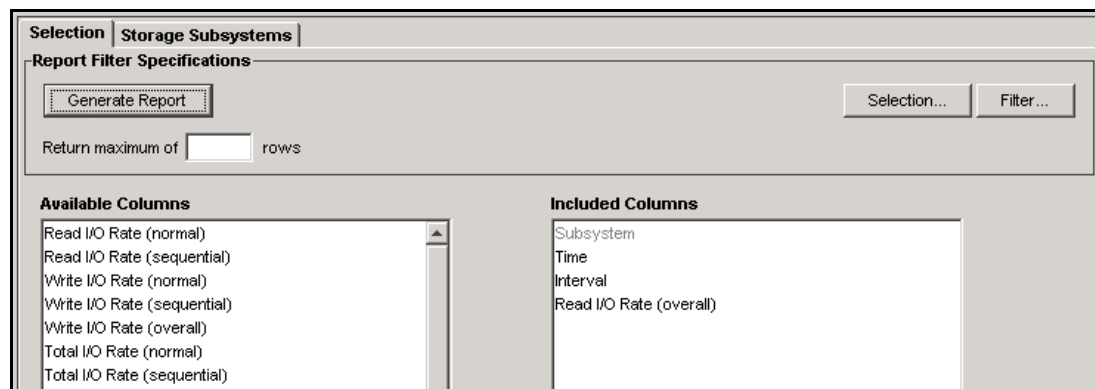


Figure 7-29 Selection of performance metrics

4. Click **Selection** on the Selection tab to select the storage subsystems for which you want to generate the report. In our case, we select **DS4300** and **DS8000**, as shown in Figure 7-30.

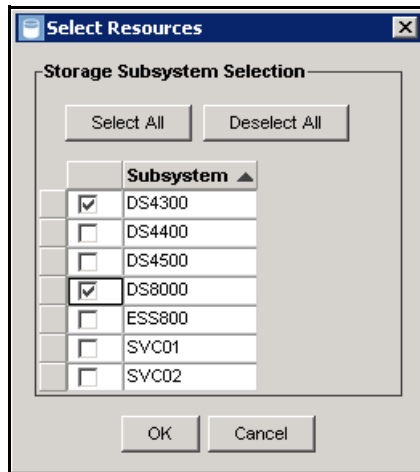


Figure 7-30 Storage subsystem selection

5. Click **Generate Report** (still on the Selection tab) and you get the result shown in Figure 7-31.

The 'Selection' tab is active, showing 'Storage Subsystems'. Below the tab, it says 'Storage Subsystem Performance: By Storage Subsystem' and 'Number of Rows: 2'. A table displays the performance data for the selected subsystems.

Subsystem	Time	Interval	Read I/O Rate (overall)
DS4300	Dec 5, 2006 9:39:45 AM	612 s	149.58 ops/s
DS8000	Dec 5, 2006 8:56:31 AM	300 s	225.33 ops/s

Figure 7-31 Generated report based on selection

6. We can now create a chart based on the generated report. Click the chart icon and select the chart type and rows of the report that you want to be included in the chart as well, as the performance metric to be charted (see Figure 7-32).

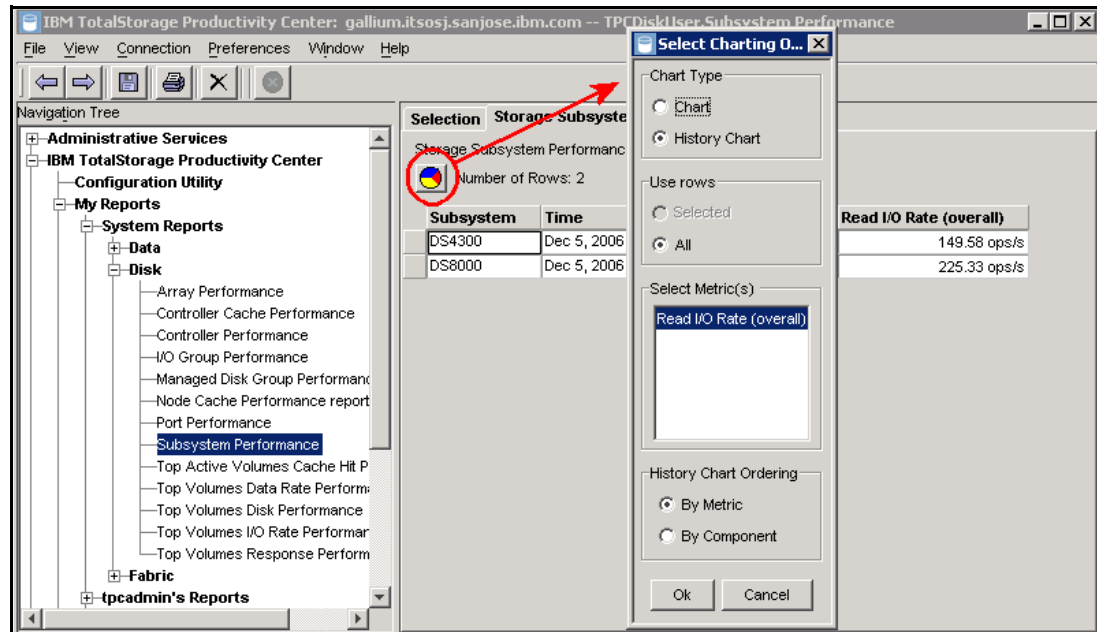


Figure 7-32 Select Charting Option

If we select the chart type as Chart in the Select Charting Option, the system generates a bar diagram for the selected storage subsystems, representing the average value of the selected metrics over the complete period for which the samples are available.

Here we select **History Chart** to generate the chart as shown in Figure 7-33.

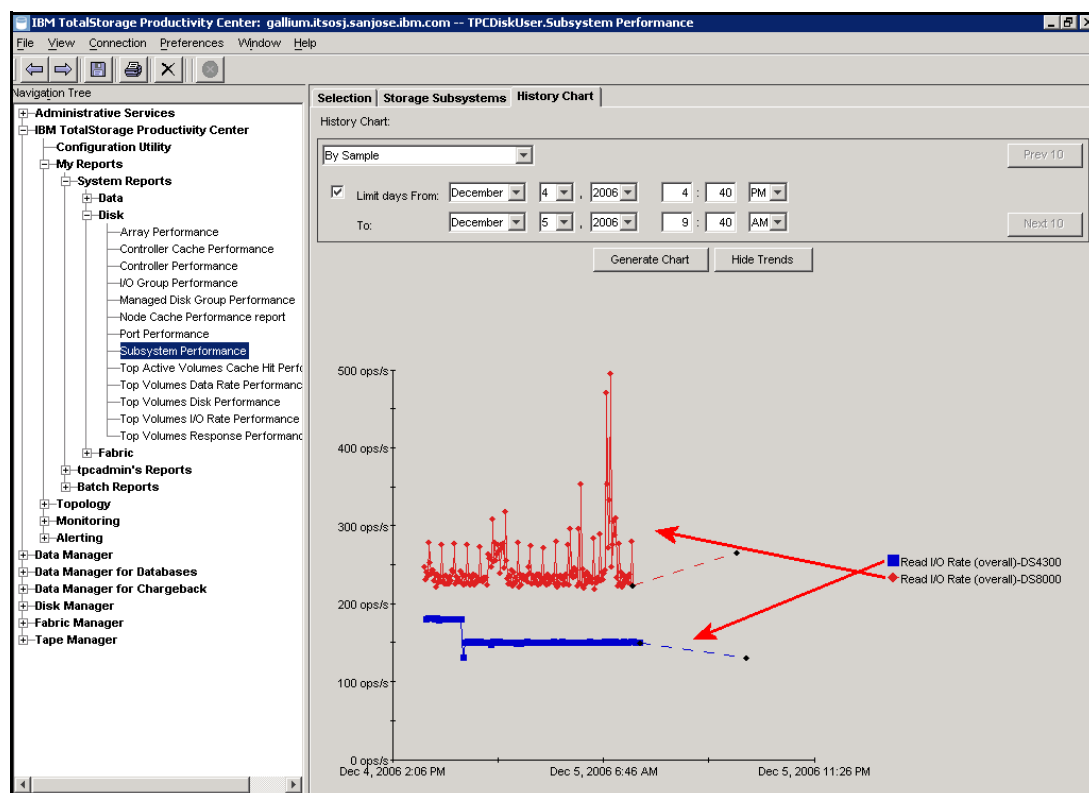


Figure 7-33 Result of historical chart

We can now graphically compare and analyze the storage system's performance for the selected metric. In this example, we see that overall read I/O rate of DS8000 is higher than for the DS4300. You can generate similar reports to compare other storage subsystems and look at different performance metrics.

Note that the generated chart also shows performance trend lines which are useful to foresee performance bottlenecks and determine appropriate measures to prevent them from occurring.

Example 2: DS4000 custom performance report

In this example we measure and compare in a TPC custom report, the overall read and write I/O rate of two volumes (logical drives) in a DS4300 storage subsystem.

Follow these steps to generate this particular custom performance report in TPC:

1. Expand **Disk Manager** → **Reporting** → **Storage Subsystem Performance**, then click **By Volume**.

- On the Selection tab, move all performance metrics in the Included Columns into the Available Columns, except for read I/O rate (overall) and write I/O rate (overall) metrics. Check the Display historic performance data check box and select the start and end dates to generate a historical report, as shown in Figure 7-34.

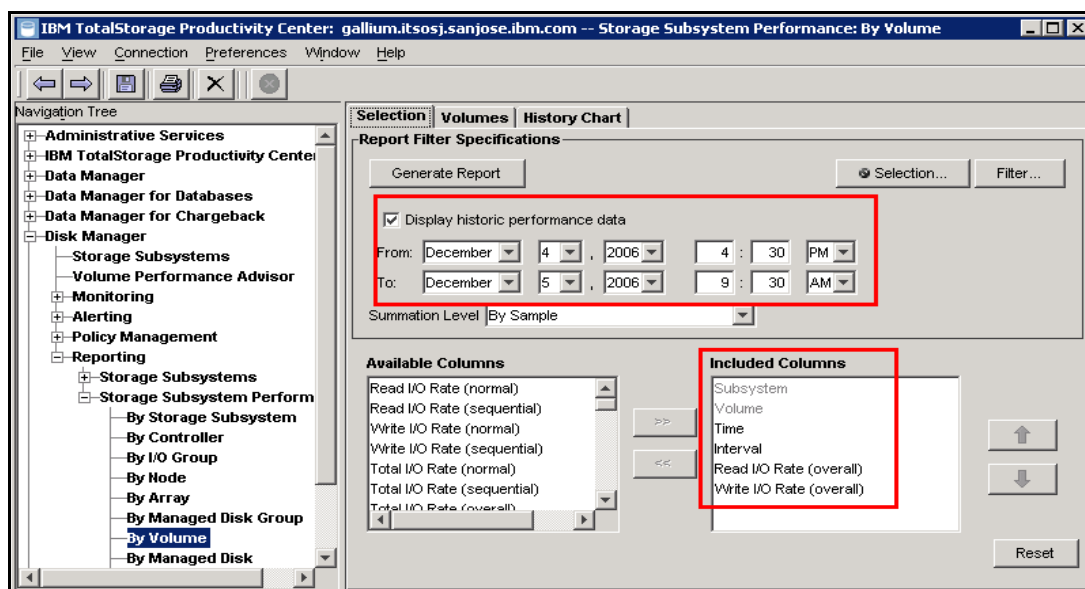


Figure 7-34 Selecting performance metrics and historical data

- Next select the storage subsystem to report on. Click **Selection** to bring up the Select Resources window. In our example, we select two volumes from a DS4300 storage subsystem, as shown in Figure 7-35.

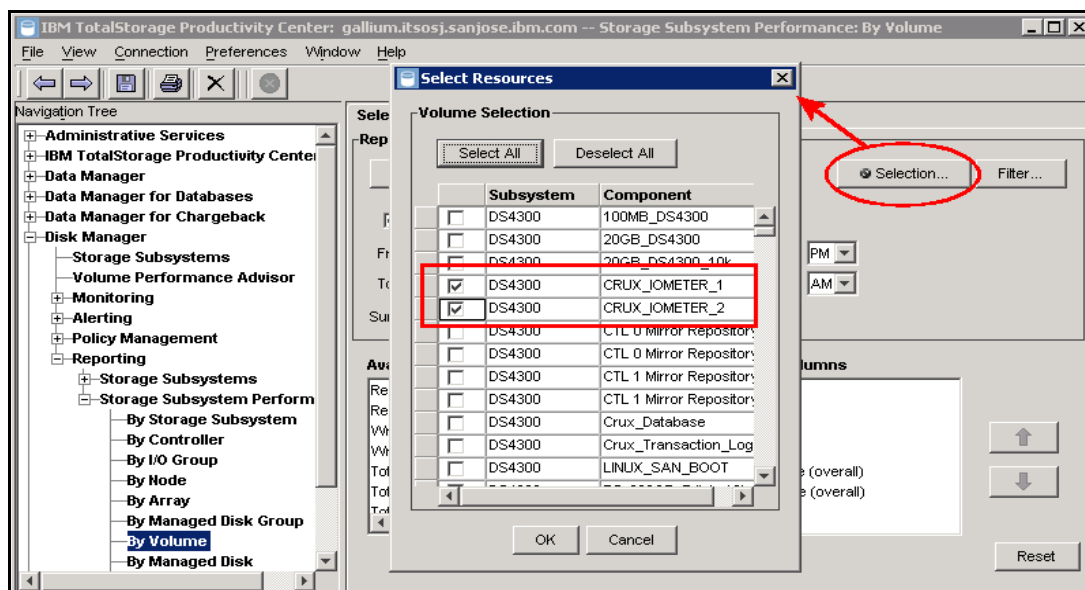


Figure 7-35 Select volumes from Volume Selection

- Click **Generate Report** to start the query. The result is displayed as shown in Figure 7-36.

Subsystem	Volume	Time	Interval	Read I/O Rate (overall)	Write I/O Rate (overall)
DS4300	CRUX_JOMETER_1	Dec 4, 2006 4:39:40 PM	612 s	125.52 ops/s	31.14 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 4:39:40 PM	612 s	54.87 ops/s	218.82 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 4:49:52 PM	612 s	125.1 ops/s	31.67 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 4:49:52 PM	612 s	54.69 ops/s	219.7 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 5:00:04 PM	612 s	125.56 ops/s	31.12 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 5:00:04 PM	612 s	55.13 ops/s	220.31 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 5:10:16 PM	613 s	125.28 ops/s	31.09 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 5:10:16 PM	613 s	54.92 ops/s	220.36 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 5:20:29 PM	612 s	125.27 ops/s	30.6 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 5:20:29 PM	612 s	54.87 ops/s	220.91 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 5:30:41 PM	613 s	124.41 ops/s	31.41 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 5:30:41 PM	613 s	54.97 ops/s	220.53 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 5:40:54 PM	612 s	125.48 ops/s	31.08 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 5:40:54 PM	612 s	54.84 ops/s	220.69 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 5:51:06 PM	611 s	123.2 ops/s	30.57 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 5:51:06 PM	611 s	54.75 ops/s	219.67 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 6:01:17 PM	613 s	124.62 ops/s	31.41 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 6:01:17 PM	613 s	55.25 ops/s	219.34 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 6:11:30 PM	612 s	124.27 ops/s	31.41 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 6:11:30 PM	612 s	55.13 ops/s	220.19 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 6:21:42 PM	611 s	124.48 ops/s	31.37 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 6:21:42 PM	611 s	54.48 ops/s	221.3 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 6:31:53 PM	613 s	124.36 ops/s	31.21 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 6:31:53 PM	613 s	55 ops/s	219.86 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 6:42:06 PM	611 s	124.38 ops/s	31.39 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 6:42:06 PM	611 s	55.19 ops/s	218.93 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 6:52:17 PM	611 s	124.86 ops/s	30.98 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 6:52:17 PM	611 s	54.71 ops/s	219.47 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 7:02:28 PM	612 s	124.75 ops/s	31.04 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 7:02:28 PM	612 s	54.58 ops/s	219.29 ops/s
DS4300	CRUX_JOMETER_1	Dec 4, 2006 7:12:40 PM	612 s	124.48 ops/s	31.15 ops/s
DS4300	CRUX_JOMETER_2	Dec 4, 2006 7:12:40 PM	612 s	54.92 ops/s	219.87 ops/s

Figure 7-36 Query result

To export the query result into a file for offline analysis, select **File** → **Export Data** from the menu bar.

- Now we generate a chart. Select all of the results and click the chart icon, then select **History Chart** and all of the metrics, as shown in Figure 7-37.

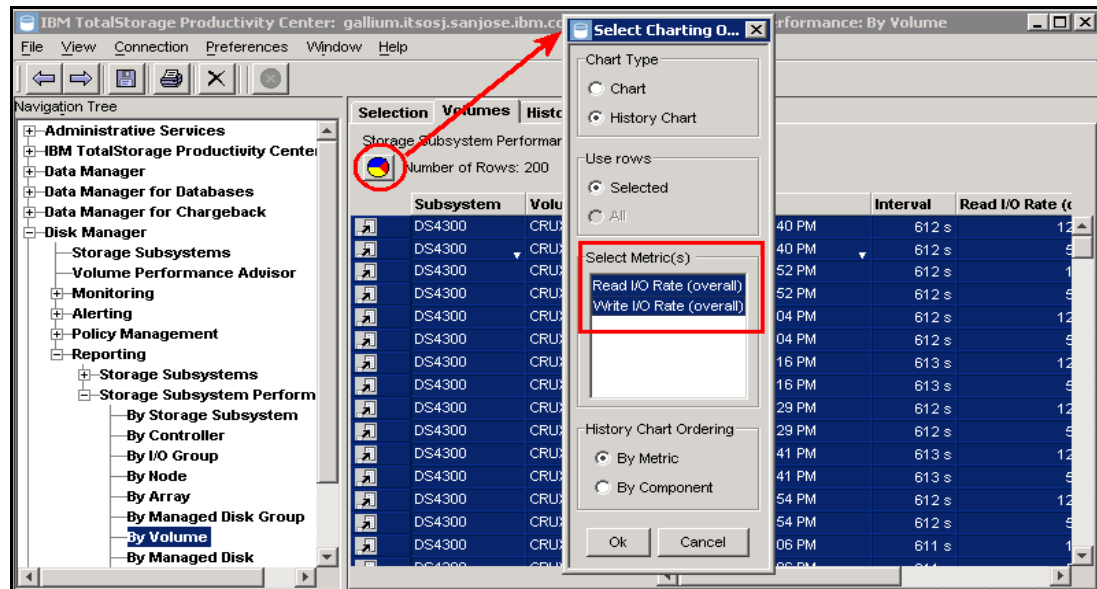


Figure 7-37 Select Charting window

The chart is displayed as shown in Figure 7-38.

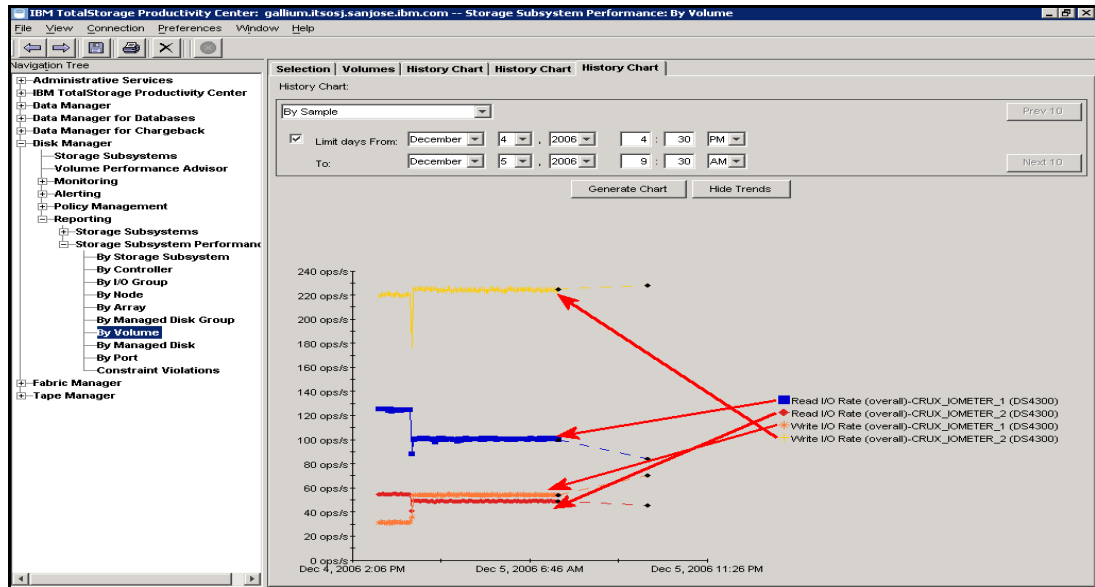


Figure 7-38 Graphical report result

From the chart result, we see that volume CRUX_IOMETER_1 has a higher read I/O rate but lower write I/O rate compared to volume CRUX_IOMETER_2, which means that volume CRUX_IOMETER_1 has a more read extensive workload, and volume CRUX_IOMETER_2 has more of a write extensive workload. As we discussed earlier in this book, this type of information can be used, for example, to do performance tuning from the application, operating system, or the storage subsystem side.

Example 3: DS4000 volume to HBA assignment report

TPC reporting is also able to generate non-performance reports. The following example shows steps required to generate a *Volume to HBA assignment* report for a DS4300 and a DS4400 that were part of our TPC environment. Since there are no predefined non performance-related reports, we need to generate a custom report.

Follow these steps to generate a Volume to HBA Assignment report By Storage Subsystem in TPC:

1. Expand **Disk Manager** → **Reporting** → **Storage Subsystems** → **Volume to HBA Assignment** → **By Storage Subsystem**. Remove unnecessary information from the Included Columns into Available Columns, and click **Selection** to select the storage subsystems for which you want to create this report. In our example, we only include Volume WWN, HBA Port WWN, and SMI-S Host Alias, and selected one DS4300 and one DS4400 among the list of storage subsystems known to our TPC environment. This is shown in Figure 7-39.

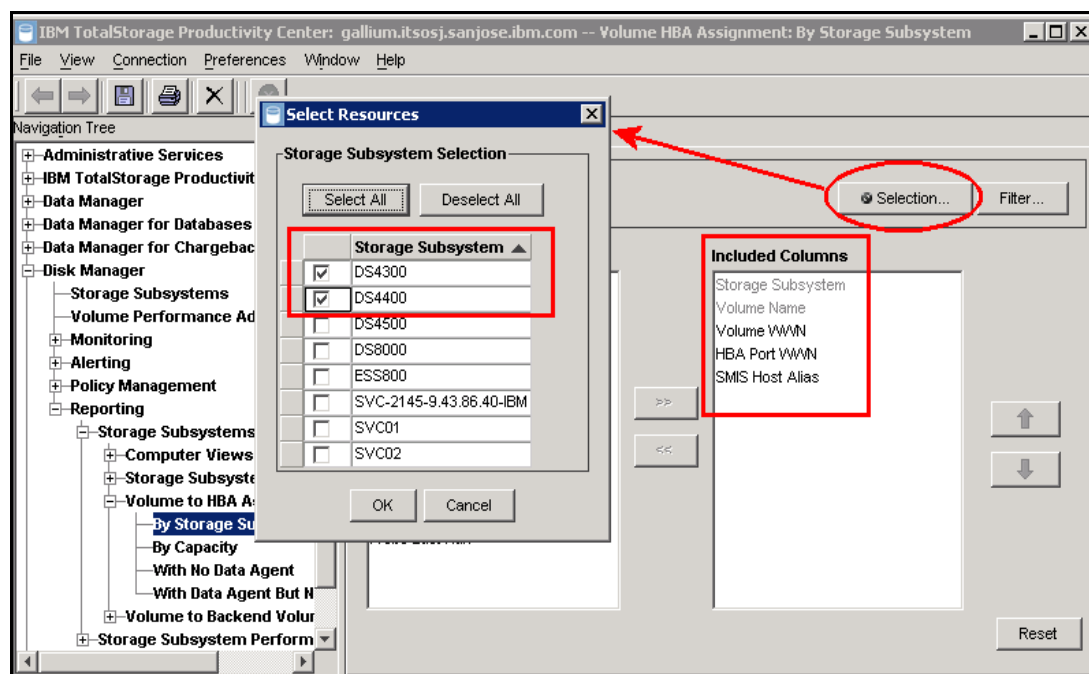


Figure 7-39 Volume to HBA assignment, select storage subsystems

2. Click **Generate Report** to generate the report, or optionally click **Filter** for a more specific query. A report example is shown in Figure 7-40.

IBM TotalStorage Productivity Center: gallium.itsosj.sanjose.ibm.com -- Volume HBA Assignment: By Storage Subsystem

File View Connection Preferences Window Help

Navigation Tree

- Administrative Services
- IBM TotalStorage Productivity Center
 - Data Manager
 - Data Manager for Databases
 - Data Manager for Chargeback
 - Disk Manager
 - Storage Subsystems
 - Volume Performance Advisor
 - Monitoring
 - Alerting
 - Policy Management
 - Reporting
 - Storage Subsystems
 - Computer Views
 - Storage Subsystem View
 - Volume to HBA Assignment
 - By Storage Subsystem (Selected)
 - By Capacity
 - With No Data Agent
 - With Data Agent But N
 - Volume to Backend Volur
 - Storage Subsystem Perform

Selection Storage Subsystems

Volume HBA Assignment: By Storage Subsystem

Number of Rows: 32

Storage S...	Volume Name	Volume WWN	HBA Port WWN	SMIS Host Alias
TOTAL =>				
DS4400	JA_LUN_2	600A0B80000CBDA	210000E08B89B9CU	SENEGAL_A
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	210000E08B18D48F	TONGA_B
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	210000E08B18FF8A	TONGA_A
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	210000E08B18D48F	TONGA_B
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	210000E08B18FF8A	TONGA_A
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	10000000C94C8C1C	ATLANTIC_FCS1
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	10000000C932A80A	ATLANTIC_FCS0
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	10000000C94C8C1C	ATLANTIC_FCS1
DS4400	KE_4400_4G_TESTVQ	600A0B80000CBCE	10000000C932A80A	ATLANTIC_FCS0
DS4300	LINUX_SAN_BOOT	600A0B8000220E9C	210000E08B0CC823	Taurus_1
DS4300	LINUX_SAN_BOOT	600A0B8000220E9C	210000E08B0C4827	Taurus_0
DS4300	R5_200GB_7disk_10k	600A0B8000220E4B	210100E08B280242	Crux_2
DS4300	R5_200GB_7disk_10k	600A0B8000220E4B	210000E08B080242	Crux_1
DS4400	TPC_REPORBE_TEST	600A0B80000CBCE	210000E08B06C50B	Colorado_HBA1
DS4400	TPC_REPROBE_2	600A0B80000CBDA	210000E08B06C50B	Colorado_HBA1

Figure 7-40 Volume to HBA assignment report

The report shows volume name, volume WWN, HBA port WWN, and SMI-S host alias for the selected storage subsystems. For a storage administrator, this report simplifies the tasks required to list volumes to servers assignments, as long as the host aliases reflect the server name.

You can click the magnifier icon to the left of each row to see more detailed information about the volume, including the RAID level of the array. In our example, we click the magnifier icon on the left of volume LINUX_SAN_BOOT, and the result is shown in Figure 7-41.

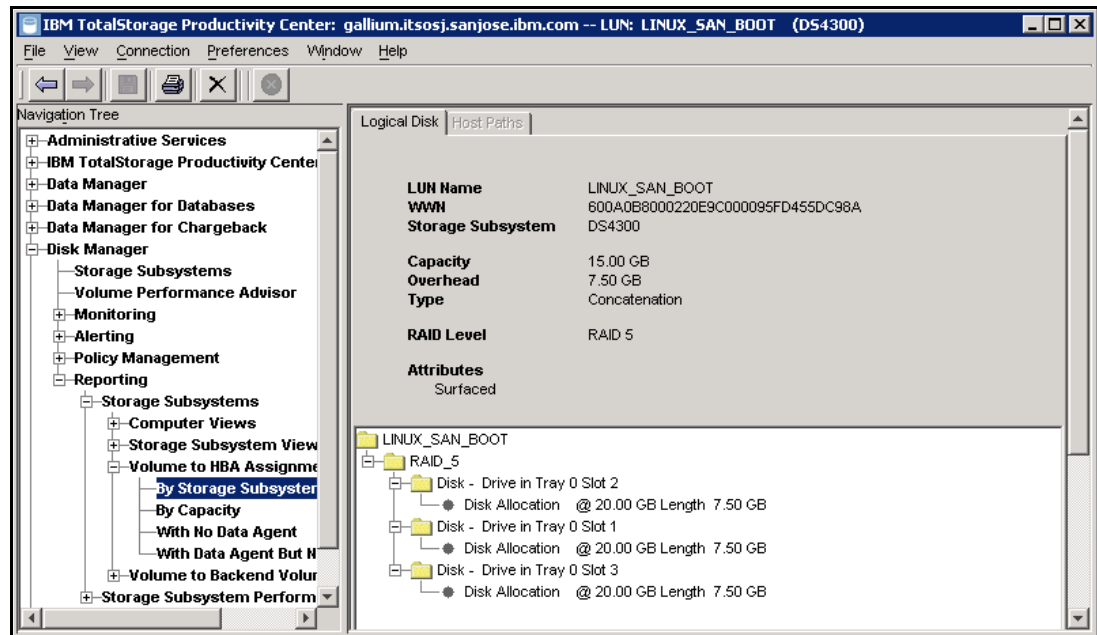


Figure 7-41 More detailed volume information



Disk Magic

This chapter describes and illustrates the use of Disk Magic, developed by the company IntelliMagic.

Disk Magic is a tool for sizing and modelling disk systems for various servers. It performs accurate performance and analysis planning for IBM DS4000, DS6000, DS8000 Storage Servers, SAN Volume Controller (SVC), and other systems. Disk Magic allows for capacity planning and performance analysis work to be undertaken prior to the purchase of new equipment.

8.1 Disk Magic overview

Disk Magic is a flexible and powerful tool to model the performance and capacity requirements of your DS4000.

Disk Magic helps you evaluate important considerations such as which disk type to use, which RAID levels are appropriate for your applications, the cache size requirements, and the utilization level of HBAs. Disk Magic shows current and expected response times, utilization levels, and throughput limits for your own installation's I/O load and server configuration.

In a Disk Magic study, you start by collecting data about your current server and storage environment. The collected data is entered (automated or manual input) in Disk Magic and is used by Disk Magic to establish a baseline. You must have Disk Magic establish a base after you have completed the initial entering of the configuration and workload data.

With the data, Disk Magic can create a simulated model of your environment. Disk Magic allows for *what-if analysis* on items such as disk upgrades, moving workloads from one disk subsystem to another, or using another RAID level. Disk Magic keeps a history of all of the configuration changes made. This allows for the restoration of the project to a known stage.

Disk Magic can produce reports and graphs showing utilization figures on CPU and disk loads. The Report function creates a report of the current step in the model, as well as the base from which this model was derived. The report can be used to verify that all data was entered correctly. It also serves as a summary of the model results. The report lists all model parameters, including the ones that Disk Magic generates internally. You can also produce various graph types.

Disk magic runs in a Microsoft Windows environment. It does not require a connection with the storage subsystem to perform any of its functions.

8.2 Information required for DS4000 modeling with Disk Magic

The importance of gathering accurate performance data for the servers and the disk subsystem cannot be stressed enough. This information should reflect the environment as a good base line in itself.

To collect the data, we recommend identifying one or two peak periods with a workload that are critical to your installation. Indeed, since you need to make sure that the disk subsystem will be able to handle peak workloads, it makes sense to feed Disk Magic with peak workload data. However, in the selection of peak periods, you should avoid extremes (that is, an I/O workload that is unusually high or unusually low). Make sure as well that the collection time is bounded to the peak period, as to not skew the statistics by periods of low activity within the same interval. Finally, make sure that your statistics are complete and include all the servers that use the disk subsystem.

You can automate the gathering of data by using *iostat* (in a Linux, AIX, or UNIX environment) or *perfmon* log files (in a Windows environment) for all of the servers that use the disk subsystem. We briefly describe below how to use *perfmon* and *iostat* to collect statistics for Disk Magic.

If these files are unavailable, you can manually enter the following: blocksize, read/write ratio, read hit ratio, and I/O rate. In fact, except for the I/O rate, Disk Magic will provide defaults for all other parameters when you do not know or cannot easily collect this data.

Perfmon and Disk Magic

Windows Performance Monitor (or perfmon) is a useful tool for gathering performance statistics for Disk Magic. Disk Magic can automatically process the performance data in a Windows perfmon file.

To start the Performance Monitor and set up the logging, go to the task bar and click **Start** → **Control Panel** → **Administrative Tools** → **Performance**. The Performance window opens, as shown in Figure 8-1.

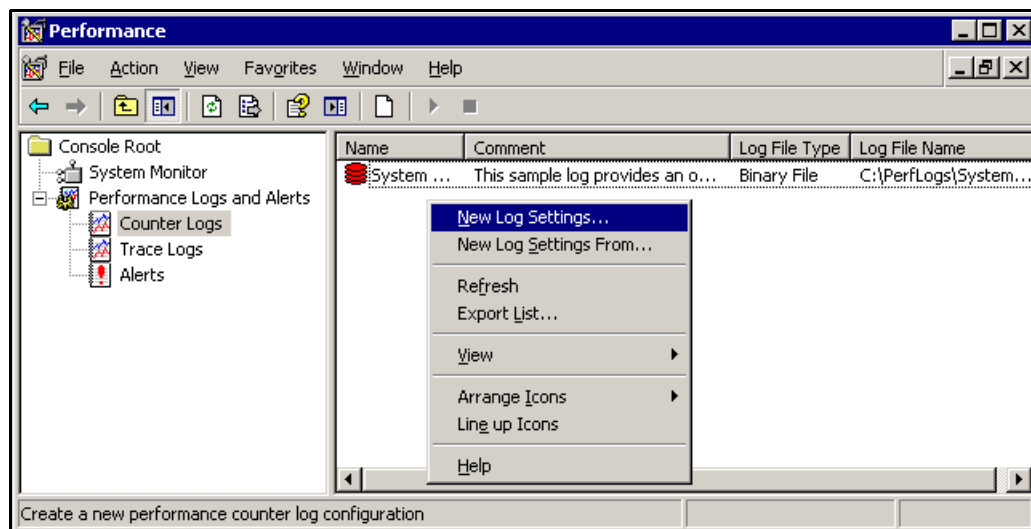


Figure 8-1 Create new log file

In the left pane select **Performance Logs and Alerts**. In the right window pane, right-click **Counter Logs** and select **New Log Settings**. Enter an appropriate name and click **OK**.

The PerformStats window, shown in Figure 8-2, opens to let you specify the log file properties. On the General tab you must define the logging interval. We recommend that you specify a logging interval that is enough to ensure that detailed accurate information is to be gathered.

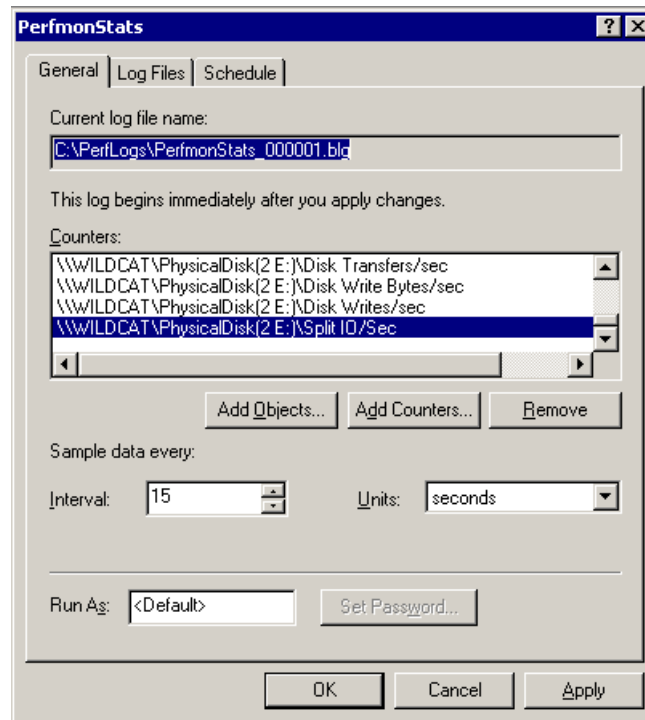


Figure 8-2 Add counters and set interval

From the General tab, click **Add Counters**, which opens the Add Counters dialog, as shown in Figure 8-2.

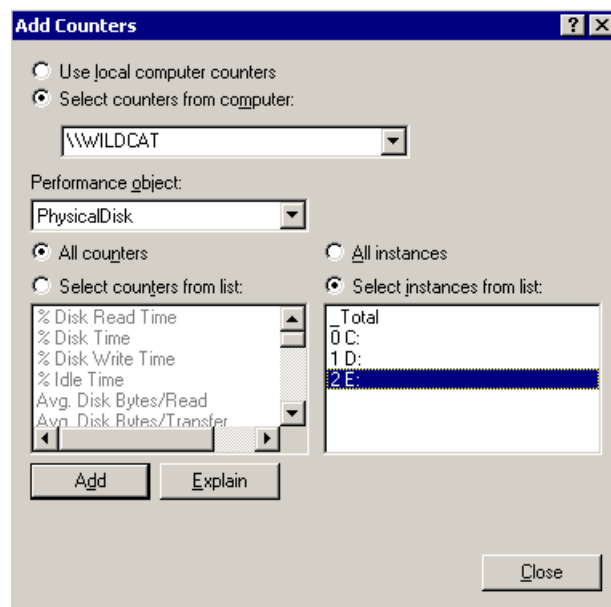


Figure 8-3 Add all counters for PhysicalDisk on the LUN

Here you select **PhysicalDisk** from the Performance object pull-down and click the **All counters** radio button.

On the right side under instances, you can select the physical disks for which the performance must be logged. These are only the disks that are connected to the disk subsystem (such as DS4000). This should normally not include the C drive, which is usually the server internal disk drive. You can select more than one disk, but only the SAN-based disks should be selected. You may also include the total, but it will be disregarded by Disk Magic since it performs its own total for the selected disks. Click **Close** to return to the PerfmonStats window.

Next, specify the logging file format by clicking the **Log Files** tab and selecting **Text File - CSV**. This will produce a comma delimited file. Take note of the location where these files are stored and their names.

Finally, you should define at what date and time the logging is to be started and stopped. This can be done on the Scheduling tab (Figure 8-4). Ensure that perfmon is set to gather I/O load statistics on each server that accesses the disk subsystem to be analyzed. Also, make sure to start and stop the perfmon procedure at the same time on each Windows server.

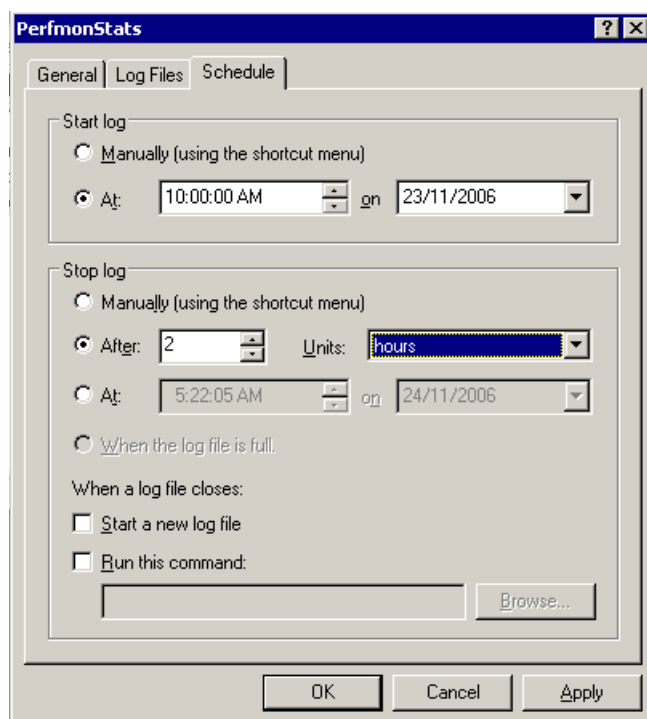


Figure 8-4 Set schedule for how long to log for

To use this perfmon file in Disk Magic, start Disk Magic and select **Windows Perfmon File** from the Open Existing File area, as shown in Figure 8-5. This will create a new project with one server and one disk system. Later, you can add more perfmon data from other servers.

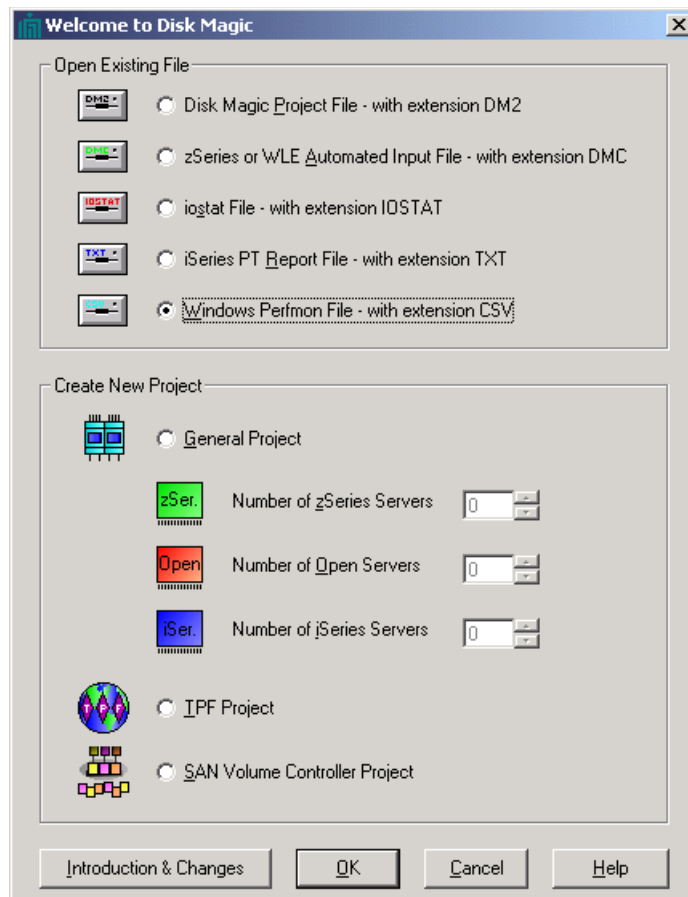


Figure 8-5 Importing Windows performance data into Disk Magic

Iostat and Disk Magic

For the Linux and UNIX environments, performance data can be captured using `iostat`. The resulting output file (report) can be processed for use with Disk Magic.

Automated input for UNIX and Linux is supported for:

- ▶ AIX
- ▶ HP UNIX
- ▶ Sun Solaris
- ▶ Linux (Redhat and SUSE)

The `iostat` command produces I/O load statistics, including MBs read and write.

Cache statistics or information about the type of disk subsystem is not included in the `iostat` report. The cache size must be entered manually into Disk Magic before Disk Magic can create the base line.

`iostat` reports do not include information that would allow Disk Magic to identify the entity of a disk subsystem, and therefore it is not possible for Disk Magic to separate the I/O load statistics by disk subsystem. Consequently, you should not create an `iostat` file that covers more than one disk subsystem.

You should make sure to run the `iostat` procedure on each server that accesses the disk subsystem to be analyzed. Make sure as well to start and stop the `iostat` procedure at the same time on each server.

The `iostat` automated input process is performed as follows:

1. Enter the command depending upon the operating system:

- For AIX

```
iostat i n > servername.iostat
```

- For HP UNIX

```
iostat i n > servername.iostat
```

- For Sun Solaris

```
iostat -xtc i n > servername.iostat or iostat -xnp il n > systemname.iostat
```

- For Linux

```
iostat -xk i n > servername.iostat
```

Where:

- *i* is the interval in seconds.

- *n* is the number of intervals. This should always be greater than 1.

- *servername.iostat* is the output file. Its file type must always be `iostat`.

A sample Linux command could be:

```
iostat -xk 600 10 > linuxserver.iostat.
```

2. When the data collection has finished, edit `servername.iostat` to insert two lines of header information:

```
os iostat system=servername interval=i  
ddmonyyyy hh:mm
```

- *os* is the operating system on which the `iostat` performance data was collected, such as AIX, HP-UX, Sun Solaris, or Linux. It does not matter whether you use upper-case or lower-case, or a mix. Disk Magic will also accept the following permutations of the UNIX / Linux names: `hp ux`, `hpux`, `sunsolaris`, `sun-solaris`, `solaris`, `sun`, `redhat`, and `suse`.

- *i* should be the same interval as in the original `iostat` command.

- *dd*, *yyyy*, *hh*, and *mm* are numerical, and 'mon' is alphabetic. They should reflect date and time of the first interval in the `iostat` gathering period.

For example, the line to add for an `iostat` Linux file could be:

```
redhat iostat system=linuxserver interval=600  
13nov2006 10:10
```

3. To use the resulting iostat file in Disk Magic, start Disk Magic and select **iostat file** from the Open Existing File area, as shown in Figure 8-6. This will create a new project with one server and one disk system. Later, you can add more iostat data from other servers.

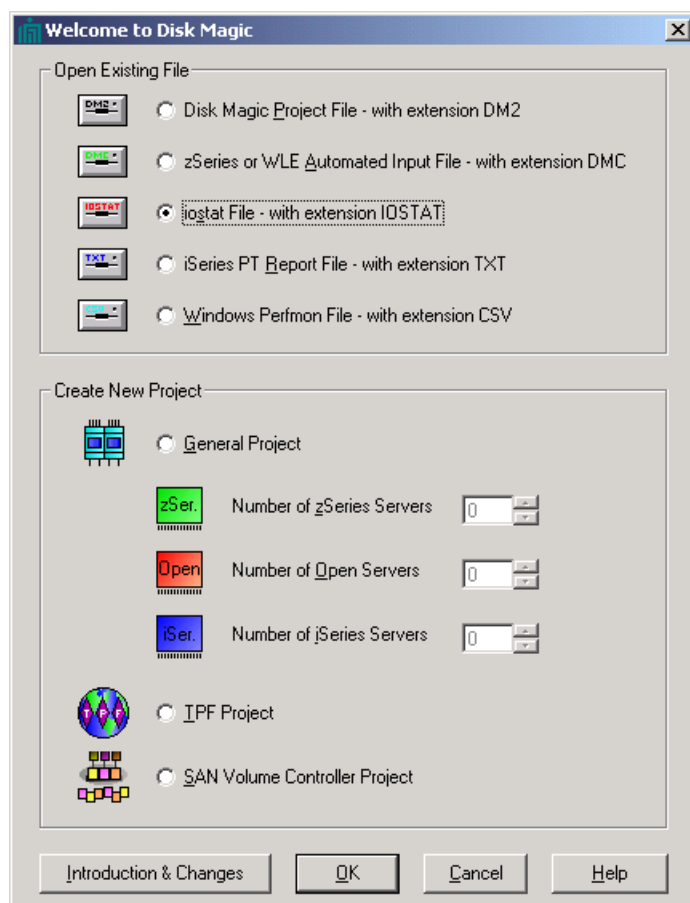


Figure 8-6 Importing an iostat file into Disk Magic

8.3 Disk Magic configuration example

To illustrate the use of Disk Magic, consider the following environment:

- ▶ A DS48000 Model 80, with EXP810 enclosures.
- ▶ All arrays need to have enclosure loss protection.
- ▶ Five different hosts (all Windows based) can access the DS48000 through SAN, and each host is equipped with two 2 Gbps HBAs.

As these are Windows hosts, the host type will be defined as *open* in Disk Magic. The statistics gained from the perfmon files are used in Disk Magic as a base line.

The hosts are as follows:

- ▶ Host 1: database server
 - It requires 350 GB of RAID 5 with 73 GB high-speed drives.
 - It has an expected workload of 100 I/Os per second with a 16K transfer size.
 - Read percentage is expected to be 63%.

- ▶ Host 2: file and print server
 - It requires at least 1 TB of storage of RAID 5 with 146 GB on 10K drives.
 - It has an expected workload of 50 I/Os per second with an 8K transfer size.
 - Read percentage is expected to be 60%.
- ▶ Host 3: database server
 - It requires 500 GB of RAID 5 with 73 GB high-speed drives.
 - It has an expected workload of 150 I/Os per second with a 4K transfer size.
 - Read percentage is expected to be 60%.
- ▶ Host 4: e-mail server
 - It requires 350 GB of RAID 10 with 73 GB high-speed drives.
 - It has 500 users, and an expected I/O load of 1 I/O per second per mailbox.
 - It's expected workload is 500 I/Os per second with a 4K transfer size.
 - Read percentage is expected to be 50%.
- ▶ Host 5: database server
 - It requires 400 GB of RAID 10 with 146 GB high-speed drives.
 - It has an expected workload of 150 I/Os per second with an 8K transfer size.
 - Read percentage is expected to be 65%.

We create a new project in Disk Magic, starting with the window shown in Figure 8-5 on page 270.

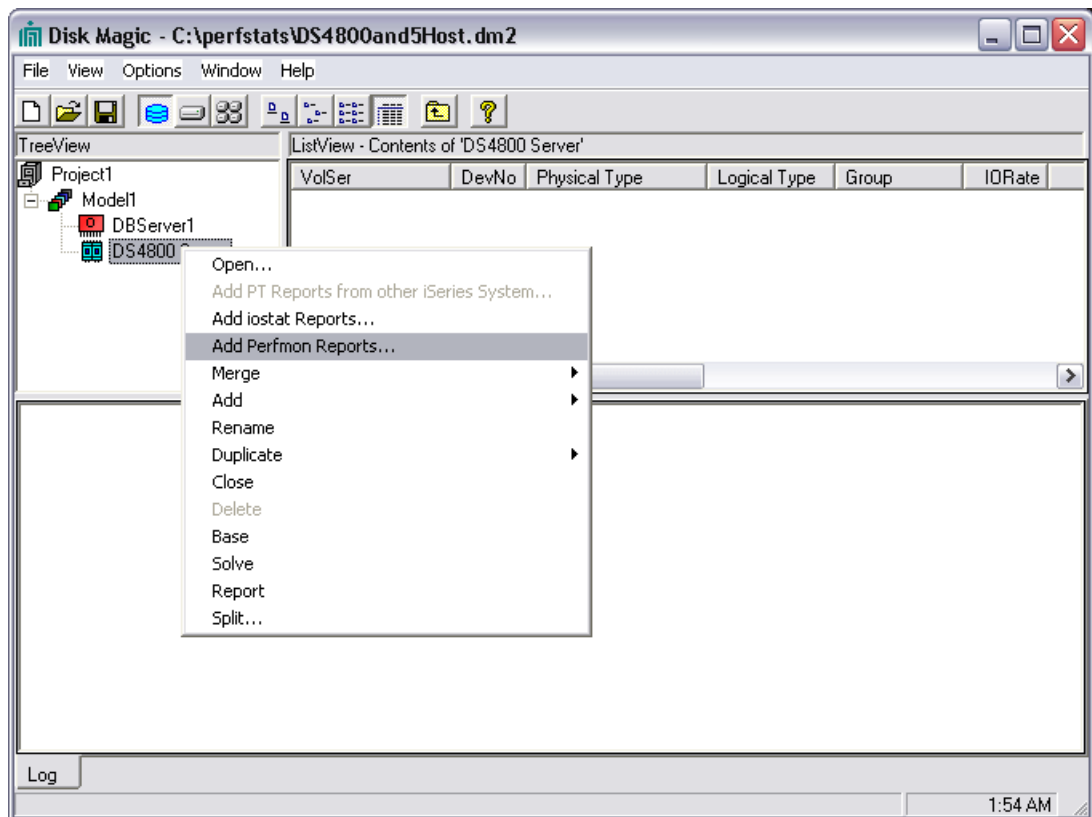


Figure 8-7 Add additional performance counters

As you can see in Figure 8-7 on page 273, the main window is divided in three panes:

- ▶ The TreeView displays the structure of a project with the servers and hosts that are included in a model. We started by specifying just one Disk Subsystem (DS4800 Server) and one open server (DBServer1).
- ▶ The ListView shows the content of any entity selected in the TreeView.
- ▶ The Browser is used to display reports and also informational and error messages.

Remaining servers and their collected performance data can be added at a later stage.

To add collected performance data, right-click the storage subsystem. From the drop-down menu, select **Add Perfmon Reports** (for Linux or UNIX systems select **Add iostat Reports**). Each time a performance report is added, the Perfmon Processing Options dialog lets you select how to use the information from the performance report (Figure 8-8).

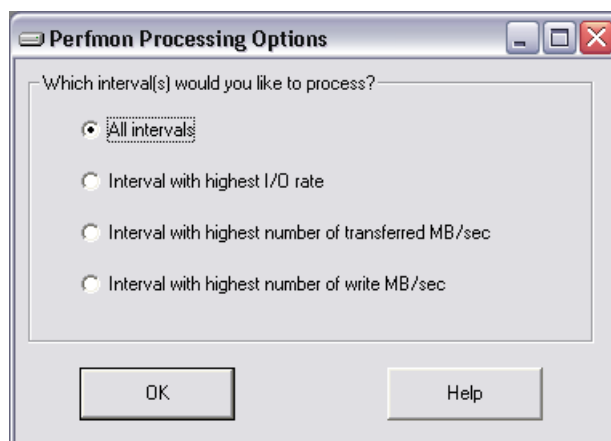


Figure 8-8 *Perfmon Processing Options*

If you double-click the storage subsystem, the Disk Subsystem dialog opens. It contains the following tabs:

- The General page allows you to enter Disk Subsystem level configuration data, such as the brand and type of subsystem and cache size. You can rename the disk subsystem and set the appropriate type, as shown in Figure 8-9.

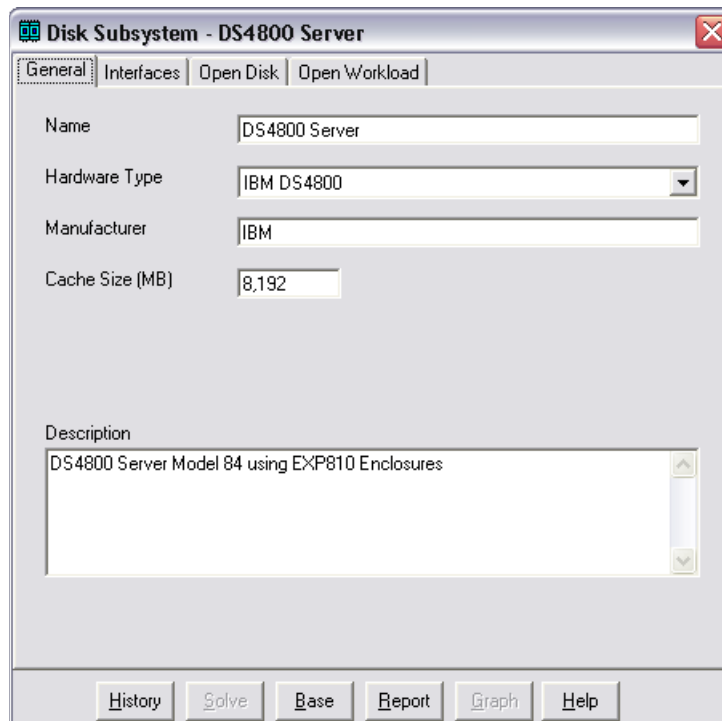


Figure 8-9 Storage subsystem General tab

- The Interfaces page is used to describe the connections from the servers and the disk subsystem. As shown in Figure 8-10, you can specify the speed (2 Gbps or 4 Gbps) and the number (Count) of interfaces.

Disk Subsystem - DS4000 Server

General | **Interfaces** | Open Disk | Open Workload

Server	Server side	DS4000 side	Count	Distance
DBServer1	Fibre 2 Gb	Fibre 2 Gb	2	0
FileServer1	Fibre 2 Gb	Fibre 2 Gb	2	0
DBServer2	Fibre 2 Gb	Fibre 2 Gb	2	0
EmailServer	Fibre 2 Gb	Fibre 2 Gb	2	0
DBServer3	Fibre 2 Gb	Fibre 2 Gb	2	0

Edit

From Disk Subsystem | **From Servers**

Remote Copy Interfaces

Remote Copy Type	Interface Type	Count	Distance
<input type="radio"/> XRC	Not supported	0	N/A
<input type="radio"/> PPRC	Not supported	0	N/A

☒ This DSS is not a Remote Copy Primary

Edit

History | Solve | Base | Report | Graph | Help

Figure 8-10 Host Interfaces

- The Open Disk page, shown in Figure 8-11, is used to describe the physical disk and logical drive (LUN) configuration for each host.

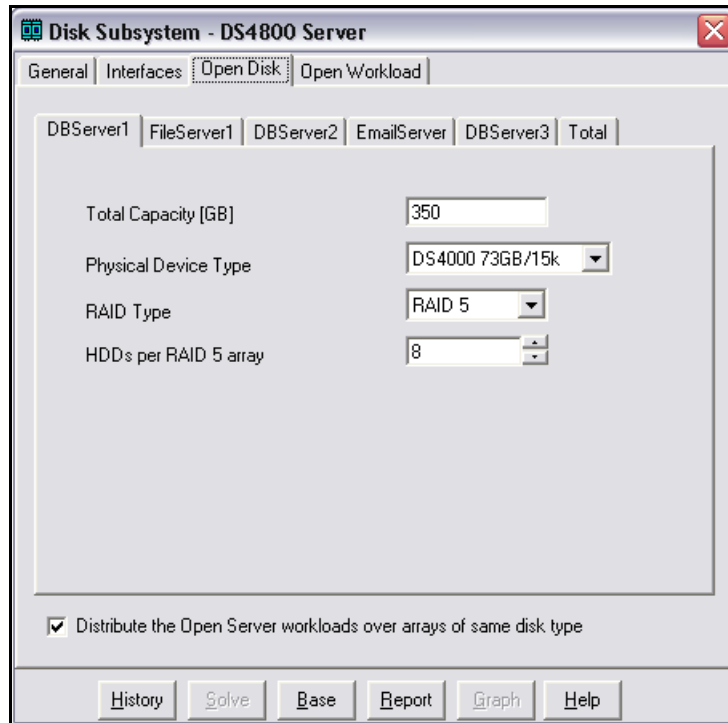


Figure 8-11 Set the disk size, number of disks, and RAID level

Note that Disk Magic will also create separate disk pages when you have storage attached to System i or System z™ servers.

As you can see in Figure 8-11, Disk Magic creates a tab in the Open Disk page for each open server that has been defined in the model.

Note that you can only define one type of physical disks and one LUN for each host.

Note: For servers with more than one LUN, you will need to define another host for each additional LUN. The workload for these pseudo servers needs to be added manually under the pseudo host name, as an automatic performance file import would add the statistics to the actual existing host.

- The Open Workload page is used to enter the I/O load statistics for each server in the model.

Note that Disk Magic will also create separate Workload pages when you have storage attached to System i or System z servers.

All of the data elements are entered at the subsystem level. I/O rate or MBps are required.

Importing the performance files will automatically populate these fields. See Figure 8-12.

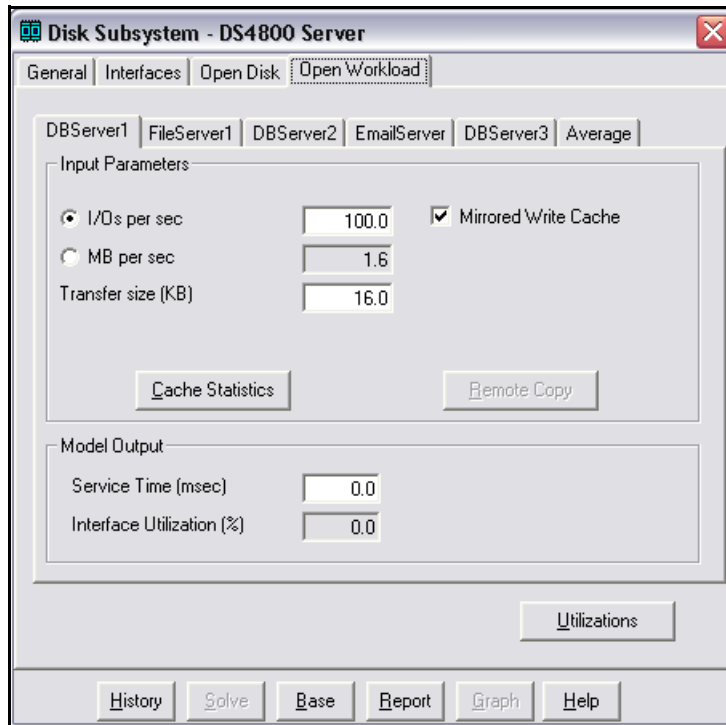


Figure 8-12 Set the workload per host

Optionally, click the **Cache Statistics** button to enter the cache statistics (see Figure 8-13).

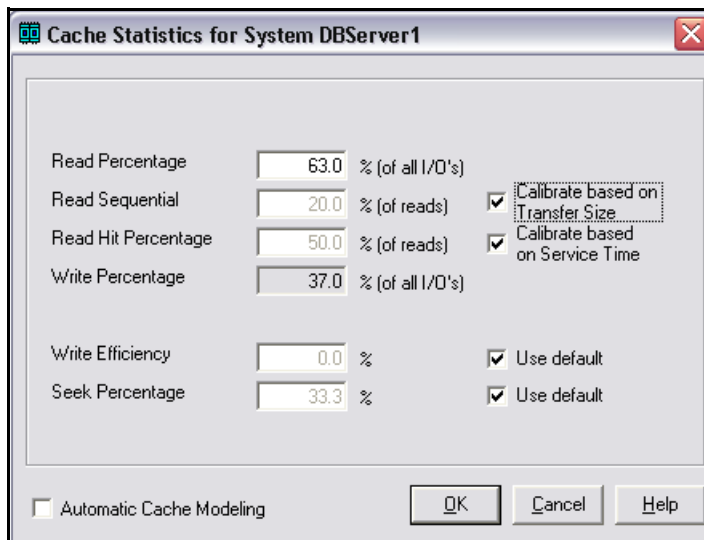


Figure 8-13 Adjust cache statistics

Once the initial data we just described has been entered into Disk Magic, click **Base** in the Disk Subsystem dialog window to create a baseline for the configuration.

A message, shown in Figure 8-14, confirms that the base was successfully created.



Figure 8-14 Base created

Once a base is created, you may examine the resource utilizations by clicking **Utilizations** on the Workload page (see Figure 8-12 on page 278). The Advanced Workload Output dialog opens, as shown in Figure 8-15.

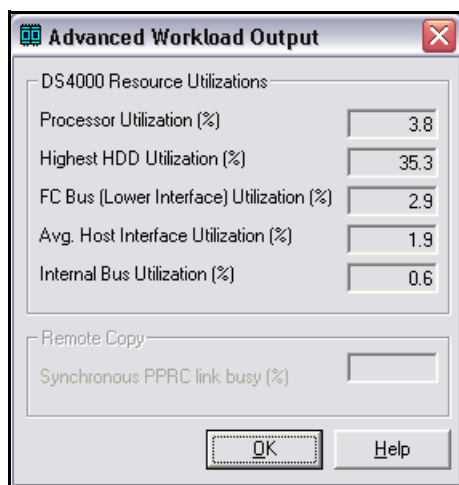


Figure 8-15 Advanced Workload Output

Disk Magic keeps a history of all of the configuration changes made. This allows for the restoration of the project to a known stage. To get the History Control panel, click **History** in the Disk Subsystem window.

In our case, Figure 8-16 shows that at this point in time only the base has been created.

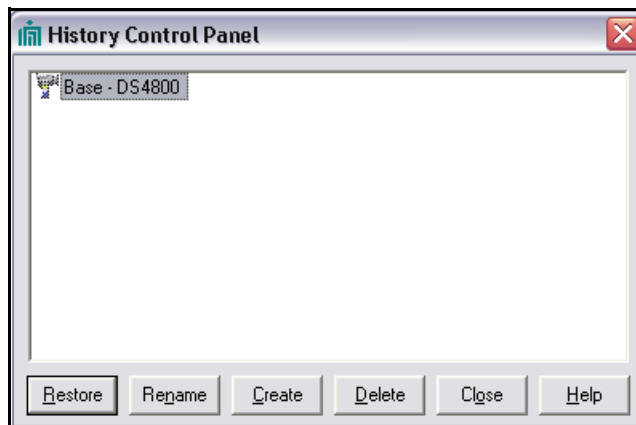


Figure 8-16 Base image history panel

Once the base is established, what-if analysis can be undertaken. For example, you can analyze the effect of:

- ▶ Changing the number of disks in an array
- ▶ Analyzing the difference in performance between 10K to 15K rpm disks
- ▶ Analyzing the difference between SATA and Fibre Channel disks
- ▶ Comparing different storage servers

As you perform your different what-if analysis and select **Solve**, additional reports appear in this History Control panel. It is a good idea to rename each entry to reflect what the changes represent.

Report

Selecting **Report** in the Disk Subsystem window creates a report of the current step in the model, as well as the base from which this model was derived. The report can be used to verify that all data was entered correctly. It also serves as a summary of the model results. The report lists all model parameters, including the ones that Disk Magic generates internally.

Example 8-1 shows a sample report.

Example 8-1 Base output from the project

Base model for this configuration

ProjectProject1

ModelModel1

Disk Subsystem name:DS4800 Server

Type:DS4800 (IBM)

Cache size:8192 Mbyte

Interfaces:4 Fibre 2 Gb FASTT, 2 Gb Server Interface Adapters

Total Capacity:2600.0 GB

approximately26 disks of type DS4000 73GB/15k

approximately8 disks of type DS4000 146GB/10k

approximately6 disks of type DS4000 146GB/15k

Remote copy type:No Remote Copy Active or Remote Copy Not Supported

Advanced FASTT Outputs (%):

Processor Utilization:3.8

Highest HDD Utilization:35.3

FC Bus (Lower Interface) Utilization:2.9

PCI Bus Utilization:0.6

Avg. Host Interface Utilization:1.9

Open Server	I/O Rate	Transfer Size (KB)	Resp Time	Read Perc	Read Hit%	Read Seq%	Write Hit%	Write Eff%	Chan Util%
Average	950	6.7	2.5	56	50	9	100	0	1
DBServer	100	16.0	2.5	63	50	20	100	0	0
FileServ	50	8.0	3.1	60	50	10	100	0	0
DBServer	150	8.0	2.2	60	50	10	100	0	1
EmailSer	500	4.0	2.6	50	50	5	100	0	2
DBServer	150	8.0	2.2	65	50	10	100	0	1

Graph

Graph solves the model and opens a dialog with options to select/define the graph to be created. Various graph types are possible, such as stacked bars and line graphs. By default, data to be charted is added to the spreadsheet and graph, allowing you to compare the results of successive modeling steps for a disk subsystem, or to compare results of different subsystems that are part of the same model.

Figure 8-17 shows the different graph options:

- ▶ Service Time in ms
- ▶ Avg Hit (Read and Write) in %
- ▶ Read Hit Percentage
- ▶ Busy Percentage of Host Interfaces
- ▶ Busy Percentage of Disk Interfaces
- ▶ I/O in MBps
- ▶ Read I/O in MBps
- ▶ Write I/O in MBps

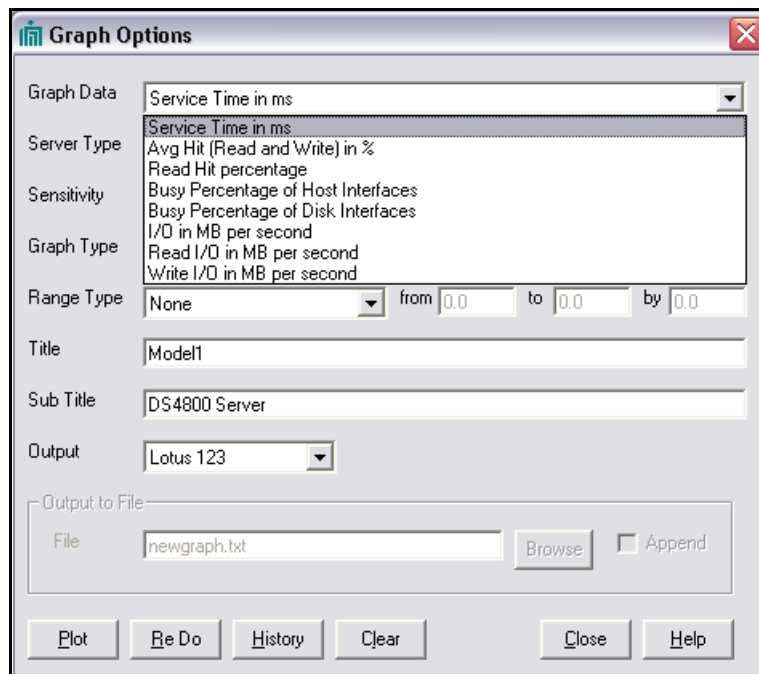


Figure 8-17 Graph Options - Graph Data

Next select a graph type, as shown in Figure 8-18.

The 'Graph Options' dialog box is shown with the following settings: Graph Data is 'Service Time in ms'; Server Type is 'Open'; Sensitivity is 'None'; Graph Type is 'Stacked Bar' (with a dropdown menu open showing 'Line', 'Stacked Bar', 'Bar by Row', 'Bar by Column', and 'Pie'); Range Type is 'Line' with values 'from 0.0 to 0.0 by 0.0'; Title is empty; Sub Title is 'DS4800 Server'; Output is 'Lotus 123'; and Output to File is 'newgraph.txt' with 'Append' checked. At the bottom are buttons for Plot, Re Do, History, Clear, Close, and Help.

Figure 8-18 Graph Options - Graph Type

Then select the range type and the range values, as shown in Figure 8-19.

The 'Graph Options' dialog box is shown with the following settings: Graph Data is 'Service Time in ms'; Server Type is 'Open'; Sensitivity is 'None'; Graph Type is 'Stacked Bar'; Range Type is 'None' (with a dropdown menu open showing 'None', 'Cache Size', 'I/O Rate with Capacity growth', 'Capacity', 'I/O Rate', 'Cache / Backstore Sensitivity', and 'Read Hit % for Open'); Title is empty; Sub Title is empty; Output is empty; and Output to File is 'newgraph.txt' with 'Append' checked. At the bottom are buttons for Plot, Re Do, History, Clear, Close, and Help.

Figure 8-19 Graph Options - Range Type

You must also select an output type as either Lotus® 123 or text file.

Finally, select **Plot** to generate the graph.

Several graph options and the resulting output are shown in Figure 8-20.

Graph Options

Graph Data: Service Time in ms

Server Type: Open ☐ By Server or iSeries ASP

Sensitivity: None ☒ Full Update

Graph Type: Line ☒ Totals on bar

Range Type: I/O Rate with Capacity grow from 100.0 to 4,950.0 by 500.0

Title: IO Rates

Sub Title: DS4800 Server

Output: Lotus 123

Output to File:
File: newgraph.txt ☐ Append

Figure 8-20 Setting for Service Time graph with range type of I/O rate with Capacity grow

The resulting graph is shown in Figure 8-21.

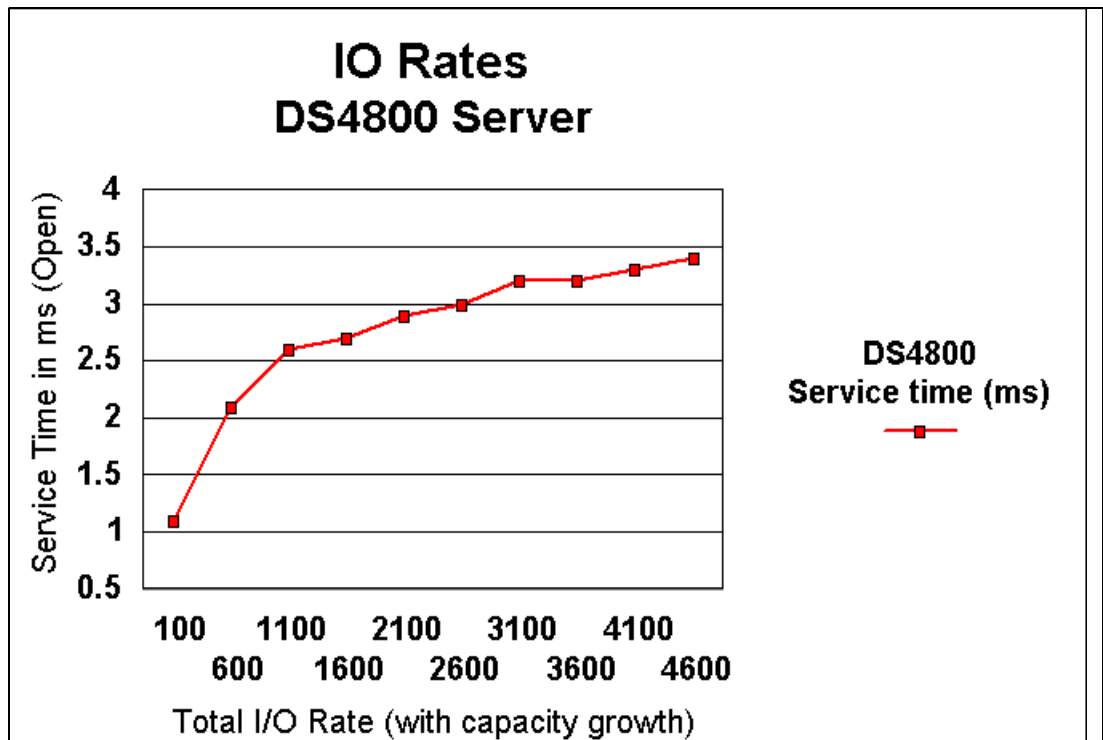


Figure 8-21 Output of Service Time graph with I/O Rate with Growth Capacity

The next graph was to compare read I/O rate to I/O rate. The range was set from 100 to 2,500 by 200. See Figure 8-22.

Graph Options

Graph Data: Read I/O in MB per second

Server Type: Open ☐ By Server or iSeries ASP

Sensitivity: None ☒ Full Update

Graph Type: Line ☒ Totals on bar

Range Type: I/O Rate from 100.0 to 2,500.0 by 200.0

Title: IO Rates

Sub Title: DS4800 Server

Output: Lotus 123

Output to File:
File: newgraph.txt ☐ Append

Figure 8-22 Settings for read I/O versus I/O rate graph

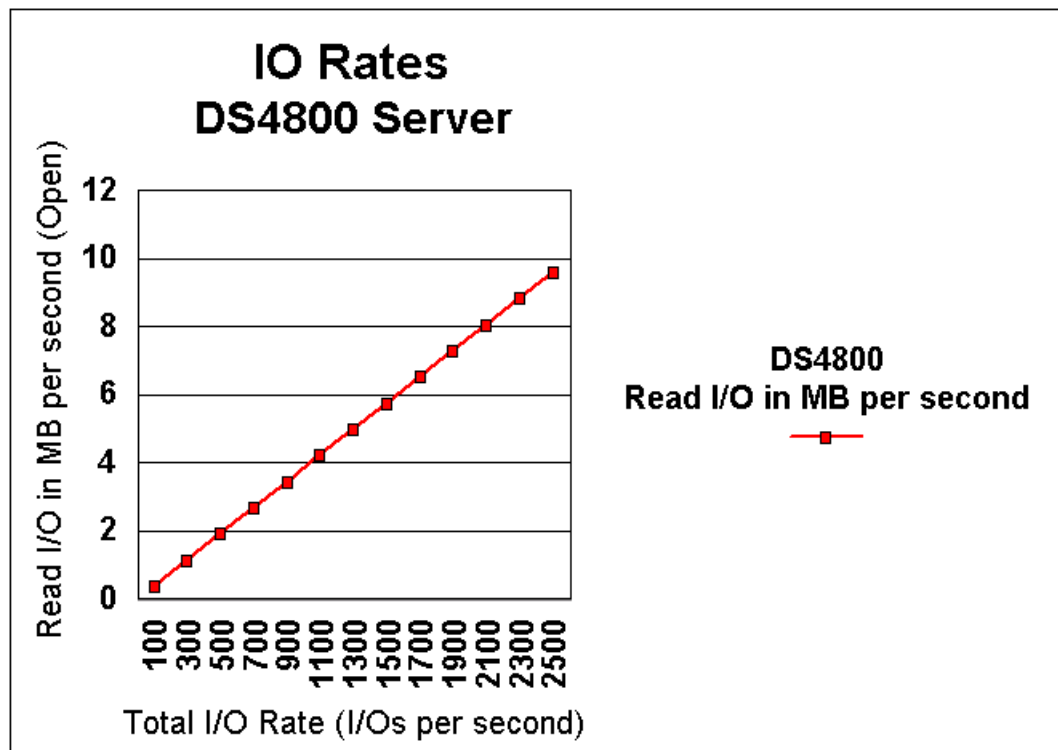


Figure 8-23 Output from read I/O versus I/O rate graph

Graph Options

Graph Data: Write I/O in MB per second

Server Type: Open ☐ By Server or iSeries ASP

Sensitivity: None ☒ Full Update

Graph Type: Line ☒ Totals on bar

Range Type: I/O Rate from 100.0 to 2,500.0 by 200.0

Title: IO Rates

Sub Title: DS4800 Server

Output: Lotus 123

Output to File:
File: newgraph.txt ☐ Append

Figure 8-24 Settings for write I/O to compare read I/O graph

With both the read and write I/O versus the I/O rate on the same scale.

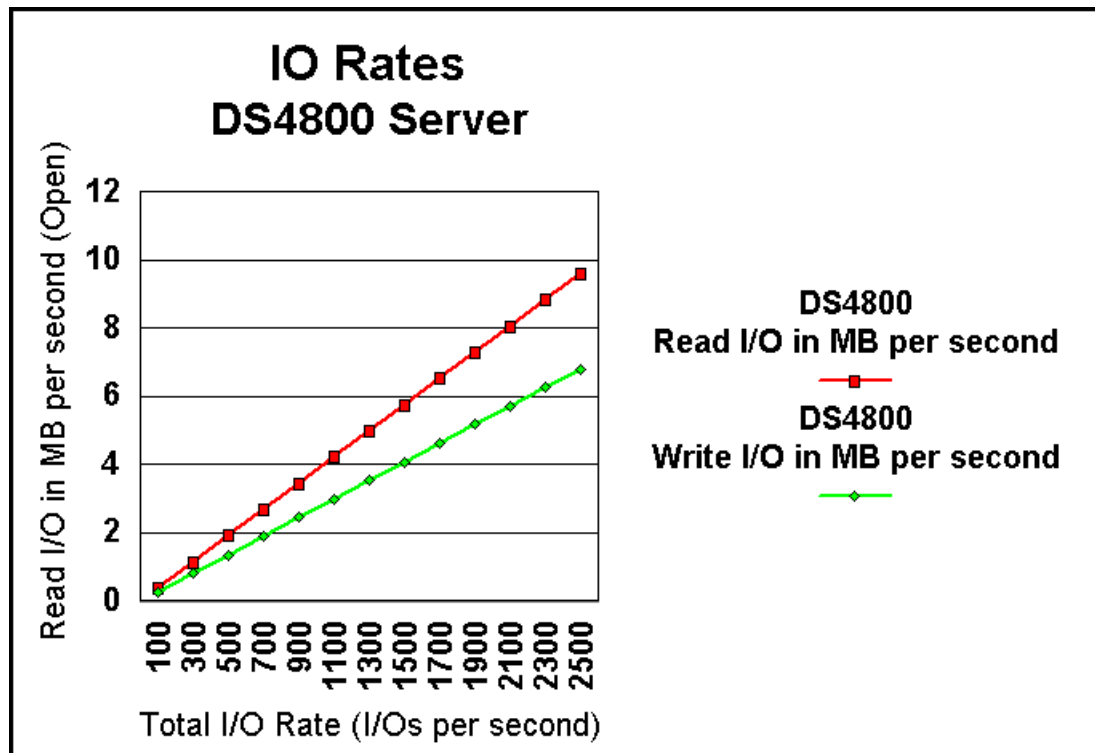


Figure 8-25 Output for read versus write I/O graph



ERM planning and implementation

Enhanced Remote Mirroring (ERM) is an important feature of the DS4000. Its implementation should be carefully planned and you must also understand its impact on the resources and performance of your environment.

This chapter gives you guidelines and takes you through the critical preparation steps required for a successful Remote Mirroring implementation. It also includes an installation checklist and describes the ERM Bandwidth Estimator Tool.

Note: The Bandwidth Estimator Tool is not directly available to customers. Ask your IBM Service Representative.

9.1 Introduction to ERM

A remote mirroring capability is an important and essential feature of a robust storage solution.

IBM DS4000 storage systems can protect critical business data through real-time mirroring of data writes to a second DS4000 storage system located at a remote site. This advanced capability, called Enhanced Remote Mirroring (ERM), enables the DS4000 to provide the underpinnings of a robust disaster recovery solution. ERM for DS4000 is a technical feature, and can only be one component of the DR solution. It is not a complete DR solution by itself.

ERM is independent of and transparent to host application servers. For each set of source and target logical drives that form a mirror pair, ERM supports a variety of capabilities and mirroring options. These features are summarized below:

- ▶ Metro mirror mode

Metro mirroring is a synchronous mirroring mode. Any host write request is written to the primary (local) storage subsystem and then transferred to the secondary (remote) storage subsystem. The host application must wait until receiving acknowledgement that the write request has been executed on both (local and remote) storage controllers.

- ▶ Global mirror mode

Global copy is an asynchronous write mode. All write requests from a host are written to the primary (local) storage subsystem and immediately reported as completed to the host system.

- ▶ Global copy mode

This mode must be used when the host application has dependent data spread across several logical drives to ensure that the dependent write requests are carried out in the same order at the remote site. This is done through the realization of *consistency groups*.

- ▶ Dynamic mode switching

This feature allows you to go back and forth between metro mirror, global copy, and global mirror modes without breaking the remote mirror.

- ▶ Read-only access to mirrored logical drives at the remote site

- ▶ Delta resynchronization after suspend and resume

A resume operation will synchronize only the data written to the primary logical drive while the mirror was stopped (resynchronization). This includes a mirror that was manually halted (suspend) or dropped due to unplanned loss of remote communication. This is made possible by a delta log that keeps track of changes to the primary volume during a communication interruption (planned or unplanned).

- ▶ Role reversal

In the event of a site failure or for maintenance, you can change the role of a logical drive in a given mirror relationship from primary to secondary and vice versa, and grant or deny (to the hosts) write access on these logical drives.

Designing a robust and effective ERM solution relies in part on appropriately selecting and using these capabilities and options, which will be reviewed in the following sections.

For a detailed description of the DS4000 ERM features refer to *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010.

ERM enhancements

ERM now supports link speeds as low as 2 Mbps. ERM will no longer time out from inadequate link speed between sites unless the effective transfer rate drops below 2 Mbps. Previously, ERM would occasionally time out if the transfer rate dropped below 10 Mbps.

ERM enables improved host performance for sequential writes.

Minimum required firmware levels for ERM enhancements

Table 9-1 shows the minimum firmware level including the above enhancements, depending on the DS4000 model.

Table 9-1 Minimum required firmware levels for ERM enhancements

	6.12.27.05	6.15.27.05	6.16.92.00
DS4100	EXP100		
DS4300	EXP100, EXP710		
DS4400	EXP100, EXP710		
DS4500	EXP100, EXP710		
DS4200			EXP420
DS4700			EXP710, 810
DS4800		EXP100, 710	EXP710, 810

9.2 ERM as part of a DR solution

The needs and reasons to mirror data are critical for many customers. Companies might be mandated through regulations to mirror data, or they want to be able to maintain a copy of data at a remote location without physically shipping tapes, or they may need a method to migrate from one host to another or from one storage subsystem to another with as little disruption as possible. In most cases, as modern business pressures increasingly require 24-hour data availability they are required to ensure that critical data is safeguard against potential disasters they need ERM as part of a Disaster Recovery or Business Continuity (BC) solution.

In the remainder of this section, we review important elements to consider when planning for ERM as part of a Disaster Recovery (DR) or Business Continuity (BC) solution. The focus of the discussion is on ERM. A complete DR solution encompasses much more than just mirroring, and is beyond the scope of this book. For comprehensive information about Disaster Recovery and Business Continuity solutions, refer to *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547, and *IBM System Storage Business Continuity: Part 2 Solutions Guide*, SG24-6548.

In any case, before starting your plans for a DR and ERM solution, make sure that you understand the cost and complexity.

The total price to implement a complete disaster recovery solution is an important test of commitment to pass before moving forward. Price may involve more than finances: it most likely will involve new technical skills and changes to business processes.

DR solutions often require extended periods of time to implement. Setting executive expectations up front is critical to successful completion. Unrealistic expectations (such as finishing the project next week or even within a month in many cases) are red flags that need

to be addressed promptly by resetting expectations. Timeframes for implementing DR solutions are typically measured in quarters of a year.

The development a company has to undergo to find, design, and implement suitable ways for an adequate disaster recovery solution typically is a project for months. Usually, the difficulty is not just the technical implementation of the disaster recovery solution. It also lies in the necessary inspection of internal processes required to find a clear and structurized way to handle actions and responsibilities across departments or between persons.

The analysis required when planing a disaster recovery solution is an endeavour that is particular to each individual company. This analysis should give the company a better understanding of its data, not just in terms of its size or nature, but also in categorizing and rating data in terms of criticality to their business. Once the values and priorities are clarified, the company can begin to work on a plan for Disaster Recovery, and IBM can provide valuable assistance with this step.

9.2.1 Planning for ERM as part of a DR solution

To prepare and design an ERM as part of a DR solution, you must understand important parameters and how they relate to each other. These are the distance between sites, the maximum IOPS, the bandwidth required, whether you will use a synchronous or asynchronous mirroring method o:

Distance and maximum IOPS

In a remote mirroring solution the effective transfer rate between sites has to be taken into account in comparison to the nominal link speed.

What does limit the effective transfer rate from the nominal link speed? The most important factor usually is latency on the link. These latencies are the result of delays introduced by the equipment (switches, protocol converters for IP, firewalls) and the transport medium itself. The latency in the transport medium itself is determined by the finite speed of the light in the glass fibre. The effective speed of light in the fibre is about 200,000 kmps. This is a fundamental

constant in an ERM (and DR) solution, and its impact gets stronger as the distance between sites increases. The chart in Figure 9-1 shows the correlation between the distance separating the sites and the latency of the link. Remember also that with a synchronous mirroring method the signal (light) must travel the distance between sites twice (there and back) for an I/O to complete.

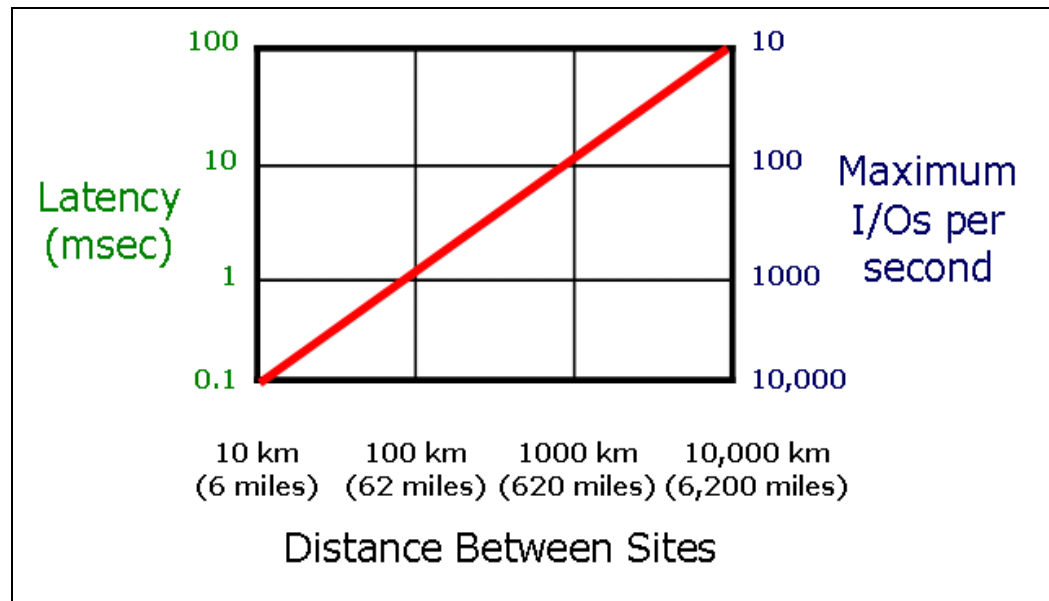


Figure 9-1 Maximum I/O by distance

The chart also displays the maximum theoretical IOPS that can be sustained. Again there is a direct, linear correlation between IOPS and the distance. To understand the correlation, think of the following: assuming a synchronous mirroring, an I/O is started at time $t=t_0$ and begins its way through the fibre until it reaches the other site after t_1 . The remote site sends an I/O completion back, and it will take again t_1 to reach the local site. The sum of time taken for $(2 \times t_1)$ represents the total I/O time, or t_{IO} (t_{IO} is also called round-trip time, RTT). Each I/O originated by a LUN at the local site will take t_{IO} to complete before the next I/O can be sent. In other words, the maximum IOPS will be $1/t_{IO}$.

For example, if there is a distance of 100km between sites, the light needs $100 \text{ km} / 200000 \text{ km/s} = 0.0005$ seconds to get to the remote site, and we have a t_{IO} time of 0.001 seconds. The maximum IOPS is $1 \text{ s} / 0.001 \text{ s} = 1000$.

The chart shows the theoretical limitations of a remote mirroring relationship of a single LUN that works in a synchronous mode. It is of general meaning and does not only reflect the theoretical limitations of a DS4000 ERM implementation.

The chart represents the theoretical, best-case maximum for a synchronous remote mirroring relationship. It does not account for other limiting elements such as switches and IP converters and the storage itself, which all introduce additional latencies.

Be aware especially of the additional latencies inherent to an IP network. An IP network is often used for long-distance mirroring. In this case the quality of the IP network becomes a critical part of the overall remote mirroring solution. Make sure that the IP network will deliver the required service level. A good, professional approach is to work with your supplier for IP converters and your telecommunication provider. They can act as agents who have the experience and know-how to measure the quality and reliability of an IP network and check the fulfillment of such guarantees.

Important: The distance between sites is an essential design point of an ERM solution.

Distance is a crucial element and essential design point of an ERM solution. As the chart implies, the distance limits IOPS in a synchronous mirroring and can substantially and negatively impact I/O-intensive applications such as transaction processing and database operations. With an asynchronous mirroring solution, longer distance and thus a smaller number of IOPS will constrain the Recovery Point Objective (RPO) of your DR solution.

Distance will dictate which mirroring mode can be implemented effectively. This is discussed in “Synchronous or asynchronous mirroring” on page 293.

Distance will also impact the implementation cost. For short distances in the range of up to 1 km, the infrastructure costs will be far below the costs of an implementation with a distance between sites measured in thousands of miles.

In DR solutions, for some industries, such as banking, government regulations govern the minimum distance between sites. For others, the cost of providing adequate network bandwidth is the primary consideration. In some cases, special region-specific conditions must be taken into account as well. For example, in southern California the recovery site should be on a different tectonic plate than the primary site. But in London the recovery site may only need to be blocks away to enable recovery from a terrorist attack or a fire. The potential for flooding and the need to put the sites on different power grids are two other factors that can dictate a minimum (or maximum) distance.

Distance and bandwidth

Another important factor in an ERM or DR implementation, especially with long distances, is the bandwidth required. The bandwidth defines how much data can be sent through the data link.

Important: Without adequate bandwidth, the mirroring solution will fail.

If the connection bandwidth is not sufficient, then the throughput required for the mirroring to function well cannot be achieved.

The data link bandwidth requirement encompasses the amount of data that needs to be replicated as well as the link speed required to provide acceptable performance. Since the amount of data that will be mirrored will likely grow (usually at a hefty rate) over time, the link speed must be dynamically adaptable to meet future requirements. Some questions that should be asked are:

- ▶ What is the minimum bandwidth to provide adequate performance for the distance between sites and the amount of data that is to be mirrored initially?
- ▶ What is the expected rate of growth of the mirrored data?
- ▶ What minimum bandwidth is recommended for the expected data growth?
- ▶ If the data is being mirrored to an existing site, how much bandwidth is available now? Is it being shared? If so, what portion is available for this mirroring? In the future?
- ▶ Are there plans to increase the bandwidth? If not, what is required to plan it?

Tip: A good ERM design will provide enough network bandwidth to minimize resynchronization time.

If the data link drops for any reason, mirroring is suspended until the link is re-established. This requires that the mirrors be resynchronized (fully mirrored again) before they can be used for recovery, which may take a significant amount of time, based on a number of factors. Make sure that you plan enough bandwidth to cope with resynchronization situations.

Prioritize and characterize the data to be mirrored

Since the amount of data to be mirrored will directly impact the bandwidth, it is important to analyze the data used by specific applications to determine whether it is required to always mirror all the data.

For most organizations, different types of data have different levels of business value, and for DR purposes, they should have different recovery point objectives and different recovery time objectives. Assuming that all data has equal value and trying to mirror it equally for disaster recovery purposes is usually not realistic. Data mirroring should be prioritized by data value (which translates to Recovery Point Objective (RPO) versus time to recover (Recovery Time Objective, or RTO). This is where the business impact analysis (BIA) becomes very important.

Types of business data typically range from simple user files and directories to complex database tables, logs, and indices. Each data type has its own set of requirements for integrity and proper recovery. It is important to understand the types of data to be mirrored and the associated requirements for successful recovery.

In particular, with databases it is usually only necessary for a DR solution to mirror the log files. Assuming that you had initially replicated the database at the remote site, you would only need to apply the logs against the database at the remote site. Only replicating what is necessary can have a substantial positive impact on the bandwidth requirements.

Refer to 9.4, “The Bandwidth Estimator Tool” on page 307, for more discussion on bandwidth and distance and how to estimate them.

Synchronous or asynchronous mirroring

Directly combined with the distance parameter is the question of whether to use synchronous or asynchronous mirroring.

Synchronous mirroring can only be used for short distances, to a maximum of 150 km (100 miles). Since control is not returned to the host application until an I/O it originated has been replicated at the secondary site and acknowledged by the primary site, the application response time becomes unacceptable with greater distances.

With asynchronous mirroring the I/O is marked as completed as soon as the primary storage has written it to disk or cache and control is returned to the application. In parallel, the I/O is written to a queue maintained by the primary storage before being sent to the secondary storage and will remain in the queue until an I/O completion from the secondary storage is received. Hence, the I/O on the mirrored (primary) LUN is not directly impacted by the mirroring.

Remember that the Mirror Repository Logical Drive can queue a number of I/O requests until the maximum allowed difference of pending write requests between primary and secondary is attained (number of unsynchronized write requests). During this process the mirrored pair state remains synchronized. If the maximum number of unsynchronized I/O requests is

exceeded, the state of the mirrored pair changes to unsynchronized. The time you have before the mirroring state changes from synchronized to unsynchronized mainly depend on two factors:

- ▶ Maximum number of unsynchronized write requests
- ▶ I/O write request per second to the primary storage subsystem

For example, if the queue is holding a maximum of 1024 I/Os and you have an impact of 2500 I/O write requests per second, the time difference between each I/O is $1\text{ s} / 2500\text{ IOPS} = 0,4\text{ ms}$.

The period of time you can hold the synchronized state with your queue can be calculated as follows:

$1024\text{ writes} * 0.4\text{ms/write} = 409.6\text{ms}$

You can assume that the connection can have a maximum latency of 200 ms (you must send the data and receive the confirmation). The theoretical value that can be assumed for an IP Network is about 100 km/ms. In this case you could theoretically have a maximum connection length of 20000 km. In reality the value is slightly less because of the unknown number of routing hops, alternate traffic volume, and other factors.

Note: The queue size is 128 outstanding I/Os per logical drive (with firmware Version 6.1x.xx.xx). For a subsystem with 32 mirrored pairs, the queue length is thus $128 \times 32 = 4096$; and for 64 pairs, it is 8192. You recommend using as many logical drives as possible for ERM long distance to keep the I/Os on one particular logical drive in a smallest possible range.

Consistency groups

In many cases, the data to be mirrored is written across multiple logical drives (perhaps for performance reasons). If this data must be mirrored using asynchronous operation, there is a possibility that data blocks may not arrive at the recovery site in the same order in which they were written at the primary site. This can lead to inconsistent data sets that may not be usable for recovery purposes.

Important: Determine which logical drives should be included in a consistency group when designing your ERM solution.

To prevent this from occurring, a DS4000 consistency group is provided with global mirror. A consistency group ensures that all data blocks from a set of logical drives are sent in the same order in which they were written on the primary system. In the event of a mirroring failure, a consistency group also ensures data integrity when related data resides on multiple logical drives. Use of the consistency group is the primary difference between DS4000 global mirror and global copy.

9.2.2 Implementation recommendations

There are several recommendations that apply to ERM in general and others that are more specific to a particular type of application, such as database. We begin with recommendations at the primary storage.

On the primary storage

On the primary storage:

- ▶ DS4000 Enhanced Remote Mirroring (ERM) reserves the last port of each controller for its own use. This reduces the overall bandwidth available between the DS4000 and the SAN. If the last port is already being used, then logical drive and channel assignments must be changed to accommodate this restriction
- ▶ Double the write I/O rates of the logical drives that will be mirrored. Mirroring requires the DS4000 to write every I/O twice (to itself and a DS4000 at the recovery site), and this takes additional resources within the storage system. The additional resources will help to compensate for the performance impact that will occur when logical drive synchronization is being performed.
- ▶ Increase the read rate of all logical drives by 25% to account for mirroring latencies, synchronization, and overhead. The data transfer rate, mirroring priority, and network latencies all influence this, but 25% is a reasonable rule of thumb.
- ▶ Increase the possible write I/O rate of the mirrored drives at the primary site by 25% if FlashCopy is integrated on the primary site. Increase the read rate by 15% for the same reason.
- ▶ If other DS4000 replication features are used while remote mirroring is in progress, they will also impact performance and should be taken into account when sizing the primary storage system. Of particular note are backups that use FlashCopy or VolumeCopy.
- ▶ Place the ERM repositories on a separate array that is not busy with data I/O. This will help to minimize the impact of ERM, especially during synchronization periods.

On the secondary storage

The remote storage system presents a more complex sizing effort than the primary storage system. Some questions to ask about the recovery storage system include:

- ▶ Is it to be used for more than recovery purposes?
- ▶ What is the performance expectation in the event of failover?
- ▶ Are more or fewer applications going to be running on it (compared to the primary storage system) after failover?
- ▶ How will post-failover backup impact it?
- ▶ Will it become a new primary storage system for remote mirroring to another site (e.g., to facilitate failback)?

If the recovery mirrors are expected to perform as well as the primary logical drives, then the same sizing for production work will apply to the remote storage system as the primary. Do not forget that one port on each controller will be reserved for ERM, and do not forget to compensate for performance impacts caused by DS4000 data replication features.

Best practice: The capacity sizing for production work at the recovery site needs to be multiplied by a factor of at least 3.2.

This is to provide additional capacity, as follows:

- ▶ The same initial capacity (1X) is for the continuous mirroring of production logical disks that takes place prior to failover.
- ▶ You need another equivalent of the initial capacity (2X) to sustain a second copy of point-in-time mirrors. This enables database roll-forward capability in case a corruption of the database at the primary site is mirrored to the recovery site.

- ▶ You need yet another equivalent of the initial capacity (3X) to replicate the mirrored logical disks for testing or online backup.
- ▶ An additional 20% (0.2X) provides incremental capacity to enable FlashCopy for the mirrored logical disks.

The 320% increase enables all of the above (that is, it provides capacity for the ERM mirrors, a point-in-time copy for fall-back, space to copy the database for testing or backup, and capacity for FlashCopy repositories).

In other words, if 10 TB of production data at the primary site is to be mirrored and put into production at the recovery site after failover, the recovery storage system should have at least 32 TB of capacity — with enough spindles (physical disks) to meet performance expectations.

It is not unusual to have a smaller storage system at the recovery site as some of the applications running at the primary side may not be necessary in the event of a site failure. This is why it is so important to determine what data will be mirrored and what applications will actually need to run at the remote site.

If, after failover, the recovery site is to mirror data to a remote site for failback or to cascade mirroring, then the recovery storage system must be sized accordingly. It will, in fact, function as a primary storage system in the new data mirroring scheme, and therefore must have its sizing altered according to the rules of thumb provided in the Primary Storage System section above.

If the data stored on the primary system is being mirrored to a bunker site instead of a recovery site, performance is not as important as price. This is the **ONLY** case where SATA drives may be considered for remote mirrors. A bunker site is typically used for data mining or centralized backup, and the appropriate sizing rules apply.

When configuring your secondary storage, remember that:

- ▶ The logical drives belonging to controller A in the primary storage server must be mirrored to the logical drives owned by controller A in the secondary subsystem. The same rule applies for the logical drives owned by controller B.
- ▶ DS4000 ERM has a limit on the number of mirrored pairs supported that varies by DS4000 model. This limit must be considered when sizing either storage system:
 - DS4100, DS4200, DS4300 Turbo and DS4700 support 32 mirror pairs
 - DS4400, DS4500 and DS4800 support 64 mirror pairs

Note: please check IBM Web site for the latest number of supported pairs.

Assess current and future storage needs

Is the data already stored on DS4000 storage systems? Will it have to be migrated from another storage platform? Are firmware levels up to date? What about future storage needs? Most likely growth will require additional capacity, but how much? Will additional DS4000 storage systems be required in the future? To keep costs down, is it better to start with DS4700s and upgrade to DS4800s as growth takes place?

Set synchronization priority

If for any reason a link drop occurs and the mirror goes into the suspended state until the link is re-established, a full resynchronization will be required and will depend upon on the available bandwidth. There is also a strong dependency on the synchronization settings: When changing the synchronization setting from medium (that is the default) to highest, a reduction in the time required for re-synchronisation of up to a factor of 10 can be observed.

We thus recommend setting the synchronization setting to the highest value when manually restarting the link after a failure.

Dynamic mode switching

One of the interesting features of the DS4000 ERM is the fact that you can dynamically switch between the different mirroring modes (from metro mirror to global copy or global mirror and vice versa). The feature is also useful if a network problem results in degraded network bandwidth (in this case, switching from metro to global mirror allows mirroring to run more efficiently while the bandwidth is substandard).

9.2.3 Network considerations

One of the most important aspects of designing a successful ERM (and DR) solution is properly sizing the network that connects the primary and recovery sites.

We have already discussed the effect of bandwidth and latency in 9.2.1, “Planning for ERM as part of a DR solution” on page 290. Refer to that section for additional details.

Bandwidth

Bandwidth defines how much data can be sent through the data link. If bandwidth is inadequate, required throughput will never be achieved.

One very important matter to consider in designing the network is the amount of time that will be required to perform initial synchronization as well as ongoing resynchronizations. Additional bandwidth may be needed for these tasks to get mirrors synchronized in a reasonable time frame. If the network is to be shared, sizing for synchronization is especially important because ERM could potentially use all available bandwidth or, worst case, not have enough bandwidth.

When sizing bandwidth, it is best to use actual performance measurements. Your IBM representative has a network bandwidth calculator (see 9.4, “The Bandwidth Estimator Tool” on page 307) that can help you size the minimum bandwidth required for a successful DR solution and set realistic performance expectations.

Important: Bandwidth is typically measured in bits per second: Kilobit (Kb), Megabit (Mb), or Gigabit (Gb) per second (Kbps, Mbps, or Gbps). Always remember that a bit is only one-eighth of a byte when performing sizing calculations. Therefore, divide the network bandwidth in bits per second by ten to get the equivalent storage rate of bytes per second. Also remember that, if the data being transmitted has been converted to TCP/IP, there is a significant increase in overhead that reduces effective bandwidth up to 50%.

Latency

Latency is the time it takes a mirroring I/O and its acknowledgment to traverse a network link. The longer the distance between primary and recovery sites, the more time it takes to send the data and receive an acknowledgment. Distance becomes a limiting factor for the number of I/Os that can be sent per second. Therefore, latency is the controlling factor for I/O rate that can be supported in a disaster recovery solution.

Important: The effects of latency can only be overcome by multiplying the number of links that carry mirroring I/Os simultaneously.

Network information

Since the network is so important to the overall solution, it is important that all required information be obtained before the design is finalized. A detailed SAN/LAN/WAN diagram should be developed so any discrepancies or problems can be discovered early on during implementation.

Required network information includes:

- ▶ Switches involved (number, layout, type)
- ▶ Actual data path to be used for the interconnect between storage systems
- ▶ Will the WAN/LAN connection be dedicated or shared?
- ▶ How much of the available bandwidth will be used by other applications or users?
- ▶ Are there any planned changes that could affect available bandwidth?
- ▶ What network monitoring tools are available? Can they be made available during implementation?
- ▶ How will performance be monitored?
- ▶ How will a network failure affect mirroring? Will mirrors be suspended? If so, how long?

Buffer credits

A Fibre Channel switch uses buffers to allow multiple data frames over a single Fibre Channel link. The more buffers your switch allows, the more frames it can process in a sequence (that is, without requiring acknowledgement between single frames). The amount of buffers is called buffer credit (BBcredit). It is limited by default to a maximum of 16 in a standard Brocade switch, but can be increased slightly.

Specifying the correct number of buffer credits ensures that the Fibre Channel link is running at optimal efficiency in terms of data delivery. When longer distances are involved, the appropriate number of buffer credits becomes important in order to prevent a networking term called *drooping*. Drooping occurs when data cannot be sent because the sending system is waiting for the opportunity to send more data through the link. For Fibre Channel, drooping can be caused by not having enough buffer credits.

The optimal number of buffer credits is a function of bandwidth and distance. A standard Fibre Channel frame contains roughly 2 KBytes of data. With a link speed of 1 Gbps, and knowing that the frame is moving at the speed of light in the fibre (200,000 km/ps), you can calculate that the frame will span over about 4 km. For a 2 Gbps link this frame will be 2 km long and 1 km only with a 4 Gbps link. To calculate the optimal number of buffer credits required to optimize mirroring performance divide distance between sites by the frame length. If the distance is increasing, then more frames can be accommodated on the link and the number of buffer credits should be increased.

In any case, consult with your SAN extension provider for details on buffer credits. Some devices like the Storage Net Edge Router from CNT/McData (now part of Brocade) can handle buffers locally with the FC-device and does not depend on any buffer credits when communicating on long-distance over IP.

Pipelining

CNT and others offer with their IP-converters some features to emulate a device locally. This may have some functionality with streaming requirements but we cannot say right now whether there is anything useful about it in a Database environment as there are consistency requirements as well in case of a disaster.

Compression

Vendors like CNT also offer hardware-based compression for dataflow. This is a way to carry more application data within the same I/O limitations imposed by the distance. However the compression-factor is dictated by the nature of data itself and there is no practical possibility to measure the compressibility factor in advance (at least we do not know of any).

Recommendations

It is useful to make a detailed diagram of the whole SAN/WAN/LAN layout including IP-addresses for management ports and data-paths, maximum I/O-rates and bandwidth on the lines and contact details to the telecommunication providers.

It is also very interesting to have a monitoring of the telecommunication lines in place. This can be supplemented with performance monitoring on the storage as well.

When contracting with a telecommunication provider it is not always that clear that the Service Level Agreements (SLAs) do give you a means for insisting on guarantees of availability and support for the network services you buy. Penalties are often not existing at all. But that is something you will really need as your business might depend on the reliability of these communication lines.

Theoretically there is also the possibility to use DWDM and dedicated fiber for mirroring. But this is from a cost perspective usually not comparable to a leased line. However the WAN is even with leased lines typically the most expensive component of a DR solution. To save money on the bandwidth could mean that the solution does not work at all. Then the whole implementation is lost money as well.

9.2.4 Application considerations

There are several decisions or actions that can be taken at the host or application level to optimize your ERM implementation.

I/O blocksize

As a rule of thumb, make I/O blocksize as large as possible, while keeping performance acceptable. Indeed, with a bigger blocksize you will use less I/O for the same throughput.

Databases and file systems are designed with standard block sizes, which result from balancing the size of a block with server memory caching efficiency. Databases offer the ability to change block size, but the change is a disruptive process.

File system/database layout

Databases that make use of a file system for file access may have additional overhead that should be considered when designing and building the ERM (and DR) solution. Specifically, a journaled file system will require additional I/Os since an I/O is staged in the file system journal before it is written to the main area on disk. A journaled file system may result in increased cost for the ERM solution if it requires additional network bandwidth.

Whenever possible, avoid file systems and use raw devices to reduce this overhead as much as possible. This may not be important with smaller databases, but it becomes increasingly important as databases grow.

Temporary/scratch files and table spaces

Whenever possible, temporary/scratch files and temporary/scratch table spaces should be assigned to a separate logical drive that is *not* mirrored. These temporary data holders are typically used for reports or data conversions and are not part of the data that needs to be

mirrored to the recovery site. If mirrored, they will translate to additional I/Os (and possibly additional bandwidth cost) with no advantage in the event of a failover. The database or system administrator should be able to determine whether this condition exists and remedy it.

Another storage object that should not be replicated is the swap file, for the same reasons. In Windows environments, the swap file needs to be moved from the normal %systemroot% device to another drive if the %systemroot% device is to be replicated. Replicating the %systemroot% device is not recommended, so this may not be an issue.

DR guidelines for databases

Here we describe the possibilities of a DR method for a database using ERM. There are several ways to mirror a database for disaster recovery. Each method has advantages and disadvantages. Understanding them will help in designing the best solution.

What needs to be mirrored in a database? The two most popular database mirroring methods are (1) mirror everything and (2) mirror log files only.

Method 1 - mirror everything

With this approach, the entire database and the log files are mirrored. The advantage is that massive database updates will normally be handled without additional procedures or resources. There are several disadvantages, however. This approach is more susceptible to problems, it leads to a more complex solution, and it requires more network bandwidth.

A typical sequence of events for mirroring a database and logs is:

1. Establish mirroring for all database logical drives.
2. On a regular basis:
 - a. Put the primary database into hot backup mode.
 - b. Suspend mirroring to the recovery site.
 - c. Create a FlashCopy of the mirrored image at the recovery site.
 - d. Resume mirroring between sites.
 - e. Exit database hot backup mode at the primary site and resume normal operation.
 - f. Using the FlashCopy, backups and data migration tasks can now be performed at the recovery site.
 - g. Do not delete the FlashCopy.

It is best to always have at least one FlashCopy available at the recovery site in case database corruption at the primary site is mirrored to the recovery site (that is, eliminate the mirror as a single point of failure). A best practice is to keep several copies of the database at the recovery site for multiple recovery points. FlashCopy enables a quick and easy point-in-time process to accomplish this.

Method 2 - mirror log files only

Using this approach, the entire database is replicated to the recovery storage system initially, and then only the log files are mirrored thereafter (until it becomes necessary to replicate the database again). The logs are applied to the database at the remote site. This reduces the bandwidth required for mirroring, but also requires a server at the recovery site to apply the logs. If a massive database change is made, the entire database should be copied over again, which takes time and additional bandwidth temporarily. One significant advantage of this approach is that the two database images are truly separate from each other. In the event of database corruption at the primary site, the database is still intact at the recovery location.

A typical sequence of events for mirroring only log files is:

1. Establish mirroring for all database logical drives.
2. Suspend or remove mirroring of the database when synchronization is complete.
3. Continue mirroring the log files.
4. Then, on a regular basis:
 - a. Put the primary database into hot backup mode.
 - b. Suspend mirroring to the recovery site.
 - c. Create a FlashCopy of the log files at the recovery site.
 - d. Resume mirroring of log files between sites.
 - e. Exit database hot backup mode at the primary site and resume normal operation.
 - f. Using the FlashCopy, apply the log files to the recovery site database.

The FlashCopys may or may not be removed once the log files have been applied. Best practice retains at least one FlashCopy image of the log files.

We also recommend that the log file space on the primary storage system be able to retain a minimum of 24 hours of logs in case problems occur at the recovery site (such as a failure that prevents the log files from being applied). It is also important to make sure that the log files are not deleted on the primary side until they have been applied at the recovery side. This can be accomplished with scripts.

RTO and RPO will also influence the DR solution design. Does the data need to be synchronized at all times or can images be sent over the data link in batch mode on a regular basis (for example, hourly or at a shift change)? Can the organization afford to lose an hour or more of work? Must the DR solution provide an RPO within a single transaction? All these questions must be addressed during the RTO/RPO evaluation and answers must be provided in the design.

Other database recommendations

Consider multiplexing the log files (if this is supported by the database). It can reduce the performance impact. Most databases allow for multiple copies of log files. If the log files are multiplexed (two or three copies of the data instead of just one) and only one log image is replicated, the primary system will have nearly the same performance with replication as it does without it.

In a database environment, the best rule is to place the database logical drives and the log files into the consistency group if *mirror everything* is the approach selected. Otherwise, place the log files into the consistency group if *mirror log files only* is chosen.

9.2.5 Other design considerations

A DR solution addresses primary site failure by providing failover and recovery at a secondary site. If processing is to be resumed at the primary site at some point in the future, a failback process may be designed into the DR solution.

However, failback introduces new requirements for the DR solution design that affect both implementation and testing of the solution.

For example, failback adds the potential of losing transactions beyond the disaster recovery window. Any data with an RPO less than zero are subject to some amount of loss during disaster recovery failover and during failback. Failback also causes a second window of application downtime since it takes a finite amount of time to restore any application. The organization must be prepared to handle these realities.

If failback is not needed, it is much easier to test the DR solution because failover testing proceeds without taking down the primary systems. Failback testing requires that the primary site be used as a recovery site. Hence, normal production cannot be continued with the primary systems while failback tests are conducted.

If failback is required, a modified approach may possibly address the testing issue discussed above. The modified approach is to treat failback as a second failover and test it only after the first failover has occurred. In other words, delay testing failback until recovery has been completed at the secondary site. This involves starting the mirroring process over again from the secondary site that has now taken on primary status. The down side of this approach is that it could take longer to fail over to the original primary site than the failover/failback approach since the new failover process will have to be implemented and tested first. If resources at the secondary site do not provide adequate performance, this extra time may be unacceptable. Planning ahead for this modified approach will minimize the time required to get the original primary site up and running again.

With either approach, there is no guarantee that the original primary site will survive the disaster such that failback or the modified approach described above can actually be carried out. Therefore, a complete DR solution should plan for the possibility of never resuming operations at the original primary site.

To safely and adequately test failover and failback, multiple servers and large amounts of additional storage are required, which adds to the cost of the overall solution. One way to lessen this cost is to use equipment allocated to another project for the tests and then deploy it for the intended purpose after the planned DR testing exercise has finished.

Keep in mind, however, that DR testing is never complete. There must be an ongoing executive commitment to keep the DR solution current and to do regular testing to ensure that it remains a viable means to business continuity.

Failover/failback

The goal of a DR solution is to provide failover of selected applications to a recovery site that is not critically affected by the disaster. To restore full business operations after the disaster, failback to the primary site or failover to a third site (depending on the status of the primary site) may be required. This is an important consideration that should be factored into the DR solution design. As mentioned previously, failback can be complicated and needs careful planning before implementation and testing begin.

The simplest and preferred way to implement failback is to reverse the pre-disaster process of mirroring. Because the data integrity of primary storage is unknown after a disaster, it is necessary to copy back all of the recovery data to the primary site. A major concern is the amount of time it will take to perform the mirroring, which, in some cases, could take days or weeks, especially if the database has increased its size substantially since the failover occurred. Regular reviews of failback time requirements will ensure that failback time objectives are met. Reducing the time to fail back may require physically moving a fresh backup copy from the recovery site to the primary site. Once the database has been restored, mirroring the logs and applying them is accomplished in a much shorter time frame.

Testing failover and failback can be a significant challenge and needs to be included when defining the DR solution. In many cases, executive management will not want to do the testing, primarily because of the time it will take to fail back to the primary site. But these procedures must be tested to verify that they will work and to develop efficiency enhancements. Simulating failover is fairly easy and can be tested without a lot of effort. However, simulating failback can be a much more complicated task and requires additional equipment (including additional storage capacity or storage systems), which is why failback is often left out of the overall solution.

In cases where a planned failover is needed (for example, evacuating an area that is in the path of an approaching hurricane), shutting down the servers at the primary site, suspending the mirrors, and then reversing the direction of mirroring is an acceptable way to fail over quickly and facilitate rapid fallback.

If failover is triggered by an unplanned outage, a complete copy of the recovery database must be restored at the primary site before resuming operations. This is because servers at the primary site will likely be running when the failure occurs, so the two images of the database will not be synchronized. There is no practical way to merge the two images and guarantee data integrity, hence one site or the other must be designated as the master and the other must have its data overwritten with the master image.

9.3 Site readiness and installation checklist

ERM (and DR) projects are typically long-term projects. The time scale is months rather than weeks. IBM has developed a checklist for ERM that can serve as an initial project plan. This Site Readiness and Installation Checklist can be used together with the tool described in 9.4, “The Bandwidth Estimator Tool” on page 307.

The checklist reveals three phases of a basic project plan:

- Identify key items.

In this phase the physical locations, applications, and logical drives are identified. Also, the corresponding RTO and RPO expectations are set.

- What locations will be used as primary and secondary sites?
- What applications are to be mirrored?
- What are the associated volumes (LUNs) to be mirrored?
- What are the DR requirements and expectations?

- Measure application workloads.

This phase requires performance measurements of the mirrored logical drives. The measurements are used to calculate the necessary bandwidth.

The calculations can be done using IBM Bandwidth Estimator Tool (refer to 9.4, “The Bandwidth Estimator Tool” on page 307).

Refer to Chapter 6, “Analyzing and measuring performance” on page 187 for an overview of different methods and tools that can be used to measure performance. Those include IOMeter, iostat, PerfMon, PERFMON, the DS4000 Performance Monitor, and the Microsoft Windows Performance Monitor. Refer also to Chapter 7, “IBM TotalStorage Productivity Center for Disk” on page 231, which allows you to easily collect measurements performed over a long time period and can help you get a good picture for differences between average and peak workloads.

- What are the average measured I/Ops of each mirrored LUN?
- What are the peak I/Ops (for each mirrored LUN)?
- What are the average measured throughputs for each mirrored LUN?
- What are the peak throughputs?
- What are the average I/O block sizes of each mirrored LUN?
- What is the read/write ratio?
- What is the distance between the mirroring sites?

- Plan critical aspects.

In this phase the detailed design for the whole DR setup is planned and implemented. Network equipment as well as SAN infrastructure must be identified, and support of all involved components must be granted by the appropriate support process (if not

supported by the compatibility matrix). Note that for Europe an RPQ is always required for ERM distances longer than 10 km.

- Are all involved components of the setup being supported?
- Does the management station have access to both the primary and secondary storage subsystem?
- Does the telecommunication provider assure the minimum bandwidth and I/O rates?
- Is the network being tested?
- Is the mirroring performance being tested and adequate?

9.3.1 Site readiness and installation checklist details

In order to have a complete draft for a project plan we include here the Installation checklist and the reader can just print it out as needed.

This checklist should be used in conjunction with IBM System Storage DS4000 Disaster Recovery Design Guide to help plan, design, implement, and test a disaster recovery solution using the Enhanced Remote Mirroring capability of IBM DS4000 storage systems. It focuses on the technical aspects of ERM and will help with tracking and monitoring milestone progress toward ERM implementation. Used properly, it will help meet IBM goal of 100% customer satisfaction.

This checklist is designed for the disaster recovery project leader and provides a template for developing a basic project plan. Since every ERM implementation project is different, additional steps and checks should be added as required. See Table 9-2.

Table 9-2 Project Information

Business Name	ABC
Project name	ERM implementation
Project start date	9/7/06
Project leader	
Project leader e-mail	
Project leader phone	

In this phase, key items that affect DS4000 ERM implementation options are identified. They include applications and associated logical drives to be mirrored, along with the recovery objectives for each application. Primary and recovery sites are also identified.

Table 9-3 Phase-one milestones

Milestone	Date completed
Identify IT project leader and alternate.	
Identify primary and recovery sites.	
Identify key applications to be mirrored for disaster recovery.	
Identify associated logical drives.	
Identify data mirroring method to be used.	
Identify which logical drives must be in the ERM consistency group.	

Milestone	Date completed
Identify Recovery Point Objective (RPO) for each application.	
Identify Recovery Time Objective (RTO) for each application.	
Identify Network Recovery Objective (NRO) for each application.	
Identify expected capacity growth rate of applications to be mirrored.	

Table 9-4 Contact information

IT project leader	
E-mail	
Phone	
Alternate IT project leader	
E-mail	
Phone	

Table 9-5 Site information

Primary site location	
Install address	
Attention	
Contact information	
Ship to address (if different)	
Attention	
Contact information	

In this phase, performance measurements of the logical drives to be mirrored are taken. These measurements are critical to sizing required network bandwidth and latency. The measurements provide input for IBM ERM Bandwidth Estimator Tool. This tool identifies key network requirements for a successful disaster recovery solution. It also calculates minimum times required for initial synchronization and subsequent resynchronizations. It is critical that actual measurements be taken in this phase to size network bandwidth accurately.

Table 9-6 Phase-two milestones

Milestone	Date completed
Average I/Os per second measured for each logical drive (measurements taken for at least 24 hours a week is even better)	
Peak I/Os per second measured for each logical drive (measurements should be taken at peak monthly or quarterly processing times)	
Peak throughput in MBps measured for each logical drive (measurements should be taken at peak processing times)	
Average read/write ratio measured for each logical drive	
Block size determined for each logical drive	
Significant weekly/monthly patterns identified	
Current capacity determined for each logical drive	
Expected future capacity projected for each logical drive	
Circuit miles/kilometers identified between mirroring sites	
Telecommunications options identified and discussed with provider	
Measurements entered into ERM Bandwidth Estimator Tool	

In this phase, a detailed design of the disaster recovery solution is completed. Along with the DS4000 storage systems that are involved, several other components must be identified. They include SAN switches, network equipment, and network bandwidth and latency. The network components must support ERM according to IBM interoperability matrix. A network planning diagram should be created that contains the following information:

- ▶ TCP/IP addresses and masking information for the primary and recovery DS4000 storage systems
- ▶ Detailed information about the Fibre Channel SAN switches that will be used
- ▶ Detailed Information on the telecommunications devices that will be used
 - DWDM - number of ports, port numbers, number of lambdas
 - FCIP - device-specific information, compression used, encryption choice
 - FCP - device-specific information, compression used, encryption choice
 - Long-wave Fibre Channel - device-specific information

The communications network should be tested for distance and quality and must be operational before ERM can function. A DS4000 management station will need to have access to both the primary and recovery storage systems in order to configure ERM. Results obtained from the ERM Bandwidth Estimator Tool will identify required bandwidth and latency effects imposed by distance between sites.

Table 9-7 Phase-three milestones

Milestone	Date completed
IBM interoperability Web site consulted to ensure that all devices in the mirroring path are supported	
Telecommunications provider consulted to assure that minimum effective bandwidth and I/O rate will be provided	
Network planning diagram created	
Networking equipment installed and operational	
Networking settings and addresses set to optimum and recorded	
Minimum effective network bandwidth and I/O rate available and operational	
Network circuit numbers, device ports and settings set to optimum and recorded	
Management access provided to primary and recovery storage systems	
Primary and recovery logical drives set to EXACTLY the same size (Note: The ERM default varies with different DS4000 models.)	
Network tested for distance and quality	
Mirroring performance tested and adequate	

9.4 The Bandwidth Estimator Tool

We have explained that a successful remote mirroring solution is really dependent upon the network (WAN) bandwidth. Calculating the bandwidth is usually a rather complex task. To facilitate the task, IBM has developed a tool to help estimate minimum bandwidth ERM requirements for a given storage configuration.

Note: The Bandwidth Estimator Tool is not directly available to customers. Ask your IBM Service Representative.

This tool, known as the ERM Bandwidth Estimator Tool, is a Microsoft Excel®-based spreadsheet. The Bandwidth Estimator Tool can additionally help you determine the best ERM mirroring mode for your environment. It takes into account distances and the corresponding maximum number of IOPS that can be achieved with a remote mirroring solution.

The spreadsheet is divided into four tabs (sheets): Site Info, Logical Drive Info, Output, and How To. The tool is laid out conveniently to work the tabs from left to right, first entering required information and then viewing the results. You can work iteratively by simply updating the input information. However, previous inputs will be lost, so you should print all three tabs before making a change.

Site Info tab

You must start providing input within the Site Info tab, as shown in Figure 9-2.

Source Storage System Type: DS4800 = 64 Max Mirrors

Destination Storage System Type: DS4800 = 64 Max Mirrors

Network round-trip time between sites:

Method of measuring round-trip time: Distance

Select "RTT" to specify the actual network round-trip time between recovery sites, or select "Distance" to use an estimated round-trip distance between the sites.

Distance: 100 km

Maximum I/O rate per logical drive (Global Copy/Global Mirror): 1000

Planning for growth:

Expected total growth in storage array I/O requirements to be factored into calculations: 30%

Site Info / Logical Drive Info / Output / How To

Figure 9-2 Site Info tab

Enter the required information about the primary and recovery computing sites on this tab. Required fields are highlighted in bright yellow under the heading Network round-trip time between sites. The fields above this heading are provided for convenience and completeness when printing the tool's output. They are not used in any calculations.

In the field labeled Method of measuring round-trip time, select either **RTT** if you have measured the average network round-trip time (round-trip latency) between the two sites (this may be a result of using the ping command or through direct measurement of an existing telecommunications link), or **Distance** if you only know the geographical distance between sites.

If you select RTT, then in the Round-trip time field, enter the average round-trip time (in milliseconds) for a network packet to make a round trip between the two sites.

If you selected Distance in Method of measuring round-trip time, then enter the distance in the Distance field. You may select either *miles* or *km* as the unit of measurement.

Based on your entries, the maximum I/O rate (IOPS) per logical drive is calculated. This number is based solely on the network distance between the two sites and indicates the maximum number of I/Os per second that any one logical drive can achieve according to the laws of physics. Note that this calculation is a theoretical, best-case calculation. Communications protocols and link inefficiencies will likely reduce this maximum. The telecommunications provider should be consulted for more accurate information regarding this limitation.

The maximum I/O rate per logical drive is not necessarily the maximum aggregate I/O rate that can be achieved with the link, but it is the maximum that any one logical drive can achieve. Assuming that multiple logical drives are being mirrored, as long as the link has sufficient bandwidth and some mirrored I/Os are independent of each other, the independent I/Os can be multiplexed over the telecommunications link so that the aggregate I/O rate of all the logical drives can exceed the maximum I/O rate per logical drive.

If the maximum I/O rate per logical drive is the bottleneck for the mirroring solution, the only way it can be increased is to reduce the distance between primary and recovery sites.

Another point to consider when sizing ERM is the expected growth. Storage capacity demand is typically growing in the range of 30–40% per year. The IOPS demand is usually not growing at the same pace, and you can assume a rate of 15% per year or 30% for two years.

Logical Drive Info tab

The next tab, with required input, is Logical Drive Info, as shown in Figure 9-3.

C	D	E	F	G	H	J
Type of data stored on logical drive	Mirroring mode (Metro Mirror, Global Mirror, or Global Copy)	Logical drive size (GB)	Average block size (KB)	Peak I/O rate (I/Os per sec)	WRITE percentage	24-hour total Megabytes written
logs	Global Mirror	0.1	4	100	50%	13,504
table spaces	Global Mirror	300	8	200	50%	54,000
other	Global Copy	400	4	300	80%	8,100
	Metro Mirror					
	Global Mirror					
	Global Copy					

Figure 9-3 Logical Drive Info tab; required input

Column headings highlighted in yellow indicate required information. Column headings highlighted in green indicate output information.

Logical drive info input

Here you provide details about all logical drives (LUNs) that need to be mirrored.

The columns Application for the logical drive and Logical drive name or description are provided to describe the logical drives to be mirrored, but are not used in any calculations. They are merely descriptive entries provided for convenience and correlation of applications with logical drives.

In the column Type of data stored on the logical drive, choose one of the following: **logs** for database logs, **table spaces** for database table spaces, or **other** for all other types of data.

The next column is labeled Mirroring Mode, and you must select the desired type of mirroring to be used for the logical drive (metro mirror, global mirror, or global copy). The selected mode is used in calculating the requirements for the WAN connection.

Enter the size of the logical drive in gigabytes in the Logical drive size column.

In the Average block size column, enter the logical drive's average write block size in kilobytes. This is a critical factor used in the calculations.

Enter the measured peak I/O rate for the logical drive, in I/O operations per second, in the Peak I/O rate column. This is a critical factor used in the calculations. If a measured value cannot be obtained, enter an estimated or anticipated peak I/O rate, but understand that the output of the tool will only be as accurate as the estimate.

In the column labeled WRITE percentage, enter the average percentage of I/O operations for the logical drive that are writes (as opposed to reads). This is a critical factor used in the calculations.

Enter the total amount of data written during one typical 24-hour period in the column labeled 24-hour total Megabytes written. This factor is used in calculations for global mirror mode

Logical drive info output

Based on your input, you can now see the resulting output (green header columns). You can see, for instance, what is the *Peak Write I/O rate* per LUN, as illustrated in Figure 9-4.

K	M	T	U
Peak write I/O rate	Maximum link I/O rate allows peak write I/O rate for this logical drive using Global Mirror?	Global Copy/Global Mirror rate of data written (Mbytes/sec) (CALCULATED)	Metro Mirror rate of data written (MBytes/sec) (CALCULATED)
65	Yes	0.25	N/A
130	Yes	1.02	N/A
312	N/A	0.12	N/A
0	N/A	N/A	N/A
0	N/A	N/A	N/A

Figure 9-4 Peak write I/O per LUN

Peak write I/O rate is the peak number of I/O operations per second that are writes for the logical drive, calculated from the peak I/O rate, the WRITE percentage, and the Planning for growth fields.

The next column is labeled “Maximum link I/O rate allows peak write I/O rate for this logical drive using Global Mirror?” If the distance between the primary site and the recovery site is too long to support the selected global mode, this field will display No. Otherwise, if the distance is within acceptable calculated limits, it will display Yes. If the selected mode is metro mirror or global mirror, then N/A will be displayed. The calculation that determines a yes or no answer is a comparison of the peak write I/O rate for the logical drive and the maximum I/O rate per logical drive value that was calculated on the Site Info tab.

“Global Copy/Global Mirror rate of data written” shows the bandwidth required for global operations for this logical drive. If global mirror mode was specified, this is the product of the peak write I/O rate and the average block size divided by 1024. If global mirror mode was chosen, this is calculated by dividing the 24-hour total Megabytes written by the number of seconds in a day.

“Metro Mirror rate of data written” shows the bandwidth required if metro mirror mode was specified for this logical drive. It is the product of the peak write I/O rate and the average block size divided by 1024.

The Notes column contains information about detected problems with the selected configuration for the logical drive.

Other output results, as shown in Figure 9-5, include:

- ▶ *Maximum I/O rate per logical drive* simply repeats the value calculated for the Site Info tab. It is provided here for quick comparison when viewing the Peak write I/O Rate column.
- ▶ *Total bandwidth required: Global Copy/Global Mirror* is the aggregate WAN bandwidth needed to mirror all the logical drives that use global operations (global mirror or global copy). The value is the sum of bandwidths calculated in the global copy/global mirror rate of data written column multiplied by 10 bits/byte. This total bandwidth is required to enable WAN multiplexing for optimal performance.

W	X	Y	Z	AA
Maximum I/O rate per logical drive (Global Copy/Global Mirror):	Total bandwidth required: Global Copy/Global Mirror (Mbit/sec): CALCULATED	Target I/Os per second for Global Mirror (CALCULATED)	Total bandwidth Required: Metro Mirror (Mbit/sec): CALCULATED	Target I/Os per second for Metro Mirror (CALCULATED)
1,000	13.91	195	0	0

Figure 9-5 Target IOPS

- ▶ *Target I/Os per second for Global Mirror* is the aggregate peak I/O rate obtained by summing the values in the Peak write I/O Rate column that apply to logical drives that will use global mirror mode. This target I/O rate can only be achieved if all I/Os generated by these logical drives can be multiplexed and none of the peak write I/O rates are limited by the distance between sites (as identified in the Maximum link I/O rate allows peak write I/O rate for this logical drive using Global Mirror? column).
- ▶ *Total bandwidth required: Metro Mirror* is the aggregate SAN bandwidth needed to mirror all of the logical drives that use metro mirror mode. The value is the sum of bandwidths calculated in the metro mirror rate of data written column multiplied by 10 bits/byte. This total bandwidth is required to enable SAN multiplexing for optimal performance.
- ▶ *Target I/Os per second for Metro Mirror* is the aggregate peak I/O rate obtained by summing the values in the Peak write I/O Rate column that apply to logical drives which will use metro mirror mode. This target I/O rate can only be achieved if all I/Os generated by these logical drives can be multiplexed and none of the peak write I/O rates are limited by the distance between sites.

Output tab

The output tab summarizes the calculated SAN and WAN link requirements for a successful mirroring solution. Note the disclaimer that applies to the tool's output.

The minimum aggregate requirements for the metro mirror links are listed as *Total Metro Mirror bandwidth* requirement and *Total Metro Mirror I/O rate* requirement. Use this information to ensure that the metro mirror infrastructure will provide sufficient bandwidth and I/O rate.

The section titled Summary of Long-Distance Network Parameters shown in Figure 9-6 recaps critical factors that determine which the proposed global mirror/global copy mirroring solution (as characterized to the tool by the input fields) is expected to be successful.

Summary of Long-Distance Network Parameters

Total required long-distance bandwidth for global operations:	13.91 Mbit/sec
Distance between sites for global operations:	100 km
Best case round-trip time:	1.00 mSec
Maximum I/O rate per logical drive (Global Copy/Global Mirror):	1000 I/Os per sec

Long-Distance Network Connection Analysis for Global Modes

Connection type	Connection description	Able to support total required long-distance bandwidth for global operations?	Able to support peak write I/O rates of logical drives using Global Mirror?	Minimum initial synchronization time for logical drives using Global Mirror (hours)	Connection allows optimal write rates for global modes?
T1	1.54 Mbit/sec (N.Amer. & Japan)	No	Yes	554.30	No
E1	2.048 Mbit/s (Europe)	No	Yes	416.81	No
DS3 or T3	44.736 Mbit/sec each (N.Amer.)	Yes	Yes	19.08	Yes
T3	32.064 Mbit/s (Japan)	Yes	Yes	26.62	Yes
E3	34.368 Mbit/s (Europe)	Yes	Yes	24.84	Yes
OC-3	155.52 Mbit/sec each	Yes	Yes	5.49	Yes
OC-12	622.08 Mbit/sec each	Yes	Yes	1.37	Yes
OC-48	2.48 Gbit/sec each	Yes	Yes	0.34	Yes
DWDM Lambda	2.48 Gbit/sec each (FC Protocol)	Yes	Yes	0.34	Yes

Figure 9-6 Tab output with total bandwidth requirement for global operations

For metro mirror it is assumed that the Links run at FC speed and therefore its bandwidth requirements are not mentioned here.

First, the minimum aggregate bandwidth requirement for all logical drives that use either global copy or global mirror is listed as Total required long-distance bandwidth for global operations. This is the same value that was calculated in the Total bandwidth required: Global Copy/Global Mirror column of the Logical Drive Info tab. The tool assumes that all global operations apply to the same recovery site. Use this information to ensure that the WAN infrastructure will provide sufficient bandwidth. The telecommunications provider should be consulted to assure that at least this much effective bandwidth will be available for the mirroring solution.

The Distance between sites for global operations, Best case round trip time and Maximum I/O rate per logical drive (which is limited by the round trip time) are shown next for convenience and reference.

The table called *Long-Distance Network Connection Analysis for Global Modes* provides information useful for deciding on the WAN connection type required for the global mirror/global copy configuration. A variety of connection types are analyzed for their suitability. For each connection type, the geographic availability and maximum bandwidth of the connection are provided.

The third column of the table indicates whether the maximum bandwidth of the connection type is sufficient to support the total required long-distance bandwidth for global operations.

The next column indicates whether the WAN is able to support all peak write I/O rates of logical drives that use global mirror mode. This is a key criterion of acceptable performance that is based on the maximum I/O rate per logical drive, which in turn is based on the distance between sites. If the connection does not support all these peak write I/O rates, then those logical drives that have a No in the Maximum link I/O rate allows peak write I/O rate for this logical drive using Global Mirror? column of the Logical Drive Info tab will have their peak write I/O rate throttled by the link I/O rate.

The next column indicates the minimum amount of time it will take to do initial synchronization of all the data contained in the logical drives that use global mirror. Initial synchronization (replication) of the data stored on a logical drive is required before mirroring is started. Note that these calculations are theoretical, best-case calculations. Communications protocols, link inefficiencies, and other factors will likely increase them. The telecommunications provider should be consulted for more accurate information regarding initial synchronization time.

The final column summarizes whether the connection allows optimal write rates for global modes. Optimal write rates are achieved when the link has sufficient bandwidth to support the Total required long-distance bandwidth for global operations and the maximum I/O rate per logical drive does not throttle peak write I/O rates of any logical drives. Connections that do not allow optimal write rates may still function, but the performance of the mirroring solution will be adversely impacted.

Now, going back to the tab *Logical Drive Info*, you can find the column called Notes (see Figure 9-7). If you find a line that reads The I/O rate for the logical drive will be throttled by the link I/O rate, it means that the link will not be able to support the peak write I/O rates of LUNs that use global mirror.

V	W
Notes	Maximum I/O rate per logical drive (Global Copy/Global Mirror):
	100
The I/O rate for this logical drive will be throttled by the link I/O rate.	

Figure 9-7 IO throttling on Logical Drive Info tab

Such a situation would shown up on the Output tab, as illustrated in Figure 9-8.

A	B	C	D	E	F
Distance between sites for global operations:				1000 km	
Best case round-trip time:				10.00 mSec	
Maximum I/O rate per logical drive (Global Copy/Global Mirror):				100 I/Os per sec	
Long-Distance Network Connection Analysis for Global Modes					
Connection type	Connection description	Able to support total required long-distance bandwidth for global operations?	Able to support peak write I/O rates of logical drives using Global Mirror?	Minimum initial synchronization time for logical drives using Global Mirror (hours)	Connection allows optimal write rates for global modes?
T1	1.54 Mbit/sec (N.Amer. & Japan)	No	No	554.30	No
E1	2.048 Mbit/s (Europe)	No	No	416.81	No
DS3 or T3	44.736 Mbit/sec each (N.Amer.)	Yes	No	19.08	No
T3	32.064 Mbit/s (Japan)	Yes	No	26.62	No
E3	34.368 Mbit/s (Europe)	Yes	No	24.84	No
OC-3	155.52 Mbit/sec each	Yes	No	5.49	No
OC-12	622.08 Mbit/sec each	Yes	No	1.37	No
OC-48	2.48 Gbit/sec each	Yes	No	0.34	No
DWDM Lambda	2.48 Gbit/sec each (FC Protocol)	Yes	No	0.34	No
<p>Note: If the connection does not allow for optimal write rates for Global Mirror, then the I/O rate for logical drives using Global Mirror will be throttled by the link I/O rate.</p>					
<p>Site Info Logical Drive Info Output How To</p>					

Figure 9-8 IO throtteling on Output tab



SVC guidelines for DS4000

This chapter very briefly introduces IBM System Storage SAN Volume Control (SVC), describing its components and concepts. It compares the SVC Copy Services with those of the DS4000. It focuses on general best practices and guidelines for the use of SVC with the DS4000 Storage Server, and includes a configuration example.

10.1 IBM System Storage SAN Volume Controller overview

The IBM System Storage SAN Volume Controller is a scalable hardware and software solution that provides block aggregation and logical drive management for different disk storage subsystems in a SAN.

SVC delivers a single view of the storage attached to the SAN. Administrators can manage, add, and migrate physical disks non disruptively even between different storage subsystems. The SAN is zoned in such a way that the application servers cannot see the back-end storage, preventing any possible conflict between SVC and the application servers both trying to manage the back-end storage.

The SAN Volume Controller provides storage virtualization by creating a pool of managed disks from attached back-end disk storage subsystems. These managed disks are then mapped to a set of virtual disks for use by various host computer systems:

- ▶ Performance, availability, and other service level agreements can be mapped by using different storage pools and advanced functions.
- ▶ Dependencies, which exist in a SAN environment with heterogeneous storage and server systems, are reduced.

In conjunction with the DS4000 Storage Server family, the SAN Volume Controller (SVC) can enhance the copy services functionality and also the flexibility of managing the DS4000 storage.

The SVC also offers an alternative for the DS4000 copy services including FlashCopy, VolumeCopy, and Enhanced Remote Mirroring. If you are planning to use the SVC with your DS4000 Storage Server, then you may not need to purchase these DS4000 premium features. The additional benefits are:

- ▶ FlashCopy works even between different storage systems.
- ▶ Metro mirror/global mirror is done on the SAN level.
- ▶ Data migration between storage subsystems can be performed without application interruption.
- ▶ Data migration from existing native storage to SVC storage can be performed with minimal interruption.

SVC is very flexible in its use. It can be used to manage all of your disk storage requirements or just part. In other words, when using SVC with the DS4000, you can still use the DS4000 Storage Manager functions to allocate part of the storage to some hosts.

SVC offers a large scalable cache.

SVC may also reduce the requirement for additional partitions. The SVC only consumes one storage partition for each storage server that connects to it. If you plan to use the SVC for all your hosts then a storage partition upgrade on the DS4000 may not be required. In addition it improves capacity utilization. Spare capacity on underlying physical disks can be reallocated non disruptively from an application server point of view irrespective of the server operating system or platform type.

SVC simplifies device driver configuration on hosts, so all hosts within your network use the same IBM device driver to access all storage subsystems through the SAN Volume Controller.

SVC is licensed by the capacity that is being managed. The features such as FlashCopy, metro mirror, and global mirror are part of the feature set included with the SVC. The licence cost is for the capacity of all storage managed by the SVC plus the capacity of the copy services maintained by the SVC. Capacity upgrades can be done at anytime during the life of the SVC by purchasing a licence for the additional capacity required.

SVC supports a wide variety of disk storage and host operating system platforms. For the latest information refer to IBM Web site:

<http://www.ibm.com/servers/storage/software/virtualization/svc/interop.html>

Next we review the basic SVC concepts and components. For details and information about SVC implementation refer to:

- ▶ *IBM System Storage SAN Volume Controller, SG24-6423*
- ▶ *IBM TotalStorage: Integration of the SAN Volume Controller, SAN Integration Server and the SAN File System, SG24-6097*

10.2 SVC components and concepts

An SVC implementation consists of both hardware and software. The hardware consists of a management console server, a minimum of two node pairs to form the SVC cluster, and a minimum of two UPSs.

Here we define some of the concepts and terminology used with SVC.

SVC cluster

When the first two nodes are installed, they form an SVC cluster. An SVC cluster can contain up to eight nodes (four pairs of nodes). All of the nodes should be located in close proximity to each other. To span longer distances it is best to create two separate clusters.

Master console

The master console is the platform on which the software used to manage the SVC runs. It is a server running Windows 2003 and has installed the SVC management console software, Storage Manager Client, Tivoli Productivity Centre (TPC), and IBM Director. This Storage Manager is the same version as used to manage the creation of LUNs and the allocation of storage from the DS4000 to the SVC.

Node

A node is a name given to the individual servers in a SAN Volume Controller cluster on which the SVC software runs. Nodes are always installed as pairs. Each node must be connected to its own UPS.

Managed disks

A managed disk (mDisk) is a SCSI disk presented by the storage (for instance, a DS4000 logical drive) and managed by the SVC. A managed disk provides usable blocks (or extents) of physical storage to the SVC cluster. The mDisks are not visible to hosts systems on the SAN.

An mDisk can be in one of three states in SVC: unmanaged, managed, or image mode.

- ▶ An unmanaged disk is one that has been assigned to the SVC but not yet managed by SVC.
- ▶ A managed disk is one that is managed by SVC.

- An image mode disk is one that has been imported into SVC and contains data. This is used when migrating existing LUNs (logical drives) with data into SVC. The data on the LUN is preserved. To extend the disk once managed by the SVC, this type of disk must be migrated to an mDisk.

Managed disk group

The managed disk group (MDG) is a collection of mDisks. Each MDG is divided into a number of *extents*, which are numbered sequentially, starting with 0. When creating a MDG you must choose an extent size. Once set, the extent size stays constant for the life of that MDG and cannot be changed. Each MDG may have a different extent size.

Extents

An extent is a fixed size unit of data that is used to manage the mapping of data between MDisks and VDisks. The extent size choices are 12, 32, 64, 128, 256, or 512 MB. The choice of extent size effects the total storage that can be managed by the SVC. A 16 MB extent size supports a maximum capacity of 64 TB.

Virtual disks

The virtual disks (VDisk) is a logical entity that represents extents contained in one or more mDisks from a MDG. VDisks are allocated in a whole number of extents. The vDisk is then presented to the host as a LUN for use.

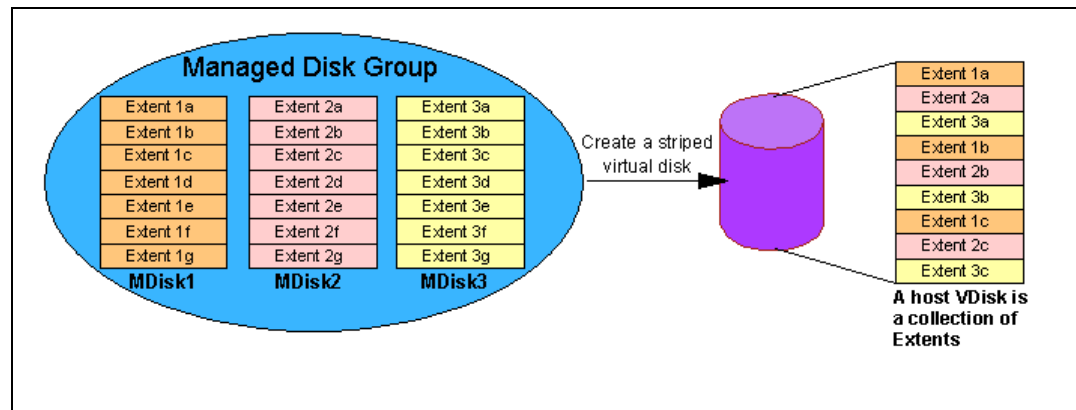


Figure 10-1 Extents used to create a vDisk

Note: The SAN Volume Controller is not a RAID controller. The disk subsystems attached to SANs that have the SAN Volume Controller provide the basic RAID setup. The SAN Volume Controller uses what is presented to it as a managed disk to create virtual disks.

I/O group

An I/O group contains two nodes. These are configured as a pair. Each node is associated with only one I/O group. The nodes in the I/O group provide access to the vDisks in the I/O group. Each vDisk is associated with exactly one I/O group. See Figure 10-2.

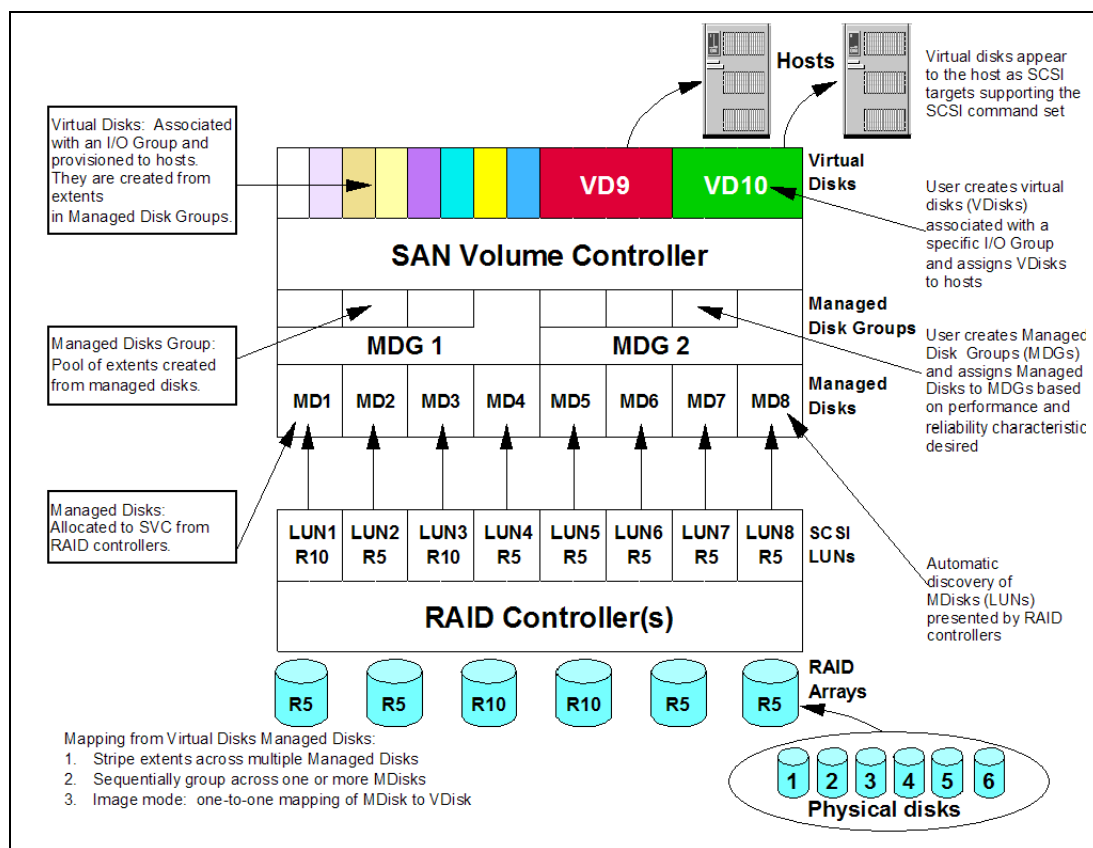


Figure 10-2 Relationships between physical and virtual disks

Consistency group

Consistency groups address the issue where the objective is to preserve data consistency across multiple vDisks, because the applications have related data which spans multiple vDisks. A requirement for preserving the integrity of data being written is to ensure that *dependent writes* are executed in the application's intended sequence.

Multipath device drivers

The multipath drivers used by the SVC are found at IBM Web site below. The drivers required are typically the Subsystem Device Driver (SDD). These drivers are available from IBM Web site.

For Windows environments there is a choice of two device drivers for SVC, SDD or Subsystem Device Driver Device Specific Module (SDDDSM) is a multipath I/O driver based on Microsoft MPIO technology. These drivers are designed to provide multipath support for hosts systems accessing logical volumes from the Enterprise Storage Server (ESS), SVC, and the DS8000 family of storage servers.

Always refer to IBM SVC Web site for the latest device drivers and the support matrix:

<http://www.ibm.com/storage/support/2145>

10.3 SVC copy services

The copy services offered by the SVC are FlashCopy, metro mirror, and global mirror. If you plan to use SVC for all of your copy services, then purchasing the additional premium features, such as FlashCopy, VolumeCopy, or Enhanced Remote Mirroring for the DS4000, may not be necessary.

The SVC copy services functions provide additional capabilities over the similar DS4000 functions. For instance:

- ▶ SVC supports consistency groups for FlashCopy, metro mirror, and global mirror.
- ▶ Consistency groups in SVC can span across underlying storage systems.
- ▶ FlashCopy source volumes that reside on one disk system can write to target volumes on another disk system.
- ▶ Metro mirror and global mirror source volumes can be copied to target volumes on a dissimilar storage system.

10.3.1 SVC FlashCopy

FlashCopy provides the capability to perform an instantaneous point-in-time (PiT) copy of one or more VDisks. This is a copy of the VDisk at that PiT. Once created, it no longer requires the source to be active or available.

FlashCopy works by defining a FlashCopy mapping consisting of a source VDisk and a target VDisk. Multiple FlashCopy mappings can be defined, and PiT consistency can be observed across multiple FlashCopy mappings using consistency groups.

Note: As is the case with the DS4000, the first step before invoking this function is to make sure that all of the application data is written to disk — in other words, that the application data is consistent. This can be achieved, for example, by quiescing a database and flushing all data buffers to disk.

When FlashCopy is started, it makes a copy of a source VDisk to a target VDisk, and the original contents of the target VDisk are overwritten.

When the FlashCopy operation is started, the target VDisk presents the contents of the source VDisk as they existed at the single point in time (PiT) the FlashCopy was started.

When a FlashCopy is started, the source and target VDisks are instantaneously available. This is so because when started, bitmaps are created to govern and redirect I/O to the source or target VDisk, respectively, depending on where the requested block is present, while the blocks are copied in the background from the source to the target VDisk.

Both the source and target VDisks are available for read and write operations, although the background copy process has not yet completed copying across the data from the source to target volumes. See Figure 10-3.

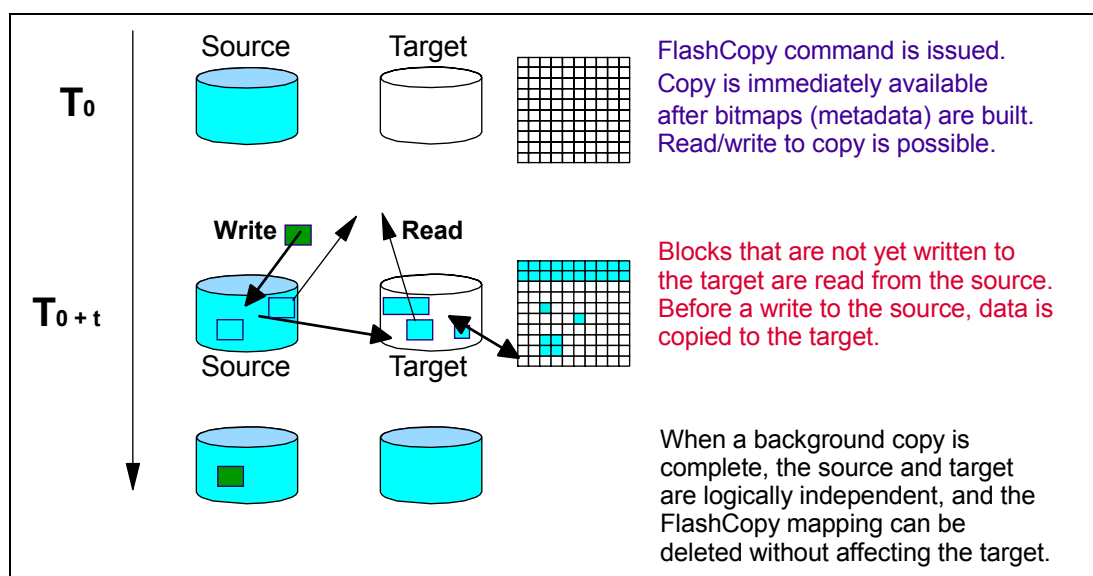


Figure 10-3 Implementation of SVC FlashCopy

Compared to the DS4000, the SVC FlashCopy is more like a combination of the DS4000 FlashCopy and VolumeCopy functions.

10.3.2 Metro mirror

The general application of metro mirror is to maintain two real-time synchronized copies of a data set. Often, the two copies are geographically dispersed on two SVC clusters, though it is possible to use metro mirror in a single cluster (within an I/O group). If the primary copy fails, the secondary copy can then be enabled for I/O operation.

Metro mirror works by defining a metro mirror relationship between VDisks of equal size. When creating the metro mirror relationship, one VDisk should be defined as the master, and the other as the auxiliary. The contents of the auxiliary VDisk that existed when the relationship was created are destroyed.

To provide management (and consistency) across a number of metro mirror relationships, consistency groups are supported (as with FlashCopy).

The SVC provides both intracluster and intercluster metro mirror, as described below.

Intracluster metro mirror

Intracluster metro mirror can be applied within any single I/O group. Metro mirror across I/O groups in the same SVC cluster is not supported.

Intercluster metro mirror

Intercluster metro mirror operations require a pair of SVC clusters that are separated by a number of moderately high bandwidth links. The two SVC clusters must be defined in an SVC partnership, which must be performed on both SVC clusters to establish a fully functional metro mirror partnership.

Using standard single mode connections, the supported distance between two SVC clusters in a metro mirror partnership is 10 km, although greater distances can be achieved by using extenders.

A typical application of this function is to set up a dual-site solution using two SVC clusters where the first site is considered the primary production site, and the second site is considered the failover site, which is activated when a failure of the first site is detected.

Metro mirror is a fully synchronous remote copy technique, which ensures that updates are committed at both primary and secondary VDisks before the application is given completion to an update. As shown in Figure 10-4, a write to the master VDisk is mirrored to the cache for the auxiliary VDisk before an acknowledge of the write is sent back to the host issuing the write.

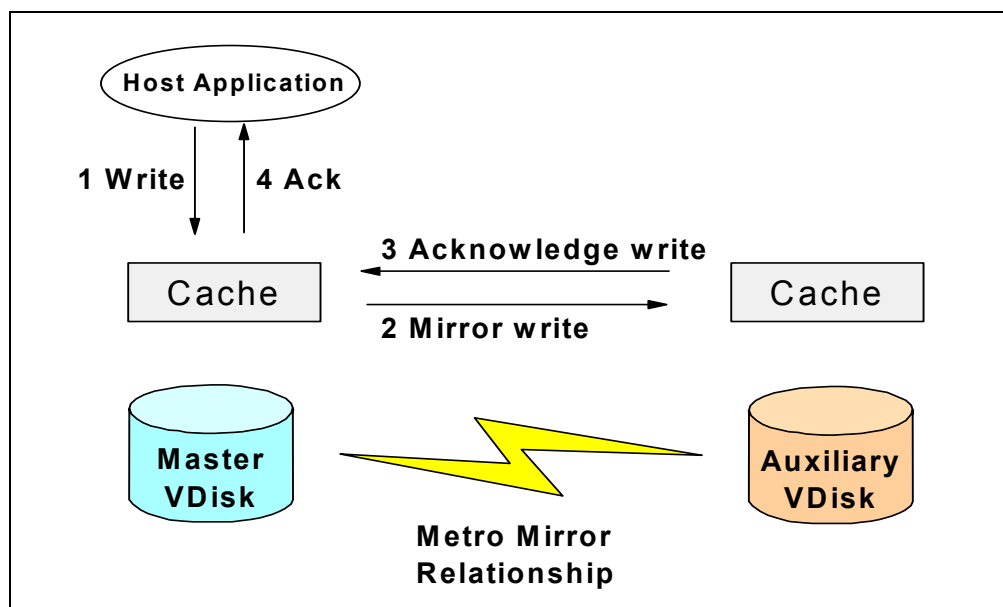


Figure 10-4 Remote Mirroring synchronous write sequence

While the metro mirror relationship is active, the secondary copy (VDisk) is not accessible for host application write I/O at any time. The SVC allows read-only access to the secondary VDisk when it contains a consistent image. To enable access to the secondary VDisk for host operations, the metro mirror relationship must first be stopped.

10.3.3 Global mirror

Global mirror (GM) works by defining a GM relationship between two VDisks of equal size and maintains the data consistency in an asynchronous manner.

When creating the global mirror relationship, one VDisk is defined as the master, and the other as the auxiliary. The relationship between the two copies is asymmetric. While the global mirror relationship is active, the secondary copy (VDisk) is not accessible for host application write I/O at any time. The SVC allows read-only access to the secondary VDisk when it contains a consistent image.

When a host writes to a source VDisk, the data is copied to the source VDisk cache. The application is given an I/O completion while the data is sent to the target VDisk cache. At that stage, the update is not necessarily committed at the secondary site yet. This provides the

capability of performing remote copy over distances exceeding the limitations of synchronous remote copy.

Figure 10-5 illustrates that a write operation to the master VDisk is acknowledged back to the host issuing the write before it is mirrored to the cache for the auxiliary VDisk.

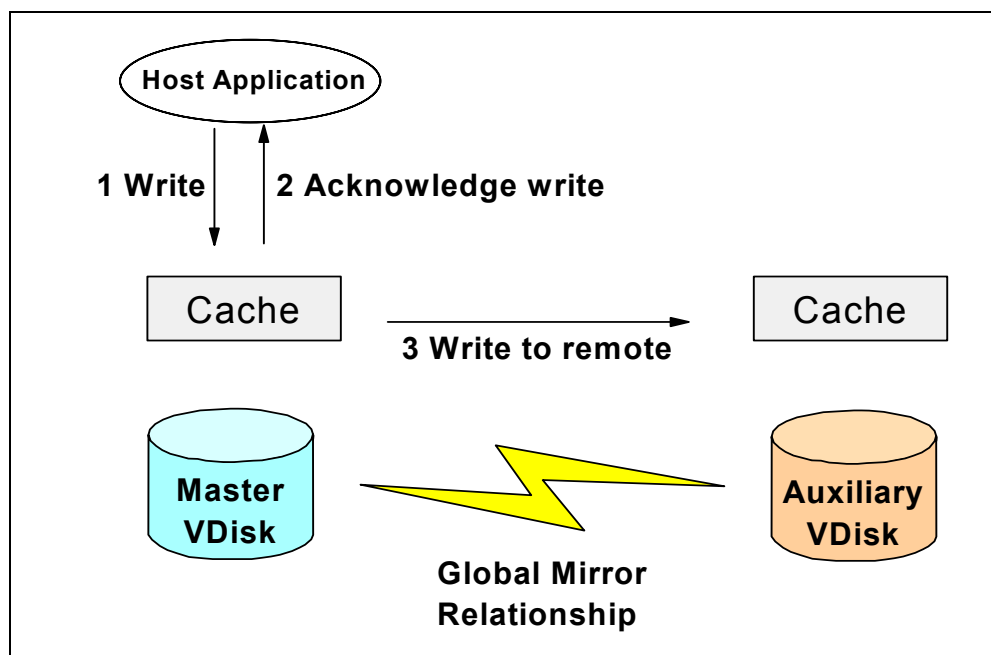


Figure 10-5 Global mirror asynchronous write sequence

To provide management (and consistency) across a number of metro mirror relationships, consistency groups are supported.

The SVC provides both intracluster and intercluster global mirror, which are described below.

Intracluster global mirror

Although global mirror is available for intracluster, it has no functional value for production use. Intracluster metro mirror provides the same capability for less overhead. However, leaving this functionality in place simplifies testing and allows experimentation and testing (for example, to validate server failover on a single test cluster).

Intercluster global mirror

Intercluster global mirror operations require a pair of SVC clusters that are commonly separated by a number of moderately high bandwidth links. The two SVC clusters must each be defined in an SVC cluster partnership to establish a fully functional global mirror relationship.

10.3.4 Differences between DS4000 and SVC copy services

This section briefly discusses the differences between the DS4000 Copy Services the DS4000 and the SVC Copy Services. Whenever a DS4000 is managed through an SVC, use the SVC copy services for the SVC managed volumes, rather than the DS4000 copy services.

FlashCopy

On the DS4000 a target FlashCopy logical drive is not a physical copy of the source logical drive. It is only the logical equivalent of a complete physical copy because only the changed blocks are copied to the target (copy on write). Consequently, if the source logical drive is damaged in any way, the FlashCopy logical drive cannot be used for recovery.

On the SVC, however, the FlashCopy creates a point-in-time image of the logical drive. This is the physical copy of the source physical drive. Once created, it no longer requires the source copy to be active or available. This image, once created, is available to be mounted on other host servers or for recovery purposes.

To get the equivalent function on the DS4000, you have to combine FlashCopy and VolumeCopy functions (that is, make a VolumeCopy of the FlashCopy target).

Metro mirror and global mirror

Both the DS4000 Copy Services and the SVC Copy Services offer a metro mirror and a global mirror. These copy services are similar in function between the two technologies, even though the underlying technology that creates and maintains these copies are different.

From a DS4000 standpoint an important difference when using SVC copy services over the DS4000 Copy services is in host port requirement. The DS4000 Mirroring requires a dedicated host port from each of the DS4000 controllers. This means that you need a DS4000 model with four host ports per controller (such as the DS4700 Model 72 or the DS4800) if you want to still have dual redundant host SAN fabrics when Remote Mirroring is implemented.

The SVC does not dedicate ports for its mirroring services (nor does it required dedicated ports on the DS4000). Rather, a zoning configuration dedicates paths between the SVC clusters.

Other advantages of SVC mirroring include the following:

- ▶ SVC maintains a control link on top of the FC link that is used for the global/metro mirror I/O. No IP connection is needed to manage both SVC clusters from one centralized administration point.
- ▶ SVC has a huge cache (depending on the number of SVC cluster nodes).
- ▶ SVC makes it possible to mirror between different IBM DS models and also between storage from different vendors. This gives a possibility for data migration between different storage.
- ▶ SVC supports intracluster mirror, where both VDisks belong to the same cluster (and I/O group).
- ▶ Intercluster and intracluster mirror can be used concurrently within a cluster for different relationships.
- ▶ SVC implements a configuration model that maintains the mirror configuration and state through major events such as failover, recovery, and resynchronization to minimize user configuration action through these events.
- ▶ SVC implements flexible resynchronization support, enabling it to re-synchronize VDisk pairs that have suffered write I/O to both disks and to resynchronize only those regions that are known to have changed.

Some disadvantages should, however, be mentioned:

- ▶ SVC is an additional layer that has to be administered. It will need resources to cover this.
- ▶ SVC will be cost intensive in a small environment.

- The same physical restrictions to database I/O over long distance in a disaster recovery environment apply to the SVC as well. There is no advantage on this point in comparison to DS4000 mirroring.

Consistency groups

The SVC can use consistency groups for all copy services. The use of consistency groups is so that the data is written in the same sequence as intended by the application. The DS4000 uses consistency groups on global mirrors only.

Premium features

With the DS4000 you must purchase premium features such as FlashCopy, VolumeCopy, Enhanced Remote Mirror Copy Services, and additional storage partitioning to connect different hosts.

With the SVC, licensing is per capacity managed (note that this capacity is the total capacity of managed volumes *and* copy services volumes). The copy services themselves are part of the product and are not licensed separately.

SVC consumes only one storage partition. All of the hosts that are managed by the SVC are virtualized behind that single storage partition. Of course, if you only partially manage the DS4000 through SVC, you still need the additional storage partitions on the DS4000 for the various hosts that need access to the logical drives not managed through SVC.

10.4 SVC maximum configuration

Table 10-1 is a summary of maximum configuration numbers supported for SVC at the time of writing. These numbers may change with subsequent hardware and software releases. Always check IBM Web site for the latest configuration information.

Table 10-1 Some of the SVC maximum configuration numbers.

Description	Maximum numbers	Comments
Nodes	8	Four pairs of nodes per cluster
I/O groups	4	One I/O group per node pair
MDisk groups	128	
MDisks	4096	Represents an average of 64 per controller
MDisks per managed disk group	128	
MDisk maximum size	2 TB	
Maximum addressable storage	2.1 PB	Using extent size of 512 MB
VDisks	4096	
Vdisks per host	512	Limit many depend upon host operating system
VDisks per I/O group	1024	
Storage systems	64	

Description	Maximum numbers	Comments
Metro mirror relationships per cluster	4096	
Metro mirror consistency groups	32	
FlashCopy mappings	2048	Up to 512 mappings per consistency group
FlashCopy consistency groups	128	
FlashCopy VDisk per I/O group	16 TB	
SAN fabrics	2	Dual fabric configurations

10.5 SVC considerations

It is important to know the following restrictions when considering SVC:

► Cluster considerations

- Two SVC clusters cannot share the same disk subsystem. Data corruption and data loss might occur when the same MDisk becomes visible on two different SVC clusters.
- All nodes in an SVC cluster should be located close to one another, within the same room or adjacent rooms for ease of service and maintenance. An SVC cluster can be connected (via the SAN fabric switches) to application hosts, disk subsystems, or other SVC clusters, via short wave only optical FC connections.

Long wave connections are no longer supported.

With short wave, distances can be of up to 150 m (short wave 4 Gbps), 300 m (short wave 2 Gbps), or 500 m (short wave 1 Gbps) between the cluster and the host, and between the cluster and the disk subsystem.

Longer distances are supported between SVC clusters when using inter cluster Metro or global mirror.

► Node considerations

- SVC nodes like the 4F2 and 8F2 always contain two host bus adapters (HBAs), each of which has two Fibre Channel (FC) ports. If an HBA fails, the configuration remains valid, but the node operates in degraded mode. If an HBA is physically removed from an SVC node, then the configuration is unsupported. The 8F4 has one HBA and four ports.
- A node will not function unless it is behind the appropriate UPS unit. That is one of the design considerations with the SVC. Even though the room may already be supplied with UPS-protected power, a dedicated UPS unit for the node is required.

► Network considerations

All nodes in a cluster must be on the same IP subnet. This is because the nodes in the cluster must be able to assume the same cluster, or service IP address.

► Fabric considerations

- SVC node ports must be connected to the Fibre Channel fabric only. Direct connections between SVC and host, or SVC and disk subsystem, are not supported.
- The Fibre Channel switch must be zoned to permit the hosts to see the SVC nodes, and the SVC nodes to see the disk subsystems. The SVC nodes within a cluster must

be zoned in such a way as to allow them to see each other, the disk subsystems, and the front-end host HBAs.

- The Fibre Channel SAN connections between the SVC node can run at either 1 Gbps, 2 Gbps, or 4 Gbps depending on the SVC model and switch hardware.
 - The 8F4 SVC nodes are 4 Gbps capable and auto negotiate the connection speed with the switch.
 - The 4F2 and 8F2 nodes are capable of a maximum of 2 Gbps, which is determined by the cluster speed.
- When a local and a remote fabric are connected together for metro mirror purposes, then the ISL hop count between a local node and a remote node cannot exceed seven hops.

Expanding disks

It is possible to expand a VDisk in the SVC cluster, even if it is mapped to a host. Some operating systems, such as Windows 2000 and Windows 2003, can handle the volumes being expanded even if the host has applications running.

However, a VDisk that is defined to be in a FlashCopy, metro mirror, or global mirror mapping on the SVC cannot get expanded unless the mapping is removed, which means that the FlashCopy, metro mirror, or global mirror on that VDisk has to be stopped before it is possible to expand the VDisk.

Multipathing

Each SVC node presents a virtual disk (VDisk) to the SAN through four paths. Because in normal operation two nodes are used to provide redundant paths to the same storage, this means that a host with two HBAs can see eight paths to each LUN presented by the SVC. We suggest using zoning to limit the pathing from a minimum of two paths to the maximum available of eight paths, depending on the kind of high availability and performance you want to have in your configuration.

The multipathing driver supported and delivered by SVC is IBM Subsystem Device Driver (SDD). The number of paths from the SVC nodes to a host must not exceed eight, even if this is not the maximum paths number handled by SDD. The maximum number of host HBA ports must not exceed four.

Zoning

A storage zone must exist that comprises all SVC ports and all ports on the disk subsystem; and there should be multiple host zones, each of which consists of one host HBA port and one port from each of the two SVC nodes.

10.6 SVC with DS4000 best practices

There are several best practices and guidelines to follow when a DS4000 is used through SVC so that it performs well with most applications.

The key is to plan how the DS4000 disks (logical disks or LUNs) will be allocated and used by SVC since LUNs (which are the actual disks seen by SVC) have to be created first on the DS4000.

The simplified process of allocating disk from the storage array to the SVC is listed below:

1. The LUN is created from the array using Storage Manager.
2. The LUN is presented to SVC by assigning the LUN to the SVCHost.
3. The LUN is discovered by the management console and then gets created as an MDisk.
4. The MDisk is assigned to an MDG. The MDG determines the extent size.
5. The MDG is assigned to an I/O group.
6. The MDisk is then used to create VDisks that are presented to the host.
7. Create a host on the SVC.
8. Map the VDisks to the host.

LUN creation

The storage arrays created on the DS4000 when defining LUNs that will be assigned to the SVC should be configured with only one LUN per array. The LUN should be sized to utilize all of the available storage space on the array.

The decision for array RAID level and which DS4000 physical disks it contains is made when creating LUNs on the DS4000 before it is defined as an SVC MDisk and mapped to an SVC VDisk. It is essential to know at that point which particular host and the nature of the host applications that will access the VDisk.

In other words, key decisions for reliability, availability, and performance are still made at the DS4000 level. You should thus still follow the recommendations given specifically for the DS4000 in previous chapters of this book.

Here we summarize some of the most important ones:

- ▶ For data protection, ensure that the array is created with enclosure loss protection so that if an enclosure fails on the DS4000, the array will still remain functional.
- ▶ For database transaction logs, RAID 10 gives maximum protection. Database transaction logs require sequential writes to the log files. These must be well protected and must deliver high performance. For database data files, RAID 5 offers a balance between protection and performance.

For best practice, the transaction logs and data files are best to be in a dedicated array:

- Create an array with one LUN of the appropriate size for the logs. Since this DS4000 LUN will then be defined as an SVC MDisk and assigned the a MDG, make sure that the LUN size is appropriate for the extent size assigned to this MDG.
 - Map the whole MDisk (all extents) and only extents from that MDisk to the VDisk. Doing so ensures that the host has dedicated use of that physical disk (DS4000 LUN). This also ensures that no other VDisks will be mapped to that MDisk and possibly cause disk thrashing.
 - For applications with small, highly random I/Os, you can stripe VDisks across multiple MDisks (from multiple DS4000 arrays or LUNS) for performance improvement. However, this is only true with applications that generate small high random I/Os. In a case of large blocksize I/Os that are mainly sequential data, using multiple MDdisks to stripe the extents across does not give a performance benefit.
- ▶ For file and print environment, RAID 5 offers a good level of performance and protection. If their requirements are not too demanding, several host applications could use the same MDisk (that is, they would use separate VDisks but carved from extents of the same MDisk). In other words, in this case several hosts would share the same DS4000 LUN through SVC. Still, you should monitor the DS4000 LUN (array) to ensure that excessive disk thrashing is not occurring. If thrashing occurs then the VDisk can be migrated to another MDG.

We also recommend that only LUNs from one disk subsystem such as a specific DS4000 form part of an MDG. This is mainly for availability reasons, since the failure of one disk subsystem will make the MDG go offline, and thereby all VDisks mapped to MDisks belonging to the MDG will go offline.

Preferred controller

The SVC attempts to follow IBM DS4000 series specified preferred ownership. You can specify which controller (A or B) is used as the preferred path to perform I/O operations to a given LUN. If the SVC can see the ports of the preferred controller and no error conditions exist, then the SVC accesses that LUN through one of the ports on that controller. Under error conditions, the preferred ownership is ignored.

VDisks and extent size

VDisks are allocated in whole numbers of extents so the VDisk size should be created as multiples to the extent size, so as not to waste storage at the end of each VDisk. The extent size is determined by the MDG. If the extent size is 16 MB then the VDisk should be created in multiples of 16 MB.

Disk thrashing

The SVC makes it simple to carve up storage and allocate it to the hosts. This also could cause problems if multiple hosts use the same physical disks and cause them to be very heavily utilized. Multiple hosts using the same array is not a problem in itself, it is the disk requirements of each host that may cause the problems. For example, you may not want a database to share the same disks as the e-mail server or a backup server. The disk requirements may be very different. Careful planning should be undertaken to ensure that the hosts that share the same array do not cause a problem due to heavy I/O operations.

Dual SAN fabric

To meet the business requirements for high availability, SAN best practices recommend the building of a dual fabric network (that is, two independent fabrics that do not connect to each other). Resources such as DS4000 Storage servers have two or more host ports per controller. These are used to connect both controllers of the storage server to each fabric. This means that controller A in the DS4000 is connected to counterpart SAN A, and controller B in the DS4000 is connected to counterpart SAN B. This improves data bandwidth performance. However, this does increase the cost, as switches and ports are duplicated.

Distance limitations

Ensure that the nodes of the cluster do not exceed any of the distance limitations. Ideally, you will want all nodes in the same room. If you wish to have SVC in separate locations that are in different buildings or farther, then create a separate cluster. For disaster recovery and business continuity purposes, using the appropriate copy services between the two clusters will provide the level of protection required.

Similarly, if you have more than one data center then you would want and SVC cluster in each data center and replicate any data between the clusters.

10.6.1 DS4000 Storage Server family and SVC configuration example

In our example we have a DS4800 Storage Server for which some LUNs are used by a host connected to SVC and some LUNs are directly mapped to other hosts (non-SVC hosts).

The SVC is a two-node cluster. Each node has two HBAs, but only ports 1 and 4 of each node are currently being used.

A host group is created in Storage Manager. The name of the group should reflect something meaningful to your environment to signify that the group is an SVC group. In the example shown in Figure 10-7 on page 333 we named the host group SVCGroup.

Next, a host is created. The host name should conform to your naming convention. In the example shown in Figure 10-7 on page 333 we called the host SVCHost. This is the only host definition required is Storage Manager for all the hosts or servers that will access the storage subsystem using SVC. The ports assigned are the ports of the SVC nodes.

The LUNs created with the Storage Manager client are then assigned to the host SVCHost. These can now be discovered with the SVC management console and assigned as SVC managed disks.

Zoning for non SVC hosts

Zoning rules for hosts that will not use the SVC remain unchanged. They should still follow the best practice where for each host, multiple zones are created such that each host HBA is paired with a specific controller on the DS4000 server. For example, for a host with two HBAs the zoning would be as follows:

- ▶ Host zone 1: HBA_1 is in a zone with DS4800 controller A.
- ▶ Host zone 2: HBA_2 is in a zone with DS4800 controller A.
- ▶ Host zone 3: HBA_1 is in a zone with DS4800 controller B.
- ▶ Host zone 4: HBA_2 is in a zone with DS4800 controller B.

Zoning for SVC and hosts that will use the SVC

All SVC nodes in the SVC cluster are connected to the same SAN, and present virtual disks to the hosts. These virtual disks are created from managed disks presented by the disk subsystems.

In our example, the zoning for SVC must be such that each node (node 1 and node 2) can address the DS4800 Storage Server.

Each SVC node has two HBAs with two fibre ports. We have only used one port from each HBA in each SVC node (ports 1 and 4).

SVC to resource zoning has to be done for the SVC to utilize the DS4800 Storage Server:

- ▶ SVC zone 1: SVC node 1 port 1, SVC node 1 port 4, SVC node 2 port 1 and SVC node 2 Port 4 is in a zone with DS4800 controller A and B.
- ▶ SVC zone 2: node 1 port 1, SVC node 1 port 4, SVC node 2 port 1 and SVC node 2 port 4 is in a zone with SVC management console.

With this zoning, each port of the SVC nodes connected to the fabric can address each DS4800 controller. It also offers many paths from the nodes to the storage for redundancy.

Host zoning has to be done for every host that will use SVC to manage the disks.

- ▶ Host zone 1: HBA_1 is in a zone with SVC node 1 port 1 and SVC node 2 port 1.
- ▶ Host zone 2: HBA_2 is in a zone with SVC node 1 port 4 and SVC node 2 port 4.

With this zoning each host HBA can see a port on each node that is connected to the fabric. It also offers many paths to the storage that the SVC will manage.

Configuring DS4800 Storage Server

The storage arrays should be created with only one LUN per array. The LUN should be sized to utilize all of the available storage space on the array. This single LUN will then be used to create smaller VDisks, which will be made available to the hosts. Choose the RAID level

required for the LUN. This RAID level is dependent upon the nature of the application that will use this LUN.

When creating the host type ensure that you select the unique host type for SVC of IBM TS SAN VCE (TotalStorage SAN Volume Controller Engine). Remember that the LUNs could be used by all types of hosts and applications attached to SVC.

In our example we created three SVC LUNs that will be managed by the SVC. These LUNs start with the letters SVC for clarity (Figure 10-6 on page 332).

- ▶ The database LUN is for SQL database files.
 - It has been created with RAID 5 and is 73 GB in size.
 - It has been created using 36 GB 15K drives.
 - It was created with a larger segment size of 128K.
 - Read ahead multiplier is set to enabled (1).
- ▶ The second LUN is the transaction logs for that SQL database.
 - It was created with RAID 10.
 - It has been created with a larger segment size of 128K.
 - It has been created using 73 GB 15K drives.
 - Read ahead multiplier is set to disabled (0).
- ▶ The third LUN is the file and print LUN.

As the file and print LUN is not as read or write intensive as the database or transaction log LUNs, it can be shared between multiple hosts. Most of the files are very small, and the environment is more a read than write.

- It has been created with a segment size of 64K.
- It has been created using 10K drives.
- Read ahead multiplier is set to enabled (1).

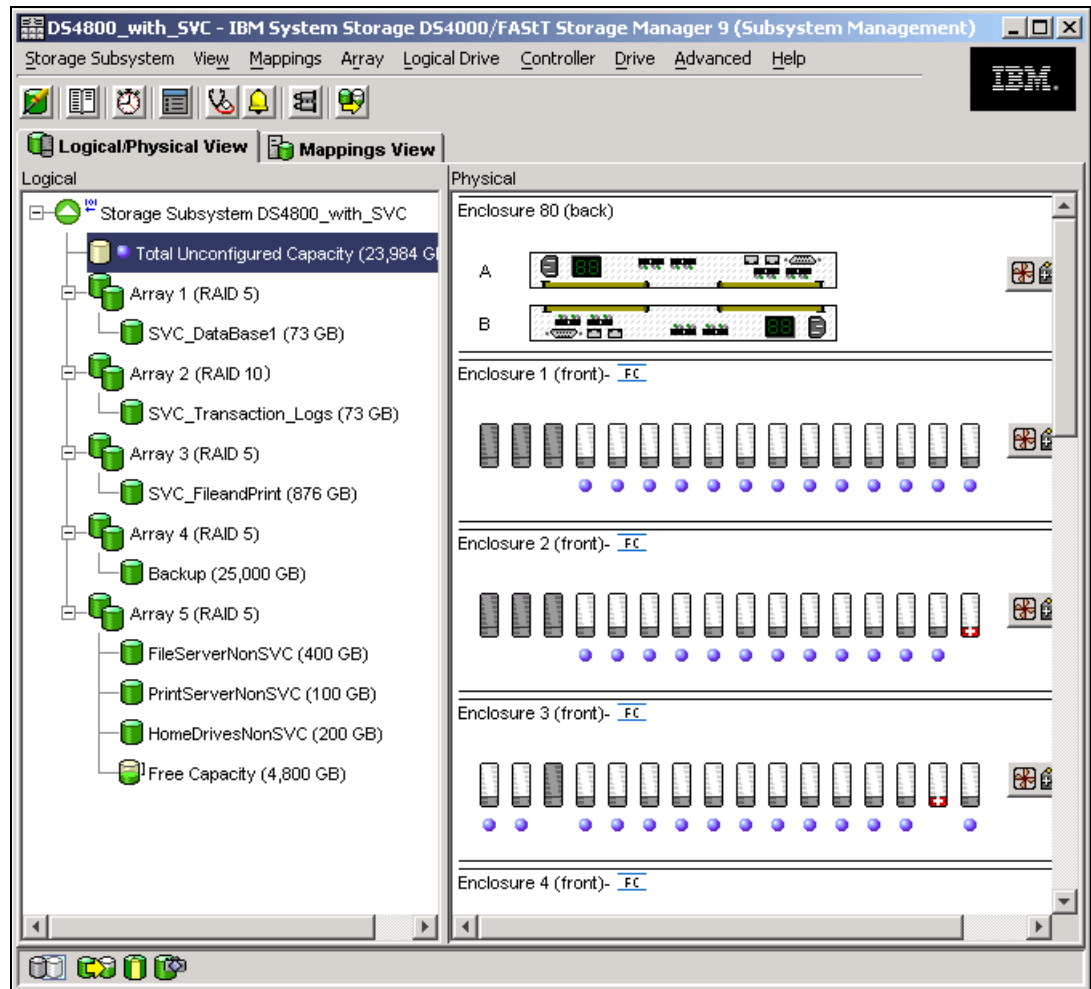


Figure 10-6 Storage Manager with LUNs created

The SVC nodes must be able to access all of the LUNs that SVC will manage. Therefore, the LUNs need to be mapped to all of the available SVC Fibre Channel host ports. The storage is mapped at the host SVCHost. The host type must be set to IBM TS SAN VCE. See Figure 10-7.

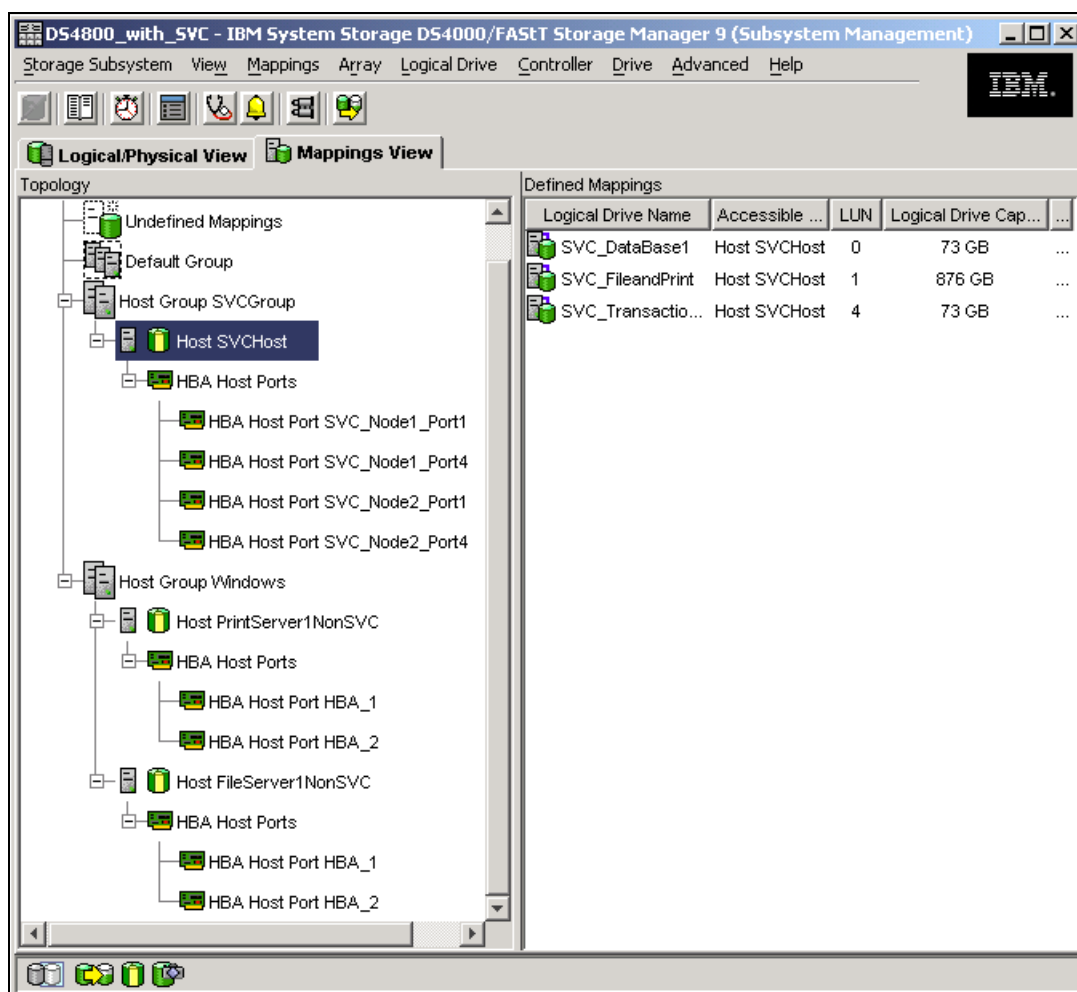


Figure 10-7 LUNs managed by the SVC are mapped to the host SVCHost

Using the LUN in SVC

The LUN can now be discovered by the management console and defined as an MDisk. The MDisk is then assigned to a MDG. Extents from the MDG are used to define a VDisk that can in turn be mapped to a SVC host. The steps are:

1. To discover the disks, log in at the SVC console with appropriate privileges.

Select **Work with Managed Disks** → **Managed Disks**. On the Viewing Managed Disks panel (Figure 10-8 on page 334), if your MDisks are not displayed, re-scan the Fibre Channel network. Select **Discover MDisks** from the list and click **Go**. Discovered disks appear as unmanaged.

If your MDisks (DS4000 LUNs) are still not visible, check that the logical unit numbers (LUNs) from the DS4000 are properly assigned to the SVC and that appropriate zoning is in place (SVC can see the disk subsystem).

2. Any unmanaged disk can be turned into a managed disk and added to a managed disk group. The MDG determines the extent size.

To create a managed disk group (MDG), select the **Work with Managed Disks** option and then the **Managed Disks Groups** link from the SVC Welcome page. The Viewing Managed Disks Groups panel opens. Select **Create an MDisk Group** from the list and click **Go**.

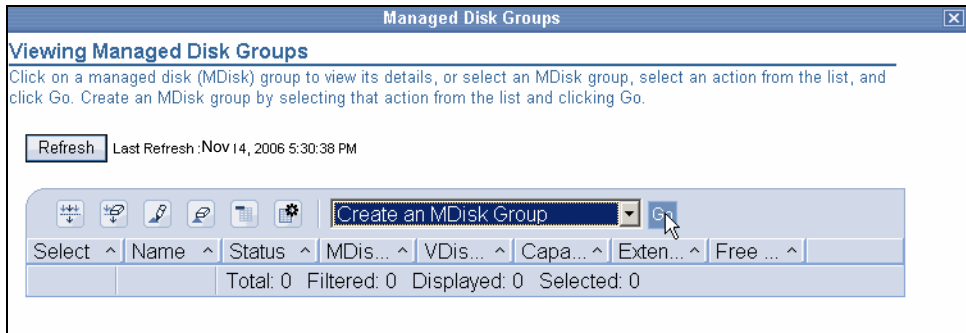


Figure 10-8 Selecting the option to create a MDG

On the Create Managed Disk Group panel, the wizard will give you an overview of what will be done. Click **Next**.

On the Name the group and select the managed disks panel, type a name for the MDG. From the MDisk Candidates box, one at a time, select the MDisks to put into the MDG. Click **Add** to move them to the Selected MDisks box. Click **Next**.

You must then specify the extent size to use for the MDG. Then click **Next**.

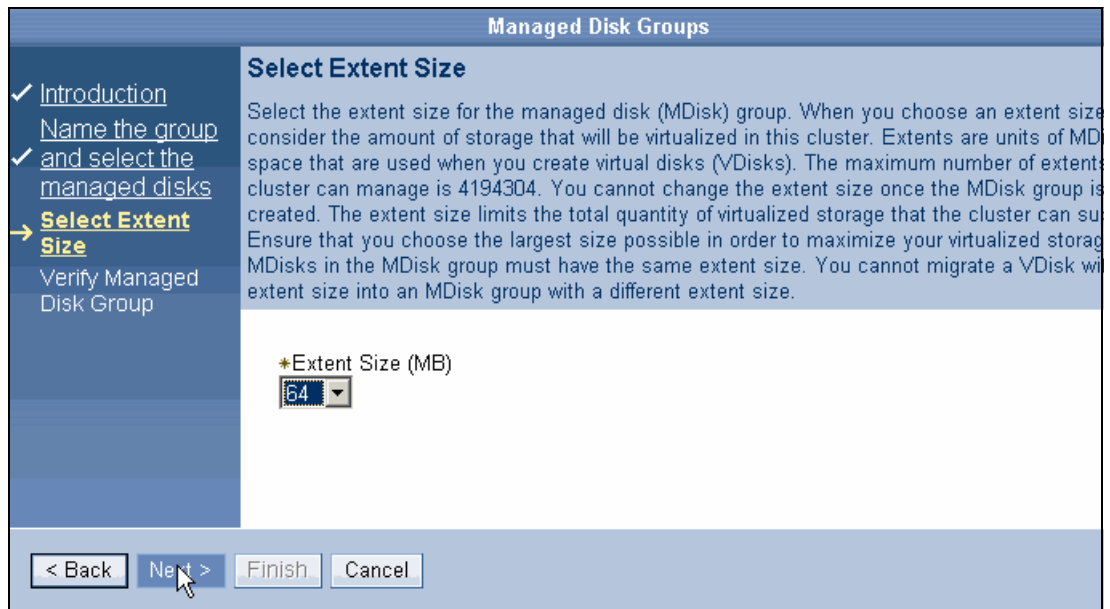


Figure 10-9 Select Extent Size panel

On the Verify Managed Disk Group panel, verify that the information specified is correct. Then click **Finish**.

Return to the Viewing Managed Disk Groups panel where the MDG is displayed. See Figure 10-10.

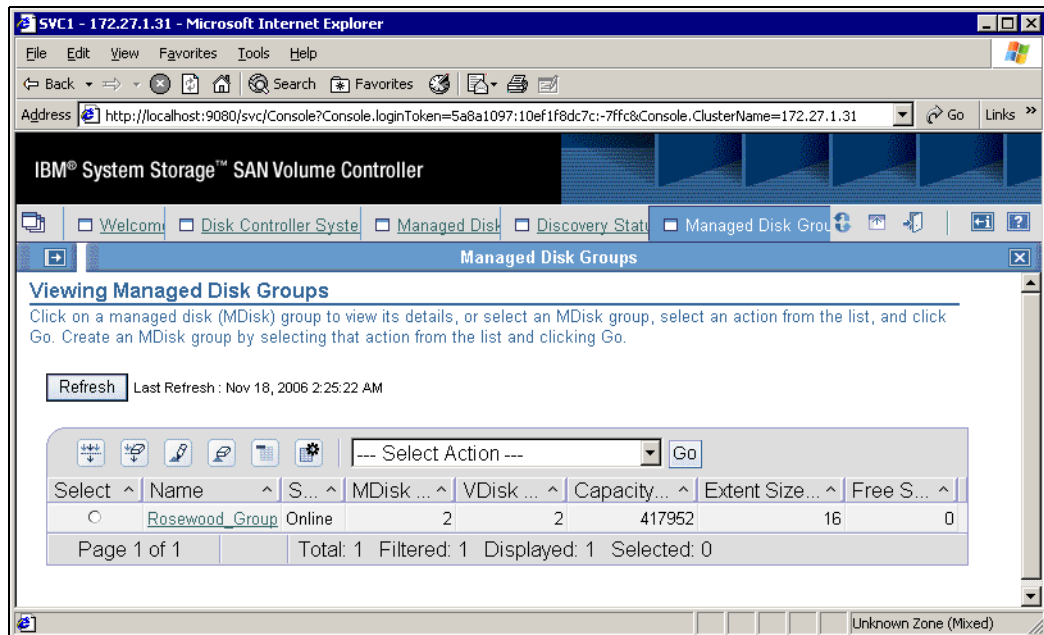


Figure 10-10 View managed disk groups

3. The MDisk is then used to create VDisks that will be presented to the host. This is done by selecting **Working with Virtual Disk** → **Virtual Disks** from the SVC Welcome page. From the drop-down menu select **Create VDisk**, as shown in Figure 10-11.

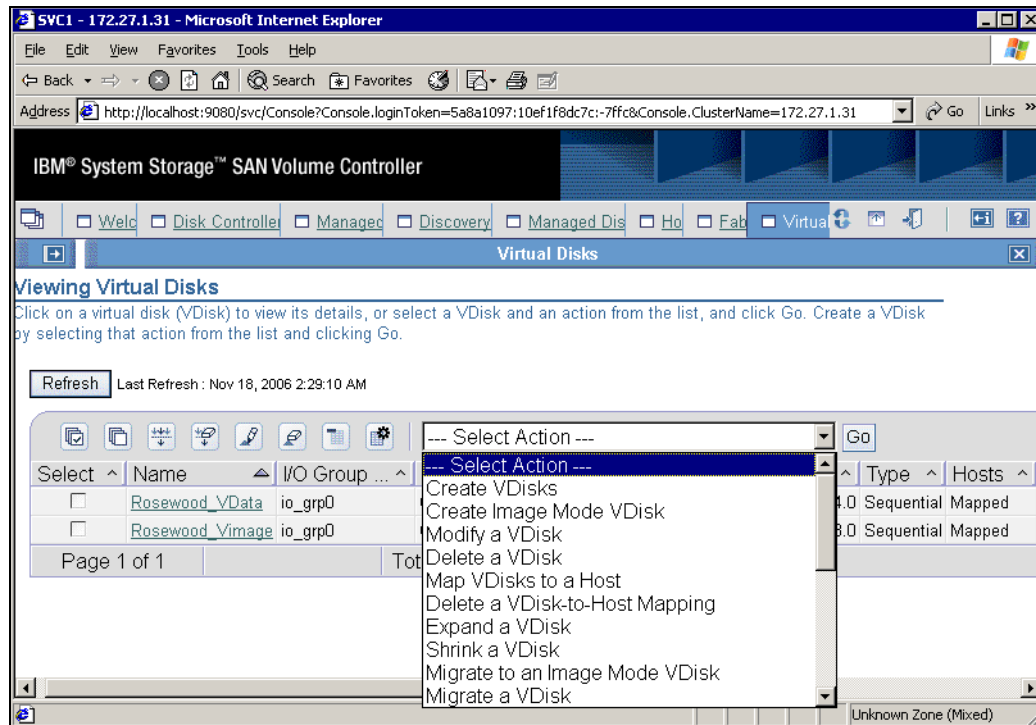


Figure 10-11 Working with virtual disks

- Before you can assign the VDisk to a host, you must define the host. To create a host select the **Working with Hosts** option and then the **Hosts** link from the SVC Welcome page.

The Viewing Hosts panel opens (see Figure 10-12). Select **Create a host** from the list and click **Go**.

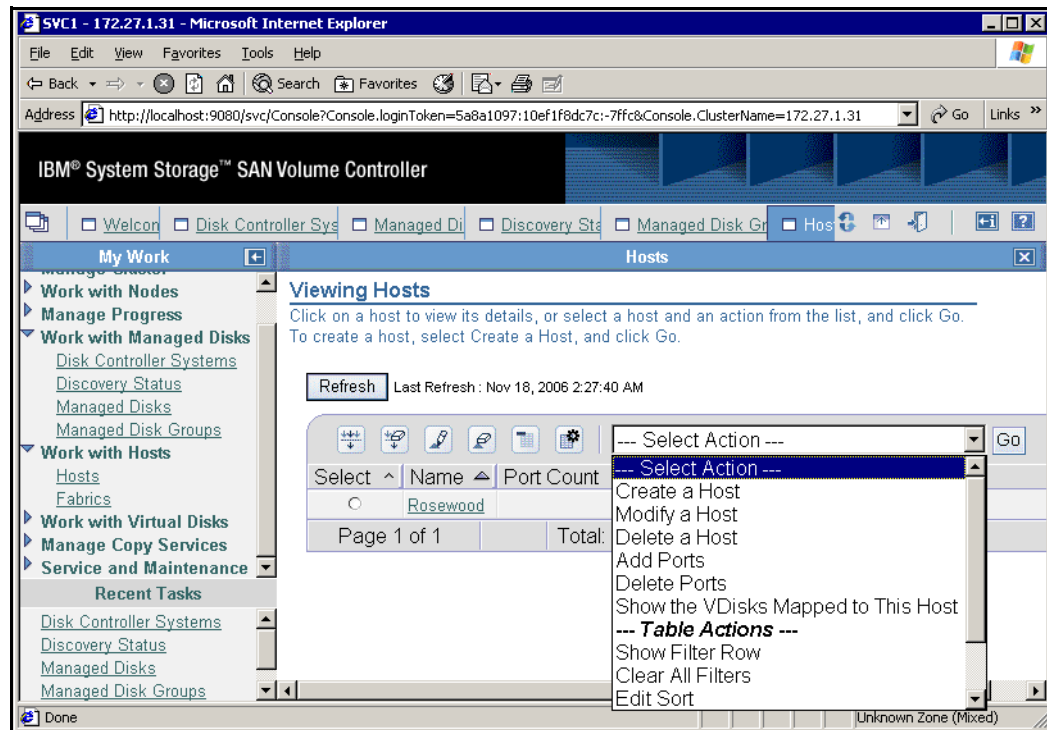


Figure 10-12 Add a host to the system

Enter details for the host, such as the name of the host, the type of host, I/O groups, and HBA ports, as shown in Figure 10-13. This panel shows all WWNs that are visible to the SVC and that have not already been defined to a host. If your WWN does not appear, check that the host has logged into the switch and that zoning in the switches is updated to allow SVC and host ports to see each other.

Creating Hosts
Type the name of the logical host object, assign World Wide Port Names (WWPNs) to it, choose the I/O groups to map to this host, and click OK. If you don't specify a name, a default name is assigned.

Host Name
Wildcat

Type
Generic

Port Mask (1 allows access, 0 denies access)
1111

*I/O Groups
io_grp0
io_grp1
io_grp2
io_grp3

Available Ports
210000E08B80356D
210100E08BA0356D

Selected Ports

Add >
< Remove

Figure 10-13 Creating Hosts panel

- Now you can map the VDisks to the host. Use **Working with Virtual Disks** and select the VDisk. From the drop-down Select Action menu, select **Map VDisk to a Host**.

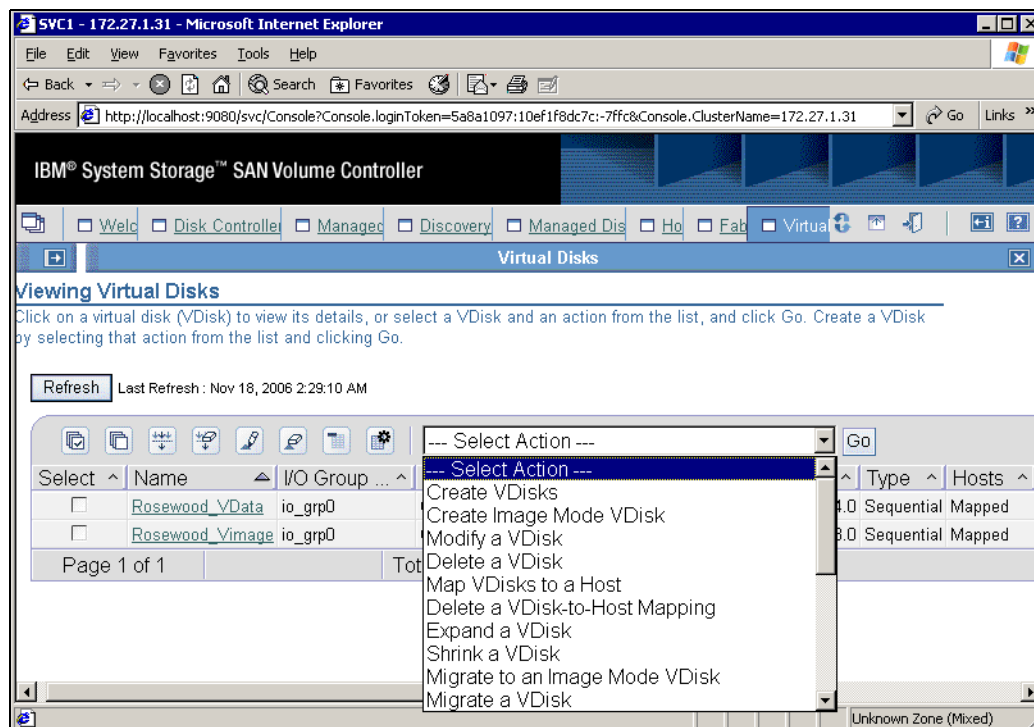


Figure 10-14 Map the VDisk to the host



DS4000 with AIX and HACMP

In this chapter, we present and discuss configuration information relevant to the DS4000 Storage Server attached to IBM System p servers and also review special considerations for High Availability Cluster Multiprocessing (HACMP) configurations in AIX.

AIX 5L™ is an award winning operating system, delivering superior scalability, reliability, and manageability. AIX 5L runs across the entire range of IBM pSeries® systems, from entry-level servers and workstations to powerful supercomputers able to handle the most complex commercial and technical workloads in the world.

In addition, AIX has an excellent history of binary compatibility, which provides assurance that your critical applications will continue to run as you upgrade to newer versions of AIX 5L.

HACMP is IBM software for building highly available clusters on a combination of System p (pSeries) systems. It is supported by a wide range of System p servers, with the new storage systems, and network types, and it is one of the highest-rated, UNIX-based clustering solutions in the industry.

11.1 Configuring DS4000 in an AIX environment

In this section we review the prerequisites and specifics for deploying the DS4000 storage server in an AIX host environment.

11.1.1 DS4000 adapters and drivers in an AIX environment

Table 11-1 lists some of the HBAs that can be used with IBM System p servers to support the DS4000 Storage Server in an AIX environment.

Table 11-1 HBAs for DS4000 and AIX

Adapter name	AIX operating system	Cluster
IBM FC 5716	AIX 5.1 5.2 5.3	HACMP 5.1 5.2 5.3
IBM FC 6227	AIX 5.1 5.2 5.3	HACMP 5.1 5.2 5.3
IBM FC 6228	AIX 5.1 5.2 5.3	HACMP 5.1 5.2 5.3
IBM FC 6239	AIX 5.1 5.2 5.3	HACMP 5.1 5.2 5.3

For detailed HBA information and to download the latest code levels, use the following link:

<http://knowledge.storage.ibm.com/servers/storage/support/hbasearch/interop/hbaSearch.do>

Verify microcode level

Always make sure that the HBA is at a supported microcode level for the model and SM firmware version installed on your DS4000.

There are two methods to check the current microcode level on the adapter.

1. The first method uses the **lscfg** command. It returns much of the information in the adapter Vital Product data (VPD). The Z9 field contains the firmware level. This method also displays the FRU number and Assembly part number, and the World Wide Name WWN:

lscfg -v1 fcsX

Where X is the number of the adapter returned by a previous **lsdev** command. The command will produce output similar to:

```
DEVICE LOCATION DESCRIPTION
fcs1 P1-I1/Q1 FC Adapter
Part Number.....80P4543
FRU Number..... 80P4544
Network Address.....10000000C932A80A
Device Specific.(Z8).....20000000C932A80A
Device Specific.(Z9).....HS1.90A4
```

2. The second method uses the **lsmcocode** command. It returns the extension of the firmware image file that is installed. This method only works with the latest adapters:

lsmcocode -d fcsX

Where X is the number of the adapter returned by the **lsdev** command:

```
DISPLAY MICROCODE LEVEL 802111
fcs3 FC Adapter
The current microcode level for fcs3 is 190104.
Use Enter to continue.
```

Always check that you have the latest supported level of the firmware and drivers. If not, download the latest level and upgrade by following the instructions found at:

<http://techsupport.services.ibm.com/server/mdownload/adapter.html>

Install the RDAC driver on AIX

You need the following file sets for the AIX device driver:

- ▶ `devices.fcp.disk.array.rte` - RDAC runtime software
- ▶ `devices.fcp.disk.array.diag` - RDAC diagnostic software
- ▶ `devices.common.IBM.fc.rte` - Common FC Software

You also need one of the following drivers, depending on your HBA:

- ▶ `devices.pci.df1000f7.com` - Feature code for 6227 and 6228 adapters require this driver.
- ▶ `devices.pci.df1000f7.rte` - Feature code 6227 adapter requires this driver.
- ▶ `devices.pci.df1000f9.rte` - Feature code 6228 adapter requires this driver.
- ▶ `devices.pci.df1080f9.rte` - Feature code 6239 adapter requires this driver.
- ▶ `devices.pci.df1000fa.rte` - Feature code 5716 adapter requires this driver.

Additional packages or even PTFs might be required, depending on the level of your AIX operating system. Before installing the RDAC driver, always check prerequisites for a list of required driver version, file sets, level of AIX system.

AIX Recommended OS Levels: AIX 5.1, 5.2 or 5.3 Multipath Driver: IBM AIX RDAC Driver (`fc.disk.array`) and FC Protocol Device Drivers PTFs:

▶ AIX 5.3

Required Maintenance Level - 5300-3
`devices.fcp.disk.array` - 5.3.0.30
`devices.pci.df1000f7.com` - 5.3.0.10
`devices.pci.df1000f7.rte` - 5.3.0.30
`devices.pci.df1000f9.rte` - 5.3.0.30
`devices.pci.df1000fa.rte` - 5.3.0.30

▶ AIX 5.2

Required Maintenance Level - 5200-7
`devices.fcp.disk.array` - 5.2.0.75
`devices.pci.df1000f7.com` - 5.2.0.75
`devices.pci.df1000f7.rte` - 5.2.0.75
`devices.pci.df1000f9.rte` - 5.2.0.75
`devices.pci.df1000fa.rte` - 5.2.0.75

▶ AIX 5.1

Required Maintenance Level - 5100-9
`devices.fcp.disk.array` - 5.1.0.66
`devices.pci.df1000f7.com` - 5.1.0.66
`devices.pci.df1000f7.rte` - 5.1.0.37
`devices.pci.df1000f9.rte` - 5.1.0.37
`devices.pci.df1000fa.rte` - Not Supported

▶ AIX 4.3

Contains no support for features beyond Storage Manager 8.3.

AIX PTF/APARs can be downloaded from:

<http://techsupport.services.ibm.com/server/aix.fdc>

The RDAC driver creates the following devices that represent the DS4000 storage subsystem configuration:

- ▶ **dar** (disk array router): Represents the entire subsystem and storage partitions.
- ▶ **dac** (disk array controller devices): Represents a controller within the storage subsystem. There are two dacs in the storage subsystem.
- ▶ **hdisk**: These devices represent individual LUNs on the array.
- ▶ **utm**: The universal transport mechanism (utm) device is used only with in-band management configurations, as a communication channel between the SMagent and the DS4000 Storage Server.

DAR

For a correct AIX host configuration, you should have DAR for every Storage Partition of every DS4000 connected to the AIX host (Example 11-1).

Example 11-1 Storage server, storage partitions and dar

One DS4000 with one Storage Partition	dar0
One DS4000 with two Storage Partition	dar0,dar1
Two DS4000 with two Storage Partition	dar0,dar1,dar2,dar3

You can verify the number of DAR configured on the system by typing the command:

```
# lsdev -C | grep dar
```

DAC

For a correct configuration, it is necessary to create an adequate zoning in such a way that every DAR shows two DACs assigned.

Attention: More than two DAC on each DAR is not supported (this would decrease the number of servers that you can connect to the DS4000 without additional benefit).

You can verify the number of DAC configured on the system by typing the command:

```
# lsdev -C | grep dac
```

Recreating the relationship between DAR and DAC

To recreate the relationships, you must first remove the DAR, DAC, and HBA definitions from AIX. Proceed as follows:

Use the **lsvg -o** command to vary off all the volume groups attached to the HBA you want to remove (make sure there is no activity on DS4000 disks).

1. Remove DAR with the following command:

```
# rmdev -dl dar 0 -R  
  
hdisk1 delete  
hdisk2 delete  
dar0 delete
```

2. Then, remove the HBA with the following command:

```
# rmdev -dl fcs1 -R  
  
dac0 delete  
fcsio delete  
fcnet0 delete  
fcs0 delete
```


3. To reconfigure the HBA, DAC, DAR and hdisks, simply run the AIX configurator manager using the following command:

```
# cfgmgr -S
```

11.1.2 Testing attachment to the AIX host

To test that the physical attachment from the AIX host to the DS4000 was done in a manner that avoids any single point of failure, unplug one fiber at a time as indicated by the X symbol in Figure 11-1.

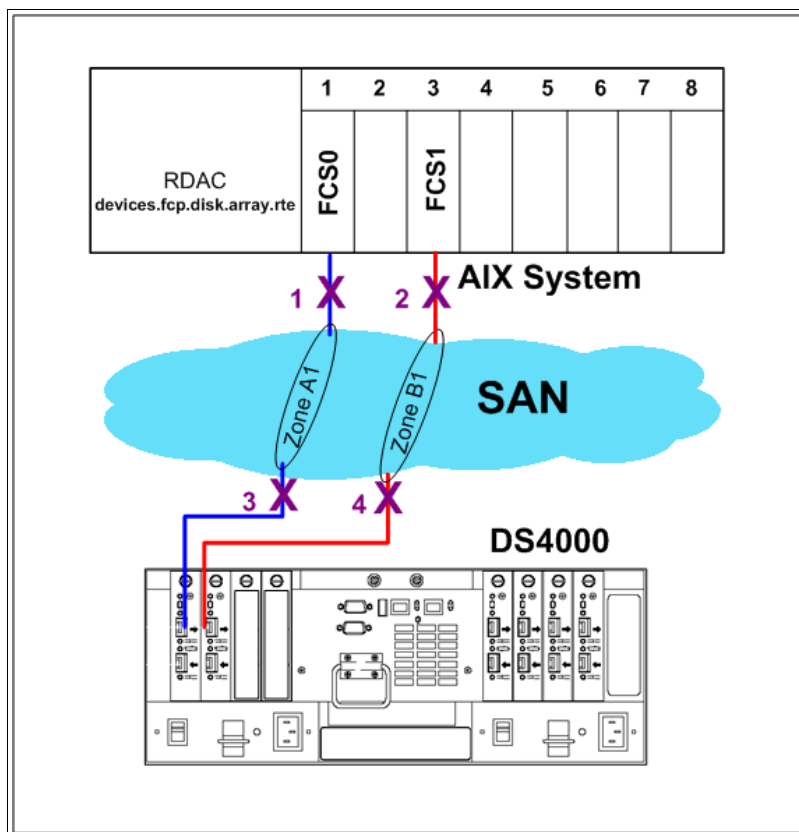


Figure 11-1 Testing attachment to the AIX host

For each cable that you unplug, verify that the AIX host still can access the DS4000 logical drives without problem.

If everything still works fine, reconnect the fibre cable that you had removed, wait a few seconds and redistribute the DS4000 logical drives using the Storage Manager option: **Advanced** → **Recovery** → **Redistribute Logical Drives**. This is necessary because when a failure event is detected on one controller, the DS4000 switches all logical drives managed by the failed controller to the other controller. Because the logical drives failback is not automatic on a DS4000, you need to manually redistribute the logical drives.

Repeat the same procedure for each cable.

Note: You can do the same test in a HACMP configuration that is up and running. HACMP does not get an event on a Fibre Channel failure and will thus not start its own failover procedure.

11.1.3 Storage partitioning in AIX

The benefit of defining storage partitions allows controlled access to the logical drives on the DS4000 storage subsystem to only those hosts also defined in the Storage Partition. Storage partitioning is defined by specifying the world wide names of the host ports.

Remember also that when you define the host ports, you specify the operating system of the attached host as well. The DS4000 uses the host type to adapt the RDAC or ADT settings for that host type. Each operating system expects slightly different settings and handles SCSI commands differently. Therefore, it is important to select the correct value. If you do not, your operating system may not boot anymore, or path failover cannot take place when required.

In the sections that follow, we show several examples of supported or unsupported storage partitioning definitions.

Storage partition with one HBA on one AIX server

This configuration has one HBA on one AIX server. This is *supported but not recommended*: with only one HBA, there is no redundancy and lesser performance (Figure 11-2).

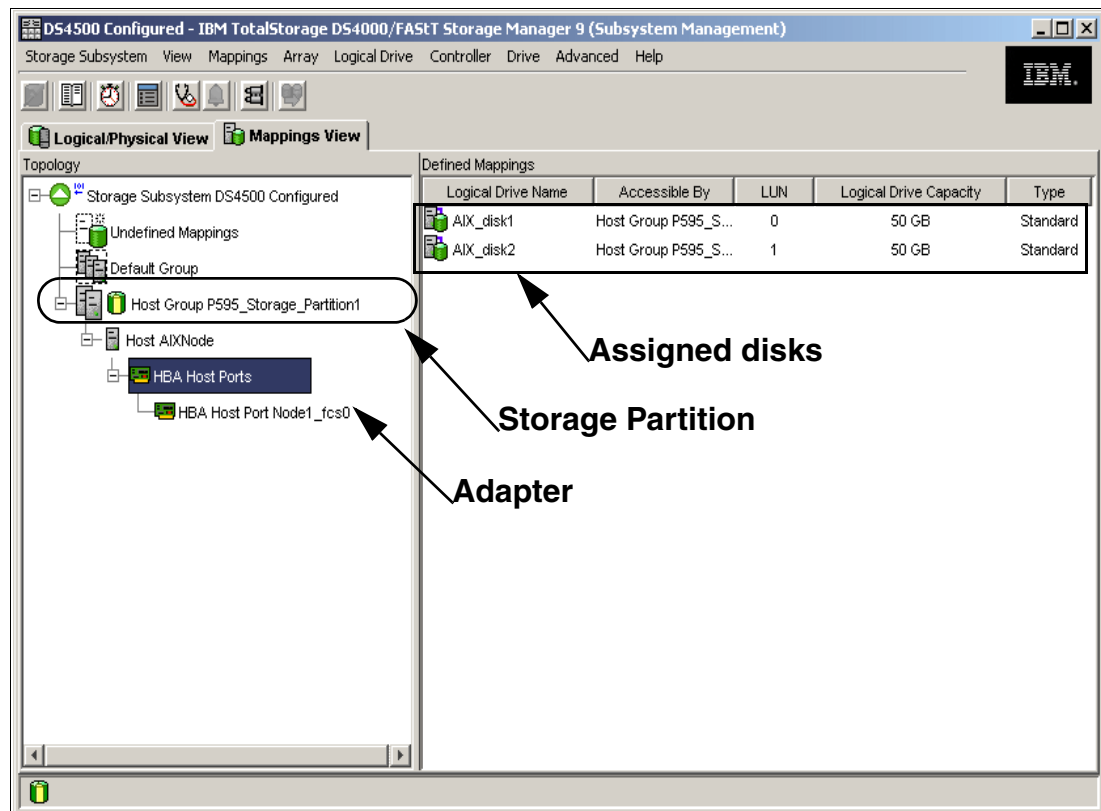


Figure 11-2 One HBA, one AIX host - not recommended

Storage partition mapping with two HBAs on one AIX server

This is the most common situation. See Figure 11-3 for this configuration.

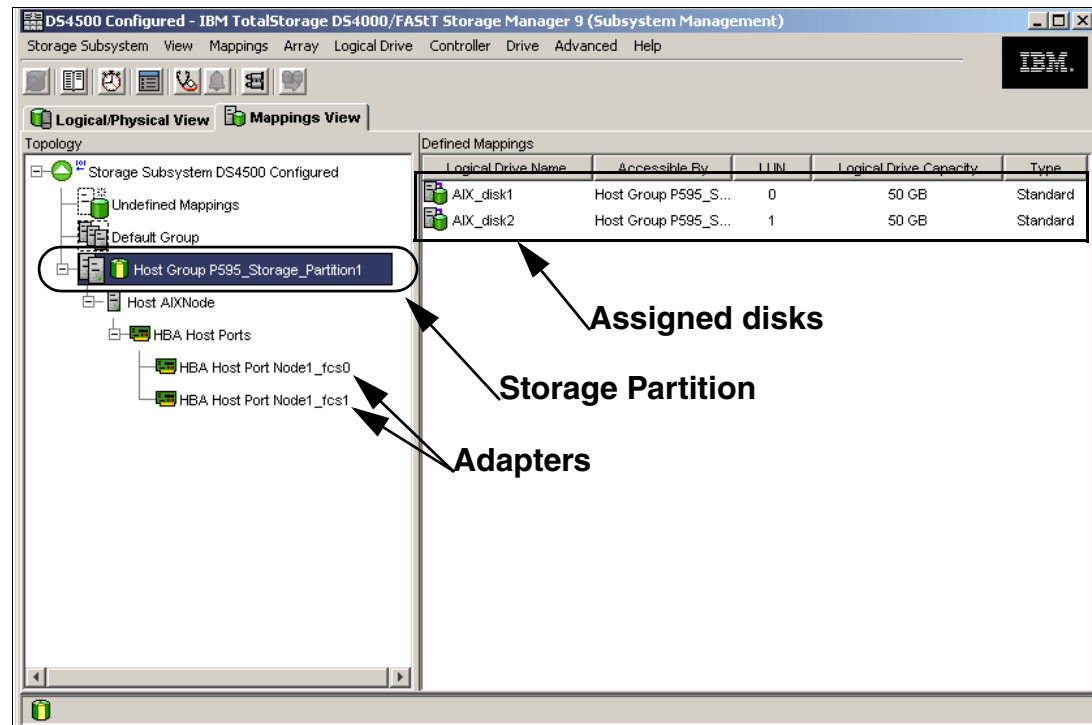


Figure 11-3 Two HBA, one AIX host

Two storage partitions with four HBAs on one AIX server

See Figure 11-4 and Figure 11-5 on page 346 for this configuration.

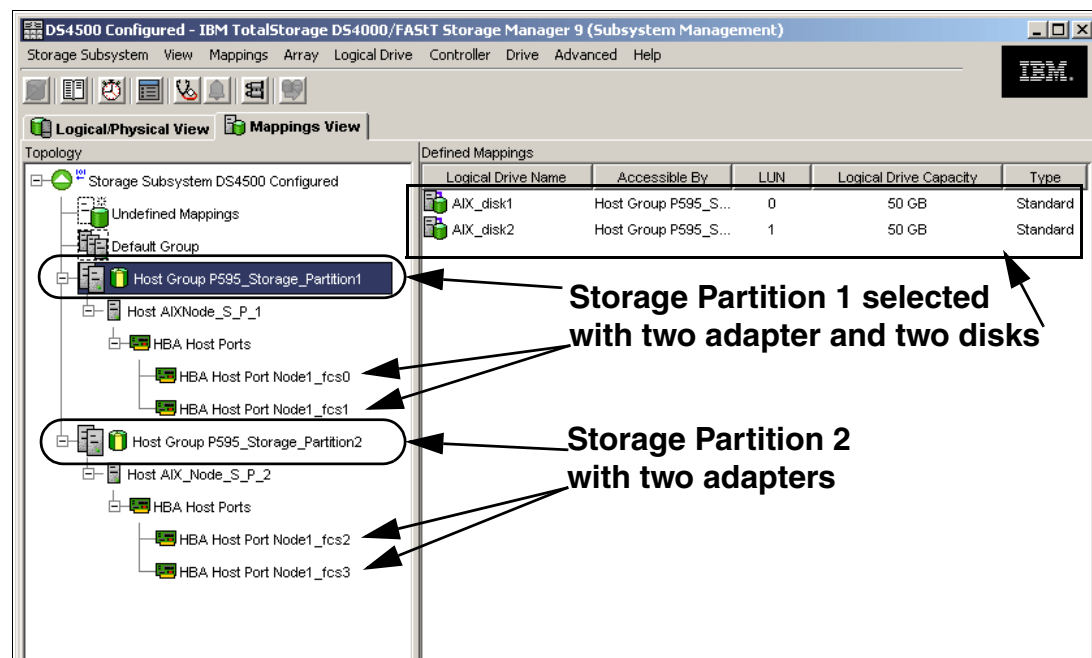


Figure 11-4 Two storage partitions - Two HBAs for each storage partition - First partition selected

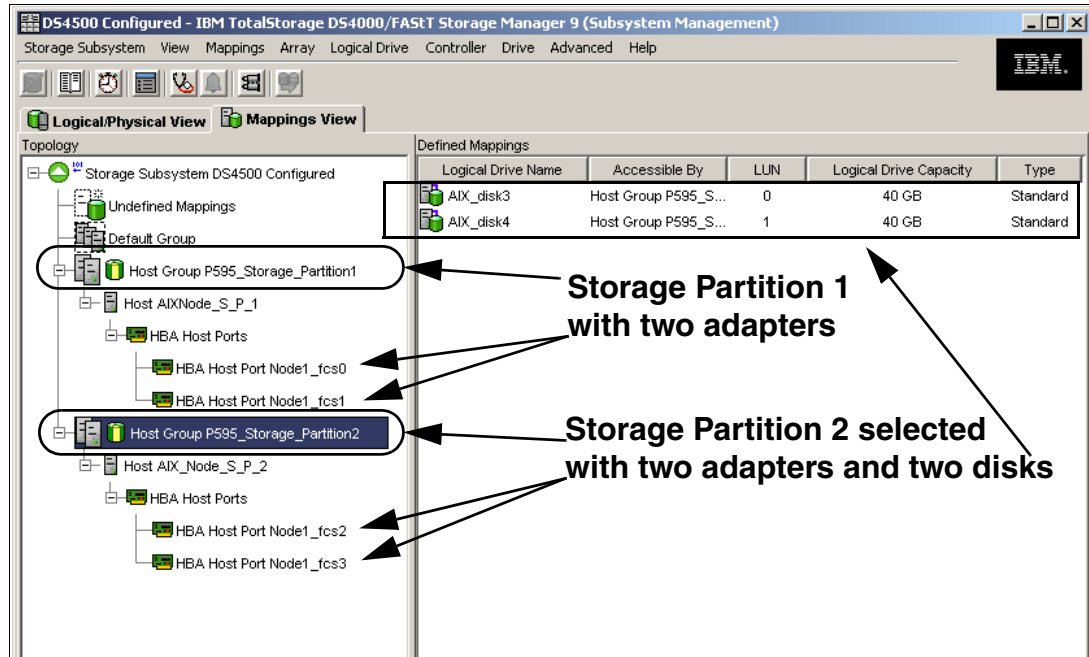


Figure 11-5 Two storage partitions - Two HBAs for each storage partition - Second partition selected

Mapping to default group (not supported)

Figure 11-6 shows the AIX disks assigned to the Default Group (no partition defined). This is *not supported*.

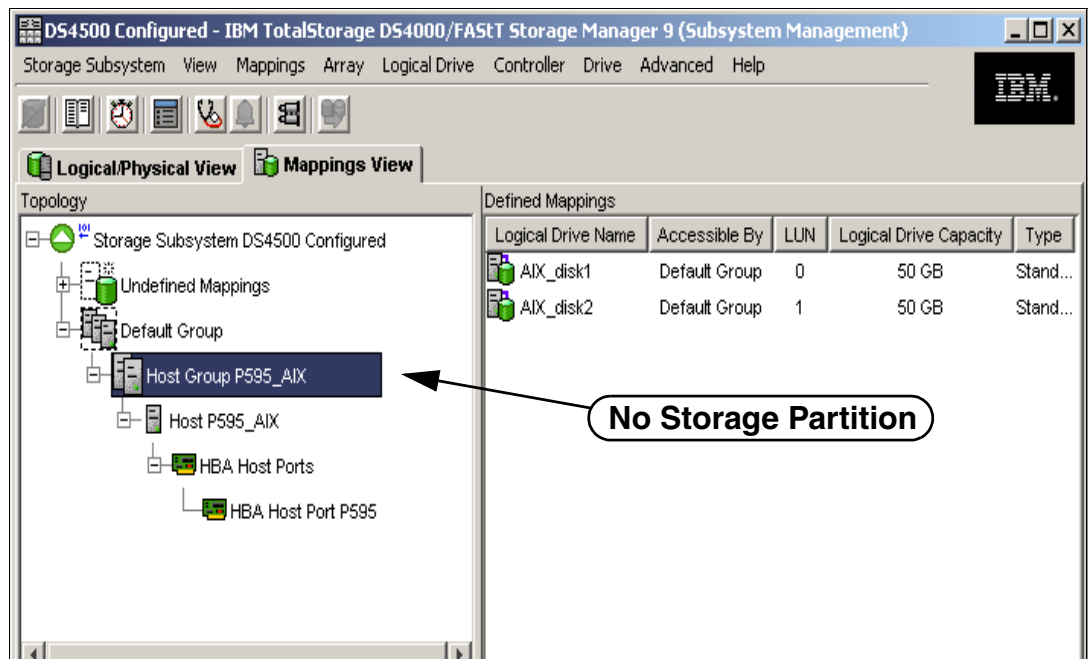


Figure 11-6 No storage partition

One storage partition with four HBAs on one AIX server (not supported)

Mapping with four HBA on one server with only one Storage Partition is *not supported* (Figure 11-7).

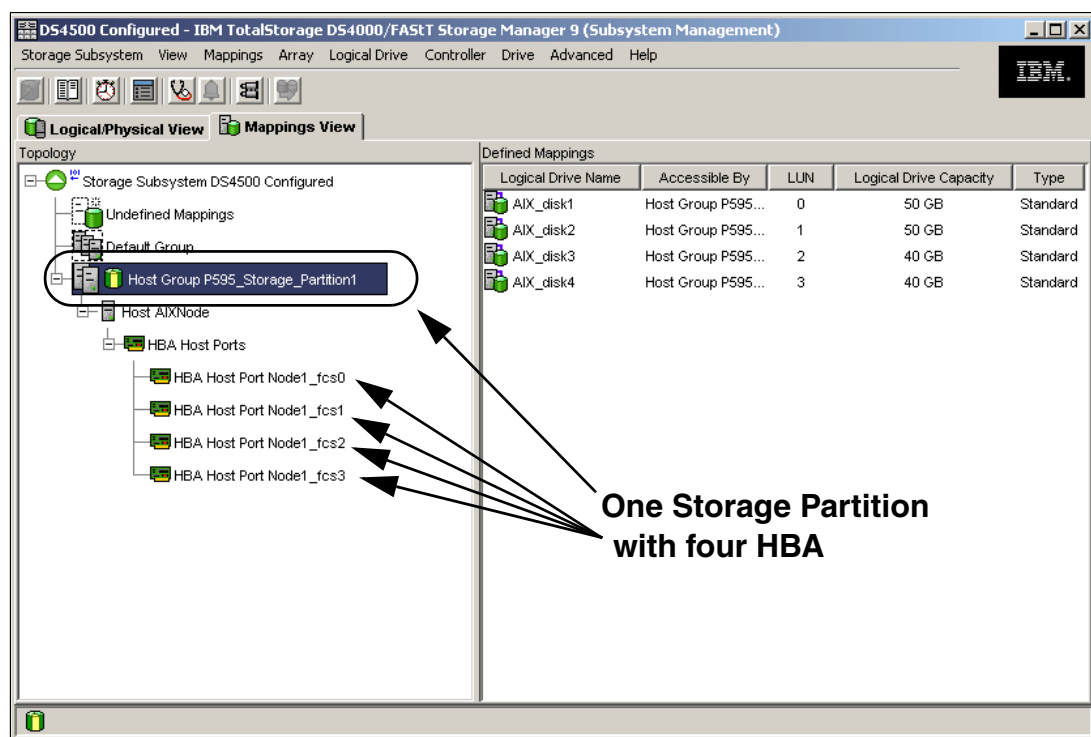


Figure 11-7 One Storage Partition with four HBA

11.1.4 HBA configurations

In this section we review the supported HBA configurations.

One HBA on host and two controllers on DS4000

One HBA and two controllers on DS4000 with appropriate zoning is *supported*, although *not recommended*.

Single HBA configurations are allowed, but require both controllers in the DS4000 to be connected to the host. In a switched environment, both controllers must be connected to the switch within the same SAN zone as the HBA. In a direct-attach configurations, both controllers must be *daisy-chained* together (Figure 11-8).

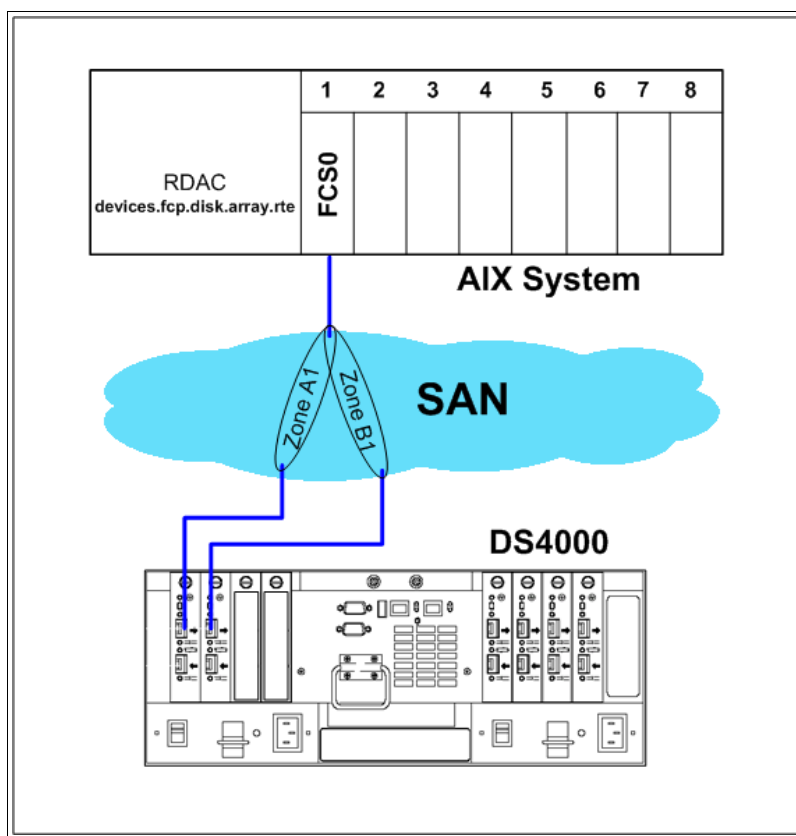


Figure 11-8 AIX system with one HBA

In this case, we have one HBA on the AIX server included in two zones for access to controller A and controller B respectively. In Storage Manager, you would define one storage partition with one host HBA. See Figure 11-2 on page 344. This configuration is *supported, but not recommended*.

Example 11-2 shows commands and expected results for verifying DAR, DAC, and the appropriate zoning.

Example 11-2 AIX system with one HBA

```
# lsdev -Cadapter | grep fcs
```

```
fcs0 Available 27-08 FC Adapter
```

```
# lsdev -C | grep dar
```

```
dar0 Available fcparray Disk Array Router
```

```
# lsdev -C | grep dac
```

```
dac0 Available 27-08-01 fcparray Disk Array Controller
```

```
dac1 Available 27-08-01 fcparray Disk Array Controller
```

Zoning

```
zone: AIX_FCS0_CTRL_A1 10:00:00:00:c9:32:a8:0a ; 20:04:00:a0:b8:17:44:32
```

```
zone: AIX_FCS0_CTRL_B1 10:00:00:00:c9:32:a8:0a ; 20:05:00:a0:b8:17:44:32
```

Configuration with two HBAs on host and two controllers on DS4000

Two HBAs and two controllers on DS4000 with appropriate zoning is *supported* (Figure 11-9).

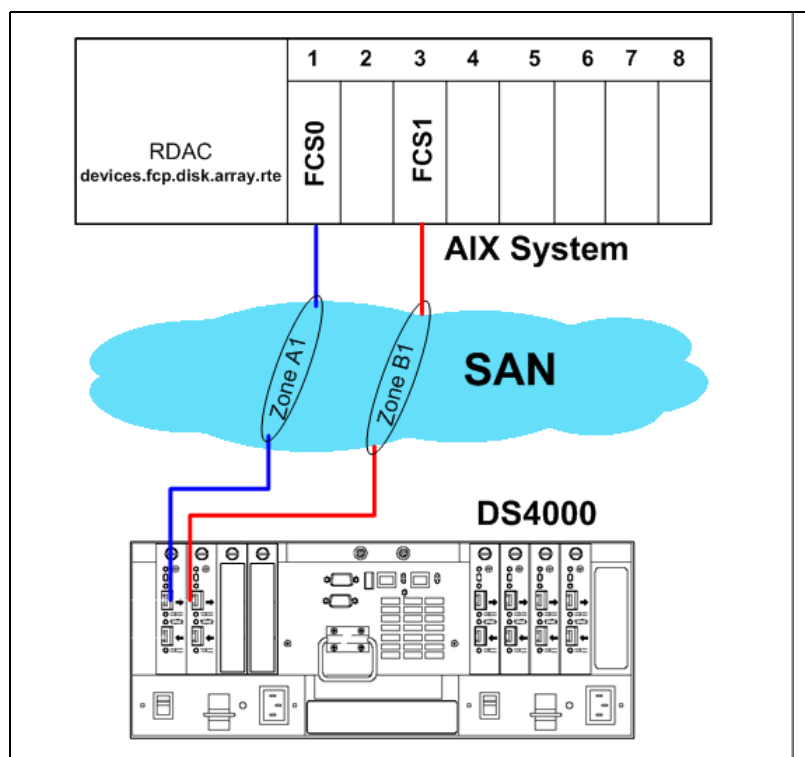


Figure 11-9 Two HBAs, two controllers

We recommend to define a zone including the first host HBA to the DS4000 controller A and another zone with second host HBA and DS4000 controller B.

In Storage Manager, create one Storage Partition with two host HBAs. Example 11-3 shows commands and expected results for verifying DAR, DAC and the appropriate zoning.

Example 11-3 Two HBAs, two controllers

```
# lsdev -Ccadapter | grep fcs
fcs0 Available 27-08 FC Adapter
fcs1 Available 34-08 FC Adapter

# lsdev -C | grep dar
dar0 Available fcparray Disk Array Router

# lsdev -C | grep dac
dac0 Available 27-08-01 fcparray Disk Array Controller
dac1 Available 34-08-01 fcparray Disk Array Controller
```

Zoning

```
zone: AIX_FCS0_CTRL_A1 10:00:00:00:c9:32:a8:0a ; 20:04:00:a0:b8:17:44:32
zone: AIX_FCS1_CTRL_B1 10:00:00:00:c9:4c:8c:1c ; 20:05:00:a0:b8:17:44:32
```

Configuration with four HBAs on host and four controllers on DS4000

A configuration with four HBAs and four controllers on DS4000 with appropriate zoning is supported (Figure 11-10).

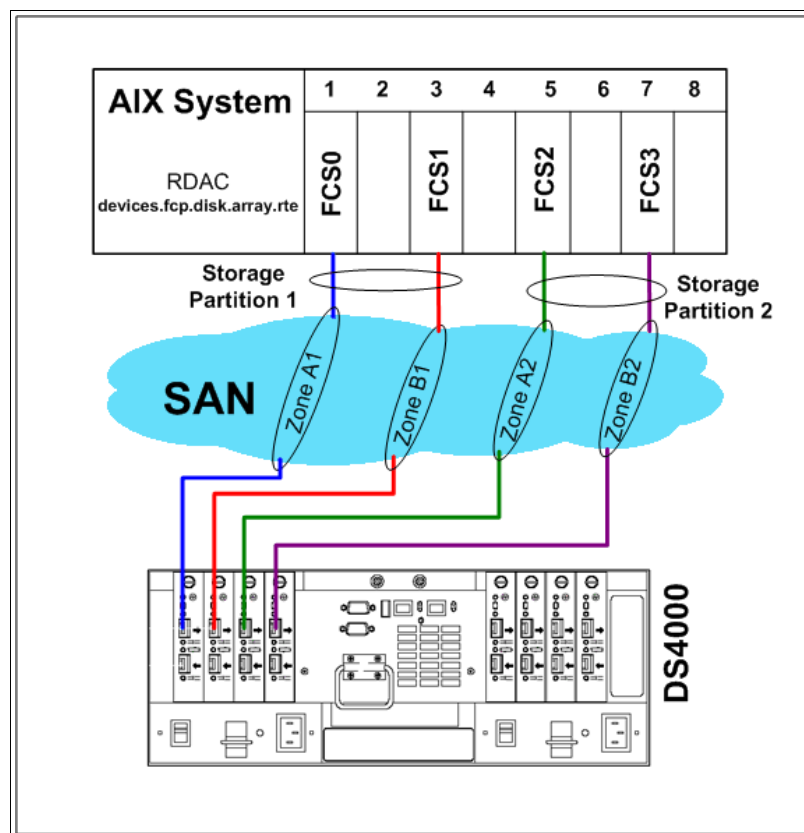


Figure 11-10 Four HBAs, two storage partitions

It is possible connect one AIX system with four adapters, but it is necessary create two storage partitions, each including two HBAs. See Figure 11-4 on page 345 and Figure 11-5 on page 346.

Example 11-4 below shows commands and expected results for verifying DAR, DAC, and the appropriate zoning.

Example 11-4 Four HBAs, two storage partitions

```
# lsdev -Ccadapter | grep fcs
fcs0 Available 27-08 FC Adapter
fcs1 Available 34-08 FC Adapter
fcs2 Available 17-08 FC Adapter
fcs3 Available 1A-08 FC Adapter

# lsdev -C | grep dar
dar0 Available fcparray Disk Array Router
dar1 Available fcparray Disk Array Router

# lsattr -El dar0
act_controller dac0,dac1 Active Controllers False
all_controller dac0,dac1 Available Controllers False

# lsattr -El dar1
act_controller dac2,dac3 Active Controllers False
all_controller dac2,dac3 Available Controllers False

# lsdev -C | grep dac
dac0 Available 27-08-01 fcparray Disk Array Controller
dac1 Available 34-08-01 fcparray Disk Array Controller
dac2 Available 17-08-01 fcparray Disk Array Controller
dac3 Available 1A-08-01 fcparray Disk Array Controller
```

Zoning

```
zone: AIX_FCS0_CTRL_A1 10:00:00:0:c9:32:a8:0a ; 20:04:00:a0:b8:17:44:32
zone: AIX_FCS1_CTRL_B1 10:00:00:0:c9:4c:8c:1c ; 20:05:00:a0:b8:17:44:32
zone: AIX_FCS2_CTRL_A2 10:00:00:0:c9:32:a7:fb ; 20:04:00:a0:b8:17:44:32
zone: AIX_FCS3_CTRL_B2 10:00:00:0:c9:32:a7:d1 ; 20:05:00:a0:b8:17:44:32
```

11.1.5 Unsupported HBA configurations

This section lists some unsupported configurations under AIX.

Configuration with one HBA and only one controller on DS4000

A configuration with one HBA and one controller on the DS4000 as depicted in Figure 11-11 is *not supported*.

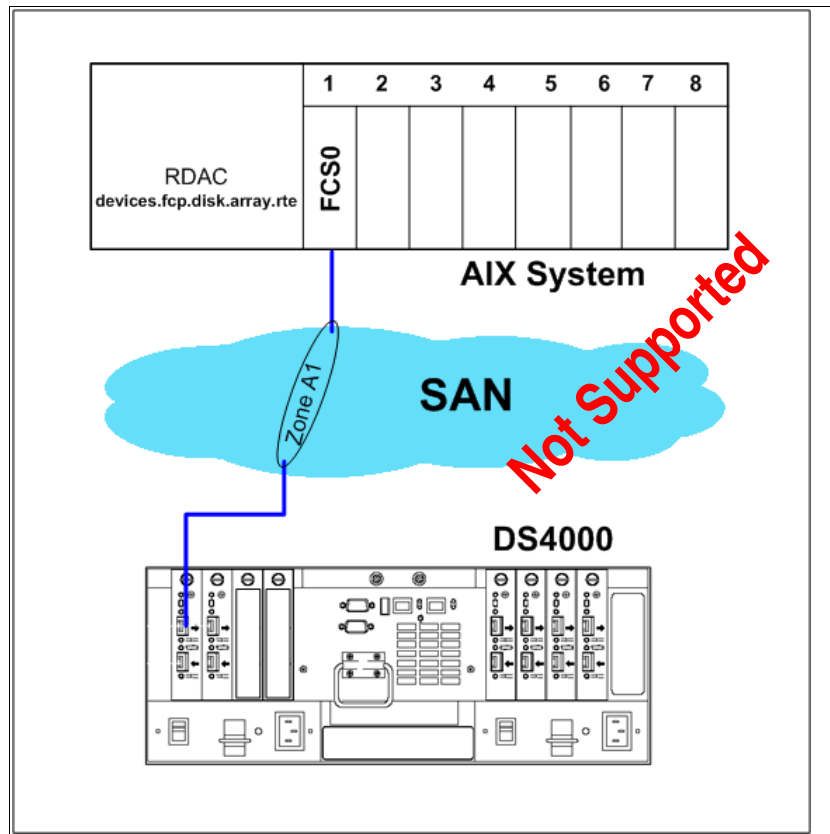


Figure 11-11 One HBA, one controller

Configuration with one HBA on host and four controllers on DS4000

One HBA and four controllers on DS4000 with zoning as depicted in Figure 11-12 is *not supported*.

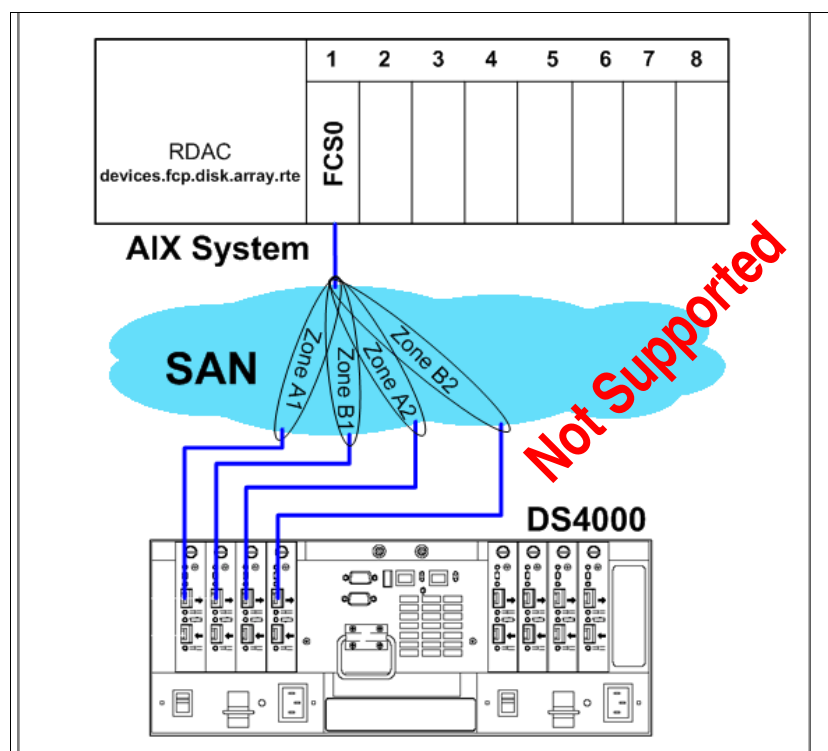


Figure 11-12 One BA, four paths to DS4000 controllers

Defining a zoning where one AIX system can have four access paths to the DS4000 controllers is *not supported*; in this configuration the AIX system uses one HBA, one DAR, and four DACs (Example 11-5).

Example 11-5 One HBA, one DAR, and four DACs

```
# lsdev -Ccadapter | grep fcs
fcs0 Available 27-08 FC Adapter

# lsdev -C | grep dar
dar0 Available fcparray Disk Array Router

# lsdev -C | grep dac
dac0 Available 27-08-01 fcparray Disk Array Controller
dac1 Available 27-08-01 fcparray Disk Array Controller
dac2 Available 27-08-01 fcparray Disk Array Controller
dac3 Available 27-08-01 fcparray Disk Array Controller
```

Zoning

```
zone: AIX_FCS0_CTRL_A1 10:00:00:00:c9:32:a8:0a ; 20:04:00:a0:b8:17:44:32
zone: AIX_FCS0_CTRL_B1 10:00:00:00:c9:32:a8:0a ; 20:05:00:a0:b8:17:44:32
zone: AIX_FCS0_CTRL_A2 10:00:00:00:c9:32:a8:0a ; 20:04:00:a0:b8:17:44:32
zone: AIX_FCS0_CTRL_B2 10:00:00:00:c9:32:a8:0a ; 20:05:00:a0:b8:17:44:32
```

Configuration with two HBA on host and four controller paths

Two HBAs and four controller paths with the zoning as depicted in Figure 11-13 is *not* supported.

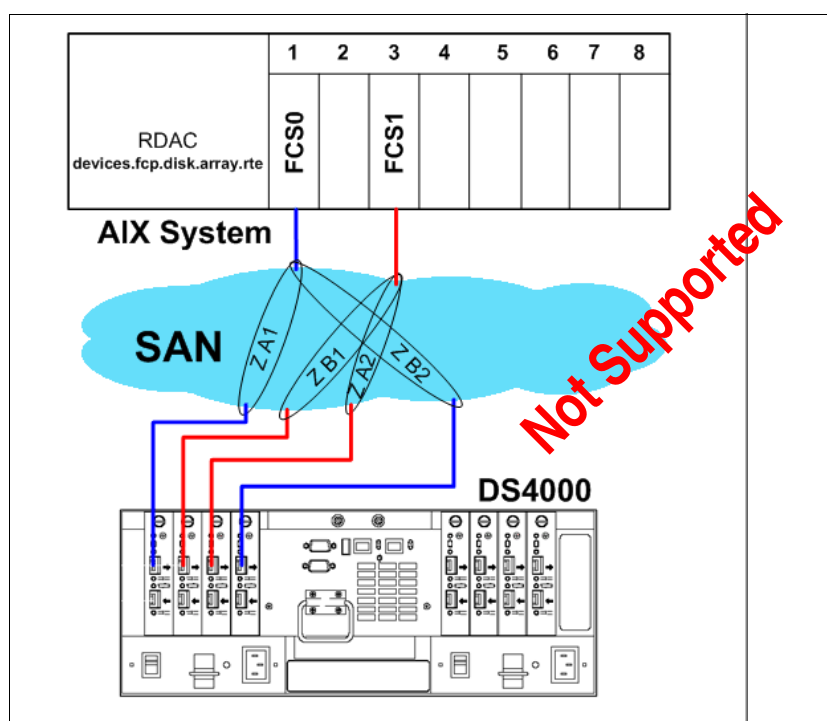


Figure 11-13 Two HBAs and four controller paths - Zoning 1

Configuration with two HBA on host and four controllers on DS4000

Two HBAs and four controllers on DS4000 with the following zoning are *not supported*.

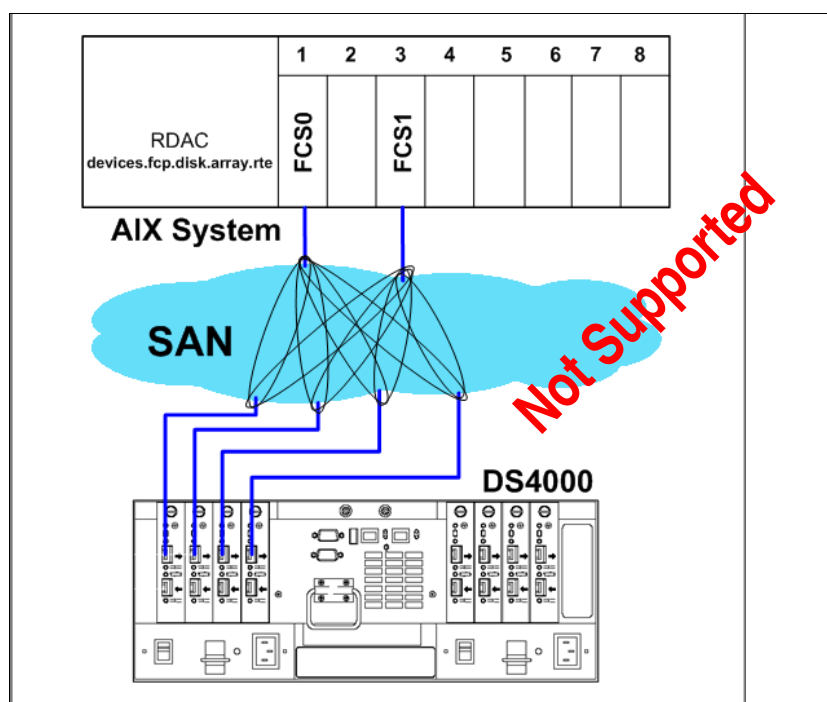


Figure 11-14 Two HBAs and four controller paths - Zoning 2

Configuration with four HBA and four controller paths

Four HBAs and four controllers on DS4000 with the zoning as depicted in Figure 11-15 are *not supported*.

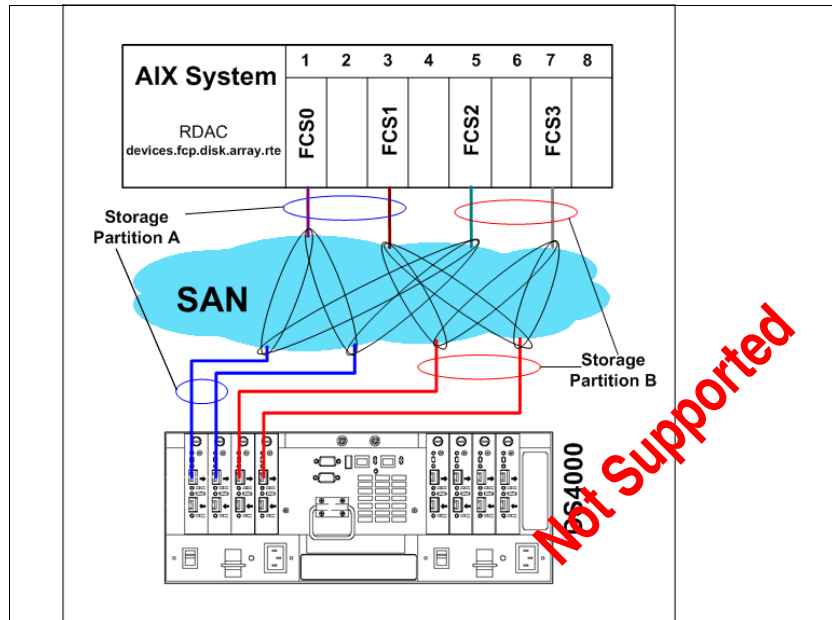


Figure 11-15 Four HBA and Four controllers - Zoning 1

One AIX system with four HBAs connected to the DS4000 is *not supported* if each HBA sees more than one DS4000 controller.

Configuration with four HBA on node and four controllers

Four HBAs and four controllers on DS4000 with this zoning are *not supported* (Figure 11-16).

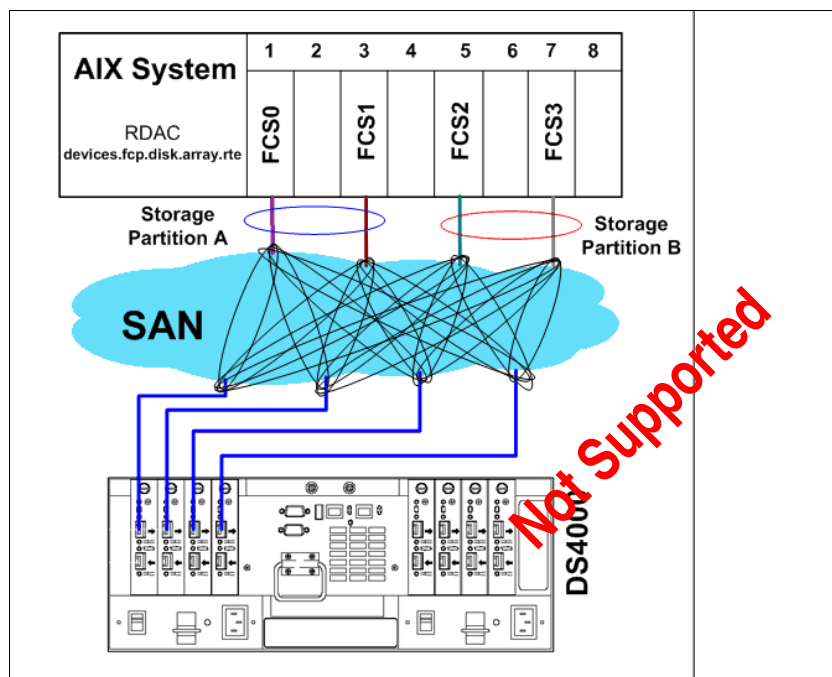


Figure 11-16 Four HBA and Four controllers - Zoning 2

11.1.6 Device drivers coexistence

In this section, we present several configurations that support, or do not support, the coexistence of different storage device drivers.

Coexistence between RDAC and SDD

RDAC and SDD (used by the ESS, DS6000, and DS8000) are supported on the same AIX host but on separate HBAs and separate zones (Figure 11-17).

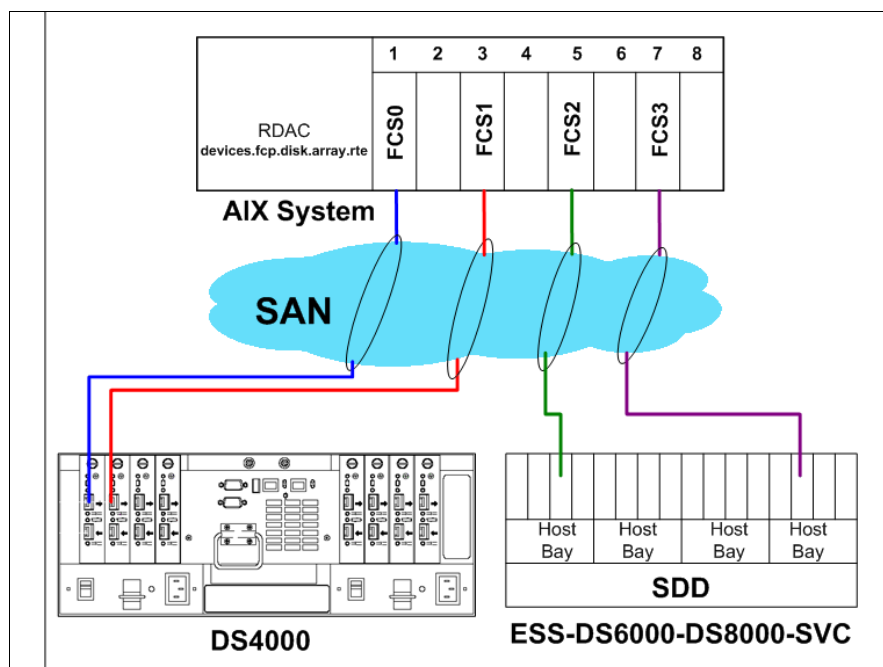


Figure 11-17 RDAC and SDD coexistence

It is possible to configure the SDD driver as used with the ESS, DS8000, DS6000 or SVC, and the RDAC driver used by the DS4000 family. However, it is necessary a pair of HBAs for access to the DS4000 and another pair of HBAs for access to the ESS or similar. You must also define separate zones for each HBA.

Therefore, the configuration with zoning as shown in Figure 11-18 is not supported.

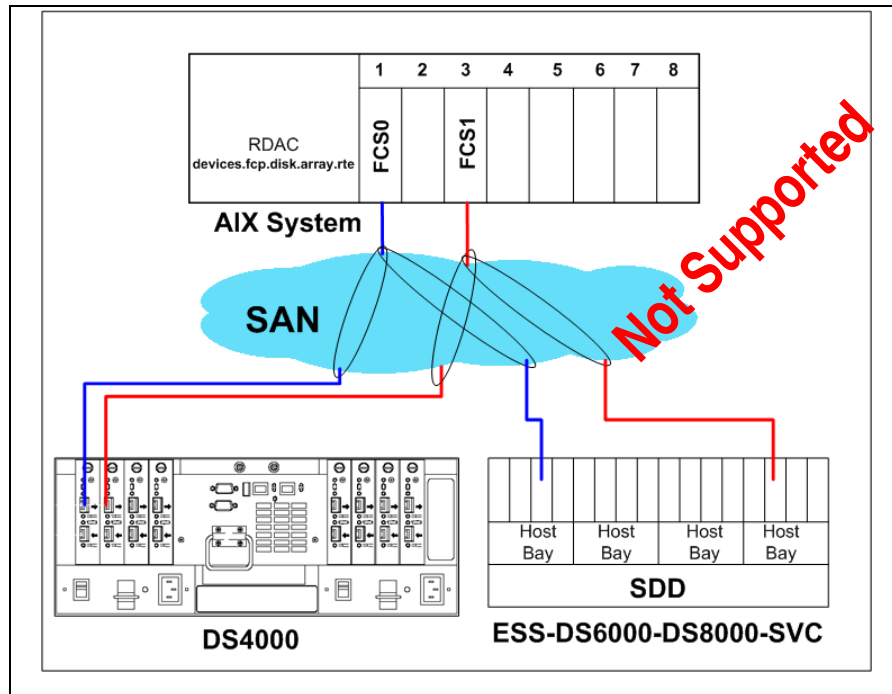


Figure 11-18 RDAC and SDD unsupported zoning configuration

Coexistence between RDAC and TAPES

Coexistence RDAC and tape devices is supported, but must have separate HBAs and separate zones (Figure 11-19).

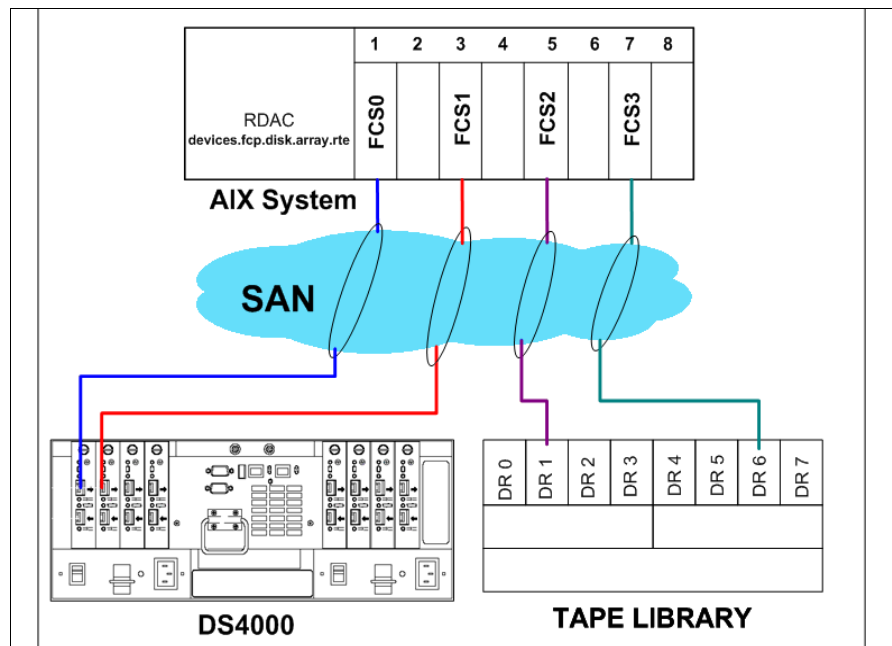


Figure 11-19 Coexistence of DS4000 and Tape - Correct zoning

Therefore, the configuration with zoning as shown in Figure 11-20 is *not supported*.

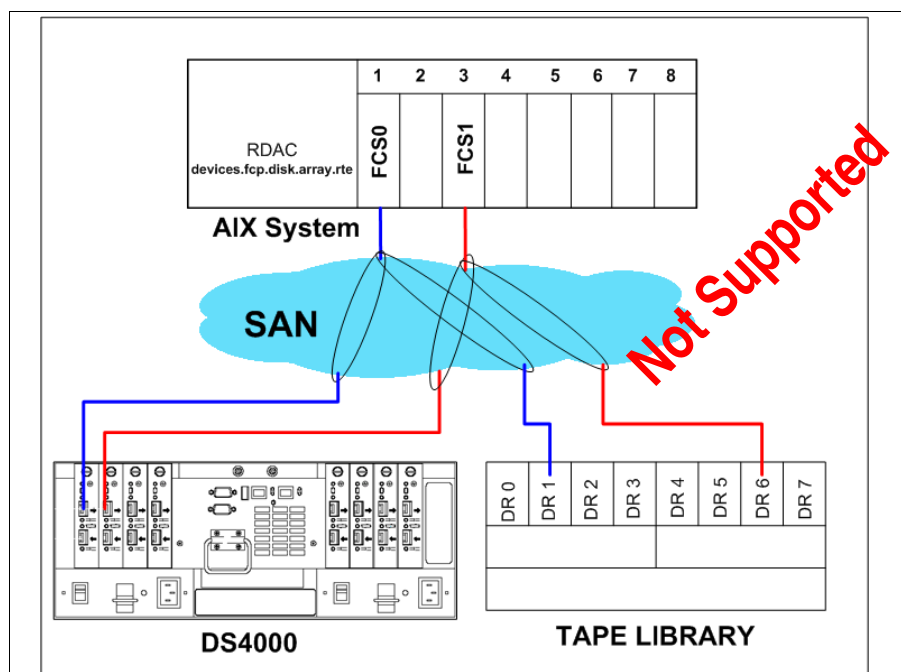


Figure 11-20 Coexistence of DS4000 and Tape - Incorrect zoning

11.1.7 Setting the HBA for best performance

There are three Fibre Channel adapter (HBA) settings that can help performance:

► **num_cmd_elems**

num_cmd_elems controls the maximum number of commands to queue to the adapter. The default value is 200. You can set it to a higher value in I/O intensive environments. The maximum of num_cmd_elems for a 2 Gb HBA is 2048. However, keep in mind that there is obviously a cost in real memory (DMA region) for a cmd_elem; this depends on the AIX version and whether it is 32 or 64 bits. In AIX 5.1 the current sizes are 232 and 288 bytes (for 32 and 64 bits, respectively), and for AIX 5.2, the sizes are 240 and 296 bytes.

There is no real recommended value. Use a performance measurement tools as discussed in Chapter 6, “Analyzing and measuring performance” on page 187 and observe the results for different values of num_cmd_elems.

► **lg_term_dma**

lg_term_dma controls the size of a DMA address region requested by the Fibre Channel driver at startup. This doesn't set aside real memory, rather it sets aside PCI DMA address space. Each device that is opened uses a portion of this DMA address region. The region controlled by lg_term_dma is not used to satisfy I/O requests.

The lg_term_dma can be set to 0x1000000 (16 MB). The default is 0x200000 (2MB). You should be able to safely reduce this to the default. The first symptom of lg_term_dma exhaustion is that disk open requests begin to fail with ENOMEM. In that case you probably wouldn't be able to vary on some VGs. Too small a value here is not likely to cause any runtime performance issue. Reducing the value will free up physical memory and DMA address space for other uses.

- **max_xfer_size**

This is the maximum I/O size that the adapter will support. The default maximum transfer size is 0x100000. Consider changing this value to 200000 or larger.

Increasing this value increases the DMA memory area used for data transfer. You should resist the urge to just set these attributes to the maximums. There is a limit to the amount of DMA region available per slot and PCI bus. Setting these values too high may result in some adapters failing to configure because other adapters on the bus have already exhausted the resources. In the other hand, if too little space is set aside here, I/Os may be delayed in the FC adapter driver waiting for previous I/Os to finish. You will generally see errors in errpt if this happens.

For more details, see 6.4.3, “Using the Performance Monitor: Illustration” on page 211.

Viewing and changing HBA settings

- To view possible attribute values of max_xfer_size for the fcs0 adapter, enter the command:

```
lsattr -Rl fcs0 -a max_xfer_size
```

- The following command changes the maximum transfer size (max_xfer_size) and the maximum number of commands to queue (num_cmd_elems) of an HBA (fcs0) upon the next system reboot:

```
chdev -l fcs0 -P -a max_xfer_size=<value> -a num_cmd_elems=<value> -P
```

This will not take effect until after the system is rebooted.

- If you want to avoid a system reboot, make sure all activity is stopped on the adapter, and issue the following command:

```
chdev -l fcs0 -a max_xfer_size=<value> -a num_cmd_elems=<value>
```

Then recreate all child devices with the **cfgmgr** command.

11.1.8 DS4000 series – dynamic functions

We know that the DS4000 offers several dynamic functions such as:

- DSS – Dynamic Segment Sizing
- DRM – Dynamic RAID Migration
- DCE – Dynamic Capacity Expansion
- DVE – Dynamic Volume Expansion

These functions are supported AIX operating environments with some exceptions (for detailed information about supported platforms and operating systems, refer to the DS4000 compatibility matrix).

For Dynamic Volume Expansion:

- DVE support for AIX requires AIX 5.2 or later.
- DVE for AIX 5.3 requires PTF U499974 installed before expanding any file systems.

In reality, in AIX there is no real Dynamic Volume Expansion because after increasing the size of the logical drive under the DS4000 Storage Manager, it is necessary, to use the additional disk space, to modify the volume group which contains that drive. A little downtime is required to perform the operation that is: stop the application, unmount all filesystems on this volume group and then varyoff the vg. At this point you need to change the characteristics of the volume group with **chvg -g vname**, varyon the volume group, mount filesystems, and restart the application.

The detailed procedure is given in the section that follows.

Increase DS4000 LUN size in AIX step by step.

This illustration assumes the following configuration:

- ▶ Two different logical drives (Kanaga_Lun0 and Kanaga_Lun1) in two separate RAID arrays on the DS4000.
- ▶ The logical drives are seen as two separate hdisk devices (hdisk3 and hdisk4) in AIX.
- ▶ In AIX LVM, we have defined one volume group (DS4000vg) with one logical volume (lvDS4000_1) striped on the two hdisks.

We show how to increase the size of the logical volume (lvDS4000-1), by increasing the size of hdisk3 and hdisk4 (thus without creating new disks).

Using the SMclient, we increase the size of the DS4000 logical drives by about 20 GB each (see Figure 11-21 and Figure 11-22).

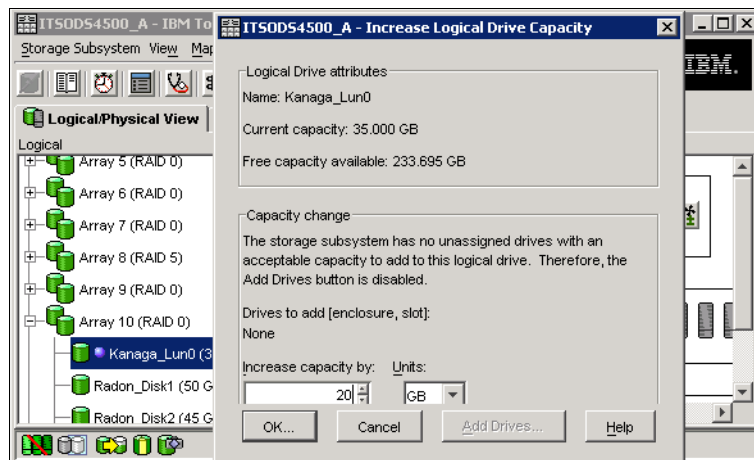


Figure 11-21 Increase the size - Kanaga_Lun0

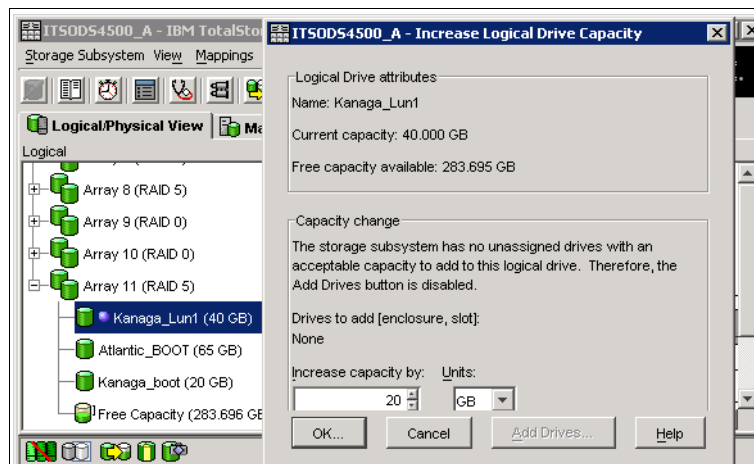


Figure 11-22 Increase the size - Kanaga_Lun1

You can verify the size of the corresponding hdisk in AIX, by typing the following command:

```
# bootinfo -s hdisk3
```

The output of the command gives the size of disk in megabyte, 61440 (that is 60 GB).

Now, stop the application, unmount all filesystems on the volume group, and varyoff the volume group.

```
varyoffvg DS4000vg
```

Modify the volume group to use the new space, with the following command:

```
chvg -g DS4000vg
```

Next, varyon the volume group (AIX informs you about the increased disk size).

```
varyonvg DS4000vg
```

```
0516-1434 varyonvg: Following physical volumes appear to be grown in size.
```

```
Run chvg command to activate the new space.
```

```
hdisk3          hdisk4
```

Finally, mount all filesystems and restart the application.

You can verify the available new space for the volume group with the command **lsvg DS4000vg**. You can now enlarge the logical volumes (Figure 11-23).

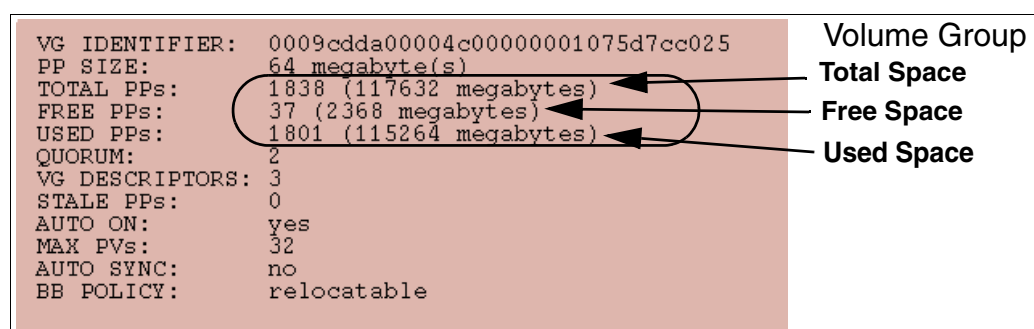


Figure 11-23 Using lsvg to see the new size

11.2 HACMP and DS4000

Clustering (of servers) is the linking of two or more computers or nodes into a single, unified resource. High-availability clusters are designed to provide continuous access to business-critical data and applications through component redundancy and application failover.

HACMP is designed to automatically detect system or network failures and eliminate a single point-of-failure by managing failover to a recovery processor with a minimal loss of end-user time. The current release of HACMP can detect and react to software failures severe enough to cause a system crash and network or adapter failures. The Enhanced Scalability capabilities of HACMP offer additional availability benefits through the use of the Reliable Scalable Cluster Technology (RSCT) function of AIX (see “HACMP/ES and ESCRM” on page 363).

HACMP makes use of redundant hardware configured in the cluster to keep an application running, restarting it on a backup processor if necessary. Using HACMP can virtually eliminate planned outages, because users, applications, and data can be moved to backup systems during scheduled system maintenance. Such advanced features as Cluster Single

Point of Control and Dynamic Reconfiguration allow the automatic addition of users, files, hardware, and security functions without stopping mission-critical jobs.

HACMP clusters can be configured to meet complex and varied application availability and recovery needs. Configurations can include mutual takeover or idle standby recovery processes. With an HACMP mutual takeover configuration, applications and their workloads are assigned to specific servers, thus maximizing application throughput and leveraging investments in hardware and software. In an idle standby configuration, an extra node is added to the cluster to back up any of the other nodes in the cluster.

In an HACMP environment, each server in a cluster is a node. Up to 32 System p servers can participate in an HACMP cluster. Each node has access to shared disk resources that are accessed by other nodes. When there is a failure, HACMP transfers ownership of shared disks and other resources based on how you define the relationship among nodes in a cluster. This process is known as node failover or node fallback.

Ultimately, the goal of any IT solution in a critical environment is to provide continuous service and data protection. The high availability is just one building block in achieving the continuous operation goal. The high availability is based on the availability of the hardware, software (operating system and its components), application, and network components.

For a high availability solution you need:

- ▶ Redundant servers
- ▶ Redundant networks
- ▶ Redundant network adapters
- ▶ Monitoring
- ▶ Failure detection
- ▶ Failure diagnosis
- ▶ Automated failover
- ▶ Automated reintegration

The main objective of the HACMP is eliminate Single Points of Failure (SPOFs) as detailed in Table 11-2.

Table 11-2 Eliminate SPOFs

Cluster object	Eliminated as a single point of failure by:
Node (servers)	Multiple nodes
Power supply	Multiple circuits or power supplies
Network adapter	Redundant network adapters
Network	Multiple networks to connect nodes
TCP/IP subsystem	A non- IP networks to back up TCP/IP
Disk adapter	Redundant disk adapters
Disk	Redundant hardware and disk mirroring or RAID technology
Application	Configuring application monitoring and backup nodes to acquire the application engine and data

Each of the items listed in Table 11-2 in the Cluster Object column is a physical or logical component that, if it fails, will result in the application being unavailable for serving clients.

The HACMP (High Availability Cluster Multi-Processing) software provides the framework and a set of tools for integrating applications in a highly available system. Applications to be integrated in a HACMP cluster require a fair amount of customization, not at the application level, but rather at the HACMP and AIX platform level. HACMP is a flexible platform that allows integration of generic applications running on AIX platform, providing for high available systems at a reasonable cost.

HACMP classic

High Availability Subsystem (HAS) uses the global Object Data Manager (ODM) to store information about the cluster configuration and can have up to eight HACMP nodes in a HAS cluster. HAS provides the base services for cluster membership, system management, and configuration integrity. Control, failover, recovery, cluster status, and monitoring facilities are also there for programmers and system administrators.

The Concurrent Resource Manager (CRM) feature optionally adds the concurrent shared-access management for the supported RAID and SSA disk subsystem. Concurrent access is provided at the raw logical volume level, and the applications that use CRM must be able to control access to the shared data. The CRM includes the HAS, which provides a distributed locking facility to support access to shared data.

Before HACMP Version 4.4.0, if there was a need for a system to have high availability on a network file system (NFS), the system had to use high availability for the network file system (HANFS). HANFS Version 4.3.1 and earlier for AIX software provides a reliable NFS server capability by allowing a backup processor to recover current NFS activity should the primary NFS server fail. The HANFS for AIX software supports only two nodes in a cluster.

Since HACMP Version 4.4.0, the HANFS features are included in HACMP, and therefore, the HANFS is no longer a separate software product.

HACMP/ES and ESCR

Scalability, support of large clusters, and therefore, large configurations of nodes and potentially disks leads to a requirement to manage “clusters” of nodes. To address management issues and take advantage of new disk attachment technologies, HACMP Enhanced Scalable (HACMP/ES) was released. This was originally only available for the SP where tools were already in place with PSSP to manage larger clusters.

ESCRM optionally adds concurrent shared-access management for the supported RAID and SSA disk subsystems. Concurrent access is provided at the raw disk level. The application must support some mechanism to control access to the shared data, such as locking. The ESCRM components includes the HACMP/ES components and the HACMP distributed lock manager.

11.2.1 Supported environment

Important: Before installing DS4000 in an HACMP environment, always read the AIX *readme* file, the DS4000 *readme* for the specific Storage Manager version and model, and the HACMP configuration and compatibility matrix information.

For up-to-date information about the supported environments for DS4000, refer to:

<http://www.ibm.com/servers/storage/disk/ds4000/ds4500/interop.html>

For HACMP, refer to the following site:

http://www.ibm.com/servers/eserver/pseries/library/hacmp_doc.html

11.2.2 General rules

The primary goal of an HACMP environment is to eliminate single points of failure. Figure 11-24 below contains a diagram of a two-node HACMP cluster (this is not a limitation; you can have more nodes) attached to a DS4000 Storage Server through a fully redundant Storage Area Network. This type of configuration eliminates a Fibre Channel (FC) adapter, switch, or cable from being a single point of failure (HACMP itself protects against a node failure).

Using only one single FC switch would be possible (with additional zoning), but would be considered a single point of failure. If the FC switch fails, you cannot access the DS4000 volumes from either HACMP cluster node. So, with only a single FC switch, HACMP would be useless in the event of a switch failure. This example would be the recommended configuration for a fully redundant production environment. Each HACMP cluster node should also contain two Fibre Channel host adapters to eliminate the adapter as a single point of failure. Notice also that each adapter in a particular cluster node goes to a separate switch (cross cabling).

DS4000 models can be ordered with more hosts ports. In the previous example, only two host attachments are needed. Buying additional mini hubs is not necessary, but can be done for performance or security reasons. Zoning on the FC switch must be done as detailed in Figure 11-1 on page 343. Every adapter in the AIX system can see only one controller (these are AIX-specific zoning restrictions, not HACMP specific).

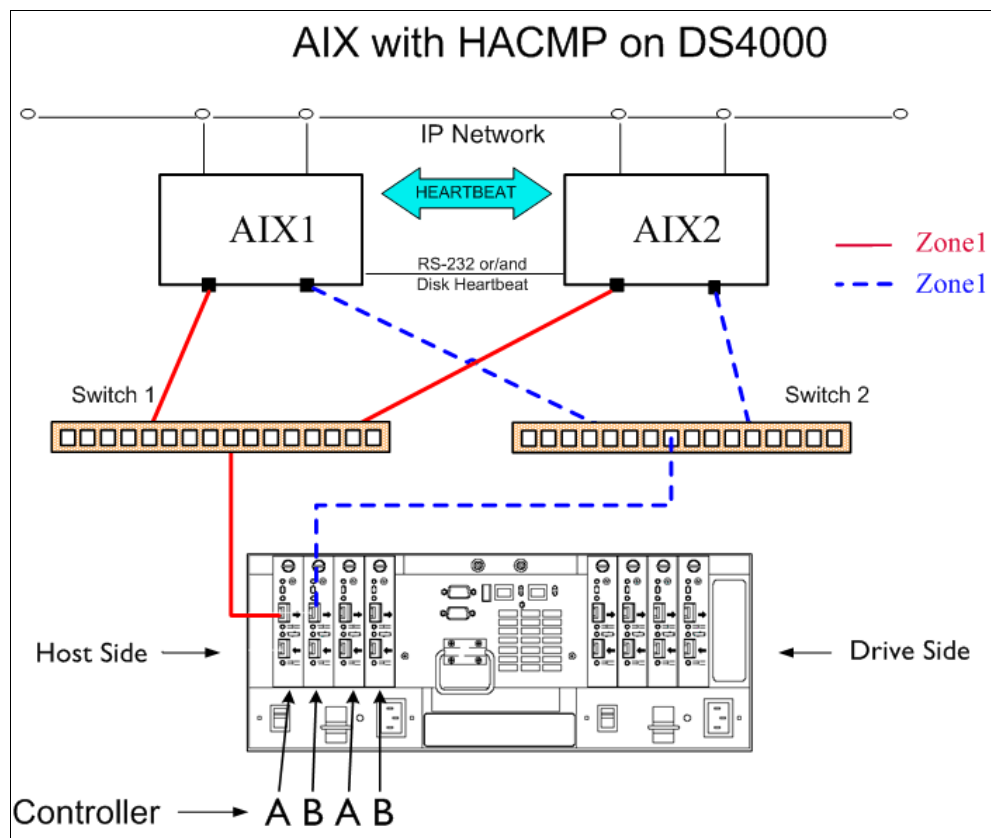


Figure 11-24 HACMP cluster and DS4000

11.2.3 Configuration limitations

When installing a DS4000 in an HACMP environment, there are some restrictions and guidelines to take into account, which we list here. It does not mean that any other configuration will fail, but it could lead to unpredictable results, making it hard to manage and troubleshoot.

Applicable pSeries and AIX limitations (not HACMP specific)

The following AIX and pSeries restrictions apply to DS4100, DS4300, DS4400, DS4500, DS4200, DS4700, and DS4800 Storage Servers:

- ▶ DS4100 and DS4300 single-controller storage subsystems are not supported with AIX hosts. (DS4100 dual-controller and DS4300 Base and Turbo dual-controller storage subsystems are supported.)
- ▶ A maximum of four HBAs per AIX host (or LPARs) can be connected to a single DS4000 storage server. You can configure up to two HBAs per partition and up to two partitions per DS4000 storage server. Additional HBAs can be added to support additional DS4000 storage servers and other SAN devices, up to the limits of your specific server platform.
- ▶ All volumes that are configured for AIX must be mapped to an AIX host group. Connecting and configuring to volumes in the default host group is not allowed.
- ▶ You cannot use dynamic volume expansion (DVE) on AIX 5.1.
- ▶ Other storage devices, such as tape devices or other disk storage, must be connected through separate HBAs and SAN zones.
- ▶ Each AIX host attaches to DS4000 Storage Servers using pairs of Fibre Channel adapters (HBAs):
 - For each adapter pair, one HBA must be configured to connect to controller A, and the other to controller B.
 - Each HBA pair must be configured to connect to a single partition in a DS4000 Storage Server or multiple DS4000 Storage Servers (fanout).
 - To attach an AIX host to a single or multiple DS4000 with two partitions, two HBA pairs must be used.
- ▶ The maximum number of DS4000 partitions (host groups) per AIX host per DS4000 storage subsystem is two.
- ▶ Zoning must be implemented. If zoning is not implemented in a proper way, devices might appear on the hosts incorrectly. Follow these rules when implementing the zoning:
 - Single-switch configurations are allowed, but each HBA and DS4000 controller combination must be in a separate SAN zone.
 - Each HBA within a host must be configured in a separate zone from other HBAs within that same host when connected to the same DS4000 controller port. In other words, only one HBA within a host can be configured with a given DS4000 controller port in the same zone.
 - Hosts within a cluster can share zones with each other.
 - For highest availability, distributing the HBA and DS4000 connections across separate FC switches minimizes the effects of a SAN fabric failure.

General limitations and restrictions for HACMP

Keep in mind the following general limitations and restrictions for HACMP:

- ▶ Switched fabric connections between the host nodes and the DS4000 storage subsystem are recommended. However, direct attachment from the host nodes to the DS4000 storage subsystem in an HACMP environment is now supported when all of the following restrictions and limitations are met:
 - Only dual-controller DS4000 storage subsystem versions are supported for direct attachment in a high-availability (HA) configuration.
 - The AIX operating system must be Version 5.2 or later.
 - The HACMP clustering software must be Version 5.1 or later.
 - All host nodes that are directly attached to the DS4000 storage subsystem must be part of the same HACMP cluster.
 - All logical drives (LUNs) that are surfaced by the DS4000 storage subsystem are part of one or more enhanced concurrent-mode volume groups.
 - The volume group varyon is in the active state only on the host node that owns the HACMP non-concurrent resource group (which contains the enhanced concurrent-mode volume group or groups). For all other host nodes in the HACMP cluster, the enhanced concurrent-mode volume group varyon is in the passive state.
 - Direct operations on the logical drives in the enhanced concurrent-mode volume groups cannot be performed from any host nodes in the HACMP cluster if the operations bypass the Logical Volume Manager (LVM) layer of the AIX operating system. For example, you cannot use a DD command while logged in as the root user.
 - Each host node in the HACMP cluster must have two Fibre Channel connections to the DS4000 storage subsystem. One direct Fibre Channel connection must be to controller A in the DS4000 storage subsystem, and the other direct Fibre Channel connection must be to controller B in the DS4000 storage system.
 - You can directly attach a maximum of two host nodes in an HACMP cluster to a DS4000 storage subsystem. Each host node must have two direct Fibre Channel connections to the storage subsystem. In a DS4400 or DS4500 storage subsystem, the two direct Fibre Channel connections from each host node must connect to independent mini-hubs. Therefore, this configuration requires that four host mini-hubs (feature code 3507) be installed in the DS4400 or DS4500 storage subsystem — two host mini-hubs for each host node in the HACMP cluster.
- ▶ HACMP Cluster-Single Point of Control (C-SPOC) cannot be used to add a DS4000 disk to AIX through the *Add a Disk to the Cluster* facility.
- ▶ HACMP C-SPOC does not support enhanced concurrent mode volume groups.
- ▶ Concurrent and non-concurrent modes are supported with HACMP Versions 5.1, 5.2, and 5.3 and DS4000 running Storage Manager Versions 8.3 or later, including Hot Standby and Mutual Take-over.
- ▶ HACMP Versions 5.1, 5.2 and 5.3 are supported on the pSeries 690 LPAR and 590 LPAR clustered configurations.
- ▶ HACMP is now supported in Heterogeneous server environments. For more information regarding a particular operating system environment, refer to the specific *Installation and Support Guide*.
- ▶ HACMP clusters can support 2-32 servers for DS4000 partition. In this environment, be sure to read and understand the AIX device drivers queue depth settings, as documented in *IBM System Storage DS4000 Storage Manager Version 9, Installation and Support Guide for AIX, HP-UX, Solaris, and Linux on POWER*, GC26-7848-02.

- ▶ Non-clustered AIX hosts can be connected to the same DS4000 that is attached to an HACMP cluster, but must be configured on separate DS4000 host partitions.
- ▶ Single HBA configurations are allowed, but each single HBA configuration requires that both controllers in the DS4000 be connected to a switch within the same SAN zone as the HBA. While single HBA configurations are supported, using a single HBA configuration is not recommended for HACMP environments, due to the fact that it introduces a single point of failure in the storage I/O path.

11.2.4 Planning considerations

When planning a high availability cluster, you should consider the sizing of the nodes, storage, network and so on, to provide the necessary resources for the applications to run properly, even in a takeover situation.

Sizing: choosing the nodes in the cluster

Before you start the implementation of the cluster, you should know how many nodes are required, and the type of the nodes that should be used. The type of nodes to be used is important in terms of the resources required by the applications.

Sizing of the nodes should cover the following aspects:

- ▶ CPU (number of CPUs and speed)
- ▶ Amount of random access memory (RAM) in each node
- ▶ Disk storage (internal)
- ▶ Number of communication and disk adapters in each node
- ▶ Node reliability

The number of nodes in the cluster depends on the number of applications to be made highly available, and also on the degree of availability desired. Having more than one spare node for each application in the cluster increases the overall availability of the applications.

HACMP V5.1, V5.2 and V5.3 support a variety of nodes, ranging from desktop systems to high-end servers. SP nodes and Logical Partitions (LPARs) are supported as well.

The cluster resource sharing is based on the applications requirements. Nodes that perform tasks that are not directly related to the applications to be made highly available and do not need to share resources with the application nodes should be configured in separate clusters for easier implementation and administration.

All nodes should provide sufficient resources (CPU, memory, and adapters) to sustain execution of all the designated applications in a fail-over situation (to take over the resources from a failing node).

We recommend using cluster nodes with a similar hardware configuration, especially when implementing clusters with applications in mutual takeover or concurrent configurations. This makes it easier to distribute resources and to perform administrative operations (software maintenance and so on).

Sizing: storage considerations

Applications to be made highly available require a shared storage space for application data. The shared storage space is used either for concurrent access, or for making the data available to the application on the takeover node (in a fail-over situation).

The storage to be used in a cluster should provide shared access from all designated nodes for each application. The technologies currently supported for HACMP shared storage are SCSI, SSA, and Fibre Channel — as is the case with the DS4000.

The storage configuration should be defined according to application requirements as non-shared (“private”) or shared storage. The private storage may reside on internal disks and is not involved in any takeover activity.

Shared storage should provide mechanisms for controlled access, considering the following reasons:

- ▶ Data placed in shared storage must be accessible from whichever node the application may be running at a point in time. In certain cases, the application is running on only one node at a time (non-concurrent), but in some cases, concurrent access to the data must be provided.
- ▶ In a non-concurrent environment, if the shared data is updated by the wrong node, this could result in data corruption.
- ▶ In a concurrent environment, the application should provide its own data access mechanism, since the storage controlled access mechanisms are by-passed by the platform concurrent software (AIX/HACMP).

11.2.5 Cluster disks setup

The following sections relate important information about cluster disk setup, and in particular describes cabling, AIX configuration, microcode loading, and configuration of DS4000 disks.

Figure 11-25 shows a simple two-node HACMP cluster and the basic cabling that we recommend. This configuration ensures redundancy and allows for possible future expansion and also could support remote mirroring because it leaves two available controllers on the DS4000.

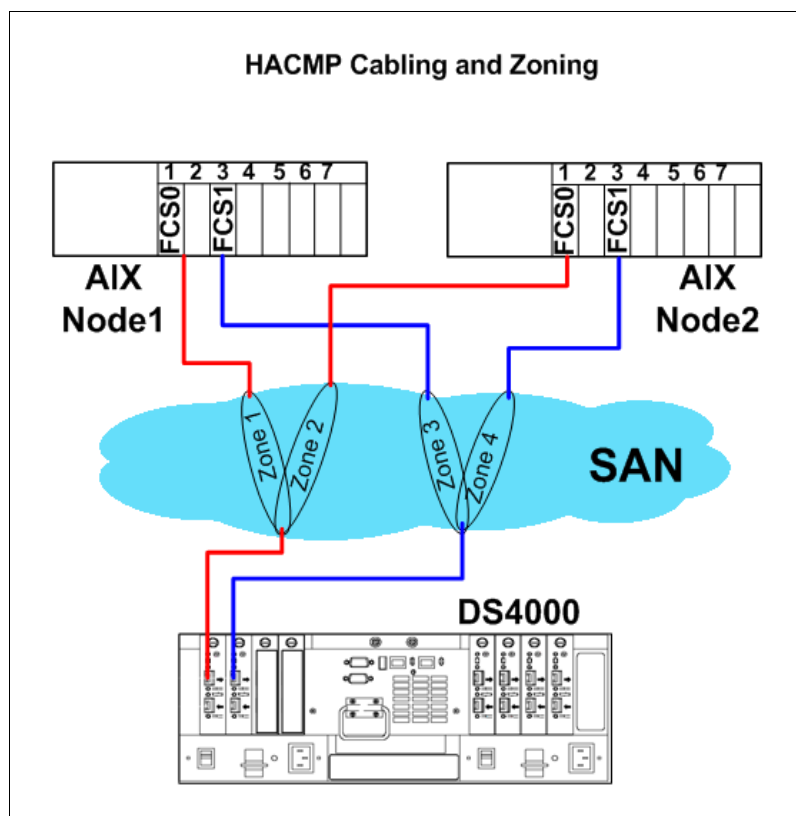


Figure 11-25 HACMP - Recommended cabling and zoning

Logon to each of the AIX nodes in the cluster and verify that you have a working configuration. You should get an output similar to what is illustrated below for the various commands:

► Node 1:

```
# lsdev -Ccadapter
```

```
fcs0    Available 1Z-08    FC Adapter
fcs1    Available 1D-08    FC Adapter
```

```
# lsdev -C | grep ^dar
```

```
dar0     Available                1742-900 (900) Disk Array Router
```

```
# lsdev -C | grep dac
```

```
dac0     Available 1Z-08-02    1742-900 (900) Disk Array Controller
dac1     Available 1D-08-02    1742-900 (900) Disk Array Controller
```

► Node 2:

```
# lsdev -Ccadapter
```

```
fcs0    Available 1Z-08    FC Adapter
fcs1    Available 1D-08    FC Adapter
```

```
# lsdev -C | grep ^dar
```

```
dar0     Available                1742-900 (900) Disk Array Router
```

```
# lsdev -C | grep dac
dac0      Available 1Z-08-02      1742-900 (900) Disk Array Controller
dac1      Available 1D-08-02      1742-900 (900) Disk Array Controller
```

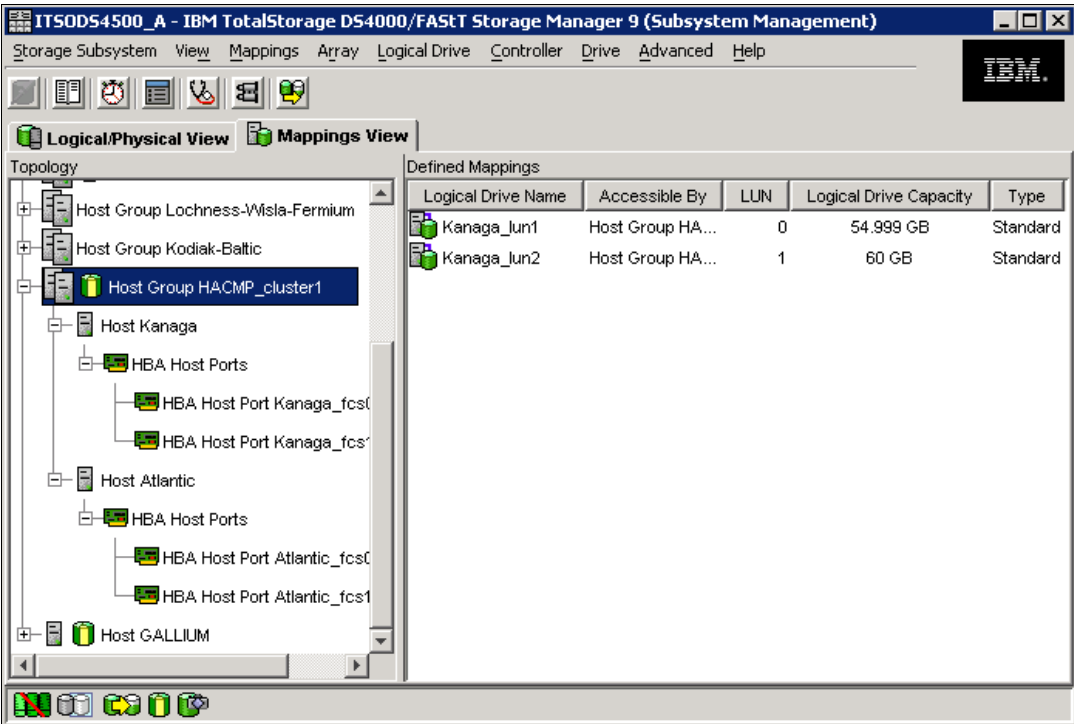


Figure 11-26 Cluster - Host Group and Mappings

Using Storage Manager, define a Host Group for the cluster and include the different hosts (nodes) and host ports as illustrated in Figure 11-26:

- Verify the disks on node 1:

```
# lsdev -Ccdisk

hdisk3 Available 1Z-08-02      1742-900 (900) Disk Array Device
hdisk4 Available 1D-08-02      1742-900 (900) Disk Array Device

# lspv

hdisk3      0009cdda4d835236      None
hdisk4      0009cdda4d835394      None
```

- Verify the disks on node 2:

```
# lsdev -Ccdisk

hdisk3 Available 1Z-08-02      1742-900 (900) Disk Array Device
hdisk4 Available 1D-08-02      1742-900 (900) Disk Array Device

lspv

hdisk3      0009cdda4d835236      None
hdisk4      0009cdda4d835394      None
```

- The zoning should look as follows:

```
Node1
zone: ATLANTIC900_0 20:04:00:a0:b8:17:44:32;10:00:00:00:c9:32:a8:0a
zone: ATLANTIC900_1 20:05:00:a0:b8:17:44:32;10:00:00:00:c9:4c:8c:1c

Node2
zone: KANAGAF900_0 20:04:00:a0:b8:17:44:32;10:00:00:00:c9:32:a7:fb
zone: KANAGAF900_1 20:05:00:a0:b8:17:44:32;10:00:00:00:c9:32:a7:d1
```

11.2.6 Shared LVM component configuration

This section describes how to define the LVM components shared by cluster nodes in an HACMP for AIX cluster environment.

Creating the volume groups, logical volumes, and file systems shared by the nodes in an HACMP cluster requires that you perform steps on all nodes in the cluster. In general, you define the components on one node (source node) and then import the volume group on the other nodes in the cluster (destination nodes). This ensures that the ODM definitions of the shared components are the same on all nodes in the cluster.

Non-concurrent access environments typically use journaled file systems to manage data, while concurrent access environments use raw logical volumes. This chapter provides different instructions for defining shared LVM components in non-concurrent access and concurrent access environments.

Creating shared VG

The following sections contain information about creating non-concurrent VGs and VGs for concurrent access.

Creating non-concurrent VG

Use the **smit mkvg** fast path to create a shared volume group. Use the default field values unless your site has other requirements. See Table 11-3 for the **smit mkvg** options.

Table 11-3 *smit mkvg options*

Options	Description
VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
Physical partition SIZE in megabytes	Accept the default.
PHYSICAL VOLUME NAMES	Specify the names of the physical volumes you want included in the volume group.
Activate volume group AUTOMATICALLY at system restart?	Set to no so that the volume group can be activated as appropriate by the cluster event scripts.
Volume Group MAJOR NUMBER	If you are not using NFS, you can use the default (which is the next available number in the valid range). If you are using NFS, you must make sure to use the same major number on all nodes. Use the lvfstmajor command on each node to determine a free major number common to all nodes.
Create VG concurrent capable?	Set this field to no (leave default)
Auto-varyon concurrent mode?	Accept the default.

Creating VG for concurrent access

The procedure used to create a concurrent access volume group varies, depending on which type of device you are using. In our case we will assume DS4000 disks.

To use a concurrent access volume group, defined on a DS4000 disk subsystem, you must create it as a concurrent-capable volume group. A concurrent-capable volume group can be activated (varied on) in either non-concurrent mode or concurrent access mode.

To define logical volumes on a concurrent-capable volume group, it must be varied on in non-concurrent mode.

Use **smit mkvg** with the options shown in Table 11-4 to build the volume group.

Table 11-4 Options for *volumn group*

Options	Description
VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
Physical partition SIZE in megabytes	Accept the default.
PHYSICAL VOLUME NAMES	Specify the names of the physical volumes you want included in the volume group.
Activate volume group AUTOMATICALLY at system restart?	Set this field to no so that the volume group can be activated, as appropriate, by the cluster event scripts.
Volume Group MAJOR NUMBER	While it is only really required when you are using NFS, it is always good practice in an HACMP cluster to have a shared volume group have the same major number on all the nodes that serve it. Use the lvfstmajor command on each node to determine a free major number common to all nodes.
Create VG concurrent capable?	Set this field to yes so that the volume group can be activated in concurrent access mode by the HACMP for AIX event scripts
Auto-varyon concurrent mode?	Set this field to no so that the volume group can be activated, as appropriate, by the cluster event scripts.

Creating shared LV and file systems

Use the **smit crjfs** fast path to create the shared file system on the source node. When you create a journaled file system, AIX creates the corresponding logical volume. Therefore, you do not need to define a logical volume. You do, however, need to later rename both the logical volume and the log logical volume for the file system and volume group (Table 11-5).

Table 11-5 *smit crjfs options*

Options	Description
Mount AUTOMATICALLY at system restart?	Make sure this field is set to no .
Start Disk Accounting	Make sure this field is set to no .

Renaming a jfslog and logical volumes on the source node

AIX assigns a logical volume name to each logical volume it creates. Examples of logical volume names are **/dev/lv00** and **/dev/lv01**. Within an HACMP cluster, the name of any shared logical volume must be unique. Also, the journaled file system log (jfslog) is a logical volume that requires a unique name in the cluster.

To make sure that logical volumes have unique names, rename the logical volume associated with the file system and the corresponding jfslog logical volume. Use a naming scheme that indicates the logical volume is associated with a certain file system. For example, **lvsharefs** could name a logical volume for the **/sharefs** file system. Follow these steps to rename the logical volumes:

1. Use the **lsvg -l volume_group_name** command to determine the name of the logical volume and the log logical volume (jfslog) associated with the shared volume groups. In the resulting display, look for the logical volume name that has type **jfs**. This is the logical

volume. Then look for the logical volume name that has type jfslog. This is the log logical volume.

2. Use the **smit chl v** fast path to rename the logical volume and the log logical volume.
3. After renaming the jfslog or a logical volume, check the /etc/filesystems file to make sure the dev and log attributes reflect the change. Check the log attribute for each file system in the volume group, and make sure that it has the new jfslog name. Check the dev attribute for the logical volume that you renamed, and make sure that it has the new logical volume name.

Importing to other nodes

The following sections cover varying off a volume group on the source node, importing it onto the destination node, changing its startup status, and varying it off on the destination nodes.

Varying off a volume group on the source node

Use the **varyoffvg** command to deactivate the shared volume group. You vary off the volume group so that it can be properly imported onto a destination node and activated as appropriate by the cluster event scripts. Enter the following command:

```
varyoffvg volume_group_name
```

Make sure that all the file systems of the volume group have been unmounted; otherwise, the **varyoffvg** command will not work.

Importing a volume group onto the destination node

To import a volume group onto destination nodes you can use the SMIT interface or the TaskGuide utility. The TaskGuide uses a graphical interface to guide you through the steps of adding nodes to an existing volume group. Importing the volume group onto the destination nodes synchronizes the ODM definition of the volume group on each node on which it is imported.

You can use the **smit importvg** fast path to import the volume group (Table 11-7 on page 375).

Table 11-6 *smit importvg options*

Options	Description
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 source node.
PHYSICAL VOLUME name	Enter the name of a physical volume that resides in the volume group. Note that a disk <i>may have</i> a different logical name on different nodes. Make sure that you use the disk name as it is defined on the destination node.
Volume Group MAJOR NUMBER	If you are not using NFS, you may use the default (which is the next available number in the valid range). If you are using NFS, you must make sure to use the same major number on all nodes. Use the lvlsmtmajor command on each node to determine a free major number common to all nodes.

Changing a volume group startup status

By default, a volume group that has just been imported is configured to automatically become active at system restart. In an HACMP for AIX environment, a volume group should be varied on as appropriate by the cluster event scripts. Therefore, after importing a volume group, use

the SMIT Change a Volume Group window to reconfigure the volume group so that it is not activated automatically at system restart.

Use the **smit chvg** fast path to change the characteristics of a volume group (Table 11-7).

Table 11-7 *smit chvg options*

Options	Description
Activate volume group automatically at system restart?	Set this field to no.
A QUORUM of disks required to keep the volume group online?	If you are using DS4000 with raid protection set this field to no.

Varying off the volume group on the destination nodes

Use the **varyoffvg** command to deactivate the shared volume group so that it can be imported onto another destination node or activated as appropriate by the cluster event scripts. Enter:

```
# varyoffvg volume_group_name
```

11.2.7 Fast disk takeover

By utilizing the Enhanced Concurrent Mode volume groups, in non-concurrent resource groups, it almost eliminates disk takeover time by removing the need for disk reserves, and breaking these reserves. The volume groups are varied online in *Active* mode on only the owning node, and all other fall over candidate nodes have it varied on in *Passive* mode. RSCT is utilized for communications to coordinate activity between the nodes so that only one node has it varied on actively.

Time for disk takeover now is fairly consistent at around 10 seconds. Now with multiple resource groups and hundreds to thousands of disks, it may be a little more — but not significantly more.

Note that while the VGs are varied on concurrently to both nodes, **1svg -o** will only show the VG as active on the node accessing the disk; however, running **1spv** will show that the VG disks are active on both nodes. Also note, that **1svg vname** will tell you if the VG is in the active or passive state.

11.2.8 Forced varyon of volume groups

For situations in which one is using LVM mirroring, and would like to survive failure of half the disks, there is a new attribute in the Resource Group smit panel to force varyon of the VG, provided that a complete copy of the LVs are available. Previously, this was accomplished by setting the HACMP_MIRROR_VARYON environment variable to yes, or via user written pre/post/recovery event scripts.

The new attribute in the Resource Group Smit panel is as follows:

Volume Groups Use forced varyon of volume groups, if necessary [false]

To take advantage of this feature, set this attribute to *true*, as the default is false.

11.2.9 Heartbeat over disks

This provides the ability to use existing shared disks, regardless of disk type, to provide serial network type connectivity. This can replace needs of using integrated serial ports or 8-port async adapters and the additional cables needed for it.

This feature utilizes a special reserve area, previously used by SSA Concurrent volume groups. Since Enhanced Concurrent does not use this space, it makes it available for use. This also means that the disk chosen for serial heartbeat can be, and probably will normally be, part of a data volume group.

The disk heartbeating code went into the 2.2.1.30 version of RSCT. Some recommended APARs bring that to 2.2.1.31. If you've got that level installed, and HACMP 5.1, you can use disk heartbeating. The relevant file to look for is `/usr/sbin/rsct/bin/hats_diskhb_nim`. Though it is supported mainly through RSCT, we recommend AIX 5.2 when utilizing disk heartbeat.

In HACMP 5.1 with AIX 5.1, enhanced concurrent mode volume groups can be used only in concurrent (or “online on all available nodes”) resource groups. At AIX 5.2, disk heartbeats can exist on an enhanced concurrent VG that resides in a non-concurrent resource group.

To use disk heartbeats, no node can issue a SCSI reserve for the disk. This is because both nodes using it for heartbeating must be able to read and write to that disk. It is sufficient that the disk be in an enhanced concurrent volume group to meet this requirement.

Creating a disk heartbeat device in HACMP v5.1 or later

This example consists of a two-node cluster (nodes Atlantic and Kanaga) with shared DS4000 disk devices. If more than two nodes exist in your cluster, you will need N number of non-IP heartbeat networks, where N represents the number of nodes in the cluster. (that is, a three node cluster requires three non-IP heartbeat networks). This creates a heartbeat ring (Figure 11-27).

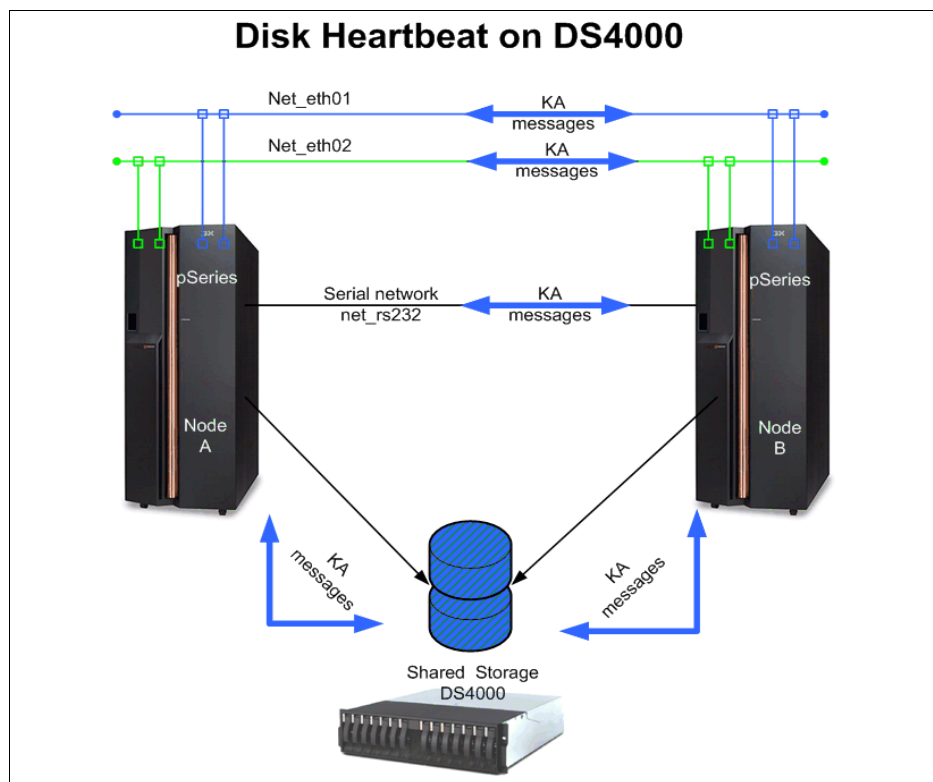


Figure 11-27 Disk heartbeat device

Prerequisites

We assumed that the shared storage devices are already made available and configured to AIX, and that the proper levels of RSCT and HACMP are already installed.

Note: Since utilizing enhanced-concurrent volume groups, it is also necessary to make sure that `bos.c1vm.enh` is installed. This is not normally installed as part of a HACMP installation via the `installp` command.

Configuring disk heartbeat

As mentioned previously, disk heartbeat utilizes enhanced-concurrent volume groups. If starting with a new configuration of disks, you will want to create enhanced concurrent volume groups by utilizing Cluster-Single Point of Control (C-SPOC).

To be able to use Cluster-Single Point of Control (C-SPOC) successfully, it is required that some basic IP based topology already exists, and that the storage devices have their PVIDs in both system's ODMs. This can be verified by running `lspv` on each system. If a PVID does not exist on each system, it is necessary to run `chdev -l <hdisk#> -a pv=yes` on each system. This will allow Cluster-Single Point of Control (C-SPOC) to match up the devices as known shared storage devices.

In the following example, since hdisk devices are being used, the following `smitty` screen paths were used.

`smitty cl_admin` → Go to **HACMP Concurrent Logical Volume Management** → **Concurrent Volume Groups** → **Create a Concurrent Volume Group**.

Choose the appropriate nodes, and then choose the appropriate shared storage devices based on pvids (hdiskx). Choose a name for the VG, desired PP size, make sure that Enhanced Concurrent Mode is set to true and press Enter. This will create the shared enhanced-concurrent vg needed for our disk heartbeat.

Attention: It is a good idea to verify via `lspv` once this has completed to make sure the device and VG is shown appropriately.

On node1 Atlantic:

```
# lspv
      hdisk7 000a7f5af78e0cf4 hacmp_hb_vg
```

On node2 Kanaga:

```
# lspv
      hdisk3 000a7f5af78e0cf4 hacmp_hb_vg
```

Creating disk heartbeat devices and network

There are two different ways to do this. Since we have already created the enhanced concurrent vg, we can use the discovery method (1) and let HA find it for us. Or we can do this manually via the Pre-defined devices method (2). Following is an example of each.

1. Creating via Discover Method

Enter `smitty hacmp` and select **Extended Configuration** → **Discover HACMP-related Information from Configured Nodes**.

This will run automatically and create a `clip_config` file that contains the information it has discovered. Once completed, go back to the Extended Configuration menu and choose **Extended Topology Configuration** → **Configure HACMP Communication Interfaces/Devices** → **Add Communication Interfaces/Devices** → **Add Discovered Communication Interface and Devices** → **Communication Devices**. Choose the appropriate devices (for example, hdisk3 and hdisk7)

- Select **Point-to-Point Pair of Discovered Communication Devices to Add**.
- Move the cursor to the desired item and press F7. Use arrow keys to scroll.
- One or more* items can be selected.
- Press Enter *after* making all selections.

```
# Node Device Pvid
> nodeAtlantic hdisk7 000a7f5af78
> nodeKanaka hdisk3 000a7f5af78
```

2. Creating via Pre-Defined Devices Method

When using this method, it is necessary to create a diskhb network first, then assign the disk-node pair devices to the network. Create the diskhb network by entering `smitty hacmp` and selecting **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Networks** → **Add a Network to the HACMP cluster**.

Choose **diskhb**.

Enter the desired network name (for example, disknet1) and press Enter. Enter smitty hacmp and select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Communication Interfaces/Devices** → **Add Communication Interfaces/Devices** → **Add Pre-Defined Communication Interfaces and Devices**.

- a. Communication Devices → Choose your diskhb Network Name
- b. Add a communication device.
- c. Type or select values in the entry fields.
- d. Press Enter *after* making all desired changes:

Device Name	[Atalantic_hboverdisk]
Network Type	diskhb
Network Name	disknet1
Device Path	[/dev/hdisk7]
Node Name	[Atlantic]

For Device Name, that is a unique name you can chose. It will show up in your topology under this name, much like serial heartbeat and ttys have in the past.

For the Device Path, you want to put in /dev/<device name>. Then choose the corresponding node for this device and device name (ex. Atlantic). Then press Enter.

You will repeat this process for the other node (for example, Kanaga) and the other device (hdisk3). This will complete both devices for the diskhb network.

Testing disk heartbeat connectivity

Once the device and network definitions have been created, test the system and make sure communications are working properly. (If the volume group is varied on in normal mode on one of the nodes, the test will probably not work, so make sure it is varied off).

To test the validity of a disk heartbeat connection, use the following command:

```
/usr/sbin/rsct/bin/dhb_read
```

The usage of dhb_read is as follows:

```
dhb_read -p devicename    //dump diskhb sector contents
dhb_read -p devicename -r  //receive data over diskhb network
dhb_read -p devicename -t  //transmit data over diskhb network
```

To test that disknet1, in our example configuration, can communicate from nodeB (Atlantic) to nodeA (Kanaga), you would run the following commands:

- On nodeA, enter:
dhb_read -p hdisk7-r
- On nodeB, enter:
dhb_read -p hdisk3 -t

If the link from nodeB to nodeA is operational, both nodes will display:

Link operating normally.

You can run this again and swap which node transmits and which one receives. To make the network active, it is necessary to sync up the cluster. Since the volume group has not been added to the resource group, we will sync up once instead of twice.

Add shared disk as a shared resource

In most cases you would have your diskhb device on a shared data volume group. It is necessary to add that VG into your resource group and synchronize the cluster.

1. Use the command `smitty hacmp` and select **Extended Configuration → Extended Resource Configuration → Extended Resource Group Configuration → Change/Show Resources and Attributes for a Resource Group**.
2. Press Enter and choose the appropriate resource group.
3. Enter the new vg (enhconcvg) into the volume group list and press Enter.
4. Return to the top of the Extended Configuration menu and synchronize the cluster.

Monitor disk heartbeat

Once the cluster is up and running, you can monitor the activity of the disk (actually all) heartbeats using `lssrc -ls topsvcs`. This command gives an output similar to the following:

```
Subsystem Group PID Status
topsvcs topsvcs 32108 active
Network Name Indx Defd Mbrs St Adapter ID Group ID
disknet1 [ 3] 2 2 S 255.255.10.0 255.255.10.1
disknet1 [ 3] hdisk3 0x86cd1b02 0x86cd1b4f
HB Interval = 2 secs. Sensitivity = 4 missed beats
Missed HBs: Total: 0 Current group: 0
Packets sent : 229 ICMP 0 Errors: 0 No mbuf: 0
Packets received: 217 ICMP 0 Dropped: 0
NIM's PID: 28724
```

Be aware that there is a grace period for heartbeats to start processing. This is normally around 60 seconds. So if you run this command quickly after starting the cluster, you may not see anything at all until heartbeat processing is started after the grace period time has elapsed.

Performance concerns with disk heartbeat

Most modern disks take somewhere around 15 milliseconds to service an I/O request, which means that they cannot do much more than 60 seeks per second. The sectors used for disk heartbeating are part of the VGDA, which is at the outer edge of the disk, and may not be near the application data.

This means that every time a disk heartbeat is done, a seek will have to be done. Disk heartbeating will typically (with the default parameters) require four (4) seeks per second. That, is each of two nodes will write to the disk and read from the disk once/second, for a total of 4 IOPS. So, if possible, a disk should be selected as a heartbeat path that does not normally do more than about 50 seeks per second. The `filemon` tool can be used to monitor the seek activity on a disk.

In cases where a disk that already has a high seek rate must be used for heartbeating, it may be necessary to change the heartbeat timing parameters to prevent long write delays from being seen as a failure.



A

DS4000 quick guide

In this appendix, we supply summarized information and pointers to DS4000 reference documentation. This is intended as a checklist and a quick guide to help you primarily during your planning and implementation of new systems.

Most of the topics summarized here have been presented and discussed in other chapters of this Best practices guide.

Pre-installation checklist

<p>Locate and review latest product documentation for DS4000: http://www.ibm.com/servers/storage/disk/ds4000/index.html</p>
<p>Download all software and firmware that is required: DS4000 Firmware, ESM, Drive firmware, HBA Firmware and drivers, and Storage Manager Decide on HBA driver type (SCSI Port vs StorPort) Decide on multipath driver (RDAC, MPIO)</p> <p>To be automatically notified of updates, register with MySupport (see 3.5.1, “Staying up-to-date with your drivers and firmware using My support” on page 120).</p>
<p>Ensure that all hardware and software is covered by the interoperability matrix: http://www3.ibm.com/servers/storage/disk/ds4000/pdf/interop-matrix.pdf Ensure that host operating systems that are planned to be connected are supported.</p>
<p>Ensure that power and environmental requirements are met: Each DS4000 and EXP unit will require two power sources.</p>
<p>Obtain 2 (or 4) IP addresses for the Storage server</p> <p>CtrlA ____·____·____·____ CtrlA service: ____·____·____·____ (DS4800 only)</p> <p>CtrlB ____·____·____·____ CtrlB service: ____·____·____·____ (DS4800 only)</p> <p>If needed, obtain IP addresses for switches or directors</p>
<p>Choose fabric design (direct connect or through switch or director).</p>
<p>Ensure that growth patterns are known and factored into design.</p>
<p>Ensure that the applications and data sets are reviewed and documented.</p>
<p>Decide on most critical application _____ Note the type of application: I/O intensive Throughput intensive</p>
<p>Plan for redundancy levels on storage (RAID Levels): It is best to map out each host with desired RAID level and redundancy required.</p>
<p>If required, plan for redundancy on host connections required (multiple HBAs with failover or load sharing)</p>
<p>Decide on premium features will be used with Storage System:</p> <ul style="list-style-type: none"> – FlashCopy – VolumeCopy – Enhanced Remote Mirroring
<p>Decide if SAN Volume Controller (SVC) is to be used.</p>
<p>Plan if there is a use or requirement of third-party management tools such as TotalStorage Productivity Center (TPC).</p>

Installation tasks

This section summarizes the installation tasks and the recommended sequence.

Rack mounting and cabling

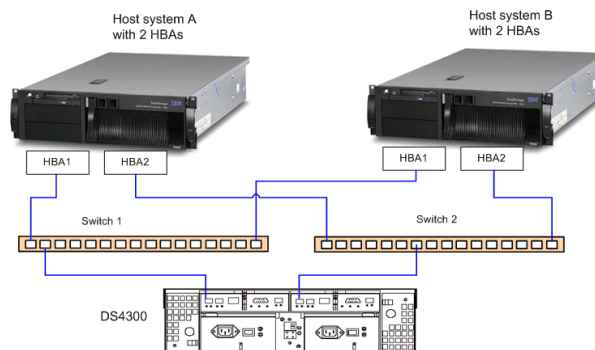
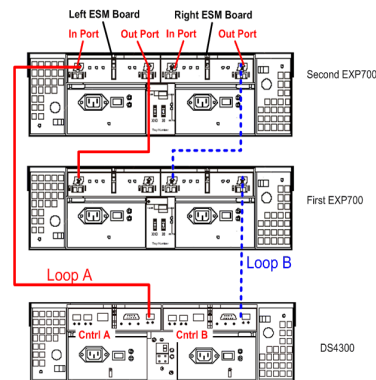
Mount server and expansion units into rack:

- Maintain 15 cm (6 in.) of clearance around your controller unit for air circulation.
- Ensure that the room air temperature is below 35°C (95°F).
- Plan the controller unit installation starting from the bottom of the rack.
- Remove the rack doors and side panels to provide easier access during installation.
- Position the template to the rack so that the edges of the template do not overlap any other devices.
- Connect all power cords to electrical outlets that are properly wired and grounded.
- Take precautions to prevent overloading the electrical outlets when you install multiple devices in a rack.

Ensure for adequate future expansion capabilities when placing the different elements in the rack.

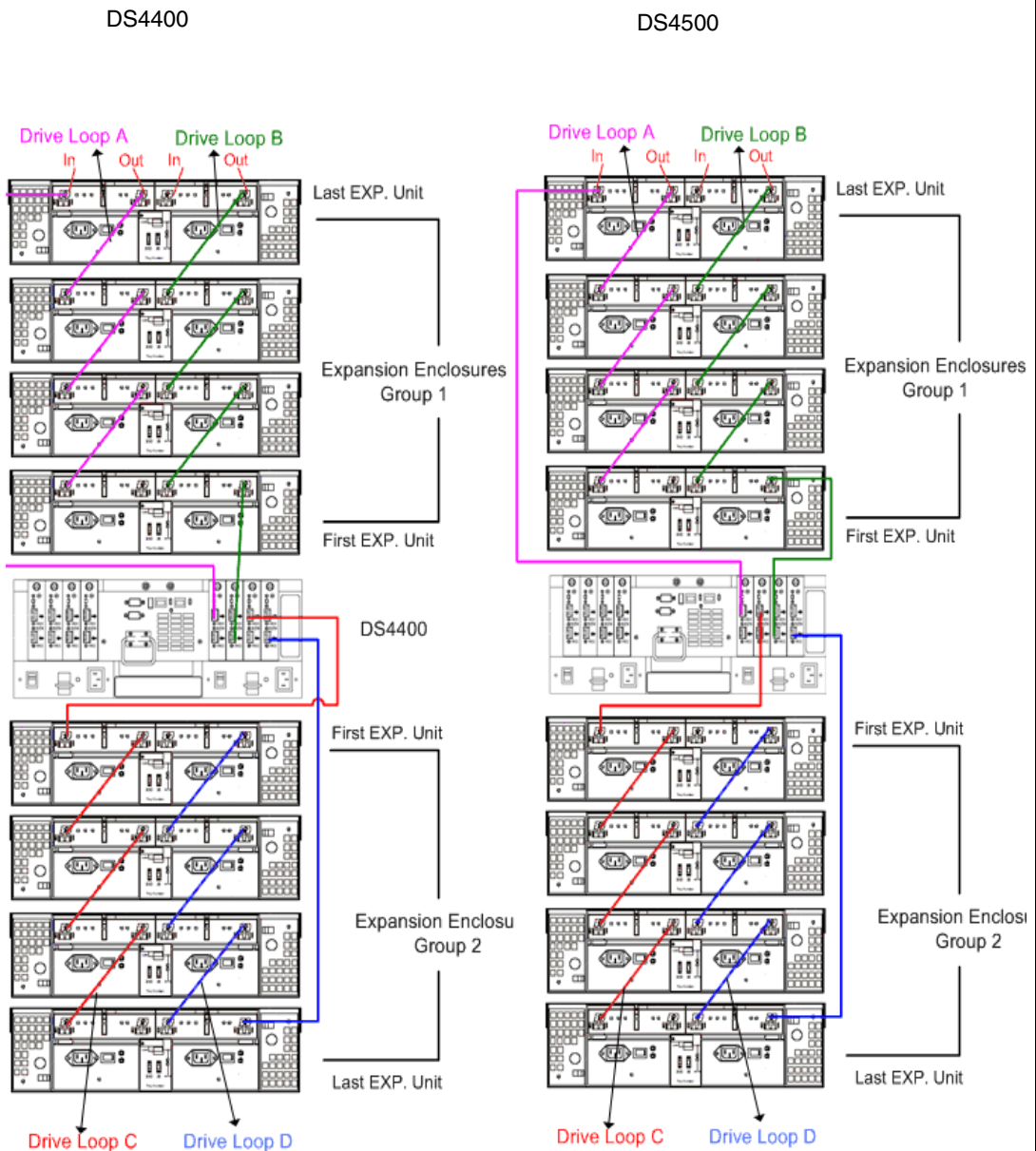
Perform drive side cabling between the DS4000 Server and Expansion enclosures. You can perform host-side cabling as well. The correct cabling depends on the DS4000 models being used:

DS4100 and DS4300 with dual controller

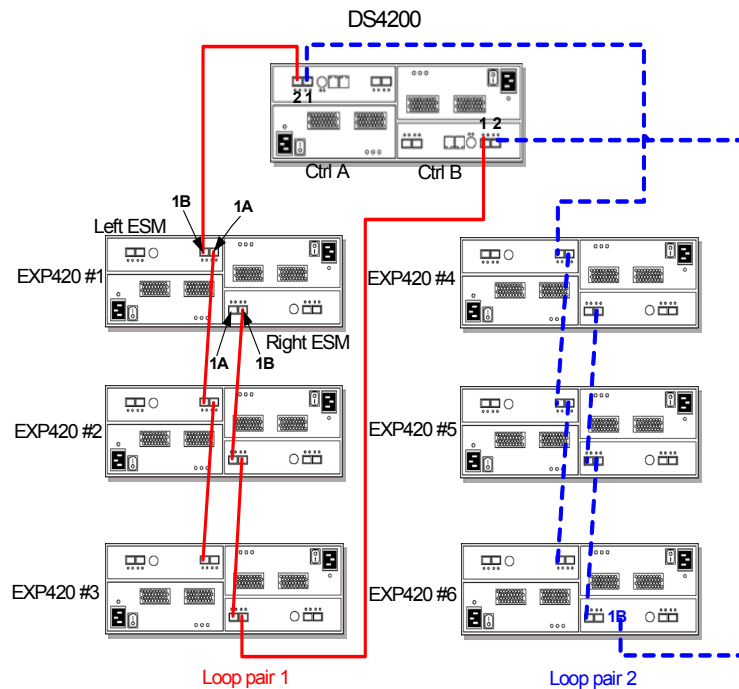
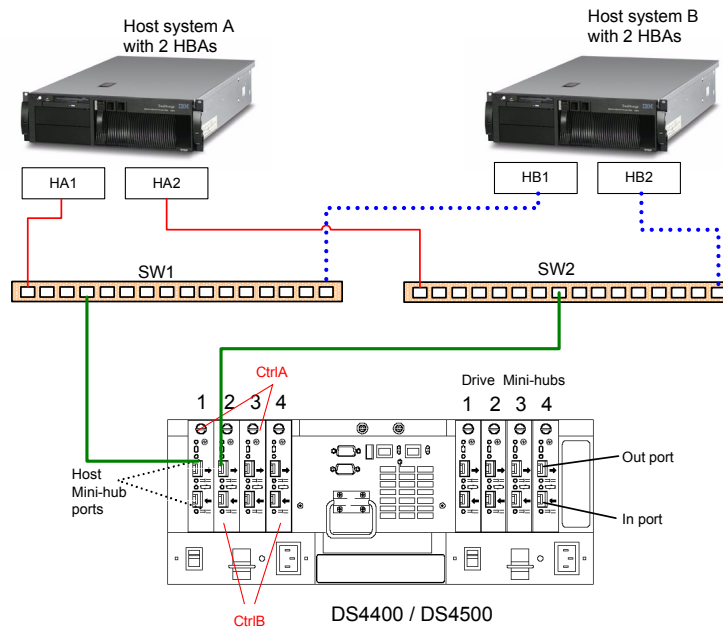


For details see 3.2.1, “DS4100 and DS4300 host cabling configuration” on page 77.

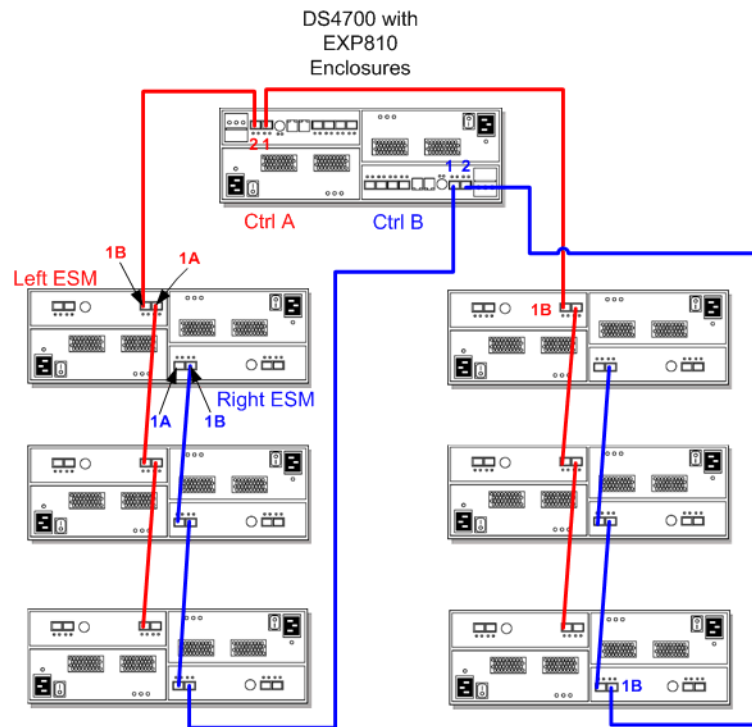
Note the difference between the DS4400 and DS4500. To prevent a drive enclosure group loss on the DS4500 YOU SHOULD PAIR MINI-HUB 1 & 3 TOGETHER TO CREATE DRIVE LOOPS A & B. PAIR MINI-HUB 2 & 4 TO CREATE DRIVE LOOPS C & D.



For details see 3.2.6, "DS4500 drive expansion cabling" on page 88.



- The DS4200 can have up to six EXP420 enclosures.
 - The DS4200 can only be expanded using EXP420 enclosures.
 - The EXP420 and DS4200 can only use the 500 GB EV-DDM drives.
- Note: The reserved ports on the EXP810 that are not used for drive cabling.
For details see 3.2.4, “DS4200 drive expansion cabling” on page 82

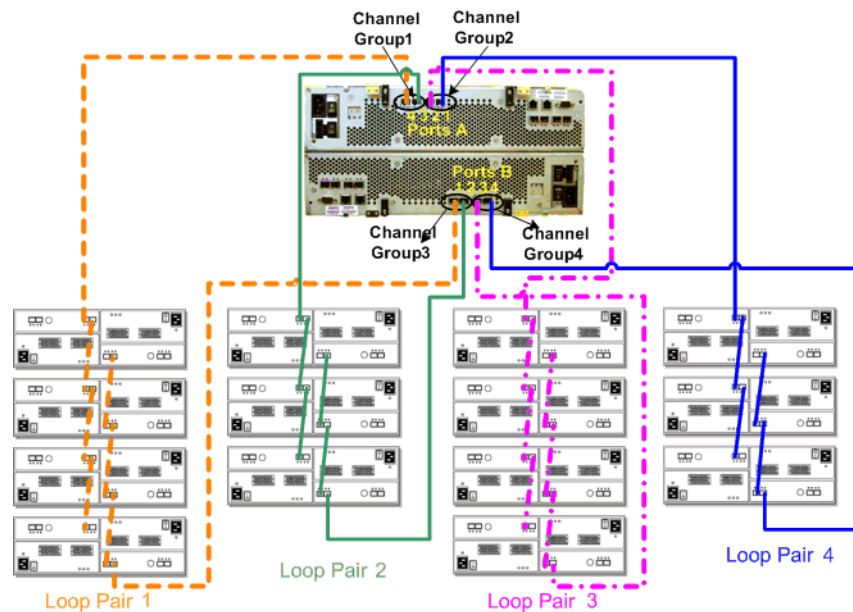


- The DS4700 with EXP810 enclosures.
- The DS4700 has 16 Internal E-DDM drives.
- The DS4700 can use both the EXP710 and the EXP810 enclosures. If utilizing both EXP710 and EXP810 enclosures, separate them on separate drive loops where possible.
- The DS4700 cannot use the EXP100 SATA enclosure. SATA intermix is only available using the EXP810 enclosure and SATA drives. At the time of writing, SATA and Fibre Channel disks cannot exist in the same enclosure.

Note: The reserved ports on the EXP810 are not to be used for drive cabling.

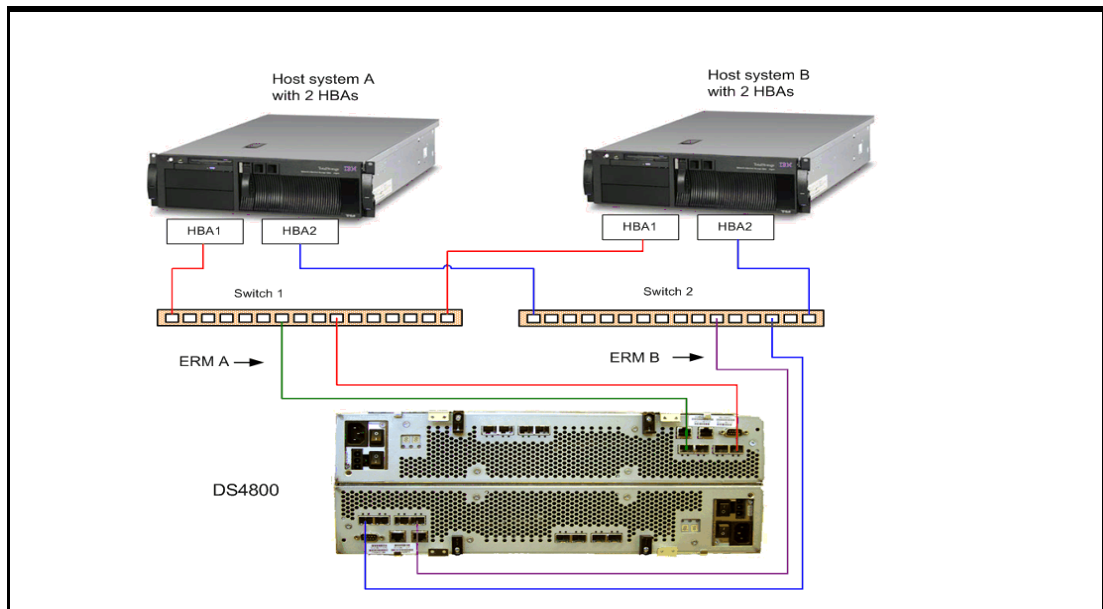
For details see 3.2.8, "DS4700 drive expansion cabling" on page 92.

For the DS4800, each drive-side Fibre Channel port shares a loop switch. When attaching enclosures, drive loops are configured as redundant pairs utilizing one port from each controller. This ensures data access in the event of a path/loop or controller failure.



- ▶ Configure DS4800 with drive trays with a maximum of four enclosures per loop.
- ▶ The DS4800 using EXP810 can only have a maximum of 14 enclosures.
- ▶ With the DS4800, EXP710 and EXP810 enclosures can be intermixed. Limitations to the number of enclosures must be followed.
- ▶ Distribute the drives equally between the drive trays.
- ▶ For each disk array controller, use four Fibre Channel loops if you have more than four expansion units.
- ▶ Based on recommended cabling from above:
 - Dedicate the drives in tray stacks 2 and 4 to disk array controller B.
 - Dedicate the drives in tray stacks 1 and 3 to disk array controller A.
 - I/O paths to each controller should be established for full redundancy and failover protection

For details see 3.2.10, “DS4800 drive expansion cabling” on page 97.



Ensure that the expansion unit IDs are not duplicated on same loop:

- ▶ DS4100 and DS4300 enclosures are always ID 0.
- ▶ Change the drive enclosures to something other than the default of '00'. New EXP drawers always come with id of '00' and this will prevent errors in the event you forget to change it before adding it to the DS4000 subsystem. Within a subsystem each drive enclosure must have a unique ID. Within a loop the subsystem IDs should be unique in the ones column. All drive trays on any given loop should have complete unique ID's assigned to them.
- ▶ When connecting the EXP100 enclosures, DO NOT use the tens digit (x10) setting. Use only the ones digit (x1) setting to set unique server IDs or enclosure IDs. This is to prevent the possibility that the controller blade has the same ALPA as one of the drives in the EXP100 enclosures under certain DS4000 controller reboot scenarios.
- ▶ For the DS4800, it is recommended to have the tens digit (x10) enclosure ID setting to distinguish between different loops and use the ones digit (x1) enclosure ID setting to distinguish storage expansion enclosures IDs within a redundant loop.

Add drives to expansion enclosures:

When about to power on a new system with expansion enclosures attached, each expansion enclosure should have a minimum of 2 drives before powering on the storage server.

Connect power to enclosures and DS4000.

Important: When powering up a new system for the first time, power up one EXP unit at a time and then add only 2 drives at a time! This means that you should pull out every drive in a new system and slowly add them into the system (2 at a time) until recognized. There can be problems with the Controller discovering large configurations all at once which can result in loss of drives, ESM, GBICs.

Recommended sequence (New system):

- Power up the 1 EXP unit with 2 drives installed.
- Power up the storage server.
- Install Storage Manager client on a workstation and connect to the DS4000 (see hereafter how to do the Network setup and DS4000 Storage Manager Setup).
- Once you have Storage Manager connected, continue adding drives (2 at a time) and EXP units. Verify with Storage Manager the DS4000 sees the drives before you continue to add units/drives.

Power on sequence (existing system):

- Start expansion enclosures and wait for the drives to be ready.
- Turn on switches.
- Power up the Storage Server.
- Power up hosts.

With firmware 05.30 and later, the controllers have a built in pause/delay to wait for the drives to stabilize, but it is still a good practice to follow the proper power up sequences to prevent any loss of data.

Connect Ethernet cables between RAID controllers and network switch:

TIP: If you have problems locating the unit over Ethernet, try the following: Make sure the Ethernet switch is set to auto sense. It does not work well with hard set ports at 100Mb. If that doesn't work, hard set the ports to 10Mb. The DS4000 controller sometimes won't work well with 100Mb or auto-sensing.

Set up IP addresses on controller ports. (See 3.1.1, "Initial setup of the DS4000 Storage Server" on page 68, for details.)

If the storage subsystem controllers have firmware version 05.30 or later, then DS4000 will have default IP settings only if NO DHCP/BOOTP server is found.

Using the Storage Manager GUI:

In 5.30 code and later, you can change the IP via the Storage Manager GUI that you install on a laptop. You'll have to change your TCP/IP setting on your laptop or workstation to an IP address that's something like 192.168.128.10...255.255.255.0. Use an Ethernet switch to connect the laptop to both controllers (required). First, discover the DS4000 and then right-click each controller.

Using the Serial port:

If you can't do this via the Ethernet, you'll have to do it through the serial port, which requires a null modem serial cable.

If switches are to be used, then update to latest supported firmware.

Set up network configuration on all switches and directors.
(Refer to manufacturer's installation and setup documentation)

Update Firmware, NVSRAM and Drive Firmware to latest supported version
Always refer to the readme file that comes with the updates about the installation instructions.

Preparing the host server

Install Host Bus Adapters (HBAs):

- ▶ Ensure that Host bus adapters (HBAs) used in hosts are updated to supported firmware and latest drivers. Decide on either SCSIPort or StorPort Drivers for the HBA. With StorPort drivers, know the controller firmware requirements.

For the latest supported host adapters, driver levels, bios, and updated readme, check:

<http://knowledge.storage.ibm.com/HBA/HBASearch>

- ▶ Configure HBA settings for operating system that will use the HBA:

Execution throttle (for Qlogic based HBAs) _____

LUNs per target _____

Default for most operating systems is 0.

(HP-UX, Netware 5.x and Solaris are limited to 32 LUNs)

NetWare 6.x with latest support packs and multi-path driver requires the setting of LUNs per target = 256).

- ▶ Ensure that additional host HBAs are installed if redundancy or load balancing is required.
- ▶ Ensure separate HBAs for Disk and Tape access:

For tape access – Enable Fibre Channel tape support.

For disk access – Disable Fibre Channel tape support.

Tips:

- Be careful not to put all the high-speed adapters on a single system bus; otherwise the computer bus becomes the performance bottleneck.
- Make a note of what slot in each server.
- Record WWPN of each HBA and what slot it is in.
- On INTEL based host platforms, Ensure that HBAs are in a higher priority slot than ServeRAID™ adapters. If not booting from the Host HBAs, it doesn't matter whether or not they are a higher PCI scan priority.

Choose the appropriate multipath driver for your requirements.

There is a choice of RDAC and MPIO come with the Storage Manager client installation.

Controller firmware of 6.19 is required for multipath drivers based on the MPIO driver architecture.

TIPS:

- A multipath driver is recommended regardless of whether or not there are multiple HBAs.
- AIX - The AIX “fcp.array” driver suite files (RDAC) are not included on the DS4000 installation CD. Either install them from the AIX Operating Systems CD, if the correct version is included, or download them from the Web site:

<http://techsupport.services.ibm.com/server/fixes>

Install required Storage Manager components.

The following components are mandatory for all DS4000 Environments:

- Multipath driver (RDAC, MPIO) - (Regardless of whether or not there are multiple paths) for Windows 2000, Windows 2003, Solaris (and Linux, when using the non-failover HBA driver)
- Client somewhere to be able to configure the solution.

The following components are optional based on the needs of the customer:

- **Agent** (All Operating Systems) - This is only needed if you wish to configure the Fibre Channel through a direct Fibre Connection. If you only want to manage the DS4000 unit over the network, it is not necessary.
- **SMxUtil** - These utilities are not required, but RECOMMENDED because they add additional functionality for troubleshooting and hot-adding devices to the OS. NOTE: In SM8.0, they are required for FlashCopy Functionality. If you plan to use SM Client through a firewall. SM Client uses Port 2463 TCP.
- Qlogic SANSurfer- Not required but RECOMMENDED because it adds Fibre Path Diagnostic capability to the system. It is recommended that customer's always install this software and leave it on the system. In addition, for Linux you need QLRemote.

Be sure you install the host-bus adapter and driver before you install the storage management software.

- For in-band management, you must install the software on the host in the following order: Note: Linux does not support in-band management.
 - a. SMclient
 - b. Multipath Driver
 - c. SMagent
 - d. SMutil

Starting with version 9.12, the Install Anywhere procedure automatically installs the selected components in the right order.

- For out-band management, you must install SM client software on a management station.

Storage Manager setup

Keep accurate and up to date documentation. Consider a change control log.

Launch Storage Manager and perform initial discovery.

When first started on a management workstation, the client software displays the Enterprise Management window and the Confirm Initial Automatic Discovery window.

Note: The Enterprise Management window can take several minutes to open. No wait cursor (such as an hourglass) is displayed.

Click Yes to begin an initial automatic discovery of hosts and storage subsystems attached to the local subnetwork.

Note: The Enterprise Management window can take up to a minute to refresh after an initial automatic discovery.

Direct Management: If the Automatic Discovery doesn't work.

- Go to EDIT > Add Device
- Enter the IP Address of controller A. Click Add.
- Enter IP Address for controller B. Click Add.
- Click Done.

Storage Controllers should appear. For new installs, it is likely that they will show "Needs Attention". This is common since the battery will be charging. Power cycle the DS4000 controller if it doesn't appear.

Rename the DS4000 storage server

- Click Storage Subsystem > Rename. The Rename Storage Subsystem window opens.
- Type the name of the storage subsystem. Then click OK.

If you have multiple controllers, it is helpful to enter the IP addresses or some other unique identifier for each subsystem controller.

Change expansion enclosure order

The EXP enclosures will likely show in the GUI different than how you have them installed into the rack. You can change the GUI to look like how the enclosures are installed. You do this by going under File>Change>Enclosure Order and move the enclosures up or down to correctly reflect how they are installed into the rack.

Check that each controller (Ctrl+A and Ctrl+B) is online and active.

Collect the Storage Server system profile.

Go to View > Storage System Profile

Click Controller Tab and Make note of NVSRAM and Firmware versions listed

Firmware version: _____

NVSRAM version: _____

Click the "Drives" tab and find the product ID and Firmware version.

HDD Firmware _____

Click the "Enclosures" tab and find the ESM Firmware version for each EXP unit (all EXP of the same model should have the same version)

ESM _____

If you need to upgrade the firmware, always refer to the documentation supplied with the firmware.

Set password on storage server and document the password in a secure area.

Set Storage Server settings.

Set start and stop cache levels as per planning document _____
Set cache block size as per planning document _____K

Create arrays and logical drives as per planning document (repeat for each array/logical drive).

Select how many drives per array: _____
Ensure alternating drive loops (odds and evens)
How many LUNs per array _____
Speed of drives in the array _____K
Choose RAID level RAID _____
Disk Type SATA _____ Disk type Fibre Channel _____

The recommendation is to do a manual selection of drives and configuration to ensure that:
Drives use both drive loops (odds and evens)

Enclosure Loss Protection

Ensure that a small amount of free space is left on each array after logical drive creation

Read ahead multiplier _____

Select segment size _____K

Select controller to ensure a balance of workload between controllers

Note: No more than 30 drives per array. Best performance around 5 to 12 drives.
Max 2 TB per LUN

Configure cache settings required for that logical drive:

Read caching enabled	Yes__	No__
Write caching enabled	Yes__	No__
Write caching without batteries	Yes__	No__
Write caching with mirroring	Yes__	No__

Configure host access to logical drive.

If using Storage Partitioning, ensure that it is enabled:

DS4100, DS4200, DS4300, DS4500, DS4700 and DS4800 come with limited number of Storage Partitions. An upgrade from the initial base partitions to 64 partitions can be achieved by the purchase of a storage partitioning premium feature.

Procedure:

Go to Storage Subsystem>Premium Features>List
You should see "Storage Partitioning Enabled"

If you do not see that it is enabled you'll have to get the feature key and enable it. Make note of the 32 digit feature key number. Call 800-IBM-SERV, enter the 4 digit machine type and tell the help desk that you need a feature key generated.

Decide how hosts will address storage

(Storage partitioning?). _____

Are there enough Storage partitions with current DS4000 configuration or require an upgrade?

If using storage partitioning, plan to use host groups.

Storage can be assigned to the host or to the group.

Host groups allow you to group like servers together in a logical group.

If you have a single server in a host group that has one or more LUNs assigned to it, it is recommended to assign the mapping to the host and not the host group.

All servers having the same host type, for example Windows servers, can be in the same group if you want, but, by mapping the storage at the host level you can define what specific server accesses which specific LUN.

If you have clusters, it is good practice to assign the LUNs at the host group, so that all of the servers in the host group have access to all the LUNs.

Refer to , “Samples of host planning documents” on page 395.

Configure zoning (keep it simple)

Minimum two zones per HBA:

- Zone one includes HBA and Controller A
- Zone two includes HBA and Controller B

Important: Make sure you have separate Tape and Disk access from HBAs.

Create hot spares

- Ensure that hot spares of each size and type are configured.
- Ensure that hot spares are on alternating drive loop (odds and evens).

TIPS:

One Hot Spare per drive tray is optimal but it'll depend on your capacity requirements. The recommendation is no less than 1 spare for every 20-30 drives. Also keep the rebuild times in mind depending upon the size of the drives installed.

EXP100 Recommendation is one hot spare per EXP100 drive expansion enclosure. One in an even slot and the other in an odd slot.

DS4100, DS4200, DS4400, DS4500 and DS4700 contain two redundant drive loops, it is recommended to put half of the hot-spares in one redundant drive loop and the rest on the other redundant drive loop.

The DS4800 has 4 drive loops so try to put at least one spare in each drive loop.

Note: a total of 15 hot-spares can be defined per DS4000 storage server configuration

Decide if you need to adjust media scan rate or leave it at 30 days.

Delete the access logical volume – (LUN 31)The DS4000 storage system will automatically create a LUN 31 for each host attached. This is used for in-band management, so if you do not plan to manage the DS4000 storage subsystem from that host, you can delete LUN 31 which will give you one more LUN to use per host. If you attached a Linux or AIX 4.3 or above to the DS4000 storage server, you need to delete the mapping of the access LUN.

Samples of host planning documents

The following are examples of documents that should be prepared when planning host attachment to the DS4000.

► Example of zoning planning document

Zone	Host Name	WWN	Cont	Switch	SW Port	Host OS
Radon- HBA1	Radon	21-00-00-E0-8B-05-4C-AA	A	SW1	5	Win2000
Radon- HBA1	Radon	21-00-00-E0-8B-05-4C-AA	B	SW1	5	Win2000
Radon- HBA2	Radon	21-00-00-E0-8B-18-62-8E	A	SW2	6	Win2000
Radon- HBA2	Radon	21-00-00-E0-8B-18-62-8E	B	SW2	6	Win2000

► Example of host and LUN settings for environment

LUN Name	Use	No# HBAs	RAID Level	# of Disks in LUN	Current Size	%Read	Growth
Radon_DB	OLTP DB	2	1	8	100Gb	63%	50%
Radon_Trans	Transaction Logs	2	1	2	50Gb	14%	50%
Nile_Backup	Backup	1	5	8	750Gb	50%	30%
AIX_Cluster	AIX Cluster	2	1	10	250Gb	67%	40%
AIX_Cluster	AIX Cluster	2	1	10	250Gb	67%	40%
Netware	File	1	5	3	200Gb	70%	35%

► Example of LUN settings for each host

Host Name	LUN	Segment Size	Write Cache	Read Cache	Write Cache with Mirroring	Read Ahead Multip.
Radon	Radon_DB	64	Enabled	Enabled	Enabled	0
Radon	Radon_Trans	128	Enabled	Enabled	Enabled	0
Nile	Nile_Backup	256	Enabled	Enabled	Disabled	1
Kanaga	AIX_Cluster	128	Enabled	Enabled	Enabled	1
Atlantic	AIX_Cluster	128	Enabled	Enabled	Enabled	1
Pacific	Netware	128	Enabled	Enabled	Enabled	1

Tuning for performance

Use Performance Monitor to view the performance of the logical drives on the system.
Adjust settings that may improve performance then monitor results, then compare results.
Settings that may be changed on a logical drive to tune for performance.

DS4800 - Performance Monitor

Devices	Total I/Os	Read Percentage	Cache Hit Percentage	Current KB/second	Maximum KB/second	Current I/O/second	Maximum I/O/second
CONTROLLER IN SLOT A	142,495	62.6	21.8	6,878.0	6,878.0	13,756.0	13,756.0
Logical Drive Host_Lun_FC1	55,140	43.8	18.1	2,833.6	3,015.0	5,667.0	6,030.0
Logical Drive Host_Lun_FC2	87,355	74.4	23.1	4,044.4	4,044.4	8,089.0	8,089.0
CONTROLLER IN SLOT B	150,764	60.8	18.3	7,206.0	7,206.0	14,412.0	14,412.0
Logical Drive Host_LUN_Sata1	150,764	60.8	18.3	7,206.0	7,206.0	14,412.0	14,412.0
STORAGE SUBSYSTEM TOT...	293,259	61.7	20.0	14,084.0	14,084.0	28,168.0	28,168.0

Segment size:

Old Segment Size _____K

New Segment Size _____K

Cache settings changes

Changed

Read caching enabled	Yes	—	No	—	—
Write caching enabled	Yes	—	No	—	—
Write caching without batteries	Yes	—	No	—	—
Write caching with mirroring	Yes	—	No	—	—

Read ahead multiplier

Old Read Ahead _____

New Read Ahead _____

If multiple logical drives exist on the same array then check for disk contention or thrashing of disks.

Use Performance Monitor to check performance statistics on the DS4000 storage server.

Look at all the statistics:

- Total I/Os
- Read Percentage
- Cache Hit Percentage
- Current KB/Sec
- Max KB/Sec
- Current I/O/Sec
- Max I/O/Sec

From these you can gauge which LUNs are high utilization LUNs and adjust settings to suit each LUN.

Use Performance Monitor to check controller balance.

Do any arrays or LUNs need to be set to another controller to even out workload?

Yes _ No _

Notes

This section is a collection of installation and configuration notes for different OS platforms. The notes apply to Version 9.19 of the Storage Manager.

Notes on Windows

In this section we discuss Windows.

Updating RDAC in a Microsoft Cluster Services configuration

In Microsoft Cluster Services (MSCS) configurations, the MSCS service must be stopped and set to manual start after server rebooting, the clusdisk driver must have to be set to offline and, then, the server must be rebooted before uninstalling the RDAC driver in the server. If the clusdisk driver is not set to offline and the MSCS service is not set to manual start, the Microsoft Cluster Service will not start after the new RDAC driver is installed because it cannot bring the Quorum disk resource online. The problem is caused by the changing of the disk signatures.

To recover from this problem, you must:

- ▶ Look up the old disk signatures of the disk that are defined as cluster disk resources. They could be found either in the registry under the registry key:

HKLM/System/CurrentControlSet/Services/Clusdisk/Parameters/Signatures

Or, in the cluster.log file.

- ▶ Look up the new disk signatures of the disks that are defined as cluster disk resources using the **dumpcfg** utility that is packaged in the Microsoft Windows resource kit.
- ▶ Compare the new and old disk signatures. If new disk signature did not match the old signature, you have to change the new disk signature to the old disk signature values by using the **dumpcfg** command. The syntax of the command is:

```
dumpcfg.exe -s <old-signature> <Disk#>
```

For example:

```
dumpcfg.exe -s 12345678 0
```

Always check in the latest documentation from Microsoft regarding this procedure.

Host type

Ensure that the host type is set correctly for the version of operating system. The Host Type setting configures how the host will access the DS4000 system. For example, do not set the host type for a Windows 2003 Non-Clustered server to Windows NT® Non-Clustered SP5 or Higher host type, otherwise this could extend the boot time (to up to two hours).

Disk alignment

Contained within Windows Server 2003 and Windows 2000 Server, the **diskpart.exe** utility is used to align the storage boundaries. This is explained in more detail in “Disk alignment” on page 144.

Extend disks

For Windows 2003 basic disks, you can extend the volume using the **extend** command in the diskpart utility. This will extend the volume to the full size of the disk. This command is dynamic and can be done while the system is in operation.

Use the Disk Management GUI to extend dynamic disks in Windows 2000 or Windows 2003.

Note that a system partition cannot be extended.

This is explained in more detail in “Using diskpart to extend a basic disk” on page 141.

Limitations of booting from the DS4000 with Windows

When the server is configured to boot the Windows operating system from the DS4000 storage server, and to have storage access and path redundancy, the following limitations apply:

- ▶ It is not possible to boot from of a DS4000 Storage Server and use it as a clustering device. This is a Microsoft Windows physical limitation.
- ▶ If there is a path failure and the host is generating I/O, the boot drive will move to the other path. However, while this transition is occurring, the system will appear to halt for up to 30 seconds.
- ▶ If the boot device (LUN 0) is not on the same path as the bootable HBA port, you will receive an INACCESSIBLE_BOOT_DEVICE error message.
- ▶ If you suffer major path problems (LIPs) or path trashing, it can hang the server indefinitely as RDAC tries to find a stable path.
- ▶ By booting from the DS4000 storage device, most of the online diagnostic strategies are effectively canceled, and path problem determination must be done from the Ctrl+Q diagnostics panel instead of FASTT MSJ.
- ▶ The internal disk devices should not be re-enabled.

If booting from the DS4000 disk on a host server that has two HBAs and the HBA that is configured as the boot device fails, use the following procedure to change the boot device from the failed HBA to the other HBA:

- ▶ During the system boot process, press Ctrl+Q at the Qlogic BIOS stage.



Figure A-1 QLogic HBA Bios

- ▶ Select the first adapter at the 2400 I/O address (the failed adapter)
- ▶ On the failed HBA, the Host Adapter BIOS should be set to disabled.
- ▶ Save and exit back to the Select Adapter menu.
- ▶ Select second adapter at the 2600 I/O address.
- ▶ The second HBA on the Host Adapter BIOS should be set to enabled.



Figure A-2 Host Adapter BIOS on Second HBA port

- From the Configurable Boot Settings panel, set the boot device to the controller's World Wide Port Name, as shown in Figure A-3 and reboot the server.

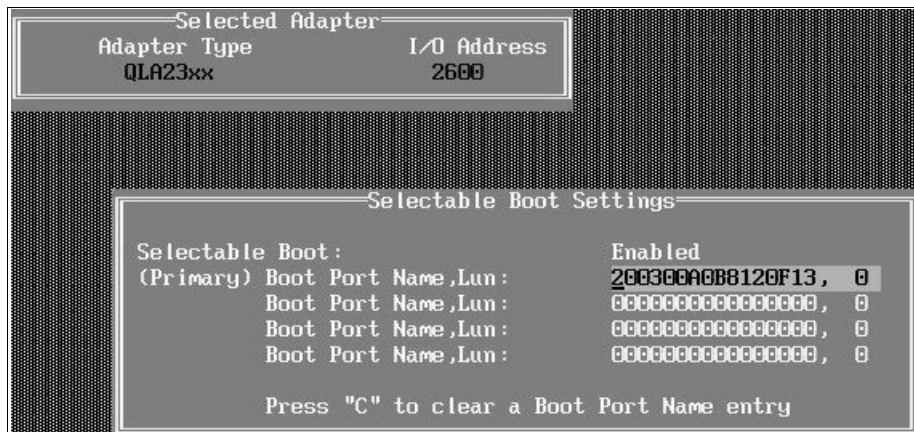


Figure A-3 Selectable boot Settings

The failed HBA should be replaced when practical, the zoning and storage mapping should be changed at that time to reflect the new HBA WWN.

Notes on Novell Netware 6.x

The following notes pertain to the Novell NetWare host platform.

Novell NetWare 6.0 with Support Pack 2 or earlier

The IBMSAN.CDM driver is the supported multi-path failover/failback driver for Novell NetWare 6.0 with Support Pack 2 or earlier. IBMSAN can be downloaded from the DS4000 Support Web site:

<http://www.ibm.com/servers/storage/support/disk/>

Please refer to IBM SAN readme included in IBM SAN package for installation instructions.

Netware 6.0 with Support Pack 3 or Netware 6.5

After Netware 6.0 support pack 3 and Netware 6.5 or later, LSIMPE.CDM became the supported multi-path driver required. Refer to IBM Web Site:

<http://www.ibm.com/servers/storage/support/disk/>

Or the Novell Support Web Site:

<http://support.novell.com>

Search for TID# 2966837.

Refer to the LSIMPE readme, included in the LSIMPE package for installation instructions.

- ▶ Disable all other multi-path support. For example, when using QLogic adapters, load the driver with '/luns /allpaths /portnames'. The 'allpaths' option disables the QLogic failover support.
- ▶ Set the LUNs per target = 256 on Qlogic HBAs.
- ▶ Set the execution throttle to correct setting.
- ▶ Ensure separate HBA for tape and disk access.
- ▶ Enable tape Fibre Channel support to HBA that will access tape drives.
- ▶ Ensure that the package multi-path and HBA drivers are copied to the correct location on the server.
- ▶ If your configuration is correct you will not need to set the failover priorities for your LUNs, the LSIMPE.CDM driver will configure these.
- ▶ Use the multi-path documentation to verify the failover devices. This documentation is available from the Web site:

<http://support.novell.com/cgi-bin/search/searchtid.cgi?/10070244.htm>

For the native NetWare Failover, you must use Novell NetWare 6.0 with Support Pack 5 or higher or NetWare 6.5 with Support Pack 2 or higher.

In addition, in the DS4000 Storage Manager Subsystem Management window, select the 'NetWare Failover' Host Type for the storage partition that the Fibre Channel HBA ports in the Netware server are defined.

Note: Do not use the drivers that are available with NetWare 6.0 Support Pack 3 or 4 or NetWare 6.5 Support Pack 1 or 1.1. Download and use either the Novell or IBM drivers and follow the installation instructions.

- ▶ Download the updated NWPA.NLM from:

<http://support.novell.com>

Search for TID# 2968190.

Then follow the installation instructions in the Novell TID.

- ▶ Download the newer MM.NLM from:

<http://support.novell.com>

Search for TID# 2968794.

Then follow the installation instructions in the Novell TID.

- ▶ After the Device Driver is installed, edit the STARTUP.NCF for the following lines:
 - Add '**SET MULTI-PATH SUPPORT = ON**'
 - Add 'Load LSIMPE.CDM' before the SCSIHD.CDM driver
 - Add the '**AEN**' command line option to SCSIHD.CDM

Your STARTUP.NCF file should look somewhat like the following example:

```
SET Multi-path Support = ON
LOAD MPS14.PSM
LOAD IDECD.CDM
LOAD IDEHD.CDM
LOAD LSIMPE.CDM
LOAD SCSIHD.CDM AEN
LOAD IDEATA.HAM SLOT=10007
LOAD ADPT160M.HAM SLOT=10010
LOAD QL2300.HAM SLOT=4 /LUNS /ALLPATHS /PORTNAMES XRETRY=400
```

Notes on Linux

There are two versions of the Linux RDAC at the time of writing. The Version 09.00.A5.22 for Linux 2.4 kernels only like Red Hat EL 3 and SuSe SLES 8 and RDAC package version 09.01.B5.35 for Linux 2.6 kernel environments such as Red Hat 4 and SuSe SLES 9.

- ▶ Make sure that you read the readme.txt files for V9.19 Linux RDAC, HBA, and Storage Manager for Linux.
- ▶ When using the Linux RDAC as the multi-pathing driver, the LNXCLVMWARE host type must be used.
- ▶ There is not a requirement that the UTM (Access LUN) must be removed from the LNXCLVMWARE Storage Partitioning partition.
- ▶ When using the Linux RDAC as the failover/failback driver, the host type should be set to LNXCLVMWARE instead of Linux. If Linux RDAC is not used, the host type of Linux must be used instead.
- ▶ The Linux RDAC driver cannot coexist with a HBA-level multi-path failover/failback driver such as the 6.06.63-fo, 7.07.61-fo and 8.01.06-fo driver. You might have to modify the driver make file for it to be compiled in the non-failover mode.
- ▶ Auto Logical Drive Transfer (ADT/AVT) mode is not supported when using Linux in a cluster, or when using RDAC. Since ADT(AVT) is automatically enabled in the Linux storage partitioning host type. It has to be disabled by selecting 'host type' of LNXCL.
- ▶ AVT is required if you are using the Qlogic Failover drivers.
- ▶ The Linux SCSI layer does not support skipped (sparse) LUNs. If the mapped LUNs are not contiguous, the Linux kernel will not scan the rest of the LUNs. Therefore, the LUNs after the skipped LUN will not be available to the host server. The LUNs should always be mapped using consecutive LUN numbers.
- ▶ Although the host server can have different FC HBAs from multiple vendors or different FC HBA models from the same vendors, only one model of FC HBAs can be connected to IBM DS4000 Storage Servers.
- ▶ If a host server has multiple HBA ports and each HBA port sees both controllers (via an un-zoned switch), the Linux RDAC driver may return I/O errors during controller failover.

- ▶ Linux SCSI device names have the possibility of changing when the host system reboots. We recommend using a utility such as **devlabel** to create user-defined device names that will map devices based on a unique identifier, called a UUID. The **devlabel** utility is available as part of the Red Hat Enterprise Linux 3 distribution, or online at:
<http://www.lerhaupt.com/devlabel/devlabel.html>
- ▶ Linux RDAC supports re-scanning to recognize a newly mapped LUN without rebooting the server. The utility program is packed with the Linux RDAC driver. It can be invoked by using either **hot_add** or **mppBusRescan** command (note that **hot_add** is a symbolic link to **mppBusRescan**). There are man pages for both commands. However, the Linux RDAC driver does not support LUN deletion. One has to reboot the server after deleting the mapped logical drives.

Related publications

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

IBM Redbooks

For information about ordering these publications, see “How to get IBM Redbooks” on page 404. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010-03
- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *Designing an IBM Storage Area Network*, SG24-5758
- ▶ *AIX 5L Performance Tools Handbook*, SG24-6039
- ▶ *IBM HACMP for AIX V5.X Certification Study Guide*, SG24-6375-00

Other publications

These publications are also relevant as further information sources:

- ▶ *IBM TotalStorage DS4000 Storage Manager Version 9 Concepts Guide*, GC26-7734-01
- ▶ *IBM DS4000 Storage Manager Version 9.19 for Windows 2000, Windows Server 2003, NetWare, ESX Server, and Linux*, GC26-7847-01
- ▶ *IBM DS4000 Storage Manager Version 9.19 for AIX, HP-UX, Solaris, and Linux on POWER*, GC26-7848-01
- ▶ *IBM TotalStorage DS4000 Storage Manager Version 9 Copy Services Guide*, GC26-7707-02
- ▶ *IBM DS4000 FC2-133 Host Bus Adapter BIOS*, GC26-7736-00
- ▶ *Fibre Channel Installation and Cabling overview*, GC26-7846-00
- ▶ *IBM TotalStorage DS4000 Hardware Maintenance Manual*, GC26-7702-00
- ▶ *IBM TotalStorage DS4000 Hard Drive and Storage Expansion Enclosure Installation and Migration Guide*, GC26-7849-00
- ▶ *IBM TotalStorage DS4000 Problem Determination Guide*, GC26-7703-01
- ▶ *IBM TotalStorage DS4000 Storage Server and Storage Expansion Enclosure Quick Start Guide*, GC26-7738-00
- ▶ *IBM TotalStorage DS4000 Fibre Channel and Serial ATA Intermix Premium Feature Installation Overview*, GC26-7713-03
- ▶ *IBM TotalStorage DS4500 Fibre Channel Storage Server Installation and Support Guide*, GC26-7727-00
- ▶ *IBM TotalStorage DS4500 Fibre Channel Storage Server User's Guide*, GC26-7726-00
- ▶ *IBM TotalStorage DS4300 Fibre Channel Storage Server Installation and User's Guide*, GC26-7722-01

- ▶ *IBM System Storage DS4800 Storage Subsystem Installation User's and Maintenance Guide*, GC26-7845-00
- ▶ *IBM System Storage DS4000 EXP810 Storage Expansion Enclosure Installation, User's and Maintenance Guide*, GC26-7798-01
- ▶ *IBM System Storage DS4700 Installation, User's and Maintenance Guide*, GC26-7843-00
- ▶ *IBM System Storage DS4200 Express Installation, User's and Maintenance Guide*, GC27-2048-00
- ▶ *IBM System Storage DS4200 Express Storage Subsystem Fibre Channel Cabling Guide*, GC27-2049-00
- ▶ *IBM TotalStorage DS4000 EXP100 Storage Expansion Unit Installation, User's and Maintenance Guide*, GC26-7694
- ▶ *IBM TotalStorage DS4100 Storage Server Installation, User's Guide, and Maintenance Guide*, GC26-7712
- ▶ *IBM TotalStorage FAStT EXP700 and EXP710 Storage Expansion Units Installation, User's and Maintenance Guide*, GC26-7647
- ▶ *Copy Services User's Guide - IBM TotalStorage DS4000 Storage Manager*
- ▶ *IBM Netfinity Rack Planning and Installation Guide*, Part Number 24L8055

Online resources

These Web sites and URLs are also relevant as further information sources:

- ▶ IBM System Storage and TotalStorage
<http://www.ibm.com/servers/storage/>
- ▶ Support for IBM System Storage and TotalStorage products
<http://www.ibm.com/servers/storage/support/disk/>
- ▶ IBM DS4000 disk systems family
<http://www.ibm.com/servers/storage/disk/ds4000/index.html>
- ▶ IBM DS4000 Storage interoperability matrix
<http://www.ibm.com/servers/storage/disk/ds4000/interop-matrix.html>

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

Numerics

4 Gbps FC technology 3

A

access logical drive 51, 112
access LUN 49, 51, 112, 394, 401
ADT 114, 344
Advanced Technology Attachment (ATA) 2, 80, 155
AIX 5.1 217, 340–341, 358, 376
AIX 5.2 341, 358–359, 376
AIX 5L
 operating system 339
AIX environment 199, 340, 374
AIX host 138, 340, 342–345, 356, 365, 367
 configuration 342
 physical attachment 343
AIX system 341, 348, 351, 353, 355
alert 113–114
 notification 114
alert delay 116
Alert Delay Period 114
alignment 137, 144, 185
allocation unit 146, 172
 size 140, 145, 172
Allocation unit size 145–147, 172
Application Program Interface (API) 224
archive log 168
array 37, 42, 45, 104–105
 configuration 43–44
 creating 106
 defragment 152
 size 42, 206
Array Support Library (ASL) 28
arrayhost.txt 242
AS (Automatic Storage) 164
ASM 170
asynchronous mirroring 59
attenuation 19
Auto Logical Drive Transfer. See ADT
Auto-Logical Drive Transfer feature (ADT) 30
Automatic Discovery 75
automatic storage management (ASM) 170
auto-negotiation 21
autonomic 2
auxiliary 322
auxiliary VDisk 323
availability reasons 329
avserv 221
AVT 32
AVT-disabled failover 30
AVT-enabled failover 32
await 221

B

backup 147
bandwidth 135
 increasing 130
baseline 217
basic disc 63–64, 140–141, 143, 397
 primary partition 64, 140
battery 38, 40
BladeCenter 16
blocksize 136, 167
BOOTP 68
bootstrap 68
boundary crossing 137
bus reset 26
Business Continuity and Disaster Recovery (BCDR) 113
Business Continuity (BC) 289

C

cable
 labeling 21–23
 length 17
 management 21–22
 routing 22
 types 18
cabling 18
 DS4000 cabling configuration 76
 mistakes 23
cache 38, 40, 53, 135, 148, 159, 323
 block size 54, 57, 154
 flushing 54, 57
 hit percentage 206
 memory 5, 206
 mirroring 54, 56
 read-ahead 136–138, 206
 read-ahead multiplier 54, 109, 159, 206
 setting 106
 write-through 56
capacity 110, 155
 unconfigured 110
 upgrade 127
chdev 138
checklist 15
 hardware 15
 software 15
CIM Agent 234, 237
CIM Client 235
CIM Managed Object 235
CIM Object Manager (CIMOM) 235
CIM-compliant 234
CIMOM agent
 register 242
clock 104
cluster 112, 146, 362, 364

- cluster.log file 397
- clustered disks 140
- command line
 - tool 7
- Command Line Interface (CLI) 209
- command-line interface 201
- Common Information Model (CIM) 234
- Concurrent Resource Manager (CRM) 363
- connectors 18
- container 166
- contention 43, 45
- controller
 - clock 104
 - firmware 7, 122–123, 126
 - network setup 68
 - ownership 45–46, 109, 206
- Controller-based 7
- copy
 - FlashCopy 14, 152–153
- copy priority 151
- copy priority rate 151
- copy services 113, 150
- copyback 153
- copy-on-write 152
- CPU utilization 171
- CRC Error 227
- Create Probe 246
- crossover 69
- Cyclic Redundancy Check (CRC) 227

D

- DACStor 42
- DACstor 43
- DACstore 68, 128, 130
- data blocks 167
- Data location 164, 167
- data pattern 134–135
- data scrubbing 51
- data striping 37–38
- database
 - log 147, 156, 164
 - RAID 168
- Database structure 164
- datafile table 166, 168
- DB2 164
 - logs 167
- DB2 object 164
- dd command 213–216
- DDK 29
- defragment 152
- Delta resynchronization 288
- de-skew 199
- de-skew window 199
 - given target 199
- destination node 374–375
 - volume group 374
- device driver 61
- Device provider 235, 237
- Device Specific Module 29
- diagnostic test 224, 226

- diagnostics 227
- direct attach 77
- Directory agent (DA) 236
- Disaster Recovery (DR) 289
- disk
 - latency 135, 157
 - mirroring 38, 40
- disk array
 - controller 342
 - router 342
- disk array controller (DAC) 25, 342–343, 349–351, 353, 369, 387
- disk array router (DAR) 342–343, 349–351
- disk capacity 35
- Disk drive 2, 33, 36, 38, 43, 48, 61, 103, 128, 148, 155–156
- disk drive 2, 6, 33
 - types 155
- disk group 64, 144
- disk heart beat 379–380
- disk heartbeat 376–377, 379–380
- Disk Management 143, 146
 - GUI 398
 - snap-in 145–146
- disk system 2
- diskhb network 378–379
- diskpar 144
- diskpart 64, 140, 142, 144–145, 186
- diskpart utility 140–142, 145–146, 397
- dispersion 18
 - modal 18
- distribute 110
- DMA 359
- DMP 28
- DMS (Database Managed Storage) 164
- drive
 - selection 35
- drive channel 97
- DRIVE Loop 34, 45, 48, 97–98, 130–131, 158, 384, 387, 393–394
- Driver Development Kit (DDK) 29
- DS family 2
 - positioning 2
- DS3000 2
- DS400 Series
 - Storage Servers 2
- DS4000 2
 - capacity upgrade 127
 - configuring 104
 - initial setup 68
 - models 149
 - overall positioning 2
 - rename 104
 - set password 104
 - software update 127
 - upgrade 121
- DS4000 Series 4
 - overall positioning 2
- DS4000 Storage
 - Server xi, 24, 67–68, 133, 135–136, 176, 190, 340,

- 365, 394, 398, 400–401
- DS4000 Storage Server xi, 1, 7, 13, 15–16, 18, 24, 67–71, 133–137, 164, 167, 169, 172, 188, 190, 202, 214, 339–340, 342, 364–365, 392, 394, 396, 398
 - management station 75
 - model 155, 157
 - product 68
- DS4000 Storage server
 - logical drive 58, 177, 228
- DS4100 4
- DS4200 3
 - cabling 86
- DS4200 Express 3
- DS4300 4
- DS4300 Turbo 4, 80
 - model 4
- DS4500 4
- DS4700 3
 - cabling 92–93
- DS4800 3
 - drive-side cabling 97
 - host cabling 90
 - partition key 68
- DS6000 2
- DS8000 2
- DSM 29
- Dynamic Capacity Expansion (DCE) 153, 359
- dynamic disc 64–65, 140, 143–144
- dynamic disk 63–64, 140–141, 143–144
- Dynamic Logical Drive Expansion (DVE) 153
- Dynamic mode switching 288
- Dynamic Multi-Pathing (DMP) 28, 49
- Dynamic RAID Migration (DRM) 153, 359
- Dynamic Reconstruction Rate (DRR) 153
- Dynamic Segment Sizing (DSS) 153, 359
- Dynamic Volume Expansion (DVE) 141, 359
- Dynamo 190

E

- e-mail 47, 113–114
- Emulex 50
- enclosure
 - ID 33, 101
 - loss protection 43, 109
 - migrating 128
- enclosure ID 34, 103
- enclosure loss protection 108
- Engenio provider
 - arrayhosts.txt 241–242
 - providerStore 242
- Enhanced Remote Mirroring (ERM) 5–6, 14, 16, 47, 57–59, 78, 82, 97, 110, 113, 150, 288, 382
- Enterprise Management 75
- entry level
 - SATA storage system 4
- environmental requirements 14, 16
- Environmental Service Module (ESM) 7, 68, 90, 382, 389, 392
- Environmental Service Modules (ESM) 89
- ERM enhancements 289

- errpt 359
- ESM 7, 122
- ESM board 33, 89–90
- ESS 2, 356
- Ethernet 121
- Ethernet network 7
- Ethernet port 68
- Ethernet switch 69, 389
- event log 47, 104
- Event Monitor 73–74, 113, 118
- Exchange
 - IOPS 179
 - mailbox 179
 - sizing 183
 - storage groups 181
- execution throttle 26, 140
- EXP unit 127, 130, 382, 389, 392
- EXP420 3
- EXP710 3, 60, 80, 101
- EXP810 3
- expansion
 - ID 101
 - intermix 101
 - supported 100
- Expansion enclosure 32, 34, 67–68, 80, 97–98, 100–101, 103, 120, 122, 130, 383, 388–389, 392, 394
- Expansion unit 90, 130, 388
- expansion unit 89
 - right ESM board switch settings 90
- extend 64, 140
- extent 167
- extent size (ES) 165, 363

F

- fabric 10
- Fabric Shortest Path First (FSPF) 10
- failed drive 41, 104–105, 129, 153
- failover 8, 47, 115
 - alert delay 115–116
- FASTT MSJ 26, 50, 187, 391, 398
- FASTT100 4
- FASTT600 4
- FASTT900 4
- FC-AL 5
- FCP 9
- fcs 138
- fcs device 139
- FC-SW 5
- feature key 58, 101, 113
- fiber
 - multi-mode 18
 - single-mode 19
- Fibre Channel 2, 6, 9
 - adapter 12, 23, 53, 88, 227, 358, 364
 - controller 16, 121, 387
 - director 10
 - hub 88, 122
 - I/O path 72
 - I/O path failover driver 7
 - loop 7, 33, 227, 387

- loop ID 103
- switch 21, 77, 364, 387
- unconfigured capacity 106
- Fibre Channel (FC) 2, 4, 7, 9–10, 12, 14, 16, 19–21, 23, 35–36, 53, 59, 72, 75, 77, 80, 88–89, 100–101, 106, 121, 128, 137, 155–156, 178, 187, 207, 227, 333, 343, 358, 364–365, 368, 387, 390–391, 400
- FICON 9
- fileplace 221–222
- filesystem 39, 41, 49, 62, 109, 139, 145–146, 153, 164, 167, 169, 172, 177, 186, 204, 207, 217–218, 222, 359, 363, 371, 373–374
- firewall 69, 244
- firmware 120–121
 - activate 126
- Firmware version 25, 120, 122, 340, 389, 392
- FlashCopy 5–6, 14, 57–58, 60, 110, 113, 151–153, 382, 391
- floor plan 17
- floor-load 16
- flushing level 54, 57
- Forced Unit Access 160
- format 153
- frame switch 10
- free capacity 106
 - logical drive 106
- free space node 153
- free-capacity 106
- FSPF 10
- ftp 117
- full stripe write 138–139, 157

G

- GBIC 20
- General Public License (GNU) 197
- gettime 198
- Gigabit Interface Converters 20
- gigabit transport 19
- given storage subsystem
 - maximum I/O transfer rates 207
- Global Copy 59, 288
- Global Mirroring 59
- GNU 197
- graphical user interface (GUI) 7
- growth 15
- GUID partition table (GPT) 143

H

- HACMP 339
- HACMP cluster
 - configuration 364, 369, 378
 - node 362, 364
- HACMP environment 362–365
- HACMP V5.1
 - disk heartbeat device 377
- HANFS 363
- HBA 7, 12, 15–16, 24, 26, 78, 96, 112, 114, 135–138, 140, 176, 186, 211, 225, 340–344, 382, 390–391, 394, 398

- sharing 16
- HBAAs 10, 16, 24–26, 50, 77–78, 82, 88, 111, 136–137, 140, 186, 215, 226, 340, 345–346, 349–350, 354–355, 365, 382, 390, 394–395, 398, 400–401
 - separate zone 365
- hdisk 342
- heterogeneous host 110
- High Availability Cluster Multi-Processing (HACMP) 363
- High Availability Subsystem (HAS) 363
- host 50, 136
 - data layout 137
 - performance 136
 - settings 136
- host agent 75, 114
- Host Bus Adapter 24
- Host Bus Adapter (HBA) 7, 15–16, 24, 26, 49–50, 112, 136–137, 176, 186, 189, 224–227, 340, 342, 344–345, 347, 349–351, 354–357, 365, 390, 394, 398–400
- host cabling 80, 87
- host computer 7, 316
- host connection 14
- Host group 110
- host group 49–50, 110–112, 365, 370, 394
 - single server 50, 112, 394
- host group. 50
- host path 135
- Host port 333
- host port 12, 49–50, 59, 77, 111–112, 344, 370
 - World Wide Names 112
- host software 72
- host system 121
- Host Type 50, 110, 112, 160, 176–177, 344, 394, 397, 400–401
- host type 112, 397, 401
- Host-based 7
- hot_add 112
- hot-scaling 130
- hot-spare 16, 48, 97, 104–105, 153
 - capacity 105
 - global 105
 - ratio 105
 - unassign 105
- hot-spare drive 104–105
- hub 10
- HyperTerm 71

I

- I/O path 110
- I/O rate 151, 206–207
- I/O request 15, 24, 27, 53–54, 144, 193, 198, 205, 207, 218
- in-band 72, 104
- in-band management 51, 72, 113, 391, 394
- in-band management (IBM) 7, 122, 342
- initial discovery 75, 392
- initial setup 68
- initialization 153
- in-order-delivery 161
- InstallAnywhere 72
- inter-disk allocation 63

- intermix 59, 88
 - feature key 101
- intermixing 101
- interoperability matrix 15
- interpolicy 139
- inter-switch link 10
- intra-disk allocation 63
- IO
 - blocksize 136
- IO rate 36, 156
- IO request 138, 165, 380
- IOD 161
- lometer 190
 - configure 191
- IOPS 6, 15, 36, 41, 53, 134, 150, 176, 179–181, 186, 202–204, 226, 380
- iostat 220
- IP address 22, 67–69, 71–72, 75–76, 114, 209, 382, 389, 392
- ISL 10

J

- Jet database 178
- JFS 139
- JFS2 139
- jfs2 filesystem 204
- journal 148, 156
- Journalized File System (JFS) 373

K

- Kb_read 218
- Kb_wrtn 218

L

- labeling 21
- latency 135, 157, 186
- LC connector 20
- LDM 63
- leaf page 165
- lg_term_dma 139, 358
- link failure 227
- Linux 50, 112, 127
- Linux RDAC 401
- list 165
- load balancing 24–25, 28, 63, 110, 170
- load distribution 28
- Load LSIMPE.CDM 401
- load sharing 24
- Logical Disk Manager (LDM) 144
- logical drive 6–7, 38, 40, 42–43, 45–47, 49, 106, 109–113, 135–139, 141, 166–167, 169, 174, 176–177, 181–182, 204, 206, 208, 210, 216, 359, 393, 396, 401
 - cache read-ahead 56
 - capacity 109–110
 - create 158
 - creating 106
 - instant notification 116
 - ownership 122

- preferred path 122
- primary 47, 113, 151
- second statistics 207
- secondary 47, 151
- segment size 207
- single drive 153
- source 113
- target 113
- logical drive transfer alert 115
- Logical Unit Number (LUN) 110
- Logical unit number (LUN) 328, 330
- Logical view 61–62, 184, 205
- logical volume 45, 61–63, 65, 139, 143, 213, 217–218, 221–222, 360–361, 363, 371–374, 378, 394
- Logical Volume Control Block 139
- Logical Volume Manager (LVM) 60, 137, 139
- long field (LF) 164
- longwave 18
- loop 10
- loopback 227
- loss protection 109
- LSIMPE.CDM 49
- lsmcocde 340
- LUN 42, 45, 49–51, 72, 110, 112–113, 128, 154, 160, 182, 213, 225, 228, 360, 393–396
 - masking 49
- LUN masking 110
- LUNs 110, 333
- LVM 63
 - conceptual view 62
- LVM component 371

M

- Major Event Log (MEL) 115, 129
- managed disk group
 - viewing 335
- management station 122
- management workstation 72
- mapping 49–50, 110–111
- Mappings View 110
- master 322
- Master Boot Record (MBR) 143–144
- master boot record (MBR) 185
- max_xfer_size 139, 213, 215–216, 359
- maximum IOs 137
- Media Scan 152
- Media scan 51
- Messaging Application Programming Interface (MAPI) 179
- Metro Mirroring 58, 288
- microcode 106, 121
 - staged upgrade 122
 - upgrade 120
- Microsoft Exchange 178
- Microsoft Management Console (MMC) 64
- Microsoft Windows
 - 2000 Resource Kit 144
 - Performance Monitor 187
 - physical limitation 398
 - resource kit 144, 397

- workstation 72
- Micrsoft SQL server 172
 - maintenance plans 175
 - transaction log 174
- migration 128
- mini hub 88
- miniport 26
- mirrored 323
- Mirrored Volume 144
- mirroring 38, 40, 140
- misalignment 137
- Mixed zone 11
- mkiv 63
- modal dispersion 18
- modification priority 152
- monitoring 45
- MPIO 14, 28, 72, 74
- MPP 182
- mpputil 228
- multi-mode fiber 18
- multi-mode fiber (MMF) 18
- Multipath 72
- multi-path driver 390, 400
- multipath driver 29
- Multipath Input/Output (MPIO) 28
- multi-path support 400–401
- multiple disk arrays 328
- multiple fabric 12
- multiple HBA 390, 401
- My support 120

N

- netCfgSet 71
- netCfgShow 71
- Netware 49
- network parameters 69
- node failover 362
- node-to-node 10
- non-concurrent access 371
 - shared LVM components 371
- Novell NetWare xii, 8, 399
- Novell TID 400
 - installation instructions 400
- NTFS 143, 172
- num_cmd_elem 138–139
- num_cmd_elems 358
- NVSRAM 7, 112, 121, 123–124, 127, 148, 160, 190, 225–226, 389, 392
- NVSRAM version 124, 392

O

- oad balancing 96
- Object Data Manager (ODM) 363, 371, 374
- offset 138, 166
- OLTP environment 164
- OnLine Transaction Processing (OLTP) 147
- online transaction processing (OLTP) 15, 147–148
- op read 201, 203
- operating system 7

- command line tool 112
- operating system (OS) 8, 25, 40–41, 43, 50, 72, 134, 180, 182, 195, 204–205, 207, 226, 341, 359, 391, 395, 397
- Oracle
 - cache memory 170
 - database 167
 - logs 169
 - volume management 170
- Oracle Database 163, 167
- OS platform 49, 72
 - configuration notes 397
- Out port 89–90
- out-of-band 72, 104
- ownership 45–46, 122, 206

P

- paging 172
- parity 157
- partition key 68
- password 104, 119
- path failure 80
- PCI bus 359
- PCI slots 24
- performance 36–37, 42, 56
- performance improvement 328
- performance monitor 110, 180, 186–187, 190, 204–205, 207–209, 211, 214–216, 396
- Performance Monitor (Windows) 228
- performance monitor job 249
- Persistent Reservation 28
- physical drive 52, 216–217, 228
- physical partitions (PPs) 61
- physical volume 61, 63, 110, 217–219, 221, 361, 372–374
- planning 13–14
- Plug and Play 28
- Point-in-Time 113
- point-to-point 10
- polling interval 207
- Port level zone 11
- preferred controller 27, 45–46, 115
- preferred path 122
- prefetch 165–166
- premium features 57
- primary 113
- probe job 245
- profile 104
- proxy 27
- PTF U499974 359
- putty 70

Q

- Qlogic 50
- Qlogic SANSurfer 224
- queue depth 23, 25–26, 36, 136, 139–140, 155, 176–177, 195, 366
- queuing 36

R

- rack 16
 - layout 17
- RAID 2, 6
 - comparison 41
 - level 41, 207
 - levels 37
 - reliability 41
- RAID 1 137
- RAID 5 137
 - definite advantage 156
- RAID controller 6, 318
- RAID Level 108
- RAID level 37–39, 41–42, 45, 110, 151, 153, 172–173, 180, 184, 206, 382, 393, 395
- RAID types 155
- range 165
- range of physical volumes 63
- raw device 138, 164, 167, 169
- RDAC 7, 14, 25, 28, 73–74, 227, 344
- RDAC driver 227, 341–342, 356, 397, 401–402
 - architecture 227
- Read cache 134, 173–175, 185, 395
- read caching 55
- read percentage 206
- read-ahead 136–138, 206
- read-ahead multiplier 54–55, 159
- read-behind-write 199
- read-head multiplier 109
- rebuild 104
- Recovery Guru 116, 121
- recovery storage group (RSG) 181, 183, 185
- recovery time objective (RTO) 113
- Redbooks Web site 404
 - Contact us xiii
- redo log 168
- Redo logs
 - RAID 169
- redundancy 80
- redundant array of independent disks 2
- Redundant Dual Active Controller (RDAC) 49, 112, 115, 121, 127, 182, 227, 341, 356–357, 382, 398
- Reliable Scalable Cluster Technology (RSCT) 361
- remote access 117
- remote login 69
- Remote Support Manager (RSM) 117
- reorgvg 222
- reservation 128
- revolutions per minute (RPM) 36
- RLOGIN 119
- rlogin 70–71
- Role Reversal 288
- Round Robin 28
- round-robin 25
- RPM 36
- RSM 117, 119
 - firewall 119
 - security 119

S

- SAN 7, 9, 14
- SAN Volume Controller (SVC) 13, 316, 356, 382
- SANSurfer 224
- SANtricity SMI Provider 237
- sar 220
- SATA 3
- SATA disk 3
- SATA drive 2, 6, 35–36, 59–60, 80, 100, 155, 173–177
- SC connector 20
- script 127
- SCSI 9
- SCSIPort 14
- SCSIport 26–27
- SDD 327
- security 104, 119
- seek time 36
- segment 52, 168
- segment size 16, 52–53, 106, 109, 136, 139, 144, 153, 157–159, 165–169, 173–177, 185, 207, 393, 395–396
- separate HBA 400
- sequential IO
 - high throughput 160
- serial connection 69
- serial port 71
- Service agent (SA) 236
- Service Alert 117
- Service Location Protocol (SLP) 236
- SFF 20
- SFP 20, 77, 88
- shell command 71
- shortwave 18
- Simple Network Management Protocol (SNMP) 73
- Simple Volume 143
- single mode fiber (SMF) 18
- single-mode fiber 19
- site planning 22
- slice 137
- SLP 237
- SM Client 68–69, 72, 75, 104, 106, 391
- SMagent 7, 72, 342
- small form factor plug (SFP) 20–21, 88
- Small Form Factor Transceivers 20
- Small Form Pluggable 20
- SMclient 72, 75
- SMI-S 236
- smit 372–375, 378
- SMS (System Managed Storage) 164
- SMTP 113–114
- SMTP queue 178, 182, 185
- SMutil 72
- SNMP 47, 73, 113–114
- software update 127
- source logical drive 113
- Spanned Volume 143
- spanned volume 65
- spanning 140
- spare 40
- SQL Server
 - logical file name 174

- operation 174
- staged microcode upgrade 122
- Standard Edition (SE) 183
- statistic
 - virtual memory 172
- statistics
 - CPU 171
 - network 172
- Storage Area Network, see SAN
- Storage Area Networks
 - server consolidation 2
- storage bus 7
- storage capacity 2
- storage group 178, 180–183, 186
 - overhead IOPS 181
 - required IOPS 181
- Storage Management Initiative - Specification (SMI-S) 236
- Storage Manager 6
 - 9.1 Agent 7
 - logical drives 343
 - warning message 76
- storage manager (SM) 33, 108
- Storage Manager 9.1
 - Client 7
 - Runtime 7
 - Utility 7
- Storage Networking Industry Association (SNIA) 236
- storage partition 49, 111, 342, 344–348, 350–351, 394, 400
- storage partitioning 45, 48–49, 110
- Storage Performance Analyzer (SPA) 382
- Storage Server
 - controller port 78
- storage server 123
 - logical drives 45, 49, 104, 135–137, 190
- storage servers 2
- storage subsystem 122, 316
 - Performance Monitor Statistics 210
- StorPort 14, 27
- Storport 26–27
- Streaming database 178
- stripe kill 41
- stripe size 138, 159
- stripe VDisks 328
- stripe width 137–138, 148, 159
- Striped Volume 144
- striped volume 64–65, 143–144
- striping 137, 140, 166
- sub-disk 64
- Subsystem Management window
 - Mappings View 111
- Subsystem management window 69, 76, 104–106, 111, 116, 123, 204, 400
- sundry drive 129
- support
 - My support 120
- surface scan 51
- switch 10, 226
 - ID 33, 103

- synchronization 151
- Synchronize Cache 160
- synchronous mirroring 58, 288
- System Performance Measurement Interface (SPMI) 224

T

- tablespace 147, 155–156, 164–167
- tape 16, 137
- target logical drive 113
- Task Assistant 75, 104, 111
- TCO 9
- Telco 4
- tempdb 173
- tempdb database 173–174
- throughput 15, 43, 53–54, 56, 134–135, 139, 148
- throughput based 147
- time server 104
- timeserver 197
- Tivoli Storage Manager 176
- Tivoli Storage Manager (TSM) 147, 176–178
- tm_act 217
- topas command 223–224
- total cost of ownership (TCO) 9
- Total I/O 206
- TotalStorage Productivity Center (TPC) 382
- TPC
 - Data agent 234
 - Data server 234
 - Device server 234
 - Fabric agent 234
 - reporting 252
- TPC for Disk 231, 237
- tps 218
- transaction 134
- transaction based 147
- Transaction log 38, 148, 172–175, 178, 180–181, 395
- transceivers 19
- tray Id 33
- trunking 10
- TSM database 176–177
- TSM instance 176–177

U

- unconfigured capacity 106
 - logical drive 106
- Universal Transport Mechanism (UTM) 342
- upgrade 128
- upgrades 121
- User agent (UA) 236
- User profile 179, 181, 186
- utilities 7

V

- Veritas Volume Manager 64
- vg 61, 139, 359, 371–373, 375
- viewing managed disk groups 335
- virtual memory 172

- vmstat 219
- volume 45
- volume group 61, 64, 170, 176–177, 213, 215, 342, 359–361, 371–377
 - available new space 361
 - file systems 371, 374
 - forced varyon 375
 - logical drives 176
- VolumeCopy 57, 113, 151
- VxVM 64

W

- Web-Based Enterprise Management (WBEM) 234
- Windows
 - basic disc 140
 - dynamic disk 140, 143
- Windows 2000 24, 49–50, 63–65, 140, 143–144, 146–147, 172–173, 187, 227–228, 391, 398
 - dynamic disks 63, 140, 398
- Windows 2003 24, 63–64, 140–142, 144–147, 172–173, 178, 227, 391, 397–398
- Windows Event Log 248
- WMI 29
- worker 191
- workload 134–135
 - throughput based 134
 - transaction based 134
- workload generator. 190
- workload type 147
- World Wide Name 111
- World Wide Name (WWN) 11–12, 50, 129, 228, 340, 344, 395, 399
- World Wide Name (WWN). See WWN
- World Wide Node Name (WWNN) 12
- World Wide Port Name (WWPN) 11–12
- write cache 136, 160
- write cache mirroring 160
- write caching 56
- write order consistency 113
- write order consistency (WOC) 113
- write-back 56
- write-through 56
- WWN 50, 111–112
- WWPN 11
- WWPN zone 11

X

- Xdd 187, 197–199, 203
 - components 197
 - de-skew 199
 - install 199
 - running 201
- xmlCIM 234

Z

- zone 10
- Zone types 11
- zoning 11, 14–15, 78, 82, 112



DS4000 Best Practices and Performance Tuning Guide



Redbooks

DS4000 Best Practices and Performance Tuning Guide

**Performance
measurement using
TPC for Disk**

**ERM guidelines and
bandwidth estimator**

**Managing and using
the DS4000 with SVC**

This IBM Redbook is intended for IBM technical professionals, Business Partners, and customers responsible for the planning, deployment, and maintenance of IBM TotalStorage DS4000 family of products. We realize that setting up a DS4000 Storage Server can be a complex task. There is no single configuration that will be satisfactory for every application or situation.

First, we provide a conceptual framework for understanding the DS4000 in a Storage Area Network. Then we offer our recommendations, hints, and tips for the physical installation, cabling, and zoning, using the Storage Manager setup tasks.

After that, we turn our attention to the performance and tuning of various components and features, including numerous recommendations. We look at performance implications for various application products such as DB2, Oracle, Tivoli Storage Manager, Microsoft SQL server, and in particular, Microsoft Exchange with a DS4000 storage server.

Then we review the various tools available to simulate workloads and to measure and collect performance data for the DS4000, including the Engenio Storage Performance Analyzer. We also consider the AIX environment, including High Availability Cluster Multiprocessing (HACMP) and General Parallel File System (GPFS). Finally, we provide a quick guide to the DS4000 Storage Server installation and configuration.

**INTERNATIONAL
TECHNICAL
SUPPORT
ORGANIZATION**

**BUILDING TECHNICAL
INFORMATION BASED ON
PRACTICAL EXPERIENCE**

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

**For more information:
ibm.com/redbooks**

SG24-6363-03

ISBN 0738486019