

# The IBM TotalStorage DS8000 Series: Concepts and Architecture

Advanced features and performance  
breakthrough with POWER5 technology

Configuration flexibility with LPAR  
and virtualization

Highly scalable solutions for  
on demand storage



Cathy Warrick	Christine O'Sullivan
Olivier Alluis	Stu S Preacher
Werner Bauer	Torsten Rothenwaldt
Heinz Blaschek	Tetsuroh Sano
Andre Fourie	Jing Nan Tang
Juan Antonio Garay	Anthony Vandewerdt
Torsten Knobloch	Alexander Warmuth
Donald C Laing	Roland Wolf

**Redbooks**





International Technical Support Organization

**The IBM TotalStorage DS8000 Series:  
Concepts and Architecture**

April 2005

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

**First Edition (April 2005)**

This edition applies to the DS8000 series per the October 12, 2004 announcement. Please note that pre-release code was used for the screen captures and command output; some details may vary from the generally available product.

**Note:** This book is based on a pre-GA version of a product and may not apply when the product becomes generally available. We recommend that you consult the product documentation or follow-on versions of this redbook for more current information.

© Copyright International Business Machines Corporation 2005. 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

<b>Notices</b> .....	xiii
Trademarks .....	xiv
<b>Preface</b> .....	xv
The team that wrote this redbook .....	xv
Become a published author .....	xix
Comments welcome .....	xix
<b>Part 1. Introduction</b> .....	1
<b>Chapter 1. Introduction to the DS8000 series</b> .....	3
1.1 The DS8000, a member of the TotalStorage DS family .....	4
1.1.1 Infrastructure Simplification .....	4
1.1.2 Business Continuity .....	4
1.1.3 Information Lifecycle Management .....	4
1.2 Overview of the DS8000 series .....	4
1.2.1 Hardware overview .....	6
1.2.2 Storage capacity .....	7
1.2.3 Storage system logical partitions (LPARs) .....	7
1.2.4 Supported environments .....	8
1.2.5 Resiliency Family for Business Continuity .....	8
1.2.6 Interoperability .....	10
1.2.7 Service and setup .....	10
1.3 Positioning .....	11
1.3.1 Common set of functions .....	11
1.3.2 Common management functions .....	12
1.3.3 Scalability and configuration flexibility .....	13
1.3.4 Future directions of storage system LPARs .....	13
1.4 Performance .....	14
1.4.1 Sequential Prefetching in Adaptive Replacement Cache (SARC) .....	14
1.4.2 IBM TotalStorage Multipath Subsystem Device Driver (SDD) .....	14
1.4.3 Performance for zSeries .....	14
1.5 Summary .....	15
<b>Part 2. Architecture</b> .....	17
<b>Chapter 2. Components</b> .....	19
2.1 Frames .....	20
2.1.1 Base frame .....	20
2.1.2 Expansion frame .....	21
2.1.3 Rack operator panel .....	21
2.2 Architecture .....	22
2.2.1 Server-based SMP design .....	24
2.2.2 Cache management .....	24
2.3 Processor complex .....	26
2.3.1 RIO-G .....	29
2.3.2 I/O enclosures .....	29
2.4 Disk subsystem .....	30
2.4.1 Device adapters .....	30

2.4.2	Disk enclosures . . . . .	31
2.5	Host adapters . . . . .	37
2.5.1	FICON and Fibre Channel protocol host adapters . . . . .	38
2.6	Power and cooling . . . . .	39
2.7	Management console network . . . . .	40
2.8	Summary . . . . .	41
<b>Chapter 3. Storage system LPARs (Logical partitions)</b> . . . . .		<b>43</b>
3.1	Introduction to logical partitioning . . . . .	44
3.1.1	Virtualization Engine technology . . . . .	44
3.1.2	Partitioning concepts . . . . .	44
3.1.3	Why Logically Partition? . . . . .	47
3.2	DS8000 and LPAR . . . . .	48
3.2.1	LPAR and storage facility images . . . . .	48
3.2.2	DS8300 LPAR implementation . . . . .	49
3.2.3	Storage facility image hardware components . . . . .	50
3.2.4	DS8300 Model 9A2 configuration options . . . . .	52
3.3	LPAR security through POWER™ Hypervisor (PHYP) . . . . .	54
3.4	LPAR and Copy Services . . . . .	55
3.5	LPAR benefits . . . . .	56
3.6	Summary . . . . .	59
<b>Chapter 4. RAS</b> . . . . .		<b>61</b>
4.1	Naming . . . . .	62
4.2	Processor complex RAS . . . . .	63
4.3	Hypervisor: Storage image independence . . . . .	66
4.3.1	RIO-G - a self-healing interconnect . . . . .	67
4.3.2	I/O enclosure . . . . .	67
4.4	Server RAS . . . . .	67
4.4.1	Metadata checks . . . . .	67
4.4.2	Server failover and failback . . . . .	68
4.4.3	NVS recovery after complete power loss . . . . .	70
4.5	Host connection availability . . . . .	71
4.5.1	Open systems host connection . . . . .	74
4.5.2	zSeries host connection . . . . .	74
4.6	Disk subsystem . . . . .	75
4.6.1	Disk path redundancy . . . . .	75
4.6.2	RAID-5 overview . . . . .	76
4.6.3	RAID-10 overview . . . . .	76
4.6.4	Spare creation . . . . .	77
4.6.5	Predictive Failure Analysis® (PFA) . . . . .	78
4.6.6	Disk scrubbing . . . . .	79
4.7	Power and cooling . . . . .	79
4.7.1	Building power loss . . . . .	80
4.7.2	Power fluctuation protection . . . . .	80
4.7.3	Power control of the DS8000 . . . . .	80
4.7.4	Emergency power off (EPO) . . . . .	80
4.8	Microcode updates . . . . .	81
4.9	Management console . . . . .	82
4.10	Summary . . . . .	82
<b>Chapter 5. Virtualization concepts</b> . . . . .		<b>83</b>
5.1	Virtualization definition . . . . .	84
5.2	Storage system virtualization . . . . .	84

5.3	The abstraction layers for disk virtualization . . . . .	85
5.3.1	Array sites . . . . .	86
5.3.2	Arrays . . . . .	87
5.3.3	Ranks . . . . .	88
5.3.4	Extent pools . . . . .	89
5.3.5	Logical volumes . . . . .	91
5.3.6	Logical subsystems (LSS). . . . .	94
5.3.7	Volume access . . . . .	96
5.3.8	Summary of the virtualization hierarchy . . . . .	98
5.3.9	Placement of data . . . . .	99
5.4	Benefits of virtualization . . . . .	100
<b>Chapter 6. IBM TotalStorage DS8000 model overview and scalability . . . . .</b>		<b>103</b>
6.1	DS8000 highlights . . . . .	104
6.1.1	Model naming conventions . . . . .	104
6.1.2	DS8100 Model 921 . . . . .	105
6.1.3	DS8300 Models 922 and 9A2 . . . . .	106
6.2	Model comparison . . . . .	108
6.3	Designed for scalability . . . . .	109
6.3.1	Scalability for capacity . . . . .	109
6.3.2	Scalability for performance . . . . .	110
6.3.3	Model upgrades . . . . .	113
<b>Chapter 7. Copy Services . . . . .</b>		<b>115</b>
7.1	Introduction to Copy Services . . . . .	116
7.2	Copy Services functions . . . . .	116
7.2.1	Point-in-Time Copy (FlashCopy). . . . .	116
7.2.2	FlashCopy options . . . . .	118
7.2.3	Remote Mirror and Copy (Peer-to-Peer Remote Copy) . . . . .	123
7.2.4	Comparison of the Remote Mirror and Copy functions. . . . .	130
7.2.5	What is a Consistency Group? . . . . .	132
7.3	Interfaces for Copy Services . . . . .	136
7.3.1	Storage Hardware Management Console (S-HMC) . . . . .	136
7.3.2	DS Storage Manager Web-based interface . . . . .	137
7.3.3	DS Command-Line Interface (DS CLI) . . . . .	138
7.3.4	DS Open application programming Interface (API). . . . .	138
7.4	Interoperability with ESS . . . . .	139
7.5	Future Plans . . . . .	139
<b>Part 3. Planning and configuration . . . . .</b>		<b>141</b>
<b>Chapter 8. Installation planning . . . . .</b>		<b>143</b>
8.1	General considerations . . . . .	144
8.2	Delivery requirements . . . . .	144
8.3	Installation site preparation . . . . .	145
8.3.1	Floor and space requirements . . . . .	145
8.3.2	Power requirements . . . . .	147
8.3.3	Environmental requirements . . . . .	149
8.4	Host attachment . . . . .	150
8.4.1	Attaching to open systems hosts . . . . .	150
8.4.2	ESCON-attached S/390 and zSeries hosts . . . . .	151
8.4.3	FICON-attached S/390 and zSeries hosts . . . . .	151
8.4.4	Where to get the updated information for host attachment. . . . .	152
8.5	Network and SAN requirements . . . . .	153

8.5.1 S-HMC network requirements . . . . .	153
8.5.2 Remote support connection requirements . . . . .	154
8.5.3 Remote power control requirements . . . . .	154
8.5.4 SAN requirements . . . . .	154
<b>Chapter 9. Configuration planning . . . . .</b>	<b>157</b>
9.1 Configuration planning overview . . . . .	158
9.2 Storage Hardware Management Console (S-HMC) . . . . .	158
9.2.1 External S-HMC . . . . .	159
9.2.2 S-HMC software components . . . . .	160
9.2.3 S-HMC network topology . . . . .	162
9.2.4 FTP Offload option . . . . .	166
9.3 DS8000 licensed functions . . . . .	167
9.3.1 Operating environment license (OEL) - required feature . . . . .	167
9.3.2 Point-in-Time Copy function (2244 Model PTC) . . . . .	168
9.3.3 Remote Mirror and Copy functions (2244 Model RMC) . . . . .	169
9.3.4 Remote Mirror for z/OS (2244 Model RMZ) . . . . .	169
9.3.5 Parallel Access Volumes (2244 Model PAV) . . . . .	170
9.3.6 Ordering licensed functions . . . . .	170
9.3.7 Disk storage feature activation . . . . .	173
9.3.8 Scenarios for managing licensing . . . . .	174
9.4 Capacity planning . . . . .	174
9.4.1 Logical configurations . . . . .	174
9.4.2 Sparing rules . . . . .	176
9.4.3 Sparing examples . . . . .	177
9.4.4 IBM Standby Capacity on Demand (Standby CoD) . . . . .	180
9.4.5 Capacity and well-balanced configuration . . . . .	181
9.5 Data migration planning . . . . .	183
9.5.1 Operating system mirroring . . . . .	184
9.5.2 Basic commands . . . . .	184
9.5.3 Software packages . . . . .	184
9.5.4 Remote copy technologies . . . . .	184
9.5.5 Migration services and appliances . . . . .	185
9.5.6 z/OS data migration methods . . . . .	185
9.6 Planning for performance . . . . .	186
9.6.1 Disk Magic . . . . .	187
9.6.2 Size of cache storage . . . . .	187
9.6.3 Number of host ports/channels . . . . .	187
9.6.4 Remote copy . . . . .	187
9.6.5 Parallel Access Volumes (z/OS only) . . . . .	187
9.6.6 I/O priority queuing (z/OS only) . . . . .	187
9.6.7 Monitoring performance . . . . .	187
9.6.8 Hot spot avoidance . . . . .	188
<b>Chapter 10. The DS Storage Manager - logical configuration . . . . .</b>	<b>189</b>
10.1 Configuration hierarchy, terminology, and concepts . . . . .	190
10.1.1 Storage configuration terminology . . . . .	190
10.1.2 Summary of the DS Storage Manager logical configuration steps . . . . .	199
10.2 Introducing the GUI and logical configuration panels . . . . .	202
10.2.1 Connecting to the DS8000 . . . . .	202
10.2.2 The Welcome panel . . . . .	203
10.2.3 Navigating the GUI . . . . .	208
10.3 The logical configuration process . . . . .	211



10.3.1	Configuring a storage complex . . . . .	211
10.3.2	Configuring the storage unit . . . . .	212
10.3.3	Configuring the logical host systems. . . . .	216
10.3.4	Creating arrays from array sites . . . . .	219
10.3.5	Creating extent pools . . . . .	221
10.3.6	Creating FB volumes from extents . . . . .	222
10.3.7	Creating volume groups . . . . .	224
10.3.8	Assigning LUNs to the hosts . . . . .	226
10.3.9	Deleting LUNs and recovering space in the extent pool . . . . .	226
10.3.10	Creating CKD LCUs . . . . .	227
10.3.11	Creating CKD volumes . . . . .	227
10.3.12	Displaying the storage unit WWNN. . . . .	228
10.4	Summary. . . . .	229
<b>Chapter 11. DS CLI . . . . .</b>		<b>231</b>
11.1	Introduction . . . . .	232
11.2	Functionality . . . . .	232
11.3	Supported environments . . . . .	233
11.4	Installation methods . . . . .	233
11.5	Command flow . . . . .	234
11.6	User security . . . . .	239
11.7	Usage concepts . . . . .	239
11.7.1	Command modes . . . . .	239
11.7.2	Syntax conventions. . . . .	241
11.7.3	User assistance . . . . .	241
11.7.4	Return codes. . . . .	242
11.8	Usage examples . . . . .	243
11.9	Mixed device environments and migration . . . . .	244
11.9.1	Migration tasks . . . . .	245
11.10	DS CLI migration example . . . . .	245
11.10.1	Determining the saved tasks to be migrated. . . . .	245
11.10.2	Collecting the task details . . . . .	246
11.10.3	Converting the saved task to a DS CLI command . . . . .	247
11.10.4	Using DS CLI commands via a single command or script . . . . .	249
11.11	Summary. . . . .	251
<b>Chapter 12. Performance considerations . . . . .</b>		<b>253</b>
12.1	What is the challenge? . . . . .	254
12.1.1	Speed gap between server and disk storage . . . . .	254
12.1.2	New and enhanced functions . . . . .	254
12.2	Where do we start? . . . . .	255
12.2.1	SSA backend interconnection. . . . .	256
12.2.2	Arrays across loops . . . . .	256
12.2.3	Switch from ESCON to FICON ports . . . . .	256
12.2.4	PPRC over Fibre Channel links . . . . .	256
12.2.5	Fixed LSS to RAID rank affinity and increasing DDM size . . . . .	256
12.3	How does the DS8000 address the challenge? . . . . .	257
12.3.1	Fibre Channel switched disk interconnection at the back end . . . . .	257
12.3.2	Fibre Channel device adapter . . . . .	260
12.3.3	New four-port host adapters . . . . .	260
12.3.4	POWER5 - Heart of the DS8000 dual cluster design . . . . .	261
12.3.5	Vertical growth and scalability. . . . .	264
12.4	Performance and sizing considerations for open systems . . . . .	264

12.4.1	Workload characteristics . . . . .	265
12.4.2	Cache size considerations for open systems . . . . .	265
12.4.3	Data placement in the DS8000 . . . . .	265
12.4.4	LVM striping . . . . .	266
12.4.5	Determining the number of connections between the host and DS8000 . . . . .	267
12.4.6	Determining the number of paths to a LUN. . . . .	268
12.4.7	Determining where to attach the host . . . . .	268
12.5	Performance and sizing considerations for z/OS . . . . .	269
12.5.1	Connect to zSeries hosts . . . . .	269
12.5.2	Performance potential in z/OS environments . . . . .	270
12.5.3	Appropriate DS8000 size in z/OS environments. . . . .	271
12.5.4	Configuration recommendations for z/OS. . . . .	274
12.6	Summary. . . . .	278

**Part 4. Implementation and management in the z/OS environment. . . . . 279**

<b>Chapter 13. zSeries software enhancements . . . . .</b>	<b>281</b>
13.1 Software enhancements for the DS8000 . . . . .	282
13.2 z/OS enhancements . . . . .	282
13.2.1 Scalability support. . . . .	282
13.2.2 Large Volume Support (LVS) . . . . .	283
13.2.3 Read availability mask support . . . . .	283
13.2.4 Initial Program Load (IPL) enhancements. . . . .	284
13.2.5 DS8000 definition to host software . . . . .	284
13.2.6 Read control unit and device recognition for DS8000. . . . .	284
13.2.7 New performance statistics. . . . .	285
13.2.8 Resource Management Facility (RMF) . . . . .	289
13.2.9 Migration considerations. . . . .	290
13.2.10 Coexistence considerations . . . . .	290
13.3 z/VM enhancements . . . . .	290
13.4 z/VSE enhancements . . . . .	290
13.5 TPF enhancements. . . . .	291

<b>Chapter 14. Data migration in zSeries environments . . . . .</b>	<b>293</b>
14.1 Define migration objectives in z/OS environments . . . . .	294
14.1.1 Consolidate storage subsystems . . . . .	294
14.1.2 Consolidate logical volumes . . . . .	295
14.1.3 Keep source and target volume size at the current size . . . . .	297
14.1.4 Summary of data migration objectives . . . . .	298
14.2 Data migration based on physical migration . . . . .	298
14.2.1 Physical migration with DFSMSdss and other storage software. . . . .	298
14.2.2 Software- and hardware-based data migration. . . . .	299
14.2.3 Hardware- or microcode-based migration. . . . .	302
14.3 Data migration based on logical migration . . . . .	307
14.3.1 Data Set Services Utility . . . . .	307
14.3.2 Hierarchical Storage Manager, DFSMSHsm. . . . .	308
14.3.3 System utilities . . . . .	308
14.3.4 Data migration within the System-managed storage environment . . . . .	308
14.3.5 Summary of logical data migration based on software utilities . . . . .	314
14.4 Combine physical and logical data migration . . . . .	314
14.5 z/VM and VSE/ESA data migration. . . . .	315
14.6 Summary of data migration . . . . .	315

**Part 5. Implementation and management in the open systems environment. . . . . 317**

<b>Chapter 15. Open systems support and software</b> . . . . .	319
15.1 Open systems support . . . . .	320
15.1.1 Supported operating systems and servers . . . . .	320
15.1.2 Where to look for updated and detailed information . . . . .	320
15.1.3 Differences to the ESS 2105 . . . . .	322
15.1.4 Boot support . . . . .	323
15.1.5 Additional supported configurations (RPQ) . . . . .	323
15.1.6 Differences in interoperability between the DS8000 and DS6000 . . . . .	323
15.2 Subsystem Device Driver . . . . .	324
15.3 Other multipathing solutions . . . . .	325
15.4 DS CLI . . . . .	325
15.5 IBM TotalStorage Productivity Center . . . . .	326
15.5.1 Device Manager . . . . .	328
15.5.2 TPC for Disk . . . . .	329
15.5.3 <b>TPC for Replication</b> . . . . .	<b>330</b>
15.6 Global Mirror Utility . . . . .	330
15.7 Enterprise Remote Copy Management Facility (eRCMF) . . . . .	331
15.8 Summary . . . . .	331
<b>Chapter 16. Data migration in the open systems environment</b> . . . . .	<b>333</b>
16.1 Introduction . . . . .	334
16.2 Comparison of migration methods . . . . .	335
16.2.1 Host operating system-based migration . . . . .	335
16.2.2 Subsystem-based data migration . . . . .	339
16.2.3 IBM Piper migration . . . . .	341
16.2.4 Other migration applications . . . . .	342
16.3 IBM migration services . . . . .	342
16.4 Summary . . . . .	342
<b>Appendix A. Open systems operating systems specifics</b> . . . . .	<b>343</b>
General considerations . . . . .	344
The DS8000 Host Systems Attachment Guide . . . . .	344
Planning . . . . .	344
UNIX performance monitoring tools . . . . .	345
IOSTAT . . . . .	345
System Activity Report (SAR) . . . . .	346
VMSTAT . . . . .	347
IBM AIX . . . . .	347
Other publications . . . . .	348
The AIX host attachment scripts . . . . .	348
Finding the World Wide Port Names . . . . .	348
Managing multiple paths . . . . .	349
LVM configuration . . . . .	352
AIX access methods for I/O . . . . .	352
Boot device support . . . . .	353
AIX on IBM iSeries . . . . .	353
Monitoring I/O performance . . . . .	354
Linux . . . . .	356
Support issues that distinguish Linux from other operating systems . . . . .	356
Existing reference material . . . . .	357
Important Linux issues . . . . .	358
Linux on IBM iSeries . . . . .	363
Troubleshooting and monitoring . . . . .	364
Microsoft Windows 2000/2003 . . . . .	366

HBA and operating system settings . . . . .	366
SDD for Windows . . . . .	366
Windows Server 2003 VDS support . . . . .	367
HP OpenVMS . . . . .	368
FC port configuration . . . . .	368
Volume configuration . . . . .	369
Command Console LUN . . . . .	370
OpenVMS volume shadowing . . . . .	370
<b>Appendix B. Using DS8000 with iSeries . . . . .</b>	<b>373</b>
Supported environment . . . . .	374
Hardware . . . . .	374
Software . . . . .	374
Logical volume sizes . . . . .	374
Protected versus unprotected volumes . . . . .	375
Changing LUN protection . . . . .	375
Adding volumes to iSeries configuration . . . . .	376
Using 5250 interface . . . . .	376
Adding volumes to an Independent Auxiliary Storage Pool . . . . .	378
Multipath . . . . .	386
Avoiding single points of failure . . . . .	386
Configuring multipath . . . . .	387
Adding multipath volumes to iSeries using 5250 interface . . . . .	388
Adding volumes to iSeries using iSeries Navigator . . . . .	390
Managing multipath volumes using iSeries Navigator . . . . .	392
Multipath rules for multiple iSeries systems or partitions . . . . .	395
Changing from single path to multipath . . . . .	396
Sizing guidelines . . . . .	396
Planning for arrays and DDMs . . . . .	397
Cache . . . . .	397
Number of iSeries Fibre Channel adapters . . . . .	398
Size and number of LUNs . . . . .	398
Recommended number of ranks . . . . .	399
Sharing ranks between iSeries and other servers . . . . .	399
Connecting via SAN switches . . . . .	400
Migration . . . . .	400
OS/400 mirroring . . . . .	400
Metro Mirror and Global Copy . . . . .	400
OS/400 data migration . . . . .	401
Copy Services for iSeries . . . . .	403
FlashCopy . . . . .	403
Remote Mirror and Copy . . . . .	403
iSeries toolkit for Copy Services . . . . .	404
AIX on IBM iSeries . . . . .	404
Linux on IBM iSeries . . . . .	405
<b>Appendix C. Service and support offerings . . . . .</b>	<b>407</b>
IBM Web sites for service offerings . . . . .	408
IBM service offerings . . . . .	408
IBM Operational Support Services - Support Line . . . . .	410
<b>Related publications . . . . .</b>	<b>413</b>
IBM Redbooks . . . . .	413
Other publications . . . . .	413

Online resources . . . . .	414
How to get IBM Redbooks . . . . .	415
Help from IBM . . . . .	415
<b>Index . . . . .</b>	<b>417</b>



# Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not 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/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM 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 illustrates 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. 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 IBM's application programming interfaces.

# Trademarks

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

@server®	DFSMSHsm™	MVS™
Redbooks (logo)  ™	DFSORT™	Notes®
ibm.com®	Enterprise Storage Server®	OS/390®
iSeries™	Enterprise Systems Connection	OS/400®
i5/OS™	Architecture®	Parallel Sysplex®
pSeries®	ESCON®	PowerPC®
xSeries®	FlashCopy®	Predictive Failure Analysis®
z/OS®	Footprint®	POWER™
z/VM®	FICON®	POWER5™
zSeries®	Geographically Dispersed Parallel	Redbooks™
AIX 5L™	Sysplex™	RMF™
AIX®	GDPS®	RS/6000®
AS/400®	Hypervisor™	S/390®
BladeCenter™	HACMP™	Seascope®
Chipkill™	IBM®	System/38™
CICS®	IMS™	Tivoli®
DB2®	Lotus Notes®	TotalStorage Proven™
DFSMS/MVS®	Lotus®	TotalStorage®
DFSMS/VM®	Micro-Partitioning™	Virtualization Engine™
DFSMSdss™	Multiprise®	VSE/ESA™

The following terms are trademarks of other companies:

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

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

Intel, Intel Inside (logos), and Pentium are trademarks of Intel Corporation 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, and service names may be trademarks or service marks of others.



# Preface

This IBM® Redbook describes the IBM TotalStorage® DS8000 series of storage servers, its architecture, logical design, hardware design and components, advanced functions, performance features, and specific characteristics. The information contained in this redbook is useful for those who need a general understanding of this powerful new series of disk enterprise storage servers, as well as for those looking for a more detailed understanding of how the DS8000 series is designed and operates.

The DS8000 series is a follow-on product to the IBM TotalStorage Enterprise Storage Server® with new functions related to storage virtualization and flexibility. This book describes the virtualization hierarchy that now includes virtualization of a whole storage subsystem. This is possible by utilizing IBM's pSeries® POWER5™-based server technology and its Virtualization Engine™ LPAR technology. This LPAR technology offers totally new options to configure and manage storage.

In addition to the logical and physical description of the DS8000 series, the fundamentals of the configuration process are also described in this redbook. This is useful information for proper planning and configuration for installing the DS8000 series, as well as for the efficient management of this powerful storage subsystem.

Characteristics of the DS8000 series described in this redbook also include the DS8000 copy functions: FlashCopy®, Metro Mirror, Global Copy, Global Mirror and z/OS® Global Mirror. The performance features, particularly the new switched FC-AL implementation of the DS8000 series, are also explained, so that the user can better optimize the storage resources of the computing center.

## The team that wrote this redbook

This redbook was produced by a team of specialists from around the world working at the Washington Systems Center in Gaithersburg, MD.

**Cathy Warrick** is a project leader and Certified IT Specialist in the IBM International Technical Support Organization. She has over 25 years of experience in IBM with large systems, open systems, and storage, including education on products internally and for the field. Prior to joining the ITSO two years ago, she developed the Technical Leadership education program for the IBM and IBM Business Partner's technical field force and was the program manager for the Storage Top Gun classes.

**Olivier Alluis** has worked in the IT field for nearly seven years. After starting his career in the French Atomic Research Industry (CEA - Commissariat à l'Energie Atomique), he joined IBM in 1998. He has been a Product Engineer for the IBM High End Systems, specializing in the development of the IBM DWDM solution. Four years ago, he joined the SAN pre-sales support team in the Product and Solution Support Center in Montpellier working in the Advanced Technical Support organization for EMEA. He is now responsible for the Early Shipment Programs for the Storage Disk systems in EMEA. Olivier's areas of expertise include: high-end storage solutions (IBM ESS), virtualization (SAN Volume Controller), SAN and interconnected product solutions (CISCO, McDATA, CNT, Brocade, ADVA, NORTEL, DWDM technology, CWDM technology). His areas of interest include storage remote copy on long-distance connectivity for business continuance and disaster recovery solutions.

**Werner Bauer** is a certified IT specialist in Germany. He has 25 years of experience in storage software and hardware, as well as S/390®. He holds a degree in Economics from the University of Heidelberg. His areas of expertise include disaster recovery solutions in enterprises utilizing the unique capabilities and features of the IBM Enterprise Storage Server, ESS. He has written extensively in various redbooks, including Technical Updates on DFSMS/MVS® 1.3, 1.4, 1.5. and Transactional VSAM.

**Heinz Blaschek** is an IT DASD Support Specialist in Germany. He has 11 years of experience in S/390 customer environments as a HW-CE. Starting in 1997 he was a member of the DASD EMEA Support Group in Mainz Germany. In 1999, he became a member of the DASD Backoffice Mainz Germany (support center EMEA for ESS) with the current focus of supporting the remote copy functions for the ESS. Since 2004 he has been a member of the VET (Virtual EMEA Team), which is responsible for the EMEA support of DASD systems. His areas of expertise include all large and medium-system DASD products, particularly the IBM TotalStorage Enterprise Storage Server.

**Andre Fourie** is a Senior IT Specialist at IBM Global Services, South Africa. He holds a BSc (Computer Science) degree from the University of South Africa (UNISA) and has more than 14 years of experience in the IT industry. Before joining IBM he worked as an Application Programmer and later as a Systems Programmer, where his responsibilities included MVS, OS/390®, z/OS, and storage implementation and support services. His areas of expertise include IBM S/390 Advanced Copy Services, as well as high-end disk and tape solutions. He has co-authored one previous zSeries® Copy Services redbook.

**Juan Antonio Garay** is a Storage Systems Field Technical Sales Specialist in Germany. He has five years of experience in supporting and implementing z/OS and Open Systems storage solutions and providing technical support in IBM. His areas of expertise include the IBM TotalStorage Enterprise Storage Server, when attached to various server platforms, and the design and support of Storage Area Networks. He is currently engaged in providing support for open systems storage across multiple platforms and a wide customer base.

**Torsten Knobloch** has worked for IBM for six years. Currently he is an IT Specialist on the Customer Solutions Team at the Mainz TotalStorage Interoperability Center (TIC) in Germany. There he performs Proof of Concept and System Integration Tests in the Disk Storage area. Before joining the TIC he worked in Disk Manufacturing in Mainz as a Process Engineer.

**Donald (Chuck) Laing** is a Senior Systems Management Integration Professional, specializing in open systems UNIX® disk administration in the IBM South Delivery Center (SDC). He has co-authored four previous IBM Redbooks™ on the IBM TotalStorage Enterprise Storage Server. He holds a degree in Computer Science. Chuck's responsibilities include planning and implementation of midrange storage products. His responsibilities also include department-wide education and cross training on various storage products such as the ESS and FAST. He has worked at IBM for six and a half years. Before joining IBM, Chuck was a hardware CE on UNIX systems for ten years and taught basic UNIX at Midland College for six and a half years in Midland, Texas.

**Christine O'Sullivan** is an IT Storage Specialist in the ATS PSSC storage benchmark center at Montpellier, France. She joined IBM in 1988 and was a System Engineer during her first six years. She has seven years of experience in the pSeries systems and storage. Her areas of expertise and main responsibilities are ESS, storage performance, disaster recovery solutions, AIX® and Oracle databases. She is involved in proof of concept and benchmarks for tuning and optimizing storage environments. She has written several papers about ESS Copy Services and disaster recovery solutions in an Oracle/pSeries environment.

**Stu Preacher** has worked for IBM for over 30 years, starting as a Computer Operator before becoming a Systems Engineer. Much of his time has been spent in the midrange area,

working on System/34, System/38™, AS/400®, and iSeries™. Most recently, he has focused on iSeries Storage, and at the beginning of 2004, he transferred into the IBM TotalStorage division. Over the years, Stu has been a co-author for many Redbooks, including “iSeries in Storage Area Networks” and “Moving Applications to Independent ASPs.” His work in these areas has formed a natural base for working with the new TotalStorage DS6000 and DS8000.

**Torsten Rothenwaldt** is a Storage Architect in Germany. He holds a degree in mathematics from Friedrich Schiller University at Jena, Germany. His areas of interest are high availability solutions and databases, primarily for the Windows® operating systems. Before joining IBM in 1996, he worked in industrial research in electron optics, and as a Software Developer and System Manager in OpenVMS environments.

**Tetsuroh Sano** has worked in AP Advanced Technical Support in Japan for the last five years. His focus areas are open system storage subsystems (especially the IBM TotalStorage Enterprise Storage Server) and SAN hardware. His responsibilities include product introduction, skill transfer, technical support for sales opportunities, solution assurance, and critical situation support.

**Jing Nan Tang** is an Advisory IT Specialist working in ATS for the TotalStorage team of IBM China. He has nine years of experience in the IT field. His main job responsibility is providing technical support and IBM storage solutions to IBM professionals, Business Partners, and Customers. His areas of expertise include solution design and implementation for IBM TotalStorage Disk products (Enterprise Storage Server, FASTT, Copy Services, Performance Tuning), SAN Volume Controller, and Storage Area Networks across open systems.

**Anthony Vandewerdt** is an Accredited IT Specialist who has worked for IBM Australia for 15 years. He has worked on a wide variety of IBM products and for the last four years has specialized in storage systems problem determination. He has extensive experience on the IBM ESS, SAN, 3494 VTS and wave division multiplexors. He is a founding member of the Australian Storage Central team, responsible for screening and managing all storage-related service calls for Australia/New Zealand.

**Alexander Warmuth** is an IT Specialist who joined IBM in 1993. Since 2001 he has worked in Technical Sales Support for IBM TotalStorage. He holds a degree in Electrical Engineering from the University of Erlangen, Germany. His areas of expertise include Linux® and IBM storage as well as business continuity solutions for Linux and other open system environments.

**Roland Wolf** has been with IBM for 18 years. He started his work in IBM Germany in second level support for VM. After five years he shifted to S/390 hardware support for three years. For the past ten years he has worked as a Systems Engineer in Field Technical Support for Storage, focusing on the disk products. His areas of expertise include mainly high-end disk storage systems with PPRC, FlashCopy, and XRC, but he is also experienced in SAN and midrange storage systems in the Open Storage environment. He holds a Ph.D. in Theoretical Physics and is an IBM Certified IT Specialist.



*Front row - Cathy, Torsten R, Torsten K, Andre, Toni, Werner, Tetsuroh. Back row - Roland, Olivier, Anthony, Tang, Christine, Alex, Stu, Heinz, Chuck.*

We want to thank all the members of John Amann's team at the Washington Systems Center in Gaithersburg, MD for hosting us. Craig Gordon and Rosemary McCutchen were especially helpful in getting us access to beta code and hardware.

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

Susan Barrett  
IBM Austin

James Cammarata  
IBM Chicago

Dave Heggen  
IBM Dallas

John Amann, Craig Gordon, Rosemary McCutchen  
IBM Gaithersburg

Hartmut Bohnacker, Michael Eggloff, Matthias Gubitz, Ulrich Rendels, Jens Wissenbach,  
Dietmar Zeller  
IBM Germany

Brian Sherman  
IBM Markham

Ray Koehler  
IBM Minneapolis

John Staubi  
IBM Poughkeepsie

Steve Grillo, Duikaruna Soepangkat, David Vaughn  
IBM Raleigh

Amit Dave, Selwyn Dickey, Chuck Grimm, Nick Harris, Andy Kulich, Joe Prisco, Jim Tuckwell,  
Joe Writz  
IBM Rochester

Charlie Burger, Gene Cullum, Michael Factor, Brian Kraemer, Ling Pong, Jeff Steffan, Pete  
Urbisci, Steve Van Gundy, Diane Williams  
IBM San Jose

Jana Jamsek  
IBM Slovenia

Gerry Cote  
IBM Southfield

Dari Durnas  
IBM Tampa

Linda Benhase, Jerry Boyle, Helen Burton, John Elliott, Kenneth Hallam, Lloyd Johnson, Carl Jones, Arik Kol, Rob Kubo, Lee La Frese, Charles Lynn, Dave Mora, Bonnie Pulver, Nicki Rich, Rick Ripberger, Gail Spear, Jim Springer, Teresa Swingler, Tony Vecchiarelli, John Walkovich, Steve West, Glenn Wightwick, Allen Wright, Bryan Wright  
IBM Tucson

Nick Clayton  
IBM United Kingdom

Steve Chase  
IBM Waltham

Rob Jackard  
IBM Wayne

Many thanks to the graphics editor, Emma Jacobs, and the editor, Alison Chandler.

## Become a published author

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

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll 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](http://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 redbook form found at:

[ibm.com/redbooks](http://ibm.com/redbooks)

- ▶ Send your comments in an email to:

[redbook@us.ibm.com](mailto:redbook@us.ibm.com)

- ▶ Mail your comments to:

IBM Corporation, International Technical Support Organization  
Dept. QXXE Building 80-E2  
650 Harry Road  
San Jose, California 95120-6099





# Part 1

# Introduction

In this part we introduce the IBM TotalStorage DS8000 series and its key features. These include:

- ▶ Product overview
- ▶ Positioning
- ▶ Performance







# Introduction to the DS8000 series

This chapter provides an overview of the features, functions, and benefits of the IBM TotalStorage DS8000 series of storage servers. The topics covered include:

- ▶ The IBM on demand marketing strategy regarding the DS8000
- ▶ Overview of the DS8000 components and features
- ▶ Positioning and benefits of the DS8000
- ▶ The performance features of the DS8000

## 1.1 The DS8000, a member of the TotalStorage DS family

IBM has a wide range of product offerings that are based on open standards and that share a common set of tools, interfaces, and innovative features. The IBM TotalStorage DS family and its new member, the DS8000, gives you the freedom to choose the right combination of solutions for your current needs and the flexibility to help your infrastructure evolve as your needs change. The TotalStorage DS family is designed to offer high availability, multiplatform support, and simplified management tools, all to help you cost effectively adjust to an on demand world.

### 1.1.1 Infrastructure Simplification

The DS8000 series is designed to break through to a *new dimension* of on demand storage, offering an extraordinary opportunity to consolidate existing heterogeneous storage environments, helping lower costs, improve management efficiency, and free valuable floor space. Incorporating IBM's first implementation of storage system Logical Partitions (LPARs) means that two independent workloads can be run on completely independent and separate virtual DS8000 storage systems, with independent operating environments, all within a single physical DS8000. This unique feature of the DS8000 series, which will be available in the DS8300 Model 9A2, helps deliver opportunities for new levels of efficiency and cost effectiveness.

### 1.1.2 Business Continuity

The DS8000 series is designed for the most demanding, mission-critical environments requiring extremely high availability, performance, and scalability. The DS8000 series is designed to avoid single points of failure and provide outstanding availability. With the additional advantages of IBM FlashCopy, data availability can be enhanced even further; for instance, production workloads can continue execution concurrent with data backups. Metro Mirror and Global Mirror business continuity solutions are designed to provide the advanced functionality and flexibility needed to tailor a business continuity environment for almost any recovery point or recovery time objective. The addition of IBM solution integration packages spanning a variety of heterogeneous operating environments offers even more cost-effective ways to implement business continuity solutions.

### 1.1.3 Information Lifecycle Management

The DS8000 is designed as the solution for data when it is at its most on demand, highest priority phase of the data life cycle. One of the advantages IBM offers is the complete set of disk, tape, and software solutions designed to allow customers to create storage environments that support optimal life cycle management and cost requirements.

## 1.2 Overview of the DS8000 series

The IBM TotalStorage DS8000 is a new high-performance, high-capacity series of disk storage systems. An example is shown in Figure 1-1 on page 5. It offers balanced performance that is up to 6 times higher than the previous IBM TotalStorage Enterprise Storage Server (ESS) Model 800. The capacity scales linearly from 1.1 TB up to 192 TB.

With the implementation of the POWER5 Server Technology in the DS8000 it is possible to create storage system logical partitions (LPARs) that can be used for completely separate production, test, or other unique storage environments.

The DS8000 is a flexible and extendable disk storage subsystem because it is designed to add and adapt to new technologies as they become available.

In the entirely new packaging there are also new management tools, like the DS Storage Manager and the DS Command-Line Interface (CLI), which allow for the management and configuration of the DS8000 series as well as the DS6000 series.

The DS8000 series is designed for 24x7 environments in terms of availability while still providing the industry leading remote mirror and copy functions to ensure business continuity.



*Figure 1-1 DS8000 - Base frame*

The IBM TotalStorage DS8000 highlights include that it:

- ▶ Delivers robust, flexible, and cost-effective disk storage for mission-critical workloads
- ▶ Helps to ensure exceptionally high system availability for continuous operations
- ▶ Scales to 192 TB and facilitates unprecedented asset protection with model-to-model field upgrades
- ▶ Supports storage sharing and consolidation for a wide variety of operating systems and mixed server environments
- ▶ Helps increase storage administration productivity with centralized and simplified management
- ▶ Provides the creation of multiple storage system LPARs, that can be used for completely separate production, test, or other unique storage environments
- ▶ Occupies 20 percent less floor space than the ESS Model 800's base frame, and holds even more capacity
- ▶ Provides the industry's first four year warranty

## 1.2.1 Hardware overview

The hardware has been optimized to provide enhancements in terms of performance, connectivity, and reliability. From an architectural point of view the DS8000 series has not changed much with respect to the fundamental architecture of the previous ESS models and 75% of the operating environment remains the same as for the ESS Model 800. This ensures that the DS8000 can leverage a very stable and well-proven operating environment, offering the optimum in availability.

The DS8000 series features several models in a new, higher-density footprint than the ESS Model 800, providing configuration flexibility. For more information on the different models see Chapter 6, “IBM TotalStorage DS8000 model overview and scalability” on page 103.

In this section we give a short description of the main hardware components.

### **POWER5 processor technology**

The DS8000 series exploits the IBM POWER5 technology, which is the foundation of the storage system LPARs. The DS8100 Model 921 utilizes the 64-bit microprocessors’ dual 2-way processor complexes and the DS8300 Model 922/9A2 uses the 64-bit dual 4-way processor complexes. Within the POWER5 servers the DS8000 series offers up to 256 GB of cache, which is up to 4 times as much as the previous ESS models.

### **Internal fabric**

DS8000 comes with a high bandwidth, fault tolerant internal interconnection, which is also used in the IBM pSeries Server. It is called RIO-2 (Remote I/O) and can operate at speeds up to 1 GHz and offers a 2 GB per second sustained bandwidth per link.

### **Switched Fibre Channel Arbitrated Loop (FC-AL)**

The disk interconnection has changed in comparison to the previous ESS. Instead of the SSA loops there is now a switched FC-AL implementation. This offers a point-to-point connection to each drive and adapter, so that there are 4 paths available from the controllers to each disk drive.

### **Fibre Channel disk drives**

The DS8000 offers a selection of industry standard Fibre Channel disk drives. There are 73 GB with 15k revolutions per minute (RPM), 146 GB (10k RPM) and 300 GB (10k RPM) disk drive modules (DDMs) available. The 300 GB DDMs allow a single system to scale up to 192 TB of capacity.

### **Host adapters**

The DS8000 offers enhanced connectivity with the availability of four-port Fibre Channel/FICON® host adapters. The 2 Gb/sec Fibre Channel/FICON host adapters, which are offered in longwave and shortwave, can also auto-negotiate to 1 Gb/sec link speeds. This flexibility enables immediate exploitation of the benefits offered by the higher performance, 2 Gb/sec SAN-based solutions, while also maintaining compatibility with existing 1 Gb/sec infrastructures. In addition, the four-ports on the adapter can be configured with an intermix of Fibre Channel Protocol (FCP) and FICON. This can help protect your investment in fibre adapters, and increase your ability to migrate to new servers. The DS8000 also offers two-port ESCON® adapters. A DS8000 can support up to a maximum of 32 host adapters, which provide up to 128 Fibre Channel/FICON ports.

## Storage Hardware Management Console (S-HMC) for the DS8000

The DS8000 offers a new integrated management console. This console is the service and configuration portal for up to eight DS8000s in the future. Initially there will be one management console for one DS8000 storage subsystem. The S-HMC is the focal point for configuration and Copy Services management, which can be done by the integrated keyboard display or remotely via a Web browser.

For more information on all of the internal components see Chapter 2, “Components” on page 19.

### 1.2.2 Storage capacity

The physical capacity for the DS8000 is purchased via disk drive sets. A disk drive set contains sixteen identical disk drives, which have the same capacity and the same revolution per minute (RPM). Disk drive sets are available in:

- ▶ 73 GB (15,000 RPM)
- ▶ 146 GB (10,000 RPM)
- ▶ 300 GB (10,000 RPM)

For additional flexibility, feature conversions are available to exchange existing disk drive sets when purchasing new disk drive sets with higher capacity, or higher speed disk drives.

In the first frame, there is space for a maximum of 128 disk drive modules (DDMs) and every expansion frame can contain 256 DDMs. Thus there is, at the moment, a maximum limit of 640 DDMs, which in combination with the 300 GB drives gives a maximum capacity of 192 TB.

The DS8000 can be configured as RAID-5, RAID-10, or a combination of both. As a price/performance leader, RAID-5 offers excellent performance for many customer applications, while RAID-10 can offer better performance for selected applications.

Price, performance, and capacity can further be optimized to help meet specific application and business requirements through the intermix of 73 GB (15K RPM), 146 GB (10K RPM) or 300 GB (10K RPM) drives.

**Note:** Initially the intermixing of DDMs in one frame is not supported. At the present time it is only possible to have an intermix of DDMs between two frames, but this limitation will be removed in the future.

### IBM Standby Capacity on Demand offering for the DS8000

Standby Capacity on Demand (Standby CoD) provides *standby* on-demand storage for the DS8000 and allows you to access the extra storage capacity whenever the need arises. With Standby CoD, IBM installs up to 64 drives (in increments of 16) in your DS8000. At any time, you can logically configure your Standby CoD capacity for use. It is a non-disruptive activity that does not require intervention from IBM. Upon logical configuration, you will be charged for the capacity.

For more information about capacity planning see 9.4, “Capacity planning” on page 174.

### 1.2.3 Storage system logical partitions (LPARs)

The DS8000 series provides *storage system* LPARs as a first in the industry. This means that you can run two completely segregated, independent, virtual storage images with differing

workloads, and with different operating environments, within a single physical DS8000 storage subsystem. The LPAR functionality is available in the DS8300 Model 9A2.

The first application of the pSeries Virtualization Engine technology in the DS8000 will partition the subsystem into two virtual storage system images. The processors, memory, adapters, and disk drives are split between the images. There is a robust isolation between the two images via hardware and the POWER5 Hypervisor™ firmware.

Initially each storage system LPAR has access to:

- ▶ 50 percent of the processors
- ▶ 50 percent of the processor memory
- ▶ Up to 16 host adapters
- ▶ Up to 320 disk drives (up to 96 TB of capacity)

With these separate resources, each storage system LPAR can run the same or different versions of microcode, and can be used for completely separate production, test, or other unique storage environments within this single physical system. This may enable storage consolidations, where separate storage subsystems were previously required, helping to increase management efficiency and cost effectiveness.

A detailed description of the LPAR implementation in the DS8000 series is in Chapter 3, “Storage system LPARs (Logical partitions)” on page 43.

## 1.2.4 Supported environments

The DS8000 series offers connectivity support across a broad range of server environments, including IBM eServer zSeries, pSeries, eServer p5, iSeries, eServer i5, and xSeries® servers, servers from Sun and Hewlett-Packard, and non-IBM Intel®-based servers. The operating system support for the DS8000 series is almost the same as for the previous ESS Model 800; there are over 90 supported platforms. This rich support of heterogeneous environments and attachments, along with the flexibility to easily partition the DS8000 series storage capacity among the attached environments, can help support storage consolidation requirements and dynamic, changing environments.

## 1.2.5 Resiliency Family for Business Continuity

*Business Continuity* means that business processes and business-critical applications need to be available at all times and so it is very important to have a storage environment that offers resiliency across both planned and unplanned outages.

The DS8000 supports a rich set of Copy Service functions and management tools that can be used to build solutions to help meet business continuance requirements. These include IBM TotalStorage Resiliency Family Point-in-Time Copy and Remote Mirror and Copy solutions that are currently supported by the Enterprise Storage Server.

**Note:** Remote Mirror and Copy was referred to as Peer-to-Peer Remote Copy (PPRC) in earlier documentation for the IBM TotalStorage Enterprise Storage Server.

You can manage Copy Services functions through the DS Command-Line Interface (CLI) called the IBM TotalStorage DS CLI and the Web-based interface called the IBM TotalStorage DS Storage Manager. The DS Storage Manager allows you to set up and manage data copy features from anywhere that network access is available.

## IBM TotalStorage FlashCopy

FlashCopy can help reduce or eliminate planned outages for critical applications. FlashCopy is designed to provide the same point-in-time copy capability for logical volumes on the DS6000 series and the DS8000 series as FlashCopy V2 does for ESS, and allows access to the source data and the copy almost immediately.

FlashCopy supports many advanced capabilities, including:

- ▶ **Data Set FlashCopy**

Data Set FlashCopy allows a FlashCopy of a data set in a zSeries environment.

- ▶ **Multiple Relationship FlashCopy**

Multiple Relationship FlashCopy allows a source volume to have multiple targets simultaneously.

- ▶ **Incremental FlashCopy**

Incremental FlashCopy provides the capability to update a FlashCopy target without having to recopy the entire volume.

- ▶ **FlashCopy to a Remote Mirror primary**

FlashCopy to a Remote Mirror primary gives you the possibility to use a FlashCopy target volume also as a remote mirror primary volume. This process allows you to create a point-in-time copy and then make a copy of that data at a remote site.

- ▶ **Consistency Group commands**

Consistency Group commands allow DS8000 series systems to hold off I/O activity to a LUN or volume until the FlashCopy Consistency Group command is issued. Consistency groups can be used to help create a consistent point-in-time copy across multiple LUNs or volumes, and even across multiple DS8000s.

- ▶ **Inband Commands over Remote Mirror link**

In a remote mirror environment, commands to manage FlashCopy at the remote site can be issued from the local or intermediate site and transmitted over the remote mirror Fibre Channel links. This eliminates the need for a network connection to the remote site solely for the management of FlashCopy.

## IBM TotalStorage Metro Mirror (Synchronous PPRC)

Metro Mirror is a remote data mirroring technique for all supported servers, including z/OS and open systems. It is designed to constantly maintain an up-to-date copy of the local application data at a remote site which is within the metropolitan area (typically up to 300 km away using DWDM). With synchronous mirroring techniques, data currency is maintained between sites, though the distance can have some impact on performance. Metro Mirror is used primarily as part of a business continuance solution for protecting data against disk storage system loss or complete site failure.

## IBM TotalStorage Global Copy (PPRC Extended Distance, PPRC-XD)

Global Copy is an asynchronous remote copy function for z/OS and open systems for longer distances than are possible with Metro Mirror. With Global Copy, write operations complete on the primary storage system before they are received by the secondary storage system. This capability is designed to prevent the primary system's performance from being affected by wait time from writes on the secondary system. Therefore, the primary and secondary copies can be separated by any distance. This function is appropriate for remote data migration, off-site backups and transmission of inactive database logs at virtually unlimited distances.

### **IBM TotalStorage Global Mirror (Asynchronous PPRC)**

Global Mirror copying provides a two-site extended distance remote mirroring function for z/OS and open systems servers. With Global Mirror, the data that the host writes to the storage unit at the local site is asynchronously shadowed to the storage unit at the remote site. A consistent copy of the data is then automatically maintained on the storage unit at the remote site. This two-site data mirroring function is designed to provide a high performance, cost effective, global distance data replication and disaster recovery solution.

### **IBM TotalStorage z/OS Global Mirror (Extended Remote Copy XRC)**

z/OS Global Mirror is a remote data mirroring function available for the z/OS and OS/390 operating systems. It maintains a copy of the data asynchronously at a remote location over unlimited distances. z/OS Global Mirror is well suited for large zSeries server workloads and can be used for business continuance solutions, workload movement, and data migration.

### **IBM TotalStorage z/OS Metro/Global Mirror**

This mirroring capability uses z/OS Global Mirror to mirror primary site data to a location that is a long distance away and also uses Metro Mirror to mirror primary site data to a location within the metropolitan area. This enables a z/OS three-site high availability and disaster recovery solution for even greater protection from unplanned outages.

### **Three-site solution**

A combination of Global Mirror and Global Copy, called Metro/Global Copy is available on the ESS 750 and ESS 800. It is a three site approach that was previously called Asynchronous Cascading PPRC. You first copy your data synchronously to an intermediate site and from there you go asynchronously to a more distant site.

**Note:** Metro/Global Copy is not available on the DS8000. According to the announcement letter IBM has issued a Statement of General Direction:

*IBM intends to offer a long-distance business continuance solution across three sites allowing for recovery from the secondary or tertiary site with full data consistency.*

For more information about Copy Services see Chapter 7, “Copy Services” on page 115.

## **1.2.6 Interoperability**

As we mentioned before, the DS8000 supports a broad range of server environments. But there is another big advantage regarding interoperability. The DS8000 Remote Mirror and Copy functions can interoperate between the DS8000, the DS6000, and ESS Models 750/800/800Turbo. This offers a dramatically increased flexibility in developing mirroring and remote copy solutions, and also the opportunity to deploy business continuity solutions at lower costs than have been previously available.

## **1.2.7 Service and setup**

The installation of the DS8000 will be performed by IBM in accordance to the installation procedure for this machine. The customer’s responsibility is the installation planning, the retrieval and installation of feature activation codes, and the logical configuration planning and application. This hasn’t changed in regard to the previous ESS model.

For maintenance and service operations, the Storage Hardware Management Console (S-HMC) is the focal point. The management console is a dedicated workstation that is



physically located (installed) inside the DS8000 subsystem and can automatically monitor the state of your system, notifying you and IBM when service is required.

The S-HMC is also the interface for remote services (call home and call back). Remote connections can be configured to meet customer requirements. It is possible to allow one or more of the following: call on error (machine detected), connection for a few days (customer initiated), and remote error investigation (service initiated). The remote connection between the management console and the IBM service organization will be done via a virtual private network (VPN) point-to-point connection over the internet or modem.

The DS8000 comes with a four year warranty on both hardware and software. This is outstanding in the industry and shows IBM's confidence in this product. Once again, this makes the DS8000 a product with a low total cost of ownership (TCO).

## 1.3 Positioning

The IBM TotalStorage DS8000 is designed to provide exceptional performance, scalability, and flexibility while supporting 24 x 7 operations to help provide the access and protection demanded by today's business environments. It also delivers the flexibility and centralized management needed to lower long-term costs. It is part of a complete set of disk storage products that are all part of the IBM TotalStorage DS Family and is the IBM disk product of choice for environments that require the utmost in reliability, scalability, and performance for mission-critical workloads.

### 1.3.1 Common set of functions

The DS8000 series supports many useful features and functions which are not limited to the DS8000 series. There is a set of common functions that can be used on the DS6000 series as well as the DS8000 series. Thus there is only one set of skills necessary to manage both families. This helps to reduce the management costs and the total cost of ownership.

The common functions for storage management include the IBM TotalStorage DS Storage Manager, which is the Web-based graphical user interface, the IBM TotalStorage DS Command-Line Interface (CLI), and the IBM TotalStorage DS open application programming interface (API).

FlashCopy, Metro Mirror, Global Copy, and Global Mirror are the common functions regarding the Advanced Copy Services. In addition to this, the DS6000/DS8000 series mirroring solutions are also compatible between IBM TotalStorage ESS 800 and ESS 750, which offers a new era in flexibility and cost effectiveness in designing business continuity solutions.

#### **DS8000 compared to ESS**

The DS8000 is the next generation of the Enterprise Storage Server, so all functions which are available in the ESS are also available in the DS8000 (with the exception of Metro/Global Copy). From a consolidation point of view, it is now possible to replace four ESS Model 800s with one DS8300. And with the LPAR implementation you get an additional consolidation opportunity because you get two storage system logical partitions in one physical machine.

Since the mirror solutions are compatible between the ESS and the DS8000 series, it is possible to think about a setup for a disaster recovery solution with the high performance DS8000 at the primary site and the ESS at the secondary site, where the same performance is not required.

## DS8000 compared to DS6000

DS6000 and DS8000 now offer an *enterprise continuum* of storage solutions. All copy functions (with the exception of Global Mirror for z/OS Global Mirror, which is only available on the DS8000) are available on both systems. You can do Metro Mirror, Global Mirror, and Global Copy between the two series. The CLI commands and the GUI look the same for both systems.

Obviously the DS8000 can deliver a higher throughput and scales higher than the DS6000, but not all customers need this high throughput and capacity. You can choose the system that fits your needs. Both systems support the same SAN infrastructure and the same host systems.

So it is very easy to have a mixed environment with DS8000 and DS6000 systems to optimize the cost effectiveness of your storage solution, while providing the cost efficiencies of common skills and management functions.

Logical partitioning with some DS8000 models is not available on the DS6000. For more information about the DS6000 refer to *The IBM TotalStorage DS6000 Series: Concepts and Architecture*, SG24-6471.

### 1.3.2 Common management functions

The DS8000 series offers new management tools and interfaces which are also applicable to the DS6000 series.

#### IBM TotalStorage DS Storage Manager

The DS Storage Manager is a Web-based graphical user interface (GUI) that is used to perform logical configurations and Copy Services management functions. It can be accessed from any location that has network access using a Web browser. You have the following options to use the DS Storage Manager:

- ▶ **Simulated (Offline) configuration**

This application allows the user to create or modify logical configurations when disconnected from the network. After creating the configuration, you can save it and then apply it to a network-attached storage unit at a later time.

- ▶ **Real-time (Online) configuration**

This provides real-time management support for logical configuration and Copy Services features for a network-attached storage unit.

#### IBM TotalStorage DS Command-Line Interface (DS CLI)

The DS CLI is a single CLI that has the ability to perform a full set of commands for logical configuration and Copy Services activities. It is now possible to combine the DS CLI commands into a script. This can enhance your productivity since it eliminates the previous requirement for you to create and save a task using the GUI. The DS CLI can also issue Copy Services commands to an ESS Model 750, ESS Model 800, or DS6000 series system.

The following list highlights a few of the specific types of functions that you can perform with the DS Command-Line Interface:

- ▶ Check and verify your storage unit configuration
- ▶ Check the current Copy Services configuration that is used by the storage unit
- ▶ Create new logical storage and Copy Services configuration settings
- ▶ Modify or delete logical storage and Copy Services configuration settings

The DS CLI is described in detail in Chapter 11, “DS CLI” on page 231.

### **DS Open application programming interface**

The DS Open application programming interface (API) is a non-proprietary storage management client application that supports routine LUN management activities, such as LUN creation, mapping and masking, and the creation or deletion of RAID-5 and RAID-10 volume spaces. The DS Open API also enables Copy Services functions such as FlashCopy and Remote Mirror and Copy.

### **1.3.3 Scalability and configuration flexibility**

With the IBM TotalStorage DS8000 you are getting the opportunity to have a linearly scalable capacity growth up to 192 TB. The architecture is designed to scale with today’s 300 GB disk technology to over 1 PB. However, the theoretical architectural limit, based on addressing capabilities, is an incredible 96 PB.

With the DS8000 series there are various choices of base and expansion models, so it is possible to configure the storage units to meet your particular performance and configuration needs. The DS8100 (Model 921) features a dual two-way processor complex and support for one expansion frame. The DS8300 (Models 922 and 9A2) features a dual four-way processor complex and support for one or two expansion frames. The Model 9A2 supports two IBM TotalStorage System LPARs (Logical Partitions) in one physical DS8000.

The DS8100 offers up to 128 GB of processor memory and the DS8300 offers up to 256 GB of processor memory. In addition, the Non-Volatile Storage (NVS) scales to the processor memory size selected, which can also help optimize performance.

Another important feature regarding flexibility is the LUN/Volume Virtualization. It is now possible to create and delete a LUN or volume without affecting other LUNs on the RAID rank. When you delete a LUN or a volume, the capacity can be reused, for example, to form a LUN of a different size. The possibility to allocate LUNs or volumes by spanning RAID ranks allows you to create LUNs or volumes to a maximum size of 2 TB.

The access to LUNs by the host systems is controlled via volume groups. Hosts or disks in the same volume group share access to data. This is the new form of LUN masking.

The DS8000 series allows:

- ▶ Up to 255 logical subsystems (LSS); with two storage system LPARs, up to 510 LSSs
- ▶ Up to 65280 logical devices; with two storage system LPARs, up to 130560 logical devices

### **1.3.4 Future directions of storage system LPARs**

IBM's plans for the future include offering even more flexibility in the use of storage system LPARs. Current plans call for offering a more granular I/O allocation. Also, the processor resource allocation between LPARs is expected to move from 50/50 to possibilities like 25/75, 0/100, 10/90 or 20/80. Not only will the processor resources be more flexible, but in the future, plans call for the movement of memory more dynamically between the storage system LPARs.

These are all features that can react to changing workload and performance requirements, showing the enormous flexibility of the DS8000 series.

Another idea designed to maximize the value of using the storage system LPARs is to have *application* LPARs. IBM is currently evaluating which kind of potential storage applications

offer the most value to the customers. On the list of possible applications are, for example, Backup/Recovery applications (TSM, Legato, Veritas, and so on).

## 1.4 Performance

The IBM TotalStorage DS8000 offers optimally balanced performance, which is up to six times the throughput of the Enterprise Storage Server Model 800. This is possible because the DS8000 incorporates many performance enhancements, like the dual-clustered POWER5 servers, new four-port 2 GB Fibre Channel/FICON host adapters, new Fibre Channel disk drives, and the high-bandwidth, fault-tolerant internal interconnections.

With all these new components, the DS8000 is positioned at the top of the high performance category.

### 1.4.1 Sequential Prefetching in Adaptive Replacement Cache (SARC)

Another performance enhancer is the new self-learning cache algorithm. The DS8000 series caching technology improves cache efficiency and enhances cache hit ratios. The patent-pending algorithm used in the DS8000 series and the DS6000 series is called Sequential Prefetching in Adaptive Replacement Cache (SARC).

SARC provides the following:

- ▶ Sophisticated, patented algorithms to determine what data should be stored in cache based upon the recent access and frequency needs of the hosts
- ▶ Pre-fetching, which anticipates data prior to a host request and loads it into cache
- ▶ Self-Learning algorithms to adaptively and dynamically learn what data should be stored in cache based upon the frequency needs of the hosts

### 1.4.2 IBM TotalStorage Multipath Subsystem Device Driver (SDD)

SDD is a pseudo device driver on the host system designed to support the multipath configuration environments in IBM products. It provides load balancing and enhanced data availability capability. By distributing the I/O workload over multiple active paths, SDD provides dynamic load balancing and eliminates data-flow bottlenecks. SDD also helps eliminate a potential single point of failure by automatically re-routing I/O operations when a path failure occurs.

SDD is provided with the DS8000 series at no additional charge. Fibre Channel (SCSI-FCP) attachment configurations are supported in the AIX, HP-UX, Linux, Microsoft® Windows, Novell NetWare, and Sun Solaris environments.

### 1.4.3 Performance for zSeries

The DS8000 series supports the following IBM performance innovations for zSeries environments:

- ▶ **FICON** extends the ability of the DS8000 series system to deliver high bandwidth potential to the logical volumes needing it, when they need it. Older technologies are limited by the bandwidth of a single disk drive or a single ESCON channel, but FICON, working together with other DS8000 series functions, provides a high-speed pipe supporting a multiplexed operation.
- ▶ **Parallel Access Volumes (PAV)** enable a single zSeries server to simultaneously process multiple I/O operations to the same logical volume, which can help to significantly

reduce device queue delays. This is achieved by defining multiple addresses per volume. With Dynamic PAV, the assignment of addresses to volumes can be automatically managed to help the workload meet its performance objectives and reduce overall queuing. PAV is an optional feature on the DS8000 series.

- ▶ **Multiple Allegiance** expands the simultaneous logical volume access capability across multiple zSeries servers. This function, along with PAV, enables the DS8000 series to process more I/Os in parallel, helping to improve performance and enabling greater use of large volumes.
- ▶ **I/O priority queuing** allows the DS8000 series to use I/O priority information provided by the z/OS Workload Manager to manage the processing sequence of I/O operations.

Chapter 12, “Performance considerations” on page 253, gives you more information about the performance aspects of the DS8000 family.

## 1.5 Summary

In this chapter we gave you a short overview of the benefits and features of the new DS8000 series and showed you why the DS8000 series offers:

- ▶ Balanced performance, which is up to six times that of the ESS Model 800
- ▶ Linear scalability up to 192 TB (designed for 1 PB)
- ▶ Integrated solution capability with storage system LPARs
- ▶ Flexibility due to dramatic addressing enhancements
- ▶ Extensibility, because the DS8000 is designed to add/adapt new technologies
- ▶ All new management tools
- ▶ Availability, since the DS8000 is designed for 24x7 environments
- ▶ Resiliency through industry-leading Remote Mirror and Copy capability
- ▶ Low long term cost, achieved by providing the industry’s first 4 year warranty, and model-to-model upgradeability

More details about these enhancements, and the concepts and architecture of the DS8000 series, are included in the remaining chapters of this redbook.





## Part 2

# Architecture

In this part we describe various aspects of the DS8000 series architecture. These include:

- ▶ Hardware components
- ▶ The LPAR feature
- ▶ RAS - Reliability, Availability, and Serviceability
- ▶ Virtualization concepts
- ▶ Overview of the models
- ▶ Copy Services







# Components

This chapter describes the components used to create the DS8000. This chapter is intended for people who wish to get a clear picture of what the individual components look like and the architecture that holds them together.

In this chapter we introduce:

- ▶ Frames
- ▶ Architecture
- ▶ Processor complexes
- ▶ Disk subsystem
- ▶ Host adapters
- ▶ Power and cooling
- ▶ Management console network

## 2.1 Frames

The DS8000 is designed for modular expansion. From a high-level view there appear to be three types of frames available for the DS8000. However, on closer inspection, the frames themselves are almost identical. The only variations are what combinations of processors, I/O enclosures, batteries, and disks the frames contain.

Figure 2-1 is an attempt to show some of the frame variations that are possible with the DS8000. The left-hand frame is a base frame that contains the processors (eServer p5 570s). The center frame is an expansion frame that contains additional I/O enclosures but no additional processors. The right-hand frame is an expansion frame that contains just disk (and no processors, I/O enclosures, or batteries). Each frame contains a frame power area with power supplies and other power-related hardware.

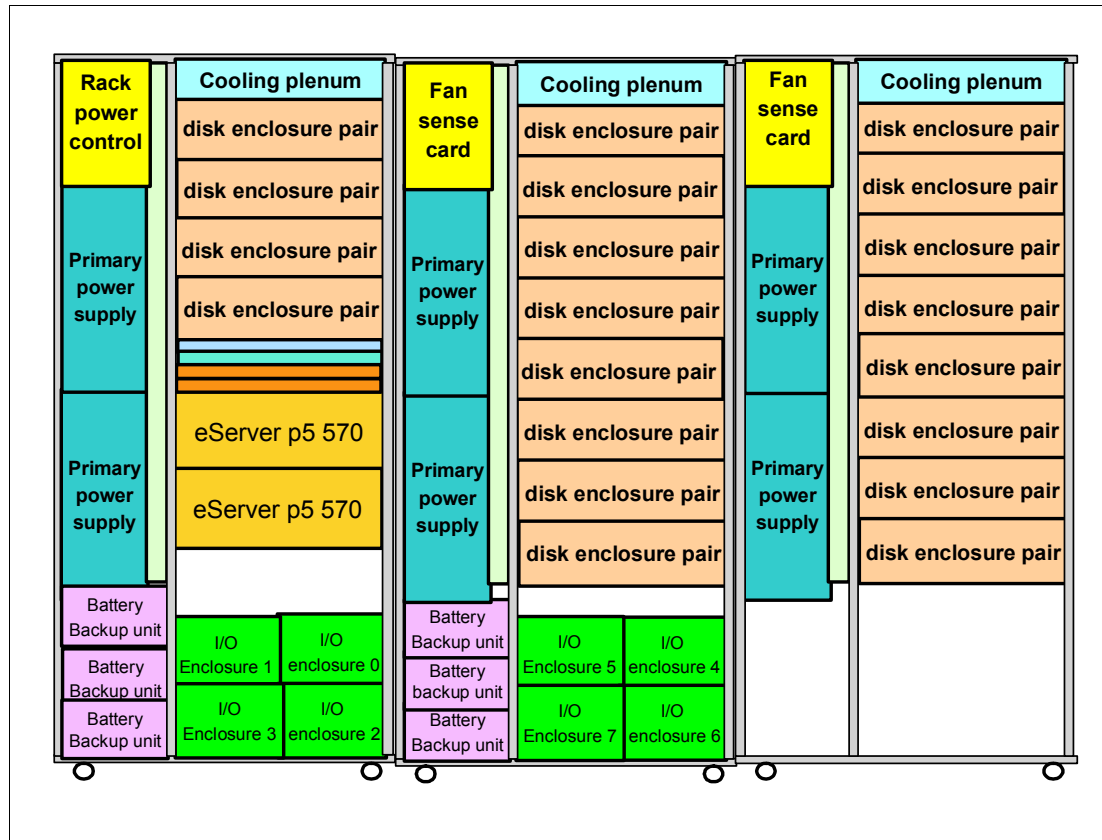


Figure 2-1 DS8000 frame possibilities

### 2.1.1 Base frame

The left-hand side of the base frame (viewed from the front of the machine) is the frame power area. Only the base frame contains rack power control cards (RPC) to control power sequencing for the storage unit. It also contains a fan sense card to monitor the fans in that frame. The base frame contains two primary power supplies (PPSs) to convert input AC into DC power. The power area also contains two or three battery backup units (BBUs) depending on the model and configuration.

The base frame can contain up to eight disk enclosures, each can contain up to 16 disk drives. In a maximum configuration, the base frame can hold 128 disk drives. Above the disk enclosures are cooling fans located in a cooling plenum.

Between the disk enclosures and the processor complexes are two Ethernet switches, a Storage Hardware Management Console (an S-HMC) and a keyboard/display module.

The base frame contains two processor complexes. These eServer p5 570 servers contain the processor and memory that drive all functions within the DS8000. In the ESS we referred to them as *clusters*, but this term is no longer relevant. We now have the ability to logically partition each processor complex into two LPARs, each of which is the equivalent of a Shark cluster.

Finally, the base frame contains four I/O enclosures. These I/O enclosures provide connectivity between the adapters and the processors. The adapters contained in the I/O enclosures can be either device or host adapters (DAs or HAs). The communication path used for adapter to processor complex communication is the RIO-G loop. This loop not only joins the I/O enclosures to the processor complexes, it also allows the processor complexes to communicate with each other.

### 2.1.2 Expansion frame

The left-hand side of each expansion frame (viewed from the front of the machine) is the frame power area. The expansion frames do not contain rack power control cards; these cards are only present in the base frame. They do contain a fan sense card to monitor the fans in that frame. Each expansion frame contains two primary power supplies (PPS) to convert the AC input into DC power. Finally, the power area may contain three battery backup units (BBUs) depending on the model and configuration.

Each expansion frame can hold up to 16 disk enclosures which contain the disk drives. They are described as *16-packs* because each enclosure can hold 16 disks. In a maximum configuration, an expansion frame can hold 256 disk drives. Above the disk enclosures are cooling fans located in a cooling plenum.

An expansion frame can contain I/O enclosures and adapters if it is the first expansion frame that is attached to either a model 922 or a model 9A2. The second expansion frame in a model 922 or 9A2 configuration cannot have I/O enclosures and adapters, nor can any expansion frame that is attached to a model 921. If the expansion frame contains I/O enclosures, the enclosures provide connectivity between the adapters and the processors. The adapters contained in the I/O enclosures can be either device or host adapters.

### 2.1.3 Rack operator panel

Each DS8000 frame features an operator panel. This panel has three indicators and an emergency power off switch (an EPO switch). Figure 2-2 on page 22 depicts the operator panel. Each panel has two line cord indicators (one for each line cord). For normal operation both of these indicators should be on, to indicate that each line cord is supplying correct power to the frame. There is also a fault indicator. If this indicator is illuminated you should use the DS Storage Manager GUI or the Storage Hardware Management Console (S-HMC) to determine why this indicator is on.

There is also an EPO switch on each operator panel. This switch is only for emergencies. Tripping the EPO switch will bypass all power sequencing control and result in immediate removal of system power. A small cover must be lifted to operate it. Do not trip this switch unless the DS8000 is creating a safety hazard or is placing human life at risk.

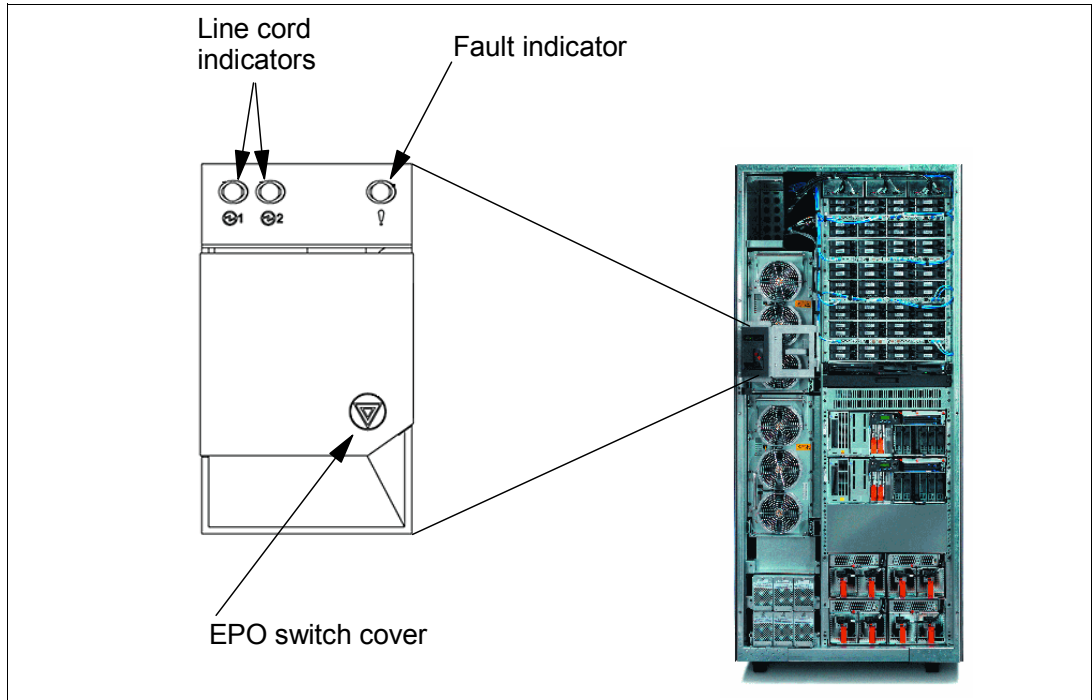


Figure 2-2 Rack operator panel

You will note that there is not a power on/off switch on the operator panel. This is because power sequencing is managed via the S-HMC. This is to ensure that all data in non-volatile storage (known as modified data) is de-staged properly to disk prior to power down. It is thus not possible to shut down or power off the DS8000 from the operator panel (except in an emergency, with the EPO switch mentioned previously).

## 2.2 Architecture

Now that we have described the frames themselves, we use the rest of this chapter to explore the technical details of each of the components. The architecture that connects these components is pictured in Figure 2-3 on page 23.

In effect, the DS8000 consists of two processor complexes. Each processor complex has access to multiple host adapters to connect to channel, FICON, and ESCON hosts. Each DS8000 can potentially have up to 32 host adapters. To access the disk subsystem, each complex uses several four-port Fibre Channel arbitrated loop (FC-AL) device adapters. A DS8000 can potentially have up to sixteen of these adapters arranged into eight pairs. Each adapter connects the complex to two separate switched Fibre Channel networks. Each switched network attaches disk enclosures that each contain up to 16 disks. Each enclosure contains two 20-port Fibre Channel switches. Of these 20 ports, 16 are used to attach to the 16 disks in the enclosure and the remaining four are used to either interconnect with other enclosures or to the device adapters. Each disk is attached to both switches. Whenever the device adapter connects to a disk, it uses a switched connection to transfer data. This means that all data travels via the shortest possible path.

The attached hosts interact with software which is running on the complexes to access data on logical volumes. Each complex will host at least one instance of this software (which is called a *server*), which runs in a logical partition (an LPAR). The servers manage all read and write requests to the logical volumes on the disk arrays. During write requests, the servers

use fast-write, in which the data is written to volatile memory on one complex and persistent memory on the other complex. The server then reports the write as complete before it has been written to disk. This provides much faster write performance. Persistent memory is also called NVS or non-volatile storage.

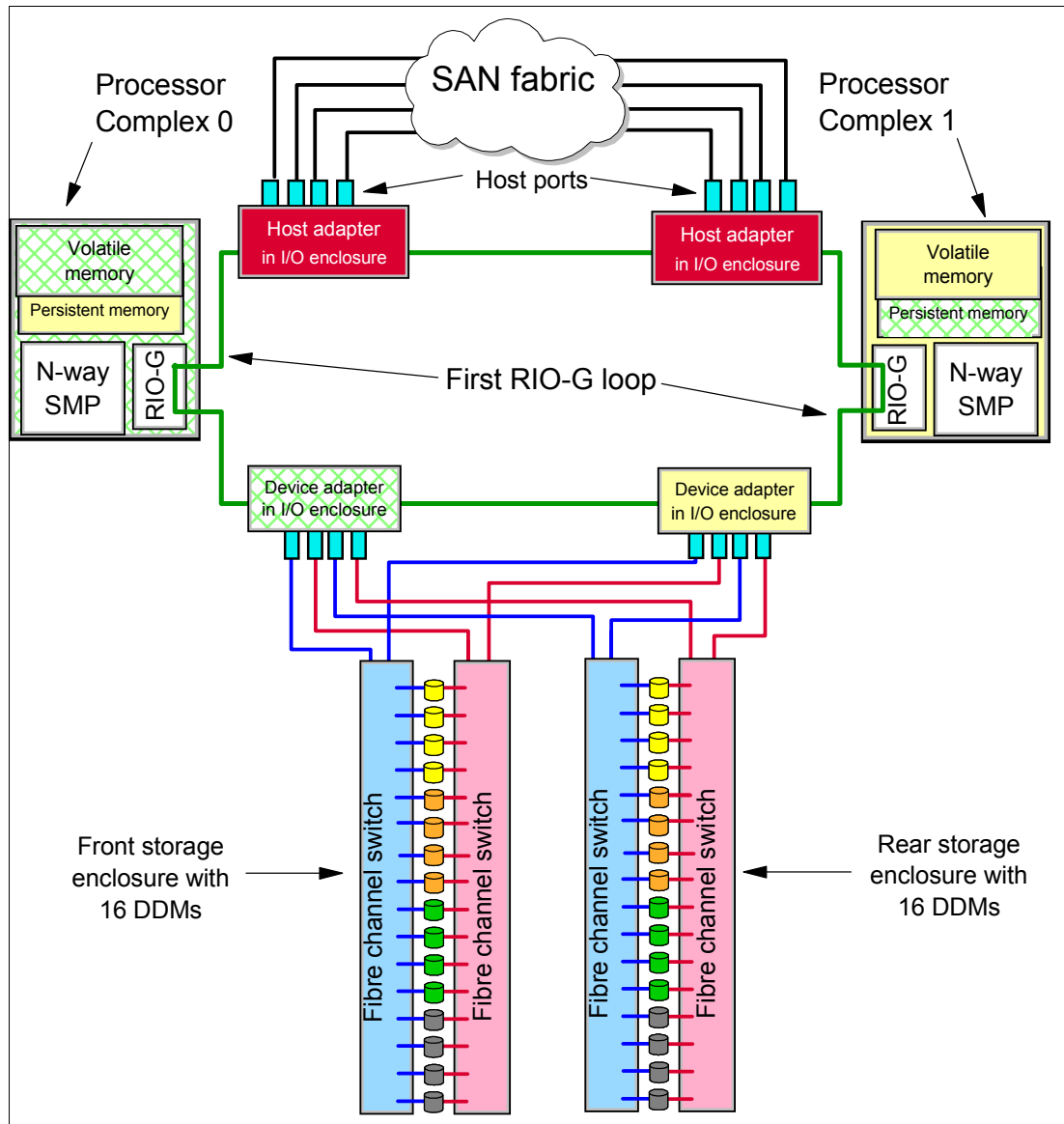


Figure 2-3 DS8000 architecture

When a host performs a read operation, the servers fetch the data from the disk arrays via the high performance switched disk architecture. The data is then cached in volatile memory in case it is required again. The servers attempt to anticipate future reads by an algorithm known as SARC (Sequential prefetching in Adaptive Replacement Cache). Data is held in cache as long as possible using this smart algorithm. If a cache hit occurs where requested data is already in cache, then the host does not have to wait for it to be fetched from the disks.

Both the device and host adapters operate on a high bandwidth fault-tolerant interconnect known as the RIO-G. The RIO-G design allows the sharing of host adapters between servers and offers exceptional performance and reliability.

If you can view Figure 2-3 on page 23 in color, you can use the colors as indicators of how the DS8000 hardware is shared between the servers (the cross hatched color is green and the lighter color is yellow). On the left side, the green server is running on the left-hand processor complex. The green server uses the N-way SMP of the complex to perform its operations. It records its write data and caches its read data in the volatile memory of the left-hand complex. For fast-write data it has a persistent memory area on the right-hand processor complex. To access the disk arrays under its management (the disks also being pictured in green), it has its own device adapter (again in green). The yellow server on the right operates in an identical fashion. The host adapters (in dark red) are deliberately not colored green or yellow because they are shared between both servers.

## 2.2.1 Server-based SMP design

The DS8000 benefits from a fully assembled, leading edge processor and memory system. Using SMPs as the primary processing engine sets the DS8000 apart from other disk storage systems on the market. Additionally, the POWER5 processors used in the DS8000 support the execution of two independent threads concurrently. This capability is referred to as *simultaneous multi-threading (SMT)*. The two threads running on the single processor share a common L1 cache. The SMP/SMT design minimizes the likelihood of idle or overworked processors, while a distributed processor design is more susceptible to an unbalanced relationship of tasks to processors.

The design decision to use SMP memory as I/O cache is a key element of IBM's storage architecture. Although a separate I/O cache could provide fast access, it cannot match the access speed of the SMP main memory. The decision to use the SMP main memory as the cache proved itself in three generations of IBM's Enterprise Storage Server (ESS 2105). The performance roughly doubled with each generation. This performance improvement can be traced to the capabilities of the completely integrated SMP, the processor speeds, the L1/L2 cache sizes and speeds, the memory bandwidth and response time, and the PCI bus performance.

With the DS8000, the cache access has been accelerated further by making the Non-Volatile Storage a part of the SMP memory.

All memory installed on any processor complex is accessible to all processors in that complex. The addresses assigned to the memory are common across all processors in the same complex. On the other hand, using the main memory of the SMP as the cache, leads to a partitioned cache. Each processor has access to the processor complex's main memory but not to that of the other complex. You should keep this in mind with respect to load balancing between processor complexes.

## 2.2.2 Cache management

Most if not all high-end disk systems have internal cache integrated into the system design, and some amount of system cache is required for operation. Over time, cache sizes have dramatically increased, but the ratio of cache size to system disk capacity has remained nearly the same.

The DS6000 and DS8000 use the patent-pending *Sequential Prefetching in Adaptive Replacement Cache (SARC)* algorithm, developed by IBM Storage Development in partnership with IBM Research. It is a self-tuning, self-optimizing solution for a wide range of workloads with a varying mix of sequential and random I/O streams. SARC is inspired by the *Adaptive Replacement Cache (ARC)* algorithm and inherits many features from it. For a detailed description of ARC see N. Megiddo and D. S. Modha, "Outperforming LRU with an adaptive replacement cache algorithm," IEEE Computer, vol. 37, no. 4, pp. 58–65, 2004.

SARC basically attempts to determine four things:

- ▶ When data is copied into the cache.
- ▶ Which data is copied into the cache.
- ▶ Which data is evicted when the cache becomes full.
- ▶ How does the algorithm dynamically adapt to different workloads.

The DS8000 cache is organized in 4K byte pages called cache pages or slots. This unit of allocation (which is smaller than the values used in other storage systems) ensures that small I/Os do not waste cache memory.

The decision to copy some amount of data into the DS8000 cache can be triggered from two policies: demand paging and prefetching. *Demand paging* means that eight disk blocks (a 4K cache page) are brought in only on a cache miss. Demand paging is always active for all volumes and ensures that I/O patterns with some locality find at least some recently used data in the cache.

*Prefetching* means that data is copied into the cache speculatively even before it is requested. To prefetch, a prediction of likely future data accesses is needed. Because effective, sophisticated prediction schemes need extensive history of page accesses (which is not feasible in real-life systems), SARC uses prefetching for sequential workloads. Sequential access patterns naturally arise in video-on-demand, database scans, copy, backup, and recovery. The goal of sequential prefetching is to detect sequential access and effectively pre-load the cache with data so as to minimize cache misses.

For prefetching, the cache management uses tracks. A *track* is a set of 128 disk blocks (16 cache pages). To detect a sequential access pattern, counters are maintained with every track to record if a track has been accessed together with its predecessor. Sequential prefetching becomes active only when these counters suggest a sequential access pattern. In this manner, the DS6000/DS8000 monitors application read-I/O patterns and dynamically determines whether it is optimal to stage into cache:

- ▶ Just the page requested
- ▶ That page requested plus remaining data on the disk track
- ▶ An entire disk track (or a set of disk tracks) which has (have) not yet been requested

The decision of when and what to prefetch is essentially made on a per-application basis (rather than a system-wide basis) to be sensitive to the different data reference patterns of different applications that can be running concurrently.

To decide which pages are evicted when the cache is full, sequential and random (non-sequential) data is separated into different lists (see Figure 2-4 on page 26). A page which has been brought into the cache by simple demand paging is added to the MRU (Most Recently Used) head of the RANDOM list. Without further I/O access, it goes down to the LRU (Least Recently Used) bottom. A page which has been brought into the cache by a sequential access or by sequential prefetching is added to the MRU head of the SEQ list and then goes in that list. Additional rules control the migration of pages between the lists so as to not keep the same pages in memory twice.

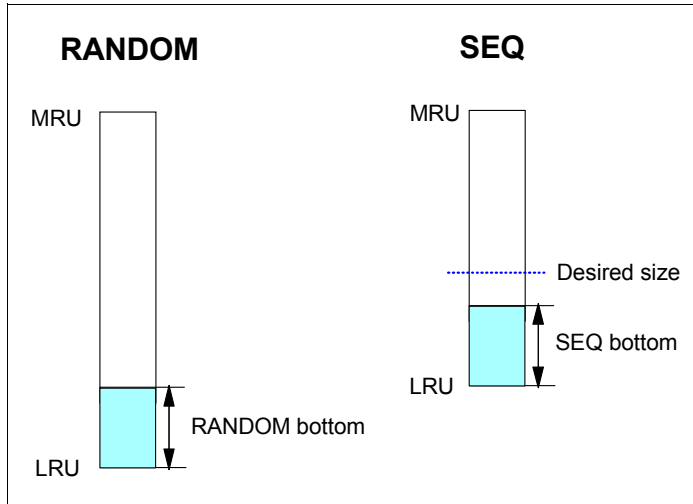


Figure 2-4 Cache lists of the SARC algorithm for random and sequential data

To follow workload changes, the algorithm trades cache space between the RANDOM and SEQ lists dynamically and adaptively. This makes SARC scan-resistant, so that one-time sequential requests do not pollute the whole cache. SARC maintains a desired size parameter for the sequential list. The desired size is continually adapted in response to the workload. Specifically, if the bottom portion of the SEQ list is found to be more valuable than the bottom portion of the RANDOM list, then the desired size is increased; otherwise, the desired size is decreased. The constant adaptation strives to make optimal use of limited cache space and delivers greater throughput and faster response times for a given cache size.

Additionally, the algorithm modifies dynamically not only the sizes of the two lists, but also the rate at which the sizes are adapted. In a steady state, pages are evicted from the cache at the rate of cache misses. A larger (respectively, a smaller) rate of misses effects a faster (respectively, a slower) rate of adaptation.

Other implementation details take into account the relation of read and write (NVS) cache, efficient de-staging, and the cooperation with Copy Services. In this manner, the DS6000 and DS8000 cache management goes far beyond the usual variants of the LRU/LFU (Least Recently Used / Least Frequently Used) approaches.

## 2.3 Processor complex

The DS8000 base frame contains two processor complexes. The Model 921 has 2-way processors while the Model 922 and Model 9A2 have 4-way processors. (2-way means that each processor complex has 2 CPUs, while 4-way means that each processor complex has 4 CPUs.)

The DS8000 features IBM POWER5 server technology. Depending on workload, the maximum host I/O operations per second of the DS8100 Model 921 is up to three times the maximum operations per second of the ESS Model 800. The maximum host I/O operations per second of the DS8300 Model 922 or 9A2 is up to six times the maximum of the ESS Model 800.



For details on the server hardware used in the DS8000, refer to *IBM p5 570 Technical Overview and Introduction*, REDP-9117, available at:

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

The symmetric multiprocessor (SMP) p5 570 system features 2-way or 4-way, copper-based, SOI-based POWER5 microprocessors running at 1.5 GHz or 1.9 GHz with 36 MB off-chip Level 3 cache configurations. The system is based on a concept of system building blocks. The p5 570 processor complexes are facilitated with the use of processor interconnect and system flex cables that enable as many as four 4-way p5 570 processor complexes to be connected to achieve a true 16-way SMP combined system. How these features are implemented in the DS8000 might vary.

One p5 570 processor complex includes:

- ▶ Five hot-plug PCI-X slots with Enhanced Error Handling (EEH)
- ▶ An enhanced blind-swap mechanism that allows hot-swap replacement or installation of PCI-X adapters without sliding the enclosure into the service position
- ▶ Two Ultra320 SCSI controllers
- ▶ One 10/100/1000 Mbps integrated dual-port Ethernet controller
- ▶ Two serial ports
- ▶ Two USB 2.0 ports
- ▶ Two HMC Ethernet ports
- ▶ Four remote RIO-G ports
- ▶ Two System Power Control Network (SPCN) ports

The p5 570 includes two 3-pack front-accessible, hot-swap-capable disk bays. The six disk bays of one IBM Server p5 570 processor complex can accommodate up to 880.8 GB of disk storage using the 146.8 GB Ultra320 SCSI disk drives. Two additional media bays are used to accept optional slim-line media devices, such as DVD-ROM or DVD-RAM drives. The p5 570 also has I/O expansion capability using the RIO-G interconnect. How these features are implemented in the DS8000 might vary.

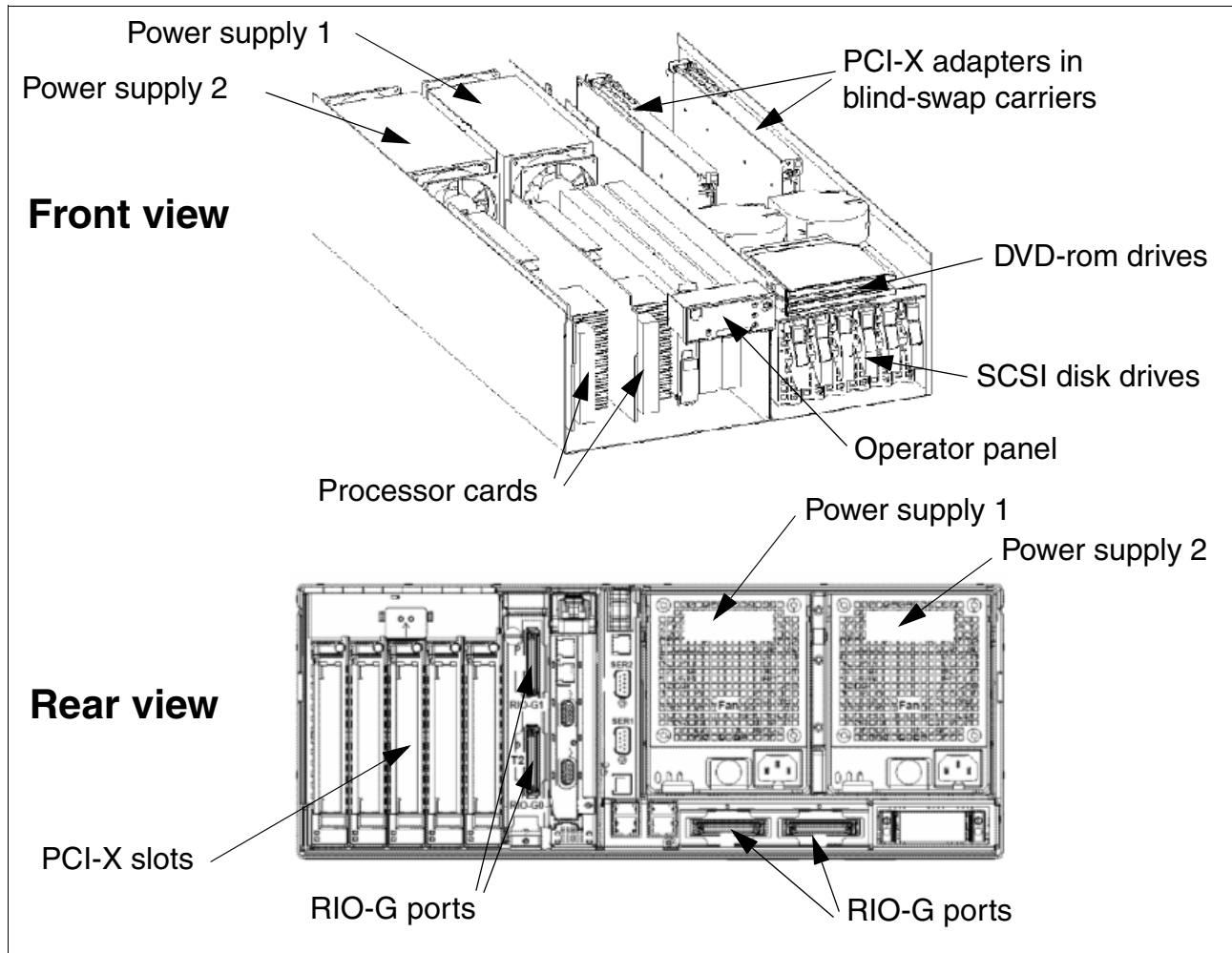


Figure 2-5 Processor complex

### Processor memory

The DS8100 Model 921 offers up to 128 GB of processor memory and the DS8300 Models 922 and 9A2 offer up to 256 GB of processor memory. Half of this will be located in each processor complex. In addition, the Non-Volatile Storage (NVS) scales to the processor memory size selected, which can also help optimize performance.

### Service processor and SPCN

The service processor (SP) is an embedded controller that is based on a PowerPC® 405GP processor (PPC405). The SPCN is the system power control network that is used to control the power of the attached I/O subsystem. The SPCN control software and the service processor software are run on the same PPC405 processor.

The SP performs predictive failure analysis based on any recoverable processor errors. The SP can monitor the operation of the firmware during the boot process, and it can monitor the operating system for loss of control. This enables the service processor to take appropriate action.

The SPCN monitors environmental conditions such as power, fans, and temperature. Environmental critical and non-critical conditions can generate Early Power-Off Warning (EPOW) events. Critical events trigger appropriate signals from the hardware to the affected components to

prevent any data loss without operating system or firmware involvement. Non-critical environmental events are also logged and reported.

### 2.3.1 RIO-G

The RIO-G ports are used for I/O expansion to external I/O drawers. RIO stands for remote I/O. The RIO-G is evolved from earlier versions of the RIO interconnect.

Each RIO-G port can operate at 1 GHz in bidirectional mode and is capable of passing data in each direction on each cycle of the port. It is designed as a high performance self-healing interconnect. The p5 570 provides two external RIO-G ports, and an adapter card adds two more. Two ports on each processor complex form a loop.

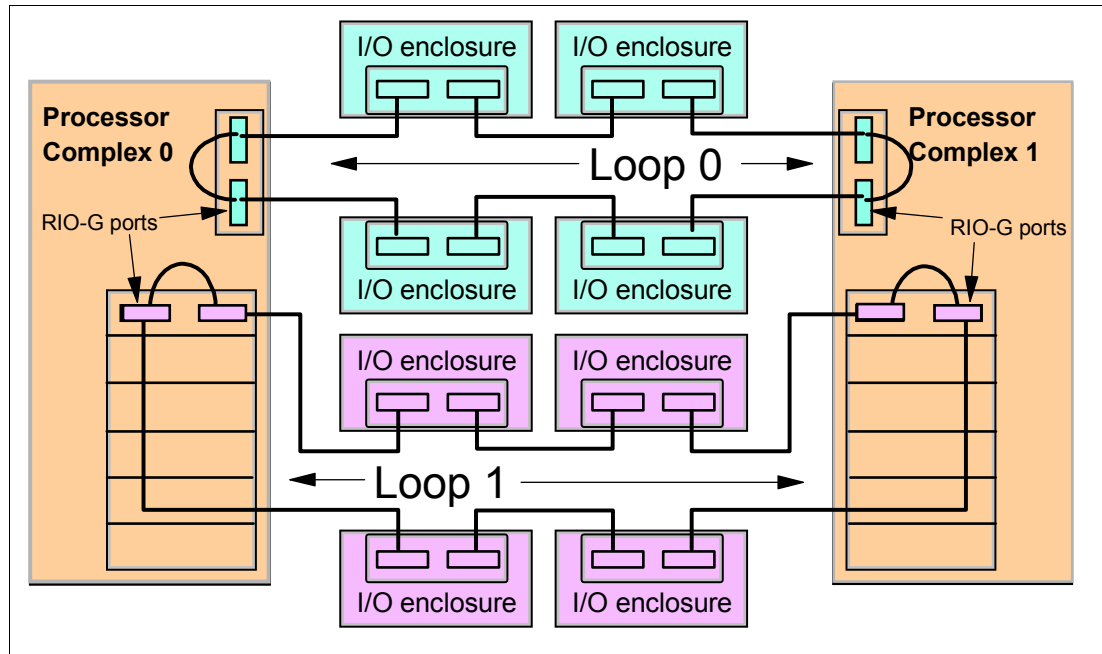


Figure 2-6 DS8000 RIO-G port layout

Figure 2-6 illustrates how the RIO-G cabling is laid out in a DS8000 that has eight I/O drawers. This would only occur if an expansion frame were installed. The DS8000 RIO-G cabling will vary based on the model. A two-way DS8000 model will have one RIO-G loop. A four-way DS8000 model will have two RIO-G loops. Each loop will support four disk enclosures.

### 2.3.2 I/O enclosures

All base models contain I/O enclosures and adapters. The I/O enclosures hold the adapters and provide connectivity between the adapters and the processors. Device adapters and host adapters are installed in the I/O enclosure. Each I/O enclosure has 6 slots. Each slot supports PCI-X adapters running at 64 bit, 133 Mhz. Slots 3 and 6 are used for the device adapters. The remaining slots are available to install up to four host adapters per I/O enclosure.

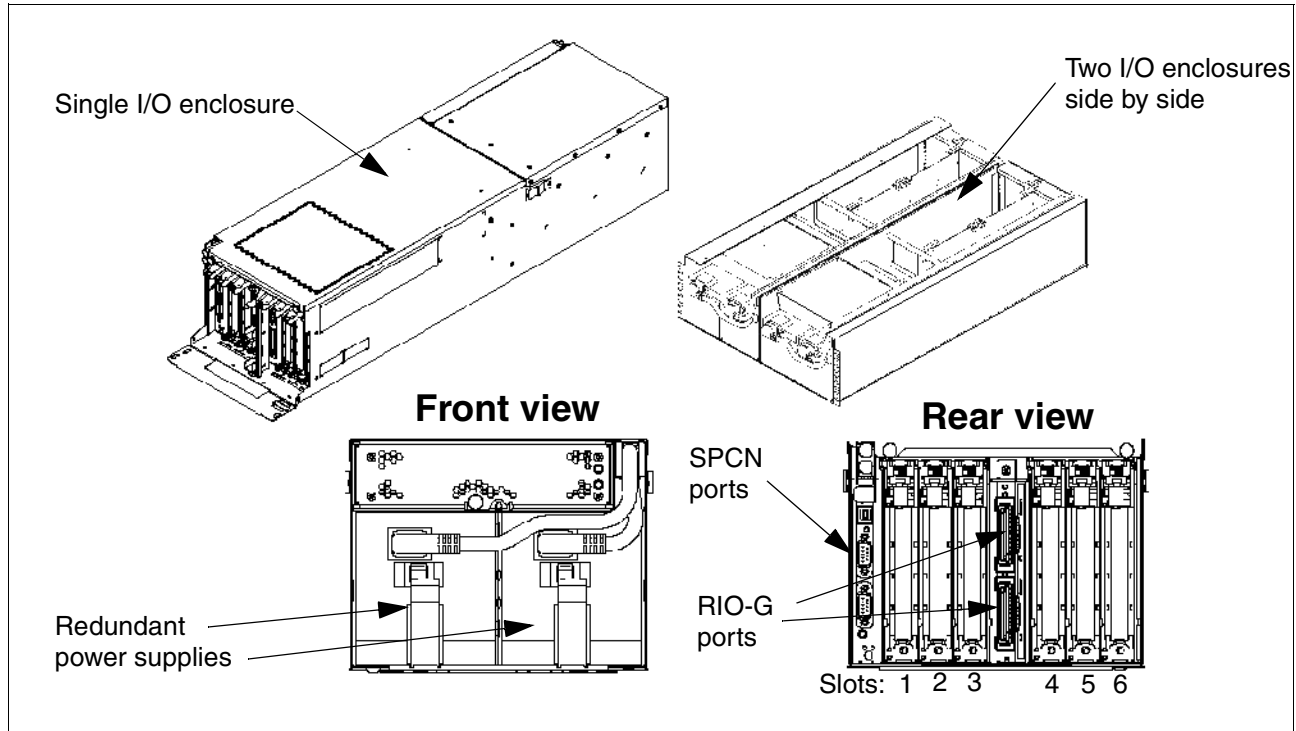


Figure 2-7 I/O enclosures

Each I/O enclosure has the following attributes:

- ▶ 4U rack-mountable enclosure
- ▶ Six PCI-X slots: 3.3 V, keyed, 133 MHz blind-swap hot-plug
- ▶ Default redundant hot-plug power and cooling devices
- ▶ Two RIO-G and two SPCN ports

## 2.4 Disk subsystem

The DS8000 series offers a selection of Fibre Channel disk drives, including 300 GB drives, allowing a DS8100 to scale up to 115.2 TB of capacity and a DS8300 to scale up to 192 TB of capacity. The disk subsystem consists of three components:

- ▶ First, located in the I/O enclosures are the device adapters. These are RAID controllers that are used by the storage images to access the RAID arrays.
- ▶ Second, the device adapters connect to switched controller cards in the disk enclosures. This creates a switched Fibre Channel disk network.
- ▶ Finally, we have the disks themselves. The disks are commonly referred to as disk drive modules (DDMs).

### 2.4.1 Device adapters

Each DS8000 device adapter (DA) card offers four 2Gbps FC-AL ports. These ports are used to connect the processor complexes to the disk enclosures. The adapter is responsible for managing, monitoring, and rebuilding the RAID arrays. The adapter provides remarkable performance thanks to a new high function/high performance ASIC. To ensure maximum data

integrity it supports metadata creation and checking. The device adapter design is shown in Figure 2-8.

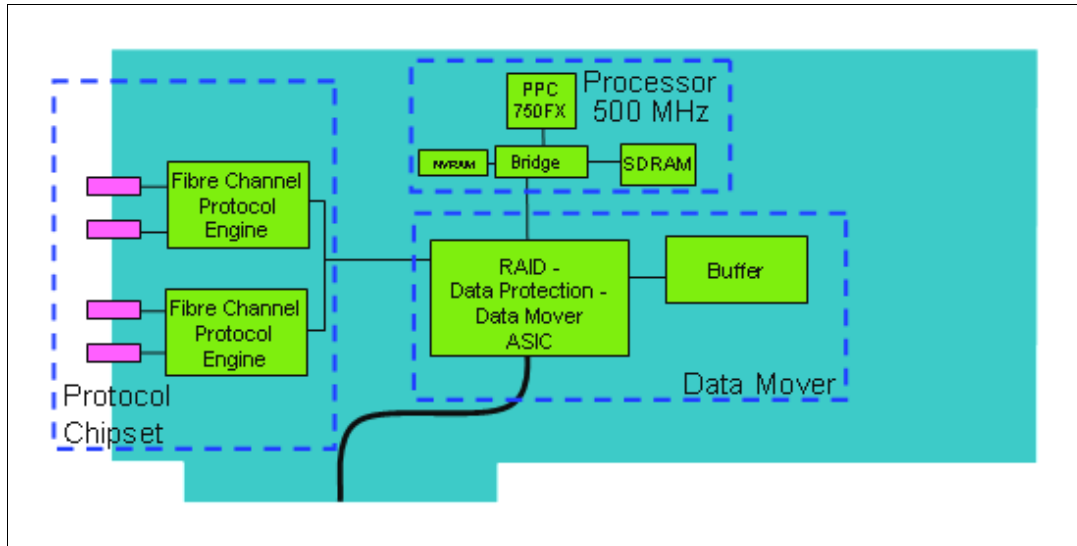


Figure 2-8 DS8000 device adapter

The DAs are installed in pairs because each storage partition requires its own adapter to connect to each disk enclosure for redundancy. This is why we refer to them as DA pairs.

## 2.4.2 Disk enclosures

Each DS8000 frame contains either 8 or 16 disk enclosures depending on whether it is a base or expansion frame. Half of the disk enclosures are accessed from the front of the frame, and half from the rear. Each DS8000 disk enclosure contains a total of 16 DDMs or dummy carriers. A dummy carrier looks very similar to a DDM in appearance but contains no electronics. The enclosure is pictured in Figure 2-9 on page 32.

**Note:** If a DDM is not present, its slot must be occupied by a dummy carrier. This is because without a drive or a dummy, cooling air does not circulate correctly.

Each DDM is an industry standard FC-AL disk. Each disk plugs into the disk enclosure backplane. The backplane is the electronic and physical backbone of the disk enclosure.

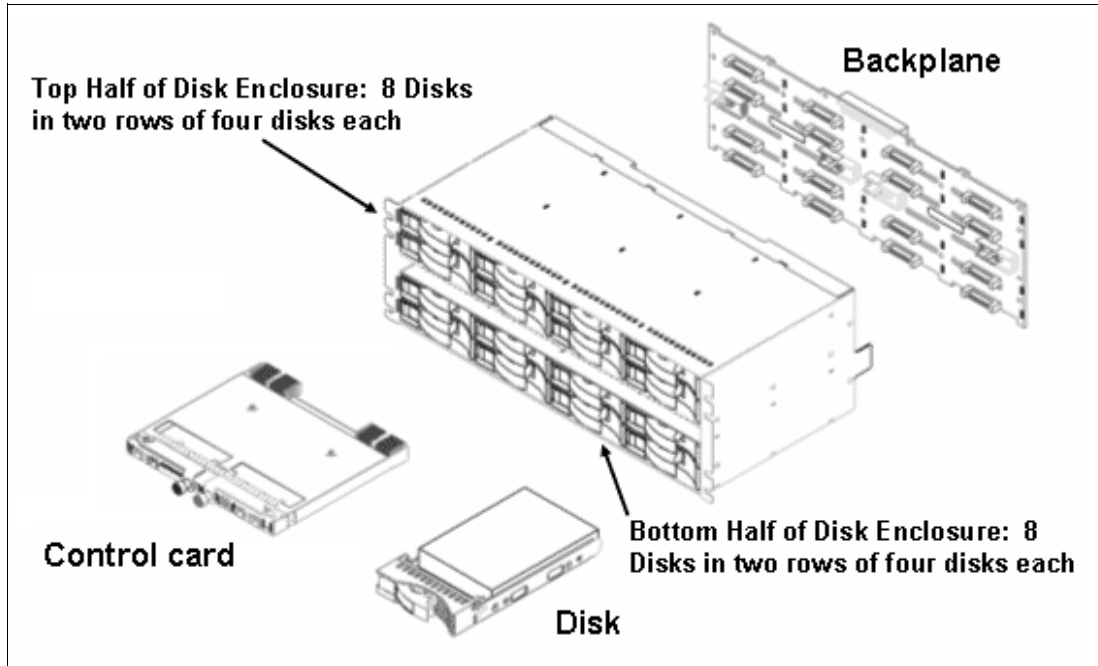


Figure 2-9 DS8000 disk enclosure

### Non-switched FC-AL drawbacks

In a standard FC-AL disk enclosure all of the disks are arranged in a loop, as depicted in Figure 2-10. This loop-based architecture means that data flows through all disks before arriving at either end of the device adapter (shown here as the *Storage Server*).

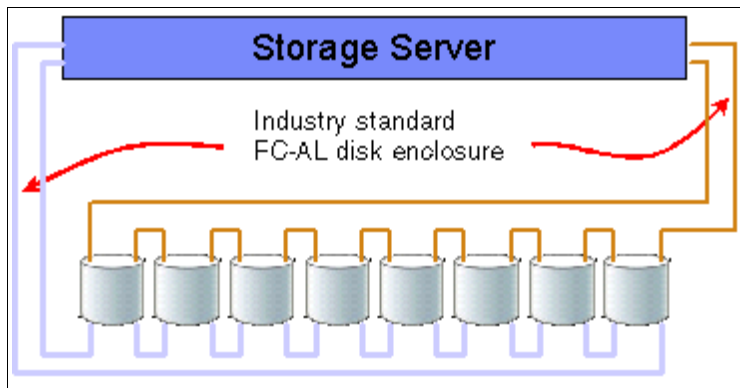


Figure 2-10 Industry standard FC-AL disk enclosure

The main problems with standard FC-AL access to DDMs are:

- ▶ The full loop is required to participate in data transfer. Full discovery of the loop via LIP (loop initialization protocol) is required before any data transfer. Loop stability can be affected by DDM failures.
- ▶ In the event of a disk failure, it can be difficult to identify the cause of a loop breakage, leading to complex problem determination.
- ▶ There is a performance dropoff when the number of devices in the loop increases.
- ▶ To expand the loop it is normally necessary to partially open it. If mistakes are made, a complete loop outage can result.

These problems are solved with the *switched* FC-AL implementation on the DS8000.

### Switched FC-AL advantages

The DS8000 uses switched FC-AL technology to link the device adapter (DA) pairs and the DDMs. Switched FC-AL uses the standard FC-AL protocol, but the physical implementation is different. The key features of switched FC-AL technology are:

- ▶ Standard FC-AL communication protocol from DA to DDMs.
- ▶ Direct point-to-point links are established between DA and DDM.
- ▶ Isolation capabilities in case of DDM failures, providing easy problem determination.
- ▶ Predictive failure statistics.
- ▶ Simplified expansion; for example, no cable re-routing is required when adding another disk enclosure.

The DS8000 architecture employs dual redundant switched FC-AL access to each of the disk enclosures. The key benefits of doing this are:

- ▶ Two independent networks to access the disk enclosures.
- ▶ Four access paths to each DDM.
- ▶ Each device adapter port operates independently.
- ▶ Double the bandwidth over traditional FC-AL loop implementations.

In Figure 2-11 each DDM is depicted as being attached to two separate Fibre Channel switches. This means that with two device adapters, we have four effective data paths to each disk, each path operating at 2Gb/sec. Note that this diagram shows one switched disk network attached to each DA. Each DA can actually support two switched networks.

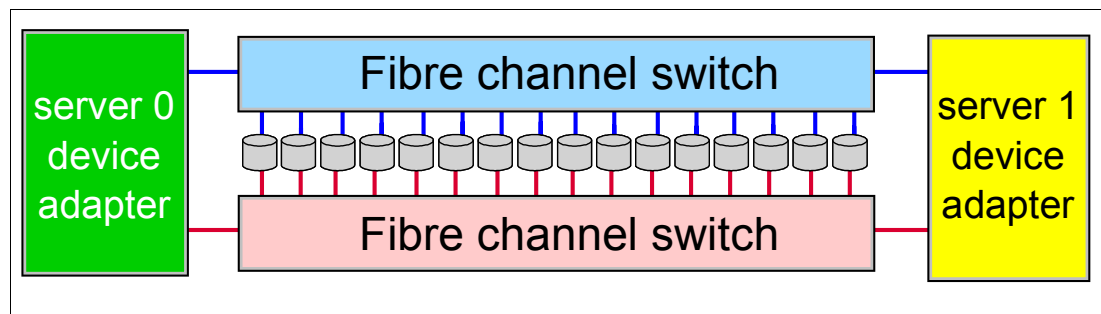


Figure 2-11 DS8000 disk enclosure

When a connection is made between the device adapter and a disk, the connection is a switched connection that uses arbitrated loop protocol. This means that a mini-loop is created between the device adapter and the disk. Figure 2-12 on page 34 depicts four simultaneous and independent connections, one from each device adapter port.

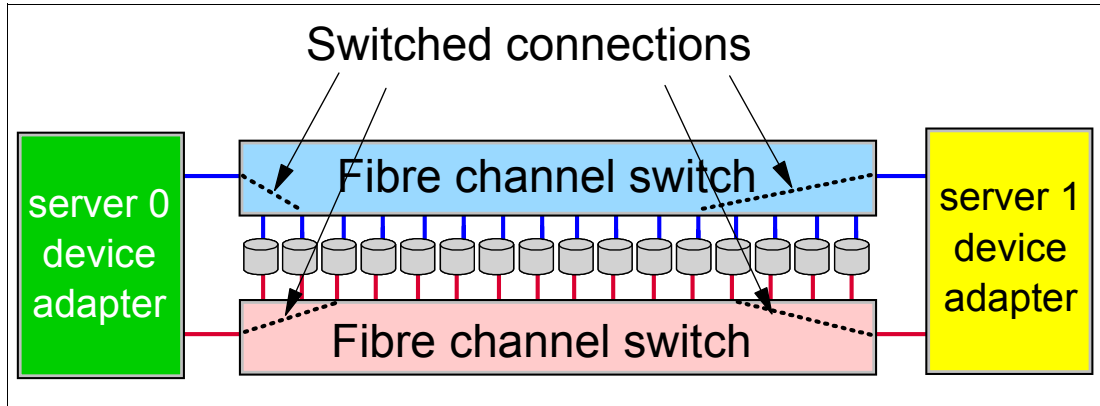


Figure 2-12 Disk enclosure switched connections

### DS8000 switched FC-AL implementation

For a more detailed look at how the switched disk architecture expands in the DS8000 you should refer to Figure 2-13 on page 35. It depicts how each DS8000 device adapter connects to two disk networks called loops. Expansion is achieved by adding enclosures to the expansion ports of each switch. Each loop can potentially have up to six enclosures, but this will vary depending on machine model and DA pair number. The front enclosures are those that are physically located at the front of the machine. The rear enclosures are located at the rear of the machine.



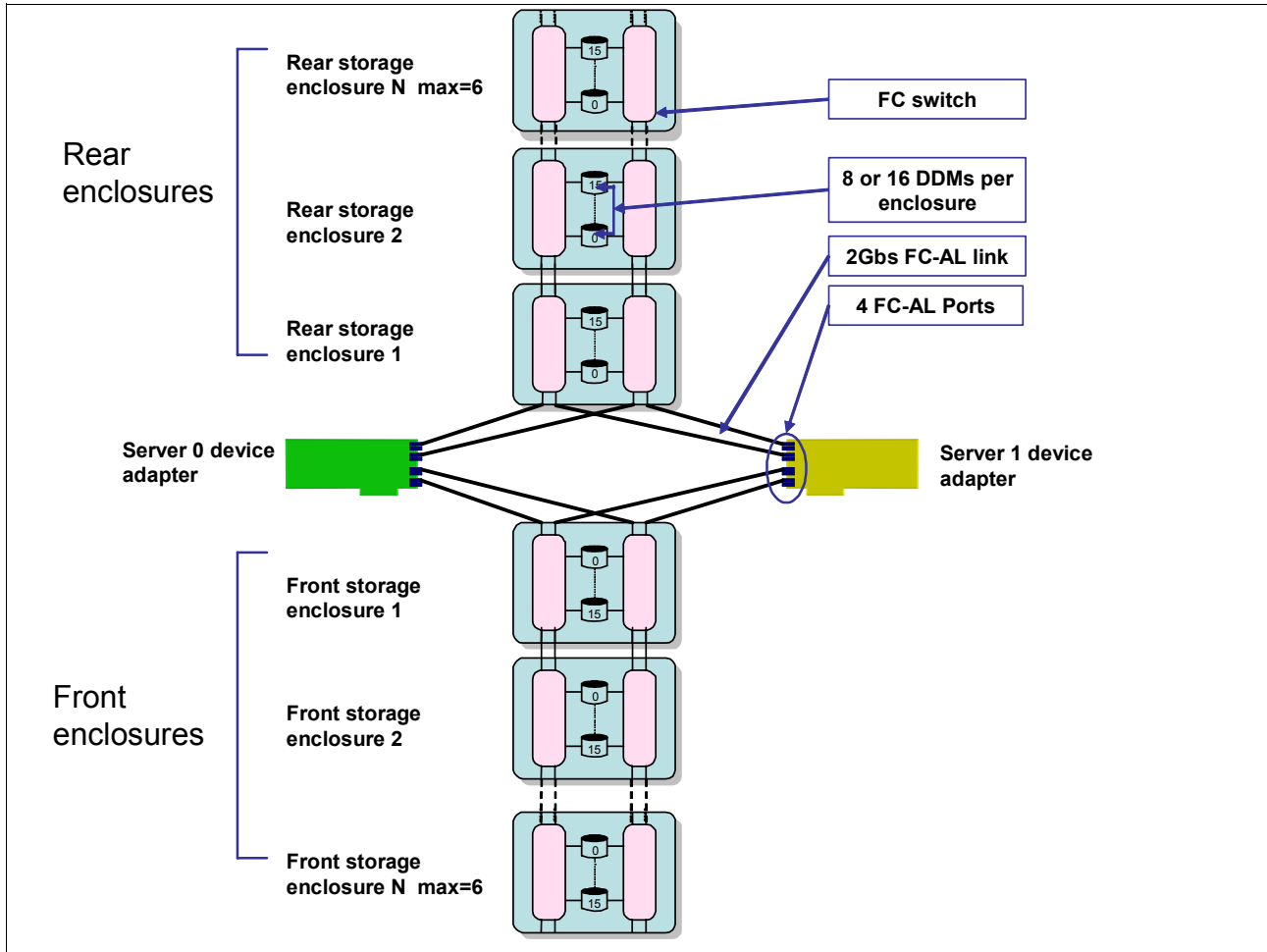


Figure 2-13 DS8000 switched disk expansion

## Expansion

Expansion enclosures are added in pairs and disks are added in groups of 16. On the ESS Model 800, the term 8-pack was used to describe an enclosure with eight disks in it. For the DS8000, we use the term 16-pack, though this term really describes the 16 DDMs found in one disk enclosure. It takes two orders of 16 DDMs to fully populate a disk enclosure pair (front and rear).

To provide an example, if a machine had six disk enclosures total, it would have three at the front and three at the rear. If all the enclosures were fully populated with disks, and an additional order of 16 DDMs was purchased, then two new disk enclosures would be added, one at the front and one at the rear. The switched networks do not need to be *broken* to add these enclosures. They are simply added to the end of the *loop*. Half of the 16 DDMs would go in the front enclosure and half would go in the rear enclosure. If an additional 16 DDMs were ordered later, they would be used to completely fill that pair of disk enclosures.

## Arrays and spares

Array sites containing eight DDMs are created as DDMs are installed. During configuration, discussed in Chapter 10, "The DS Storage Manager - logical configuration" on page 189, the user will have the choice of creating a RAID-5 or RAID-10 array by choosing one array site. The first four array sites created on a DA pair each contribute one DDM to be a spare. So at least four spares are created per DA pair, depending on the disk intermix.

The intention is to only have four spares per DA pair, but this number may increase depending on DDM intermix. We need to have four DDMs of the largest capacity and at least two DDMs of the fastest RPM. If all DDMs are the same size and RPM, then four spares will be sufficient.

### Arrays across loops

Each array site consists of eight DDMs. Four DDMs are taken from the front enclosure in an enclosure pair, and four are taken from the rear enclosure in the pair. This means that when a RAID array is created on the array site, half of the array is on each enclosure. Because the front enclosures are on one switched loop, and the rear enclosures are on a second switched loop, this splits the array across two loops. This is called *array across loops* (AAL).

To better understand AAL refer to Figure 2-14 and Figure 2-15. To make the diagrams clearer, only 16 DDMs are shown, eight in each disk enclosure. When fully populated, there would be 16 DDMs in each enclosure. Regardless, the diagram represents a valid configuration.

Figure 2-14 is used to depict the device adapter pair layout. One DA pair creates two switched loops. The front enclosures populate one loop while the rear enclosures populate the other loop. Each enclosure places two switches onto each loop. Each enclosure can hold up to 16 DDMs. DDMs are purchased in groups of 16. Half of the new DDMs go into the front enclosure and half go into the rear enclosure.

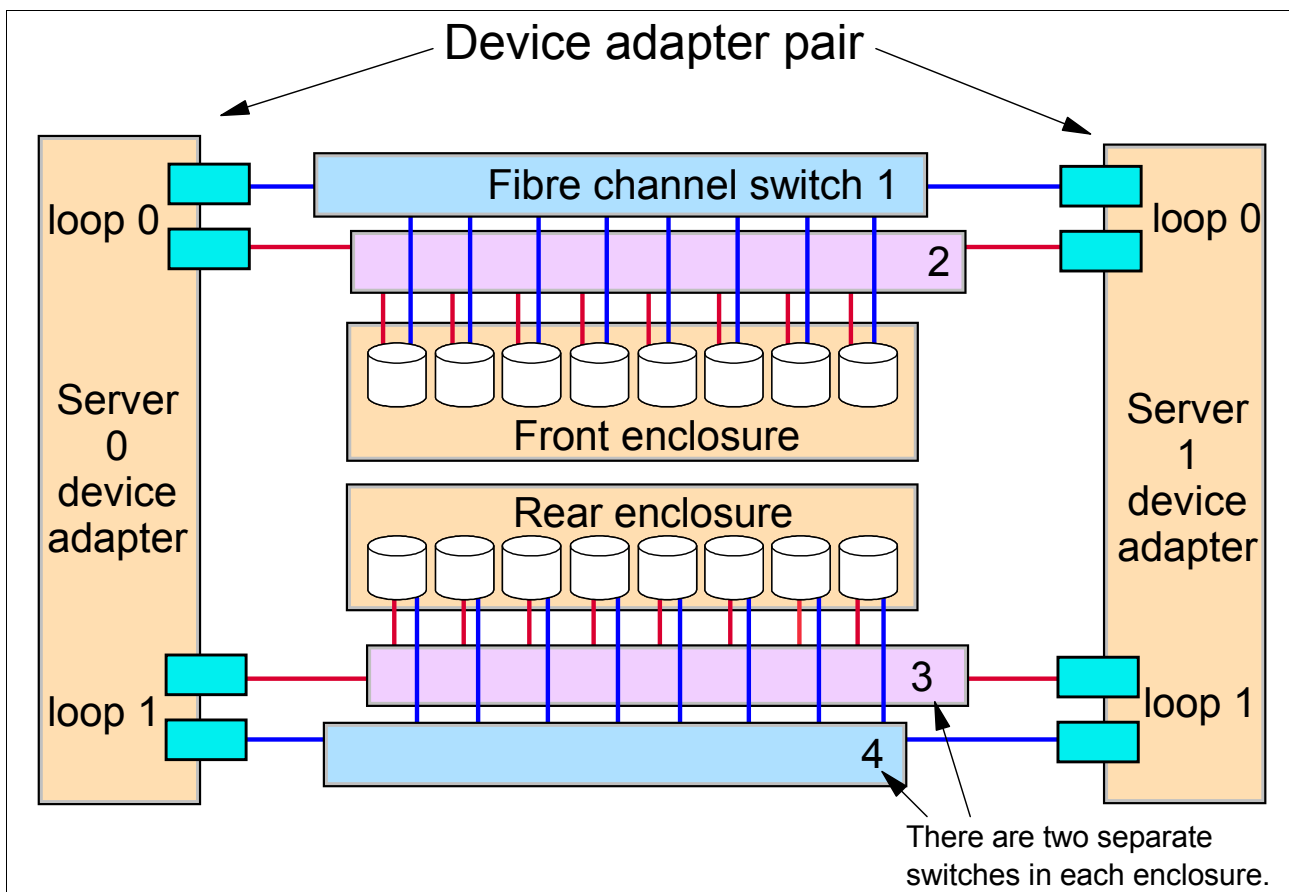


Figure 2-14 DS8000 switched loop layout

Having established the physical layout, the diagram is now changed to reflect the layout of the array sites, as shown in Figure 2-15 on page 37. Array site 0 in green (the darker disks) uses the four left-hand DDMs in each enclosure. Array site 1 in yellow (the lighter disks), uses the four right-hand DDMs in each enclosure. When an array is created on each array site, half of

the array is placed on each loop. If the disk enclosures were fully populated with DDMs, there would be four array sites.

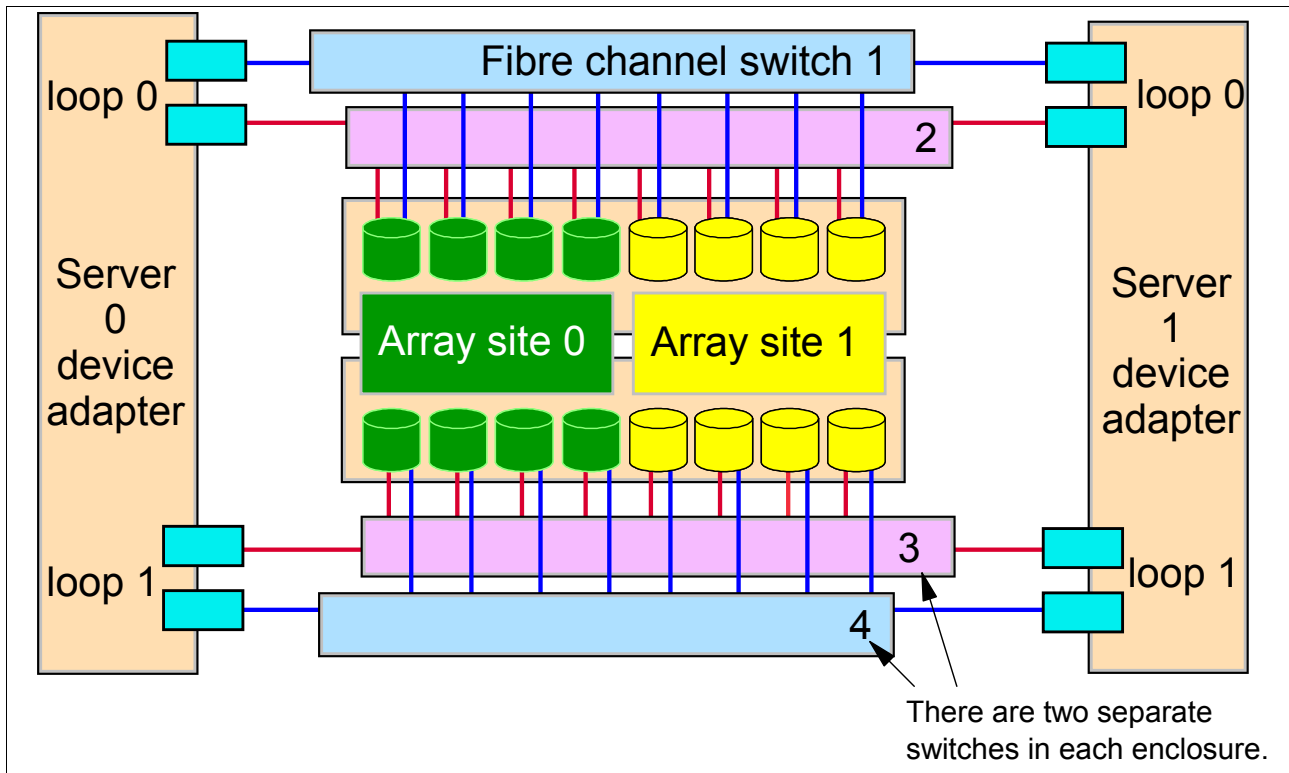


Figure 2-15 Array across loop

### AAL benefits

AAL is used to increase performance. When the device adapter writes a stripe of data to a RAID-5 array, it sends half of the write to each switched loop. By splitting the workload in this manner, each loop is worked evenly, which improves performance. If RAID-10 is used, two RAID-0 arrays are created. Each loop hosts one RAID-0 array. When servicing read I/O, half of the reads can be sent to each loop, again improving performance by balancing workload across loops.

### DDMs

Each DDM is hot pluggable and has two indicators. The green indicator shows disk activity while the amber indicator is used with light path diagnostics to allow for easy identification and replacement of a failed DDM.

At present the DS8000 allows the choice of three different DDM types:

- ▶ 73 GB, 15K RPM drive
- ▶ 146 GB, 10K RPM drive
- ▶ 300 GB, 10K RPM drive

## 2.5 Host adapters

The DS8000 supports two types of host adapters: ESCON and Fibre Channel/FICON. It does not support SCSI adapters.

The ESCON adapter in the DS8000 is a dual ported host adapter for connection to older zSeries hosts that do not support FICON. The ports on the ESCON card use the MT-RJ type connector.

### **Control units and logical paths**

ESCON architecture recognizes only 16 3990 logical control units (LCUs) even though the DS8000 is capable of emulating far more (these extra control units can be used by FICON). Half of the LCUs (even numbered) are in server 0, and the other half (odd-numbered) are in server 1. Because the ESCON host adapters can connect to both servers, each adapter can address all 16 LCUs.

An ESCON link consists of two fibers, one for each direction, connected at each end by an ESCON connector to an ESCON port. Each ESCON adapter card supports two ESCON ports or links, and each link supports 64 logical paths.

### **ESCON distances**

For connections without repeaters, the ESCON distances are 2 km with 50 micron multimode fiber, and 3 km with 62.5 micron multimode fiber. The DS8000 supports all models of the IBM 9032 ESCON directors that can be used to extend the cabling distances.

### **Remote Mirror and Copy with ESCON**

The initial implementation of the ESS 2105 Remote Mirror and Copy function (better known as PPRC or Peer-to-Peer Remote Copy) used ESCON adapters. This was known as PPRC Version 1. The ESCON adapters in the DS8000 do not support any form of Remote Mirror and Copy. If you wish to create a remote mirror between a DS8000 and an ESS 800 or another DS8000 or DS6000, you must use Fibre Channel adapters. You cannot have a remote mirror relationship between a DS8000 and an ESS E20 or F20 because the E20/F20 only support Remote Mirror and Copy over ESCON.

### **ESCON supported servers**

ESCON is used for attaching the DS8000 to the IBM S/390 and zSeries servers. The most current list of supported servers is at this Web site:

<http://www.storage.ibm.com/hardsoft/products/DS8000/supserver.htm>

This site should be consulted regularly because it has the most up-to-date information on server attachment support.

## **2.5.1 FICON and Fibre Channel protocol host adapters**

Fibre Channel is a technology standard that allows data to be transferred from one node to another at high speeds and great distances (up to 10 km and beyond). The DS8000 uses Fibre Channel protocol to transmit SCSI traffic inside Fibre Channel frames. It also uses Fibre Channel to transmit FICON traffic, which uses Fibre Channel frames to carry zSeries I/O.

Each DS8000 Fibre Channel card offers four 2 Gbps Fibre Channel ports. The cable connector required to attach to this card is an LC type. Each port independently auto-negotiates to either 2 Gbps or 1 Gbps link speed. Each of the 4 ports on one DS8000 adapter can also independently be either Fibre Channel protocol (FCP) or FICON, though the ports are initially defined as switched point to point FCP. Selected ports will be configured to FICON automatically based on the definition of a FICON host. Each port can be either FICON or Fibre Channel protocol (FCP). The personality of the port is changeable via the DS Storage Manager GUI. A port cannot be both FICON and FCP simultaneously, but it can be changed as required.

The card itself is PCI-X 64 Bit 133 MHz. The card is driven by a new high function, high performance ASIC. To ensure maximum data integrity, it supports metadata creation and checking. Each Fibre Channel port supports a maximum of 509 host login IDs. This allows for the creation of very large storage area networks (SANs). The design of the card is depicted in Figure 2-16.

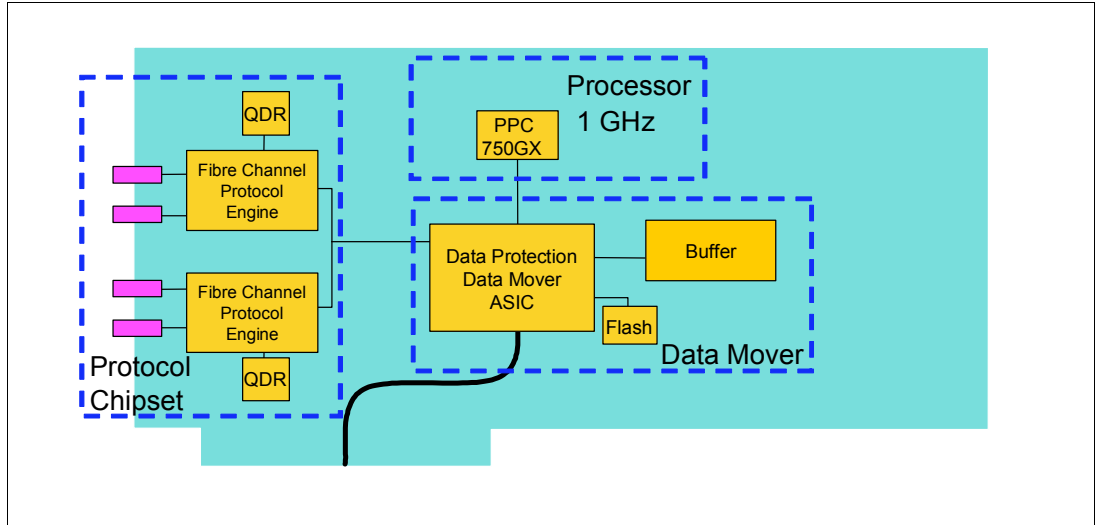


Figure 2-16 DS8000 FICON/FCP host adapter

### Fibre Channel supported servers

The current list of servers supported by the Fibre Channel attachment is at this Web site:

<http://www.storage.ibm.com/hardsoft/products/DS8000/supserver.htm>

This document should be consulted regularly because it has the most up-to-date information on server attachment support.

### Fibre Channel distances

There are two types of host adapter cards you can select: long wave and short wave. With long-wave laser, you can connect nodes at distances of up to 10 km (non-repeated). With short wave you are limited to a distance of 300 to 500 metres (non-repeated). All ports on each card must be either long wave or short wave (there can be no mixing of types within a card).

## 2.6 Power and cooling

The DS8000 power and cooling system is highly redundant.

### Rack Power Control cards (RPC)

The DS8000 has a pair of redundant RPC cards that are used to control certain aspects of power sequencing throughout the DS8000. These cards are attached to the Service Processor (SP) card in each processor, which allows them to communicate both with the Storage Hardware Management Console (S-HMC) and the storage facility image LPARs. The RPCs also communicate with each primary power supply and indirectly with each rack's fan sense cards and the disk enclosures in each frame.

## **Primary power supplies**

The DS8000 primary power supply (PPS) converts input AC voltage into DC voltage. There are high and low voltage versions of the PPS because of the varying voltages used throughout the world. Also, because the line cord connector requirements vary widely throughout the world, the line cord may not come with a suitable connector for your nation's preferred outlet. This may need to be replaced by an electrician once the machine is delivered.

There are two redundant PPSs in each frame of the DS8000. Each PPS is capable of powering the frame by itself. The PPS creates 208V output power for the processor complex and I/O enclosure power supplies. It also creates 5V and 12V DC power for the disk enclosures. There may also be an optional booster module that will allow the PPSs to temporarily run the disk enclosures off battery, if the extended power line disturbance feature has been purchased (see Chapter 4, "RAS" on page 61, for a complete explanation as to why this feature may or may not be necessary for your installation).

Each PPS has internal fans to supply cooling for that power supply.

## **Processor and I/O enclosure power supplies**

Each processor and I/O enclosure has dual redundant power supplies to convert 208V DC into the required voltages for that enclosure or complex. Each enclosure also has its own cooling fans.

## **Disk enclosure power and cooling**

The disk enclosures do not have separate power supplies since they draw power directly from the PPSs. They do, however, have cooling fans located in a plenum above the enclosures. They draw cooling air through the front of each enclosure and exhaust air out of the top of the frame.

## **Battery backup assemblies**

The backup battery assemblies help protect data in the event of a loss of external power. The model 921 contains two battery backup assemblies while the model 922 and 9A2 contain three of them (to support the 4-way processors). In the event of a complete loss of input AC power, the battery assemblies are used to allow the contents of NVS memory to be written to a number of DDMs internal to the processor complex, prior to power off.

The FC-AL DDMs are not protected from power loss unless the extended power line disturbance feature has been purchased.

## **2.7 Management console network**

All base models ship with one Storage Hardware Management Console (S-HMC), a keyboard and display, plus two Ethernet switches.

### **S-HMC**

The S-HMC is the focal point for configuration, Copy Services management, and maintenance activities. It is possible to order two management consoles to act as a redundant pair. A typical configuration would be to have one internal and one external management console. The internal S-HMC will contain a PCI modem for remote service.

## **Ethernet switches**

In addition to the Fibre Channel switches installed in each disk enclosure, the DS8000 base frame contains two 16-port Ethernet switches. Two switches are supplied to allow the creation of a fully redundant management network. Each processor complex has multiple connections to each switch. This is to allow each server to access each switch. This switch cannot be used for any equipment not associated with the DS8000. The switches get power from the internal power bus and thus do not require separate power outlets.

## **2.8 Summary**

This chapter has described the various components that make up a DS8000. For additional information, there is documentation available at:

<http://www-1.ibm.com/servers/storage/support/disk/index.html>







## Storage system LPARs (Logical partitions)

This chapter provides information about storage system Logical Partitions (LPARs) in the DS8000.

The following topics are discussed in detail:

- ▶ Introduction to LPARs
- ▶ DS8000 and LPARs
  - LPAR and storage facility images (SFIs)
  - DS8300 LPAR implementation
  - Hardware components of a storage facility image
  - DS8300 Model 9A2 configuration options
- ▶ LPAR security and protection
- ▶ LPAR and Copy Services
- ▶ LPAR benefits

## 3.1 Introduction to logical partitioning

*Logical partitioning* allows the division of a single server into several completely independent *virtual* servers or partitions.

IBM began work on logical partitioning in the late 1960s, using S/360 mainframe systems with the precursors of VM, specifically CP40. Since then, logical partitioning on IBM mainframes (now called IBM zSeries) has evolved from a predominantly physical partitioning scheme based on hardware boundaries to one that allows for virtual and shared resources with dynamic load balancing. In 1999 IBM implemented LPAR support on the AS/400 (now called IBM iSeries) platform and on pSeries in 2001. In 2000 IBM announced the ability to run the Linux operating system in an LPAR or on top of VM on a zSeries server, to create thousands of Linux instances on a single system.

### 3.1.1 Virtualization Engine technology

IBM Virtualization Engine is comprised of a suite of system services and technologies that form key elements of IBM's on demand computing model. It treats resources of individual servers, storage, and networking products as if in a single pool, allowing access and management of resources across an organization more efficiently. Virtualization is a critical component in the on demand operating environment. The system technologies implemented in the POWER5 processor provide a significant advancement in the enablement of functions required for operating in this environment.

LPAR is one component of the POWER5 system technology that is part of the IBM Virtualization Engine.

Using IBM Virtualization Engine technology, selected models of the DS8000 series can be used as a single, large storage system, or can be used as multiple storage systems with logical partitioning (LPAR) capabilities. IBM LPAR technology, which is unique in the storage industry, allows the resources of the storage system to be allocated into separate logical storage system partitions, each of which is totally independent and isolated. Virtualization Engine (VE) delivers the capabilities to simplify the infrastructure by allowing the management of heterogeneous partitions/servers on a single system.

### 3.1.2 Partitioning concepts

It is appropriate to clarify the terms and definitions by which we classify these mechanisms.

**Note:** The following sections discuss partitioning concepts in general and not all are applicable to the DS8000.

#### **Partitions**

When a multi-processor computer is subdivided into multiple, independent operating system images, those independent operating environments are called partitions. The resources on the system are allocated to specific partitions.

#### **Resources**

Resources are defined as a system's processors, memory, and I/O slots. I/O slots can be populated by different adapters, such as Ethernet, SCSI, Fibre Channel or other device controllers. A disk is allocated to a partition by assigning it the I/O slot that contains the disk's controller.

## **Building block**

A building block is a collection of system resources, such as processors, memory, and I/O connections.

## **Physical partitioning (PPAR)**

In physical partitioning, the partitions are divided along hardware boundaries. Each partition might run a different version of the same operating system. The number of partitions relies on the hardware. Physical partitions have the advantage of allowing complete isolation of operations from operations running on other processors, thus ensuring their availability and uptime. Processors, I/O boards, memory, and interconnects are not shared, allowing applications that are business-critical or for which there are security concerns to be completely isolated. The disadvantage of physical partitioning is that machines cannot be divided into as many partitions as those that use logical partitioning, and users can't consolidate many lightweight applications on one machine.

## **Logical partitioning (LPAR)**

A logical partition uses hardware and firmware to logically partition the resources on a system. LPARs logically separate the operating system images, so there is not a dependency on the hardware building blocks.

A logical partition consists of processors, memory, and I/O slots that are a subset of the pool of available resources within a system, as shown in Figure 3-1 on page 46. While there are configuration rules, the granularity of the units of resources that can be allocated to partitions is very flexible. It is possible to add just a small amount of memory, if that is all that is needed, without a dependency on the size of the memory controller or without having to add more processors or I/O slots that are not needed.

LPAR differs from physical partitioning in the way resources are grouped to form a partition. Logical partitions do not need to conform to the physical boundaries of the building blocks used to build the server. Instead of grouping by physical building blocks, LPAR adds more flexibility to select components from the entire pool of available system resources.

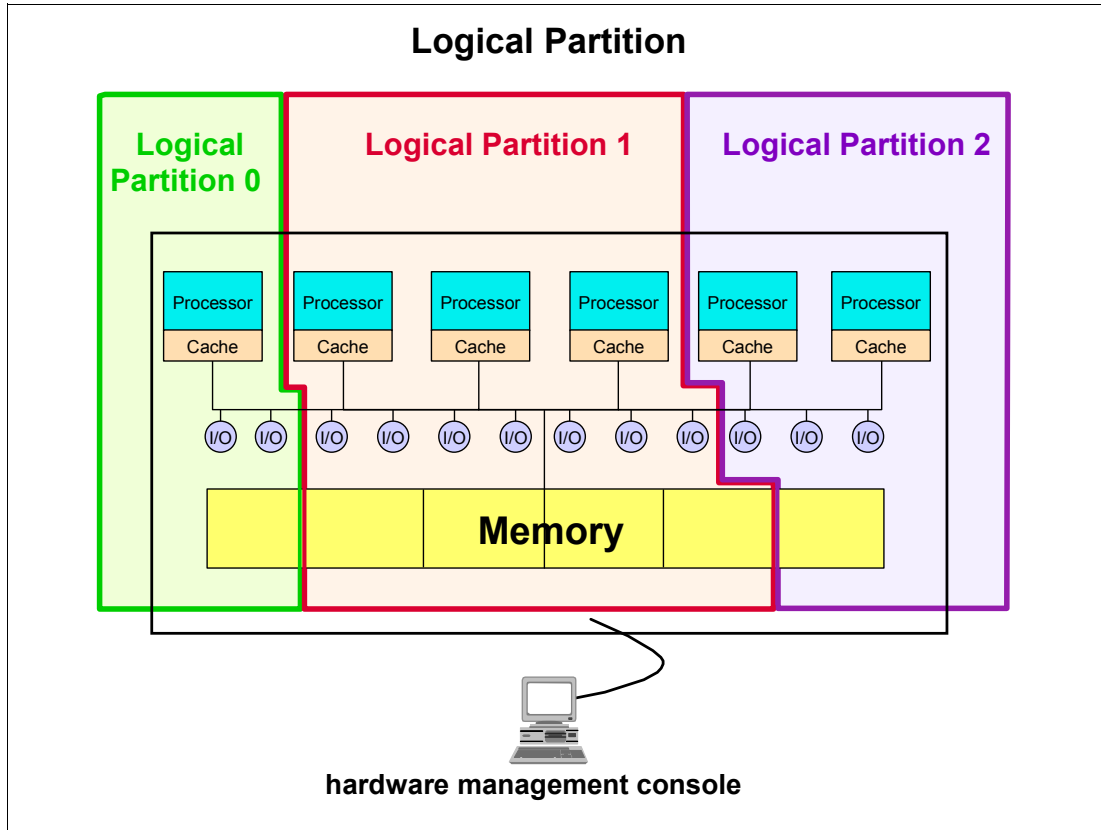


Figure 3-1 Logical partition

### Software and hardware fault isolation

Because a partition hosts an independent operating system image, there is strong *software isolation*. This means that a job or software crash in one partition will not effect the resources in another partition.

### Dynamic logical partitioning

Starting from AIX 5L™ Version 5.2, IBM supports dynamic logical partitioning (also known as DLPAR) in partitions on several logical partitioning capable IBM pSeries server models.

The dynamic logical partitioning function allows resources, such as CPUs, memory, and I/O slots, to be added to or removed from a partition, as well as allowing the resources to be moved between two partitions, without an operating system reboot (*on the fly*).

### Micro-Partitioning™

With AIX 5.3, partitioning capabilities are enhanced to include sub-processor partitioning, or Micro-Partitioning. With Micro-Partitioning it is possible to allocate less than a full physical processor to a logical partition.

The benefit of Micro-Partitioning is that it allows increased overall utilization of system resources by automatically applying only the required amount of processor resource needed by each partition.

### Virtual I/O

On POWER5 servers, I/O resources (disks and adapters) can be shared through Virtual I/O. Virtual I/O provides the ability to dedicate I/O adapters and devices to a virtual server,

allowing the on-demand allocation of those resources to different partitions and the management of I/O devices. The physical resources are owned by the Virtual I/O server.

### 3.1.3 Why Logically Partition?

There is a demand to provide greater flexibility for high-end systems, particularly the ability to subdivide them into smaller partitions that are capable of running a version of an operating system or a specific set of application workloads.

The main reasons for partitioning a large system are as follows:

#### **Server consolidation**

A highly reliable server with sufficient processing capacity and capable of being partitioned can address the need for server consolidation by logically subdividing the server into a number of separate, smaller systems. This way, the application isolation needs can be met in a consolidated environment, with the additional benefits of reduced floor space, a single point of management, and easier redistribution of resources as workloads change. Increasing or decreasing the resources allocated to partitions can facilitate better utilization of a server that is exposed to large variations in workload.

#### **Production and test environments**

Generally, production and test environments should be isolated from each other. Without partitioning, the only practical way of performing application development and testing is to purchase additional hardware and software.

Partitioning is a way to set aside a portion of the system resources to use for testing new versions of applications and operating systems, while the production environment continues to run. This eliminates the need for additional servers dedicated to testing, and provides more confidence that the test versions will migrate smoothly into production because they are tested on the production hardware system.

#### **Consolidation of multiple versions of the same OS or applications**

The flexibility inherent in LPAR greatly aids the scheduling and implementation of normal upgrade and system maintenance activities. All the preparatory activities involved in upgrading an application or even an operating system could be completed in a separate partition. An LPAR can be created to test applications under new versions of the operating system prior to upgrading the production environments. Instead of having a separate server for this function, a minimum set of resources can be temporarily used to create a new LPAR where the tests are performed. When the partition is no longer needed, its resources can be incorporated back into the other LPARs.

#### **Application isolation**

Partitioning isolates an application from another in a different partition. For example, two applications on one symmetric multi-processing (SMP) system could interfere with each other or compete for the same resources. By separating the applications into their own partitions, they cannot interfere with each other. Also, if one application were to hang or crash the operating system, this would not have an effect on the other partitions. Also, applications are prevented from consuming excess resources, which could starve other applications of resources they require.

#### **Increased hardware utilization**

Partitioning is a way to achieve better hardware utilization when software does not scale well across large numbers of processors. Where possible, running multiple instances of an

application on separate smaller partitions can provide better throughput than running a single large instance of the application.

### **Increased flexibility of resource allocation**

A workload with resource requirements that change over time can be managed more easily within a partition that can be altered to meet the varying demands of the workload.

## **3.2 DS8000 and LPAR**

In the first part of this chapter we discussed the LPAR features in general. In this section we provide information on how the LPAR functionality is implemented in the DS8000 series.

The DS8000 series is a server-based disk storage system. With the integration of the POWER5 eServer p5 570 into the DS8000 series, IBM offers the first implementation of the *server* LPAR functionality in a disk storage system.

The storage system LPAR functionality is currently supported in the DS8300 Model 9A2. It provides two virtual storage systems in one physical machine. Each storage system LPAR can run its own level of licensed internal code (LIC).

The resource allocation for processors, memory, and I/O slots in the two storage system LPARs on the DS8300 is currently divided into a fixed ratio of 50/50.

**Note:** The allocation for resources will be more flexible. According to the announcement letter IBM has issued a Statement of General Direction:

*IBM intends to enhance the Virtualization Engine partitioning capabilities of selected models of the DS8000 series to provide greater flexibility in the allocation and management of resources between images.*

Between the two storage facility images there exists a robust isolation via hardware; for example, separated RIO-G loops, and the POWER5 Hypervisor, which is described in more detail in section 3.3, “LPAR security through POWER™ Hypervisor (PHYP)” on page 54.

### **3.2.1 LPAR and storage facility images**

Before we start to explain how the LPAR functionality is implemented in the DS8300, we want to clarify some terms and naming conventions. Figure 3-2 on page 49 illustrates these terms.

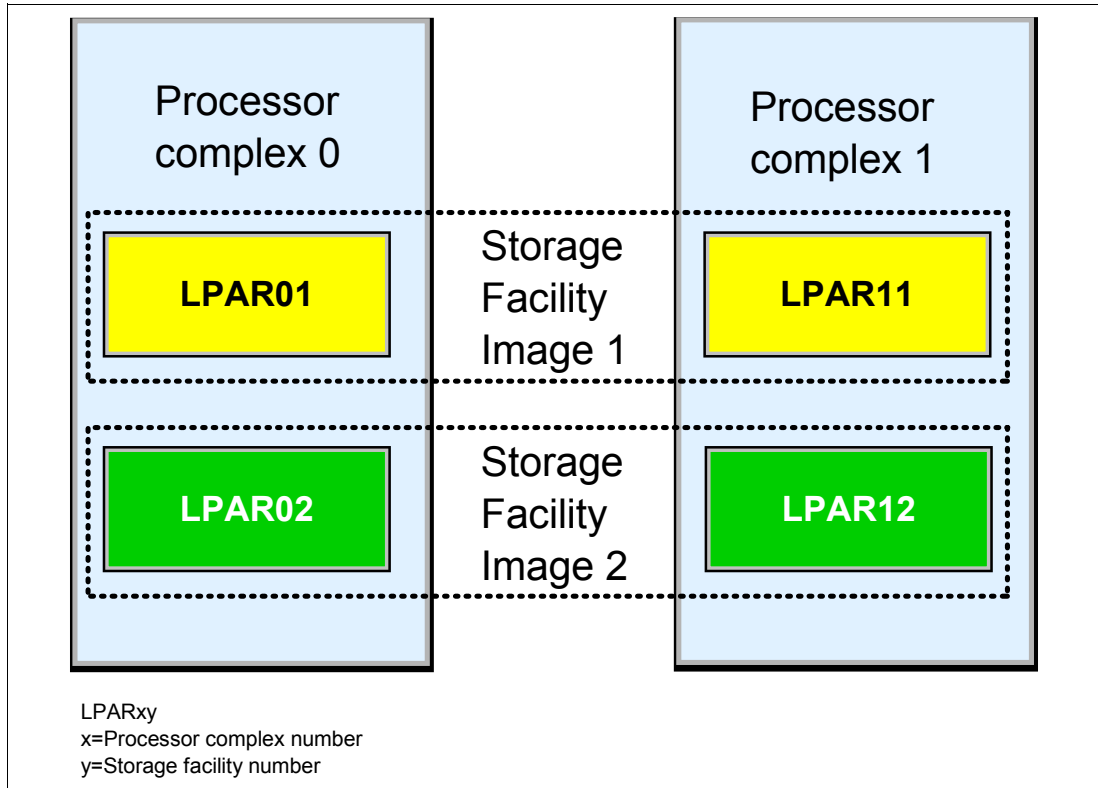


Figure 3-2 DS8300 Model 9A2 - LPAR and storage facility image

The DS8300 series incorporates two eServer p5 570s. We call each of these a *processor complex*. Each processor complex supports one or more LPARs. Currently each processor complex on the DS8300 is divided into two LPARs. An *LPAR* is a set of resources on a processor complex that support the execution of an operating system. The *storage facility image* is built from a pair of LPARs, one on each processor complex.

Figure 3-2 shows that LPAR01 from processor complex 0 and LPAR11 from processor complex 1 instantiate storage facility image 1. LPAR02 from processor complex 0 and LPAR12 from processor complex 1 instantiate the second storage facility image.

**Important:** It is important to understand that an LPAR in a processor complex is *not* the same as a storage facility image in the DS8300.

### 3.2.2 DS8300 LPAR implementation

Each storage facility image will use the machine type/model number/serial number of the DS8300 Model 9A2 base frame. The frame serial number will end with *0*. The last character of the serial number will be replaced by a number in the range *one* to *eight* that uniquely identifies the DS8000 image. Initially, this character will be a *1* or a *2*, because there are only two storage facility images available. The serial number is needed to distinguish between the storage facility images in the GUI, CLI, and for licensing and allocating the licenses between the storage facility images.

The first release of the LPAR functionality in the DS8300 Model 9A2 provides a split between the resources in a 50/50 ratio as depicted in Figure 3-3 on page 50.

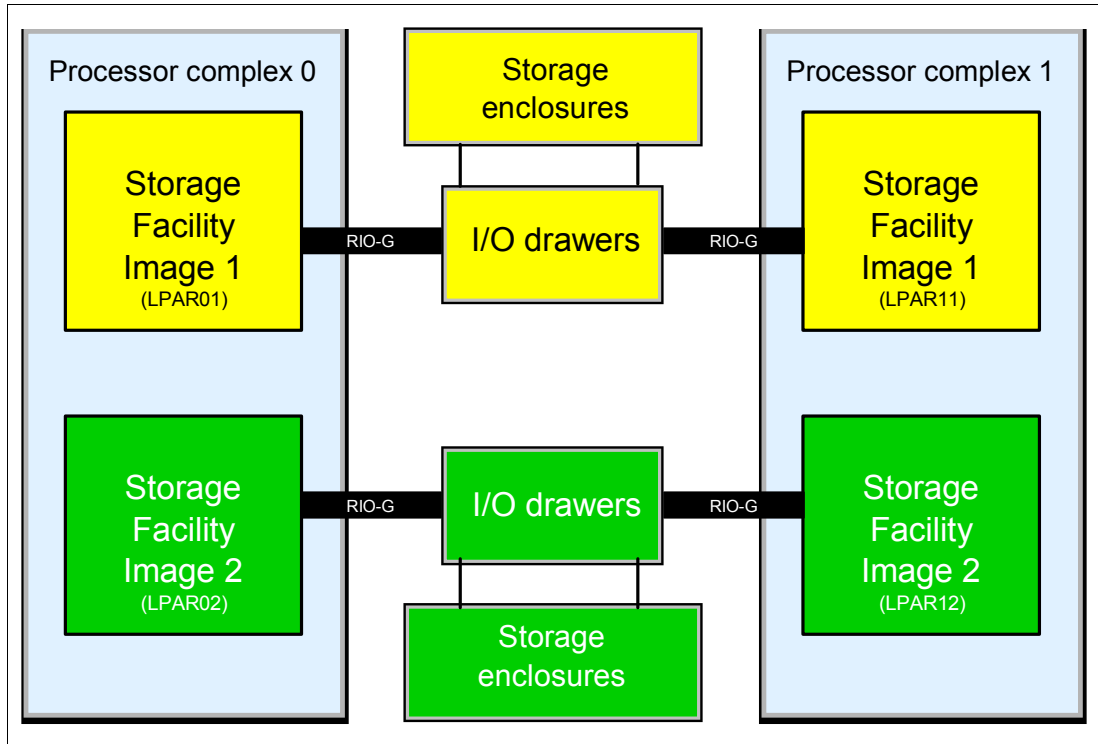


Figure 3-3 DS8300 LPAR resource allocation

Each storage facility image has access to:

- ▶ 50 percent of the processors
- ▶ 50 percent of the processor memory
- ▶ 1 loop of the RIO-G interconnection
- ▶ Up to 16 host adapters (4 I/O drawers with up to 4 host adapters)
- ▶ Up to 320 disk drives (up to 96 TB of capacity)

### 3.2.3 Storage facility image hardware components

In this section we explain which hardware resources are required to build a storage facility image.

The management of the resource allocation between LPARs on a pSeries is done via the Storage Hardware Management Console (S-HMC). Because the DS8300 Model 9A2 provides a fixed split between the two storage facility images, there is no management or configuration necessary via the S-HMC. The DS8300 comes pre-configured with all required LPAR resources assigned to either storage facility image.

Figure 3-4 on page 51 shows the split of all available resources between the two storage facility images. Each storage facility image has 50% of all available resources.



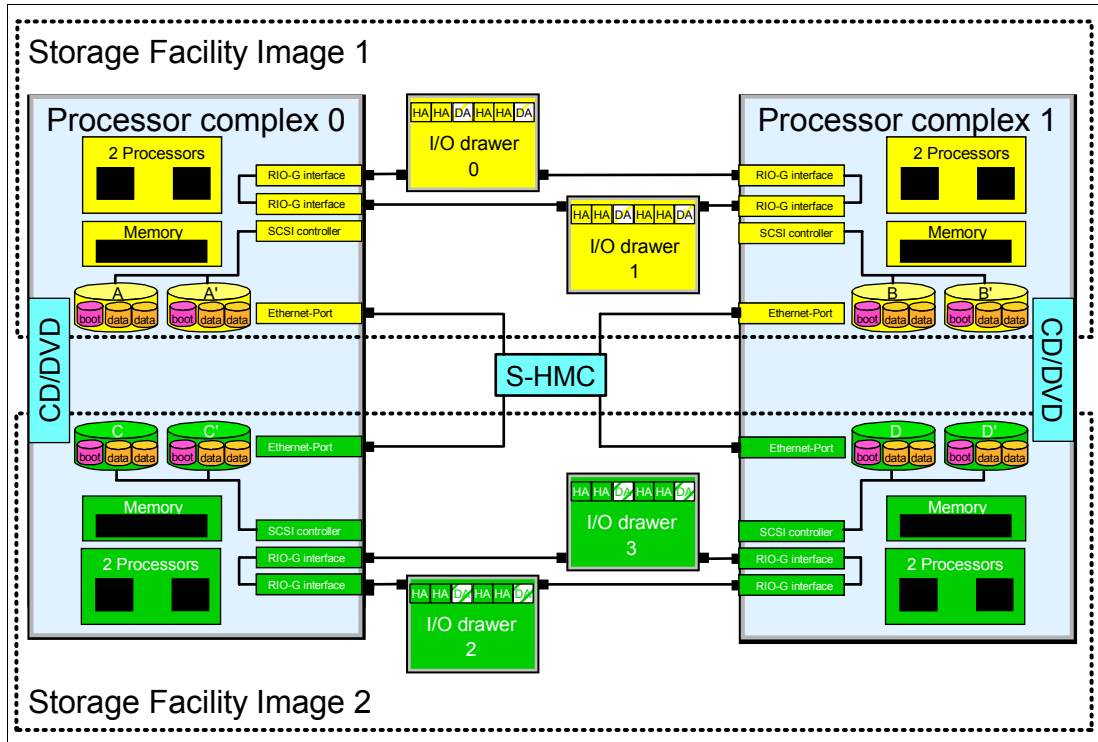


Figure 3-4 Storage facility image resource allocation in the processor complexes of the DS8300

## I/O resources

For one storage facility image, the following hardware resources are required:

- ▶ 2 SCSI controllers with 2 disk drives each
- ▶ 2 Ethernet ports (to communicate with the S-HMC)
- ▶ 1 Thin Device Media Bay (for example, CD or DVD; can be shared between the LPARs)

Each storage facility image will have two physical disk drives in each processor complex. Each disk drive will contain three logical volumes, the boot volume and two logical volumes for the memory save dump function. These three logical volumes are then mirrored across the two physical disk drives for each LPAR. In Figure 3-4, for example, the disks A/A' are mirrors. For the DS8300 Model 9A2, there will be four drives total in one physical processor complex.

## Processor and memory allocations

In the DS8300 Model 9A2 each processor complex has four processors and up to 128 GB memory. Initially there is also a 50/50 split for processor and memory allocation.

Therefore, every LPAR has two processors and so every storage facility image has four processors.

The memory limit depends on the total amount of available memory in the whole system. Currently there are the following memory allocations per storage facility available:

- ▶ 32 GB (16 GB per processor complex, 16 GB per storage facility image)
- ▶ 64 GB (32 GB per processor complex, 32 GB per storage facility image)
- ▶ 128 GB (64 GB per processor complex, 64 GB per storage facility image)
- ▶ 256 GB (128 GB per processor complex, 128 GB per storage facility image)

## RIO-G interconnect separation

Figure 3-4 on page 51 depicts that the RIO-G interconnection is also split between the two storage facility images. The RIO-G interconnection is divided into 2 loops. Each RIO-G loop is dedicated to a given storage facility image. All I/O enclosures on the RIO-G loop with the associated host adapters and drive adapters are dedicated to the storage facility image that owns the RIO-G loop.

As a result of the strict separation of the two images, the following configuration options exist:

- ▶ Each storage facility image is assigned to one dedicated RIO-G loop; if an image is offline, its RIO-G loop is not available.
- ▶ All I/O enclosures on a given RIO-G loop are dedicated to the image that owns the RIO-G loop.
- ▶ Host adapter and device adapters on a given loop are dedicated to the associated image that owns this RIO-G loop.
- ▶ Disk enclosures and storage devices behind a given device adapter pair are dedicated to the image that owns the RIO-G loop.
- ▶ Configuring of capacity to an image is managed through the placement of disk enclosures on a specific DA pair dedicated to this image.

### 3.2.4 DS8300 Model 9A2 configuration options

In this section we explain which configuration options are available for the DS8300 Model 9A2.

The Model 9A2 (base frame) has:

- ▶ 32 to 128 DDMs
  - Up to 64 DDMs per storage facility image, in increments of 16 DDMs
- ▶ System memory
  - 32, 64, 128, 256 GB (half of the amount of memory is assigned to each storage facility image)
- ▶ Four I/O bays
  - Two bays assigned to storage facility image 1 and two bays assigned to storage facility image 2
  - Each bay contains:
    - Up to 4 host adapters
    - Up to 2 device adapters
- ▶ S-HMC, keyboard/display, and 2 Ethernet switches

The first Model 9AE (expansion frame) has:

- ▶ An additional four I/O bays
  - Two bays are assigned to storage facility image 1 and two bays are assigned to storage facility image 2.
- ▶ Each bay contains:
  - Up to 4 host adapters
  - Up to 2 device adapters

- ▶ An additional 256 DDMs
  - Up to 128 DDMs per storage facility image

The second Model 9AE (expansion frame) has:

- ▶ An additional 256 DDMs
  - Up to 128 drives per storage facility image

A fully configured DS8300 with storage facility images has one base frame and two expansion frames. The first expansion frame (9AE) has additional I/O drawers and disk drive modules (DDMs), while the second expansion frame contains additional DDMs.

Figure 3-5 provides an example of how a fully populated DS8300 might be configured. The disk enclosures are assigned to storage facility image 1 (yellow, or lighter if not viewed in color) or storage facility image 2 (green, or darker). When ordering additional disk capacity, it can be allocated to either storage facility image 1 or storage facility image 2. The cabling is pre-determined and in this example there is an empty pair of disk enclosures assigned for the next increment of disk to be added to storage facility image 2.

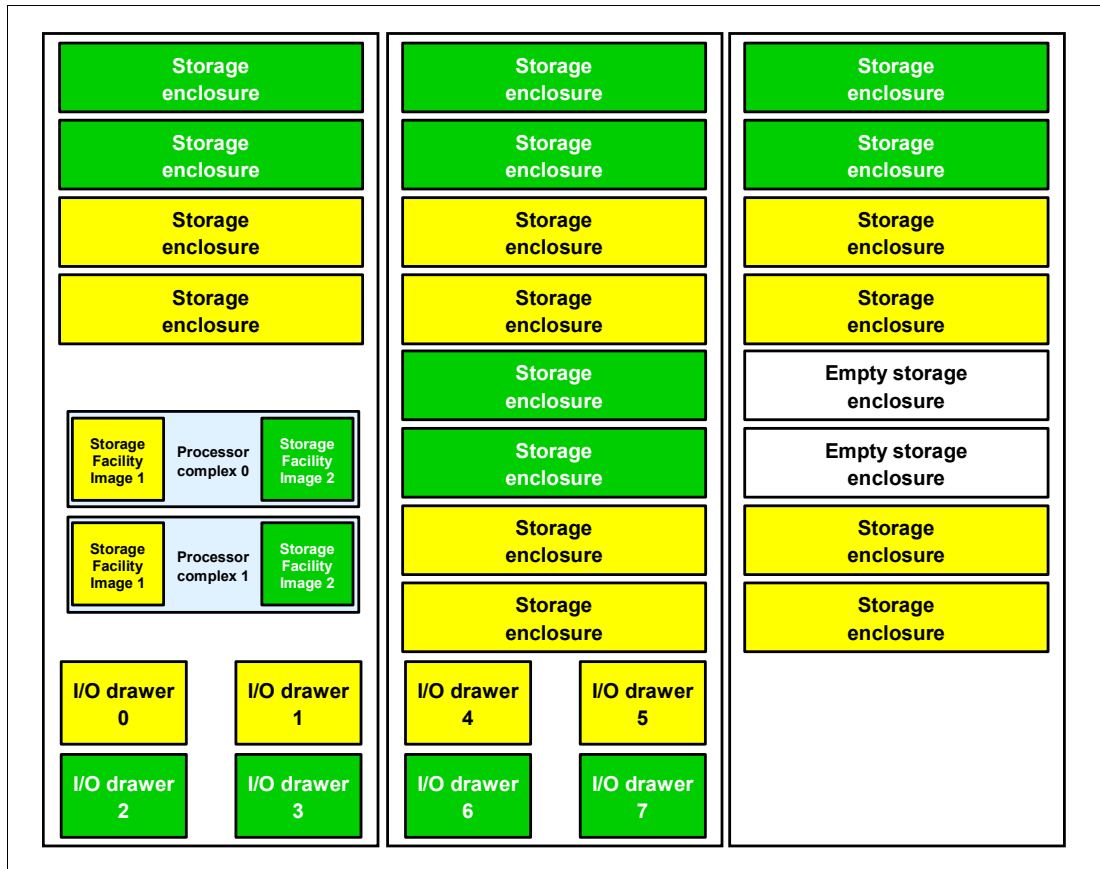


Figure 3-5 DS8300 example configuration

### Model conversion

The Model 9A2 has a fixed 50/50 split into two storage facility images. However, there are various model conversions available. For example, it is possible to switch from Model 9A2 to a full system machine, which is the Model 922. Table 3-1 shows all possible model conversions regarding the LPAR functionality.

Table 3-1 Model conversions regarding LPAR functionality

From Model	To Model
921 (2-way processors without LPAR)	9A2 (4-way processors with LPAR)
922 (4-way processors without LPAR)	9A2 (4-way processors with LPAR)
9A2 (4-way processors with LPAR)	922 (4-way processors without LPAR)
92E (expansion frame without LPAR)	9AE (expansion frame with LPAR)
9AE (expansion frame with LPAR)	92E (expansion frame without LPAR)

**Note:** Every model conversion is a disruptive operation.

### 3.3 LPAR security through POWER™ Hypervisor (PHYP)

The DS8300 Model 9A2 provides two storage facility images. This offers a number of desirable business advantages. But it also can raise some concerns about security and protection of the storage facility images in the DS8000 series. In this section we explain how the DS8300 delivers robust isolation between the two storage facility images.

One aspect of LPAR protection and security is that the DS8300 has a dedicated allocation of the hardware resources for the two facility images. There is a clear split of processors, memory, I/O slots, and disk enclosures between the two images.

Another important security feature which is implemented in the pSeries server is called the POWER Hypervisor (PHYP). It enforces partition integrity by providing a security layer between logical partitions. The POWER Hypervisor is a component of system firmware that will always be installed and activated, regardless of the system configuration. It operates as a hidden partition, with no processor resources assigned to it.

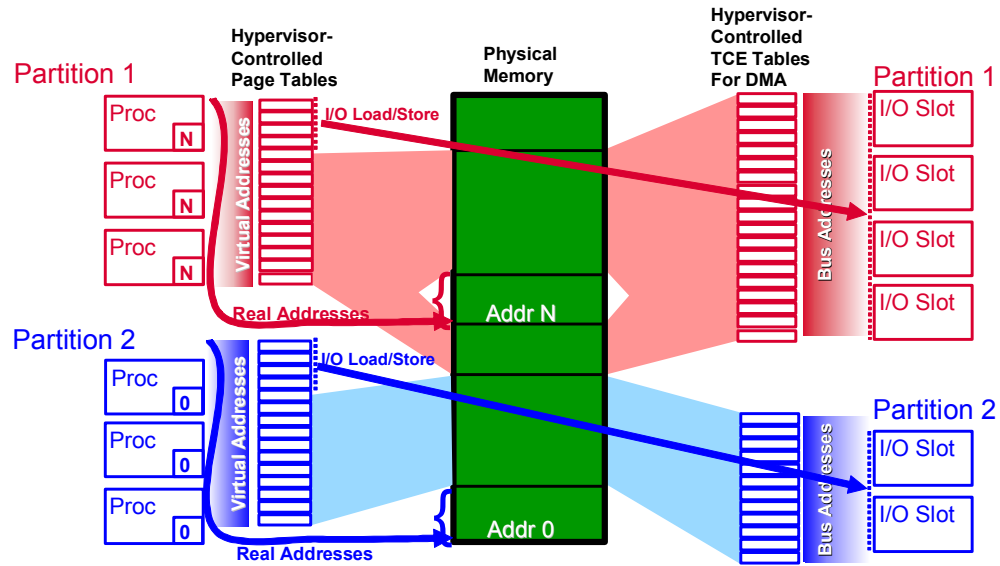
Figure 3-6 on page 55 illustrates a set of address mapping mechanisms which are described in the following paragraphs.

In a partitioned environment, the POWER Hypervisor is loaded into the first Physical Memory Block (PMB) at the physical address zero and reserves the PMB. From then on, it is not possible for an LPAR to access directly the physical memory. Every memory access is controlled by the POWER Hypervisor.

Each partition has its own exclusive page table, which is also controlled by the POWER Hypervisor. Processors use these tables to transparently convert a program's virtual address into the physical address where that page has been mapped into physical memory.

In a partitioned environment, the operating system uses hypervisor services to manage the translation control entry (TCE) tables. The operating system communicates the desired I/O bus address to logical mapping, and the hypervisor translates that into the I/O bus address to physical mapping within the specific TCE table. The hypervisor needs a dedicated memory region for the TCE tables to translate the I/O address to the partition memory address, then the hypervisor can perform direct memory access (DMA) transfers to the PCI adapters.

## LPAR Protection in IBM POWER5™ Hardware



The Hardware and Hypervisor manage the real to virtual memory mapping to provide robust isolation between partitions

Figure 3-6 LPAR protection - POWER Hypervisor

### 3.4 LPAR and Copy Services

In this section we provide some specific information about the Copy Services functions related to the LPAR functionality on the DS8300. An example for this can be seen in Figure 3-7 on page 56.

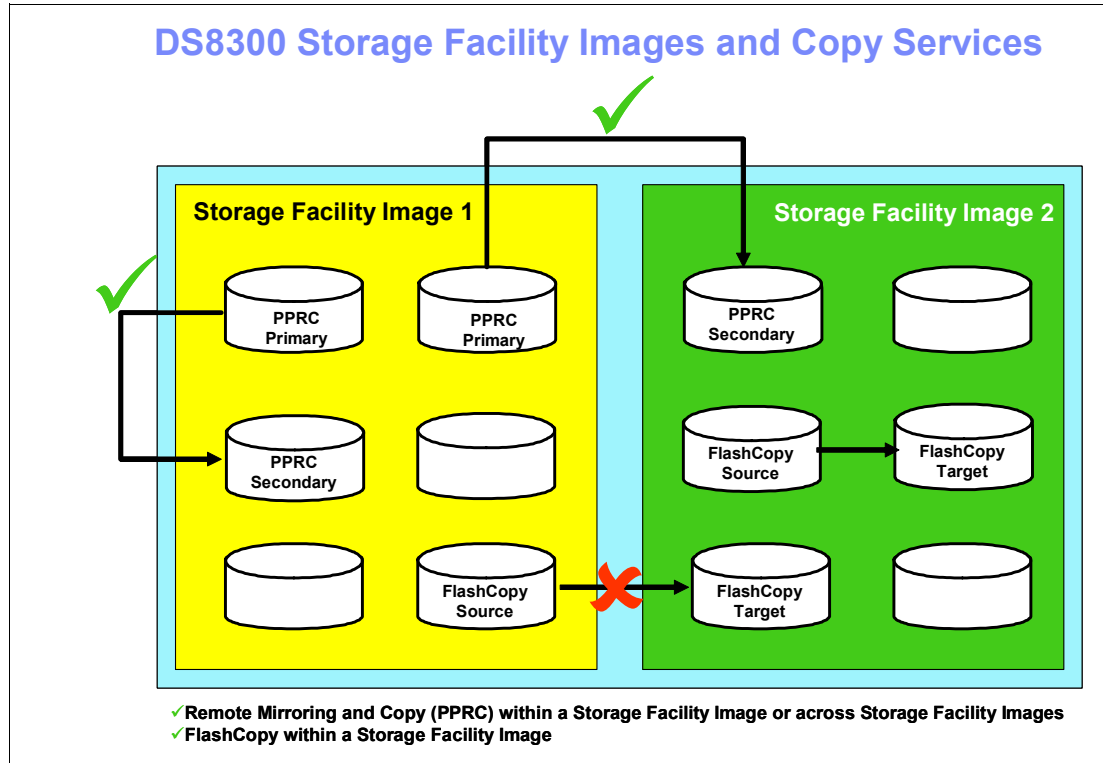


Figure 3-7 DS8300 storage facility images and Copy Services

### FlashCopy

The DS8000 series fully supports the FlashCopy V2 capabilities that the ESS Model 800 currently provides. One function of FlashCopy V2 was the ability to have the source and target of a FlashCopy relationship reside anywhere within the ESS (commonly referred to as cross LSS support). On a DS8300 Model 9A2, the source and target must reside within the *same storage facility image*.

A source volume of a FlashCopy located in one storage facility image cannot have a target volume in the second storage facility image, as illustrated in Figure 3-7.

### Remote mirroring

A Remote Mirror and Copy relationship is supported across storage facility images. The primary server could be located in one storage facility image and the secondary in another storage facility image within the same DS8300.

For more information about Copy Services refer to Chapter 7, “Copy Services” on page 115.

## 3.5 LPAR benefits

The exploitation of the LPAR technology in the DS8300 Model 9A2 offers many potential benefits. You get a reduction in floor space, power requirements, and cooling requirements through consolidation of multiple stand-alone storage functions.

It helps you to simplify your IT infrastructure through a reduced system management effort. You also can reduce your storage infrastructure complexity and your physical asset management.

The hardware-based LPAR implementation ensures data integrity. The fact that you can create dual, independent, completely segregated virtual storage systems helps you to optimize the utilization of your investment, and helps to segregate workloads and protect them from one another.

The following are examples of possible scenarios where storage facility images would be useful:

- ▶ Two production workloads

The production environments can be split, for example, by operating system, application, or organizational boundaries. For example, some customers maintain separate physical ESS 800s with z/OS hosts on one and open hosts on the other. A DS8300 could maintain this isolation within a single physical storage system.

- ▶ Production and development partitions

It is possible to separate the production environment from a development partition. On one partition you can develop and test new applications, completely segregated from a mission-critical production workload running in another storage facility image.

- ▶ Dedicated partition resources

As a service provider you could provide dedicated resources to each customer, thereby satisfying security and service level agreements, while having the environment all contained on one physical DS8300.

- ▶ Production and data mining

For database purposes you can imagine a scenario where your production database is running in the first storage facility image and a copy of the production database is running in the second storage facility image. You can perform analysis and data mining on it without interfering with the production database.

- ▶ Business continuance (secondary) within the same physical array

You can use the two partitions to test Copy Services solutions or you can use them for multiple copy scenarios in a production environment.

- ▶ Information Lifecycle Management (ILM) partition with fewer resources, slower DDMs

One storage facility image can utilize, for example, only fast disk drive modules to ensure high performance for the production environment, and the other storage facility image can use fewer and slower DDMs to ensure Information Lifecycle Management at a lower cost.

Figure 3-8 on page 58 depicts one example for storage facility images in the DS8300.

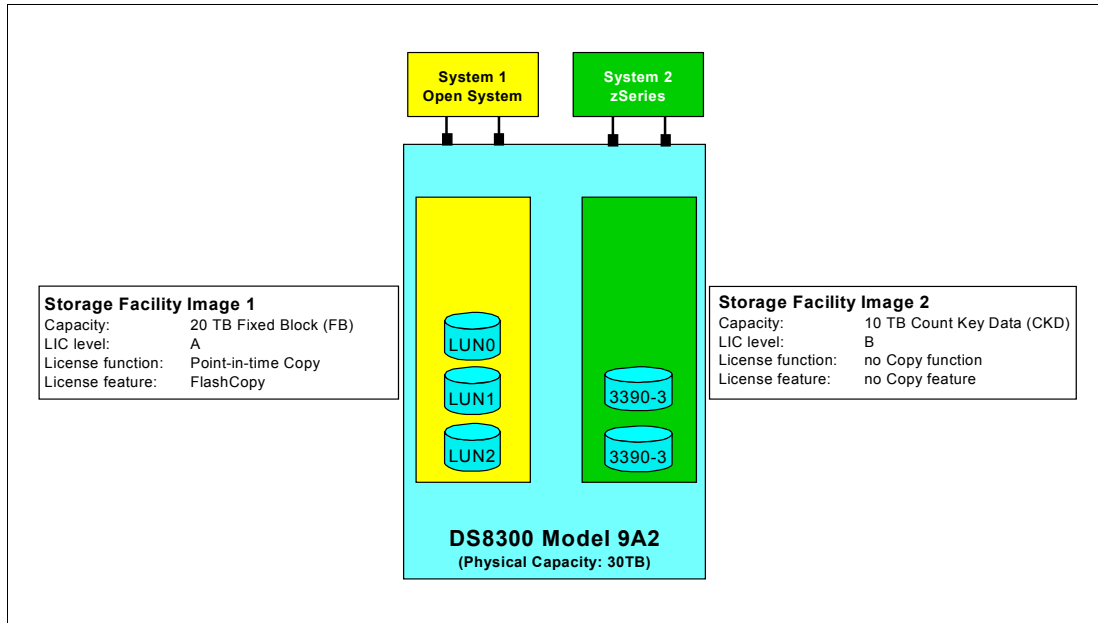


Figure 3-8 Example of storage facility images in the DS8300

This example shows a DS8300 with a total physical capacity of 30 TB. In this case, a minimum Operating Environment License (OEL) is required to cover the 30 TB capacity. The DS8300 is split into two storage facility images. Storage facility image 1 is used for an Open System environment and utilizes 20 TB of fixed block data. Storage facility image 2 is used for a zSeries environment and uses 10 TB of count key data.

To utilize FlashCopy on the entire capacity would require a 30 TB FlashCopy license. However, as in this example, it is possible to have a FlashCopy license for storage facility image 1 for 20 TB only. In this example for the zSeries environment, no copy function is needed, so there is no need to purchase a Copy Services license for storage facility image 2. You can find more information about the licensed functions in 9.3, “DS8000 licensed functions” on page 167.

This example also shows the possibility of running two different licensed internal code (LIC) levels in the storage facility images.

### Addressing capabilities with storage facility images

Figure 3-9 on page 59 highlights the enormous enhancements of the addressing capabilities that you get with the DS8300 in LPAR mode in comparison to the previous ESS Model 800.



## DS8300 addressing capabilities

	ESS 800	DS8300	DS8300 with LPAR
<b>Max Logical Subsystems</b>	<b>32</b>	<b>255</b>	<b>510</b>
<b>Max Logical Devices</b>	<b>8K</b>	<b>63.75K</b>	<b>127.5K</b>
<b>Max Logical CKD Devices</b>	<b>4K</b>	<b>63.75K</b>	<b>127.5K</b>
<b>Max Logical FB Devices</b>	<b>4K</b>	<b>63.75K</b>	<b>127.5K</b>
<b>Max N-Port Logins/Port</b>	<b>128</b>	<b>509</b>	<b>509</b>
<b>Max N-Port Logins</b>	<b>512</b>	<b>8K</b>	<b>16K</b>
<b>Max Logical Paths/FC Port</b>	<b>256</b>	<b>2K</b>	<b>2K</b>
<b>Max Logical Paths/CU Image</b>	<b>256</b>	<b>512</b>	<b>512</b>
<b>Max Path Groups/CU Image</b>	<b>128</b>	<b>256</b>	<b>256</b>

Figure 3-9 Comparison with ESS Model 800 and DS8300 with and without LPAR

## 3.6 Summary

The DS8000 series delivers the first use of the POWER5 processor IBM Virtualization Engine logical partitioning capability. This storage system LPAR technology is designed to enable the creation of two completely separate storage systems, which can run the same or different versions of the licensed internal code. The storage facility images can be used for production, test, or other unique storage environments, and they operate within a single physical enclosure. Each storage facility image can be established to support the specific performance requirements of a different, heterogeneous workload. The DS8000 series robust partitioning implementation helps to isolate and protect the storage facility images. These storage system LPAR capabilities are designed to help simplify systems by maximizing management efficiency, cost effectiveness, and flexibility.





# RAS

This chapter describes the RAS (reliability, availability, serviceability) characteristics of the DS8000. It will discuss:

- ▶ Naming
- ▶ Processor complex RAS
- ▶ Hypervisor: Storage image independence
- ▶ Server RAS
- ▶ Host connection availability
- ▶ Disk subsystem
- ▶ Power and cooling
- ▶ Microcode updates
- ▶ Management console

## 4.1 Naming

It is important to understand the naming conventions used to describe DS8000 components and constructs in order to fully appreciate the discussion of RAS concepts.

### Storage complex

This term describes a group of DS8000s managed by a single Management Console. A storage complex may consist of just a single DS8000 storage unit.

### Storage unit

A storage unit consists of a single DS8000 (including expansion frames). If your organization has one DS8000, then you have a single storage complex that contains a single storage unit.

### Storage facility image

In ESS 800 terms, a storage facility image (SFI) is the entire ESS 800. In a DS8000, an SFI is a union of two logical partitions (LPARs), one from each processor complex. Each LPAR hosts one server. The SFI would have control of one or more device adapter pairs and two or more disk enclosures. Sometimes an SFI might also be referred to as just a storage image.

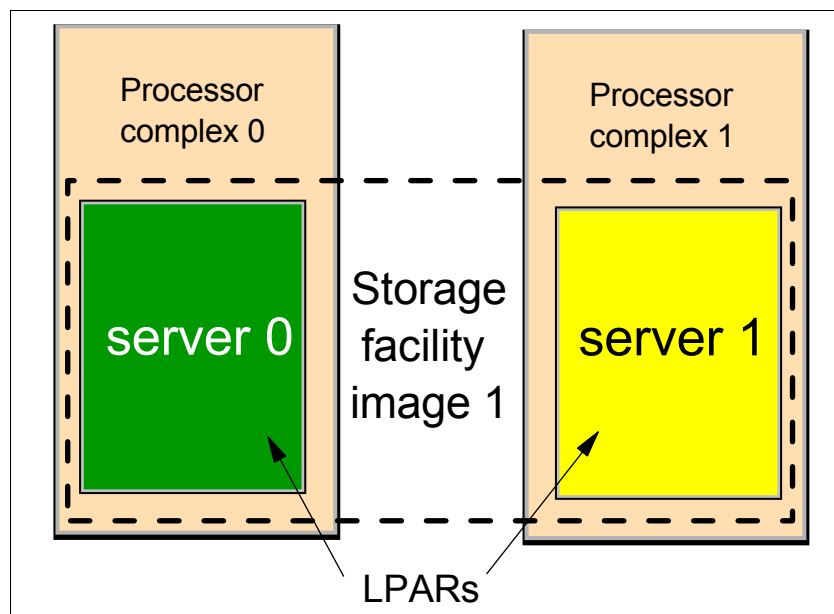


Figure 4-1 Single image mode

In Figure 4-1 server 0 and server 1 create storage facility image 1.

### Logical partitions and servers

In a DS8000, a server is effectively the software that uses a logical partition (an LPAR), and that has access to a percentage of the memory and processor resources available on a processor complex. At GA, this percentage will be either 50% (model 9A2) or 100% (model 921 or 922). In ESS 800 terms, a server is a cluster. So in an ESS 800 we had two servers and one storage facility image per storage unit. However, with a DS8000 we can create logical partitions (LPARs). This allows the creation of four servers, two on each processor complex. One server on each processor complex is used to form a storage image. If there are four servers, there are effectively two separate storage subsystems existing inside one DS8000 storage unit.

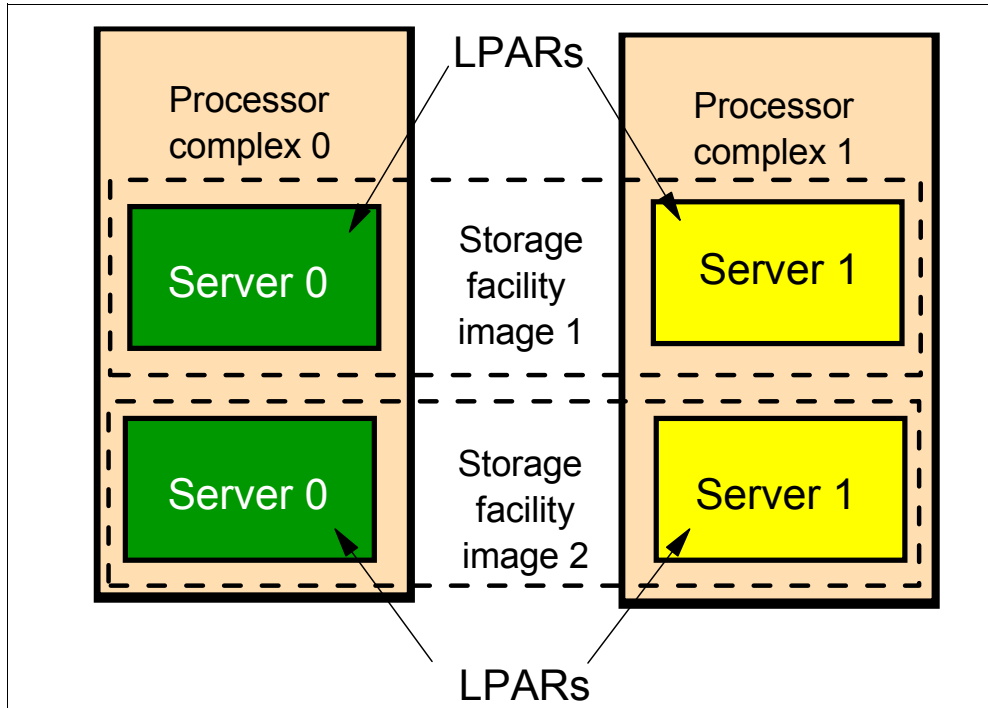


Figure 4-2 Dual image mode

In Figure 4-2 we have two storage facility images (SFIs). The upper server 0 and upper server 1 form SFI 1. The lower server 0 and lower server 1 form SFI 2. In each SFI, server 0 is the darker color (green) and server 1 is the lighter color (yellow). SFI 1 and SFI 2 may share common hardware (the processor complexes) but they are completely separate from an operational point of view.

**Note:** You may think that the lower server 0 and lower server 1 should be called server 2 and server 3. While this may make sense from a numerical point of view (for example, there are four servers so why not number them from 0 to 3), but each SFI is not aware of the other's existence. Each SFI must have a server 0 and a server 1, regardless of how many SFIs or servers there are in a DS8000 storage unit.

### Processor complex

A processor complex is one p5 570 pSeries system unit. Two processor complexes form a redundant pair such that if either processor complex fails, the servers on the remaining processor complex can continue to run the storage image. In an ESS 800, we would have referred to a processor complex as a cluster.

## 4.2 Processor complex RAS

The p5 570 is an integral part of the DS8000 architecture. It is designed to provide an extensive set of reliability, availability, and serviceability (RAS) features that include improved fault isolation, recovery from errors without stopping the processor complex, avoidance of recurring failures, and predictive failure analysis.

## **Reliability, availability, and serviceability**

Excellent quality and reliability are inherent in all aspects of the IBM Server p5 design and manufacturing. The fundamental objective of the design approach is to minimize outages. The RAS features help to ensure that the system performs reliably, and efficiently handles any failures that may occur. This is achieved by using capabilities that are provided by both the hardware, AIX 5L, and RAS code written specifically for the DS8000. The following sections describe the RAS leadership features of IBM Server p5 systems in more detail.

### ***Fault avoidance***

POWER5 systems are built to keep errors from ever happening. This quality-based design includes such features as reduced power consumption and cooler operating temperatures for increased reliability, enabled by the use of copper chip circuitry, SOI (silicon on insulator), and dynamic clock-gating. It also uses mainframe-inspired components and technologies.

### ***First Failure Data Capture***

If a problem should occur, the ability to diagnose it correctly is a fundamental requirement upon which improved availability is based. The p5 570 incorporates advanced capability in start-up diagnostics and in run-time First Failure Data Capture (FFDC) based on strategic error checkers built into the chips.

Any errors that are detected by the pervasive error checkers are captured into Fault Isolation Registers (FIRs), which can be interrogated by the service processor (SP). The SP in the p5 570 has the capability to access system components using special-purpose service processor ports or by access to the error registers.

The FIRs are important because they enable an error to be uniquely identified, thus enabling the appropriate action to be taken. Appropriate actions might include such things as a bus retry, ECC (error checking and correction), or system firmware recovery routines. Recovery routines could include dynamic deallocation of potentially failing components.

Errors are logged into the system non-volatile random access memory (NVRAM) and the SP event history log, along with a notification of the event to AIX for capture in the operating system error log. Diagnostic Error Log Analysis (diagela) routines analyze the error log entries and invoke a suitable action, such as issuing a warning message. If the error can be recovered, or after suitable maintenance, the service processor resets the FIRs so that they can accurately record any future errors.

The ability to correctly diagnose any pending or firm errors is a key requirement before any dynamic or persistent component deallocation or any other reconfiguration can take place.

### **Permanent monitoring**

The SP that is included in the p5 570 provides a way to monitor the system even when the main processor is inoperable. The next subsection offers a more detailed description of the monitoring functions in the p5 570.

#### ***Mutual surveillance***

The SP can monitor the operation of the firmware during the boot process, and it can monitor the operating system for loss of control. This enables the service processor to take appropriate action when it detects that the firmware or the operating system has lost control. Mutual surveillance also enables the operating system to monitor for service processor activity and can request a service processor repair action if necessary.

#### ***Environmental monitoring***

Environmental monitoring related to power, fans, and temperature is performed by the System Power Control Network (SPCN). Environmental critical and non-critical conditions

generate Early Power-Off Warning (EPOW) events. Critical events (for example, a Class 5 AC power loss) trigger appropriate signals from hardware to the affected components to prevent any data loss without operating system or firmware involvement. Non-critical environmental events are logged and reported using Event Scan. The operating system cannot program or access the temperature threshold using the SP.

Temperature monitoring is also performed. If the ambient temperature goes above a preset operating range, then the rotation speed of the cooling fans can be increased. Temperature monitoring also warns the internal microcode of potential environment-related problems. An orderly system shutdown will occur when the operating temperature exceeds a critical level.

Voltage monitoring provides warning and an orderly system shutdown when the voltage is out of operational specification.

### ***Self-healing***

For a system to be self-healing, it must be able to recover from a failing component by first detecting and isolating the failed component. It should then be able to take it offline, fix or isolate it, and then reintroduce the fixed or replaced component into service without any application disruption. Examples include:

- ▶ Bit steering to redundant memory in the event of a failed memory module to keep the server operational
- ▶ Bit scattering, thus allowing for error correction and continued operation in the presence of a complete chip failure (Chipkill™ recovery)
- ▶ Single-bit error correction using ECC without reaching error thresholds for main, L2, and L3 cache memory
- ▶ L3 cache line deletes extended from 2 to 10 for additional self-healing
- ▶ ECC extended to inter-chip connections on fabric and processor bus
- ▶ Memory scrubbing to help prevent soft-error memory faults
- ▶ Dynamic processor deallocation

### ***Memory reliability, fault tolerance, and integrity***

The p5 570 uses Error Checking and Correcting (ECC) circuitry for system memory to correct single-bit memory failures and to detect double-bit. Detection of double-bit memory failures helps maintain data integrity. Furthermore, the memory chips are organized such that the failure of any specific memory module only affects a single bit within a four-bit ECC word (bit-scattering), thus allowing for error correction and continued operation in the presence of a complete chip failure (Chipkill recovery).

The memory DIMMs also utilize memory scrubbing and thresholding to determine when memory modules within each bank of memory should be used to replace ones that have exceeded their threshold of error count (dynamic bit-steering). Memory scrubbing is the process of reading the contents of the memory during idle time and checking and correcting any single-bit errors that have accumulated by passing the data through the ECC logic. This function is a hardware function on the memory controller chip and does not influence normal system memory performance.

### ***N+1 redundancy***

The use of redundant parts, specifically the following ones, allows the p5 570 to remain operational with full resources:

- ▶ Redundant spare memory bits in L1, L2, L3, and main memory
- ▶ Redundant fans
- ▶ Redundant power supplies

### ***Fault masking***

If corrections and retries succeed and do not exceed threshold limits, the system remains operational with full resources and no client or IBM Service Representative intervention is required.

### ***Resource deallocation***

If recoverable errors exceed threshold limits, resources can be deallocated with the system remaining operational, allowing deferred maintenance at a convenient time.

Dynamic deallocation of potentially failing components is non-disruptive, allowing the system to continue to run. Persistent deallocation occurs when a failed component is detected; it is then deactivated at a subsequent reboot.

Dynamic deallocation functions include:

- ▶ Processor
- ▶ L3 cache lines
- ▶ Partial L2 cache deallocation
- ▶ PCI-X bus and slots

Persistent deallocation functions include:

- ▶ Processor
- ▶ Memory
- ▶ Deconfigure or bypass failing I/O adapters
- ▶ L3 cache

Following a hardware error that has been flagged by the service processor, the subsequent reboot of the server invokes extended diagnostics. If a processor or L3 cache has been marked for deconfiguration by persistent processor deallocation, the boot process will attempt to proceed to completion with the faulty device automatically deconfigured. Failing I/O adapters will be deconfigured or bypassed during the boot process.

### ***Concurrent Maintenance***

Concurrent Maintenance provides replacement of the following parts while the processor complex remains running:

- ▶ Disk drives
- ▶ Cooling fans
- ▶ Power Subsystems
- ▶ PCI-X adapter cards

## **4.3 Hypervisor: Storage image independence**

A logical partition (LPAR) is a set of resources on a processor complex that supply enough hardware to support the ability to boot and run an operating system (which we call a server). The LPARs created on a DS8000 processor complex are used to form storage images. These LPARs share not only the common hardware on the processor complex, including CPUs, memory, internal SCSI disks and other media bays (such as DVD-RAM), but also hardware common between the two processor complexes. This hardware includes such things as the I/O enclosures and the adapters installed within them.



A mechanism must exist to allow this sharing of resources in a seamless way. This mechanism is called the *hypervisor*.

The hypervisor provides the following capabilities:

- ▶ Reserved memory partitions allow the setting aside of a certain portion of memory to use as cache and a certain portion to use as NVS.
- ▶ Preserved memory support allows the contents of the NVS and cache memory areas to be protected in the event of a server reboot.
- ▶ The sharing of I/O enclosures and I/O slots between LPARs within one storage image.
- ▶ I/O enclosure initialization control so that when one server is being initialized it doesn't initialize an I/O adapter that is in use by another server.
- ▶ Memory block transfer between LPARs to allow messaging.
- ▶ Shared memory space between I/O adapters and LPARs to allow messaging.
- ▶ The ability of an LPAR to power off an I/O adapter slot or enclosure or force the reboot of another LPAR.
- ▶ Automatic reboot of a frozen LPAR or hypervisor.

### 4.3.1 RIO-G - a self-healing interconnect

The RIO-G interconnect is also commonly called RIO-2. Each RIO-G port can operate at 1 GHz in bidirectional mode and is capable of passing data in each direction on each cycle of the port. This creates a redundant high-speed interconnect that allows servers on either storage complex to access resources on any RIO-G loop. If the resource is not accessible from one server, requests can be routed to the other server to be sent out on an alternate RIO-G port.

### 4.3.2 I/O enclosure

The DS8000 I/O enclosures use hot-swap PCI-X adapters. These adapters are in blind-swap hot-plug cassettes, which allow them to be replaced concurrently. Each slot can be independently powered off for concurrent replacement of a failed adapter, installation of a new adapter, or removal of an old one.

In addition, each I/O enclosure has N+1 power and cooling in the form of two power supplies with integrated fans. The power supplies can be concurrently replaced and a single power supply is capable of supplying DC power to an I/O drawer.

## 4.4 Server RAS

The DS8000 design is built upon IBM's highly redundant storage architecture. It also has the benefit of more than five years of ESS 2105 development. The DS8000 thus employs similar methodology to the ESS to provide data integrity when performing write operations and server failover.

### 4.4.1 Metadata checks

When application data enters the DS8000, special codes or metadata, also known as redundancy checks, are appended to that data. This metadata remains associated with the application data as it is transferred throughout the DS8000. The metadata is checked by various internal components to validate the integrity of the data as it moves throughout the

disk system. It is also checked by the DS8000 before the data is sent to the host in response to a read I/O request. Further, the metadata also contains information used as an additional level of verification to confirm that the data being returned to the host is coming from the desired location on the disk.

## 4.4.2 Server failover and failback

To understand the process of server failover and failback, we have to understand the logical construction of the DS8000. To better understand the contents of this section, you may want to refer to Chapter 10, “The DS Storage Manager - logical configuration” on page 189.

In short, to create logical volumes on the DS8000, we work through the following constructs:

- ▶ We start with DDMs that are installed into pre-defined array sites.
- ▶ These array sites are used to form RAID-5 or RAID-10 arrays.
- ▶ These RAID arrays then become members of a rank.
- ▶ Each rank then becomes a member of an extent pool. Each extent pool has an affinity to either server 0 or server 1. Each extent pool is either open systems FB (fixed block) or zSeries CKD (count key data).
- ▶ Within each extent pool we create logical volumes, which for open systems are called LUNs and for zSeries, 3390 volumes. LUN stands for *logical unit number*, which is used for SCSI addressing. Each logical volume belongs to a logical subsystem (LSS).

For open systems the LSS membership is not that important (unless you are using Copy Services), but for zSeries, the LSS is the logical control unit (LCU) which equates to a 3990 (a zSeries disk controller which the DS8000 emulates). What is important, is that LSSs that have an even identifying number have an affinity with server 0, while LSSs that have an odd identifying number have an affinity with server 1. When a host operating system issues a write to a logical volume, the DS8000 host adapter directs that write to the server that *owns* the LSS of which that logical volume is a member.

If the DS8000 is being used to operate a single storage image then the following examples refer to two servers, one running on each processor complex. If a processor complex were to fail then one server would fail. Likewise, if a server itself were to fail, then it would have the same effect as the loss of the processor complex it runs on.

If, however, the DS8000 is divided into two storage images, then each processor complex will be hosting two servers. In this case, a processor complex failure would result in the loss of two servers. The effect on each server would be identical. The failover processes performed by each storage image would proceed independently.

### Data flow

When a write is issued to a volume, this write normally gets directed to the server that owns this volume. The data flow is that the write is placed into the cache memory of the owning server. The write data is also placed into the NVS memory of the alternate server.

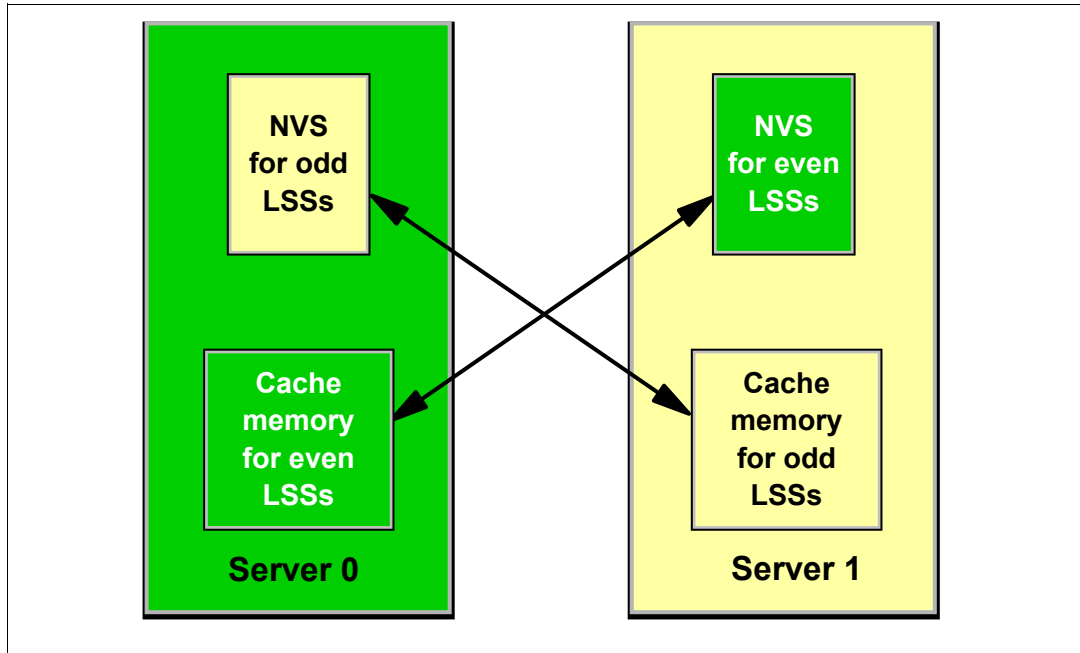


Figure 4-3 Normal data flow

Figure 4-3 illustrates how the cache memory of server 0 is used for all logical volumes that are members of the even LSSs. Likewise, the cache memory of server 1 supports all logical volumes that are members of odd LSSs. But for every write that gets placed into cache, another copy gets placed into the NVS memory located in the alternate server. Thus the normal flow of data for a write is:

1. Data is written to cache memory in the owning server.
2. Data is written to NVS memory of the alternate server.
3. The write is reported to the attached host as having been completed.
4. The write is destaged from the cache memory to disk.
5. The write is discarded from the NVS memory of the alternate server.

Under normal operation, both DS8000 servers are actively processing I/O requests. This section describes the failover and failback procedures that occur between the DS8000 servers when an abnormal condition has affected one of them.

### Failover

In the example depicted in Figure 4-4 on page 70, server 0 has failed. The remaining server has to take over all of its functions. The RAID arrays, because they are connected to both servers, can be accessed from the device adapters used by server 1.

From a data integrity point of view, the real issue is the un-destaged or modified data that belonged to server 1 (that was in the NVS of server 0). Since the DS8000 now has only one copy of that data (which is currently residing in the cache memory of server 1), it will now take the following steps:

1. It destages the contents of its NVS to the disk subsystem.
2. The NVS and cache of server 1 are divided in two, half for the odd LSSs and half for the even LSSs.
3. Server 1 now begins processing the writes (and reads) for *all* the LSSs.

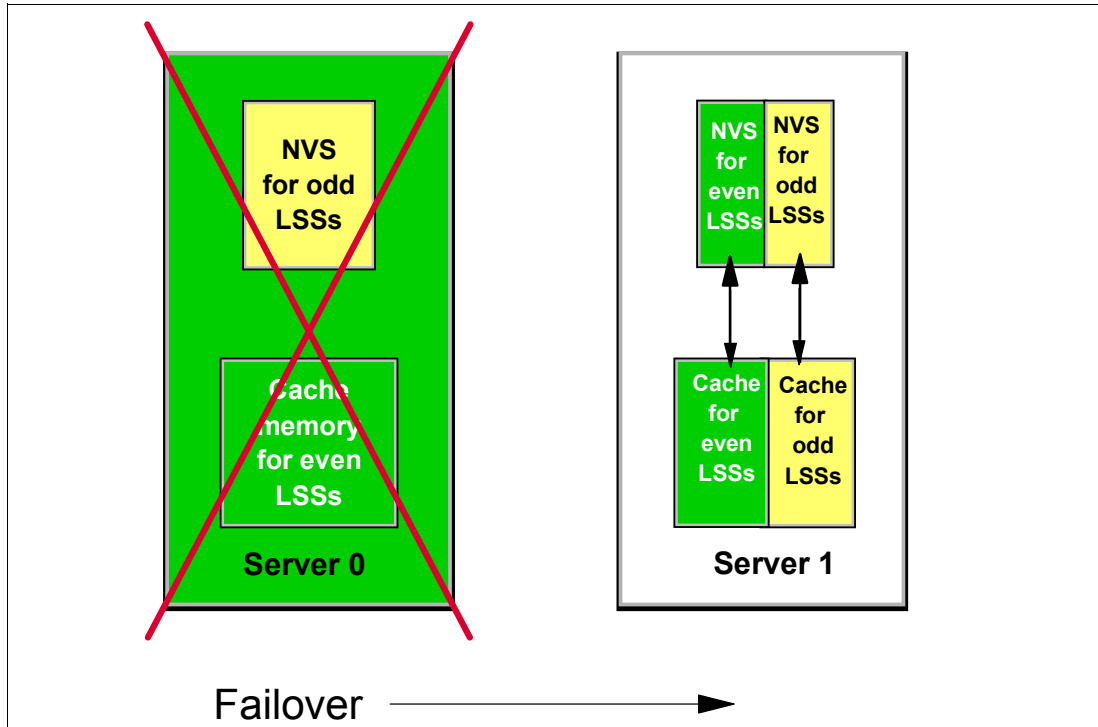


Figure 4-4 Server 0 failing over its function to server 1

This entire process is known as a *failover*. After failover the DS8000 now operates as depicted in Figure 4-4. Server 1 now owns all the LSSs, which means all reads and writes will be serviced by server 1. The NVS inside server 1 is now used for both odd and even LSSs. The entire failover process should be invisible to the attached hosts, apart from the possibility of some temporary disk errors.

### Failback

When the failed server has been repaired and restarted, the failback process is activated. Server 1 starts using the NVS in server 0 again, and the ownership of the even LSSs is transferred back to server 0. Normal operations with both controllers active then resumes. Just like the failover process, the failback process is invisible to the attached hosts.

In general, recovery actions on the DS8000 do not impact I/O operation latency by more than 15 seconds. With certain limitations on configurations and advanced functions, this impact to latency can be limited to 8 seconds. On logical volumes that are not configured with RAID-10 storage, certain RAID-related recoveries may cause latency impacts in excess of 15 seconds. If you have real time response requirements in this area, contact IBM to determine the latest information on how to manage your storage to meet your requirements,

### 4.4.3 NVS recovery after complete power loss

During normal operation, the DS8000 preserves fast writes using the NVS copy in the alternate server. To ensure these fast writes are not lost, the DS8000 contains battery backup units (BBUs). If all the batteries were to fail (which is extremely unlikely since the batteries are in an N+1 redundant configuration), the DS8000 would lose this protection and consequently that DS8000 would take all servers offline. If power is lost to a single primary power supply this does not affect the ability of the other power supply to keep all batteries charged, so all servers would remain online.

The single purpose of the batteries is to preserve the NVS area of server memory in the event of a complete loss of input power to the DS8000. If both power supplies in the base frame were to stop receiving input power, the servers would be informed that they were now running on batteries and immediately begin a shutdown procedure. Unless the power line disturbance feature has been purchased, the BBUs are not used to keep the disks spinning. Even if they do keep spinning, the design is to not move the data from NVS to the FC-AL disk arrays. Instead, each processor complex has a number of internal SCSI disks which are available to store the contents of NVS. When an on-battery condition related shutdown begins, the following events occur:

1. All host adapter I/O is blocked.
2. Each server begins copying its NVS data to internal disk. For each server, two copies are made of the NVS data in that server.
3. When the copy process is complete, each server shuts down AIX.
4. When AIX shutdown in each server is complete (or a timer expires), the DS8000 is powered down.

When power is restored to the DS8000, the following process occurs:

1. The processor complexes power on and perform power on self tests.
2. Each server then begins boot up.
3. At a certain stage in the boot process, the server detects NVS data on its internal SCSI disks and begins to destage it to the FC-AL disks.
4. When the battery units reach a certain level of charge, the servers come online.

An important point is that the servers will not come online until the batteries are fully charged. In many cases, sufficient charging will occur during the power on self test and storage image initialization. However, if a complete discharge of the batteries has occurred, which may happen if multiple power outages occur in a short period of time, then recharging may take up to two hours.

Because the contents of NVS are written to the internal SCSI disks of the DS8000 processor complex and not held in battery protected NVS-RAM, the contents of NVS can be preserved indefinitely. This means that unlike the DS6000 or ESS800, you are not held to a fixed limit of time before power must be restored.

## 4.5 Host connection availability

Each DS8000 Fibre Channel host adapter card provides four ports for connection either directly to a host, or to a Fibre Channel SAN switch.

### Single or multiple path

Unlike the DS6000, the DS8000 does not use the concept of preferred path, since the host adapters are shared between the servers. To show this concept, Figure 4-5 on page 72 depicts a potential machine configuration. In this example, a DS8100 Model 921 has two I/O enclosures (which are enclosures 2 and 3). Each enclosure has four host adapters: two Fibre Channel and two ESCON. I/O enclosure slots 3 and 6 are not depicted because they are reserved for device adapter (DA) cards. If a host were to only have a single path to a DS8000 as shown in Figure 4-5, then it would still be able to access volumes belonging to all LSSs because the host adapter will direct the I/O to the correct server. However, if an error were to occur either on the host adapter (HA), host port (HP), or I/O enclosure, then all connectivity would be lost. Clearly the host bus adapter (HBA) in the attached host is also a single point of failure.

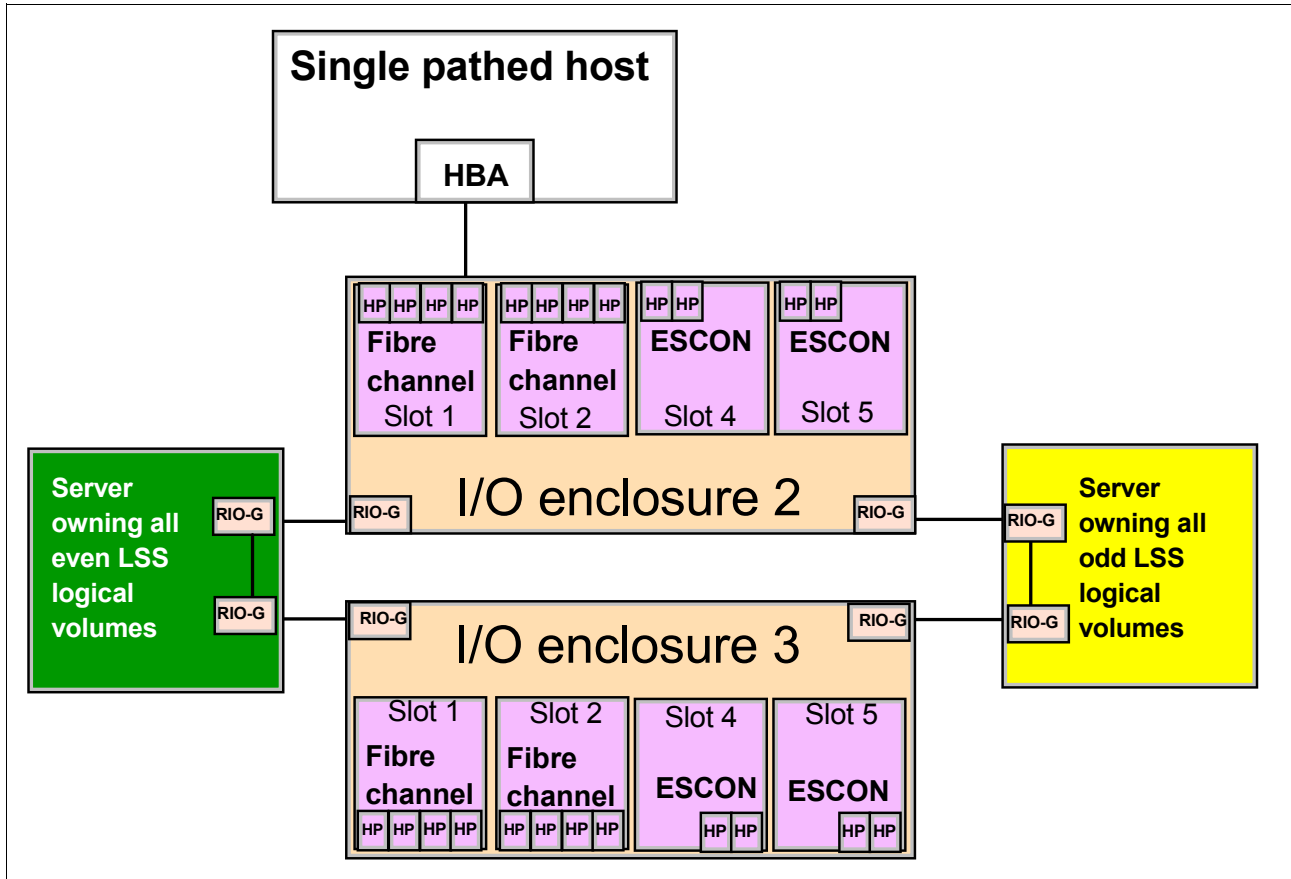


Figure 4-5 Single pathed host

It is always preferable that hosts that access the DS8000 have at least two connections to separate host ports in separate host adapters on separate I/O enclosures, as depicted in Figure 4-6 on page 73. In this example, the host is attached to different Fibre Channel host adapters in different I/O enclosures. This is also important because during a microcode update, an I/O enclosure may need to be taken offline. This configuration allows the host to survive a hardware failure on any component on either path.

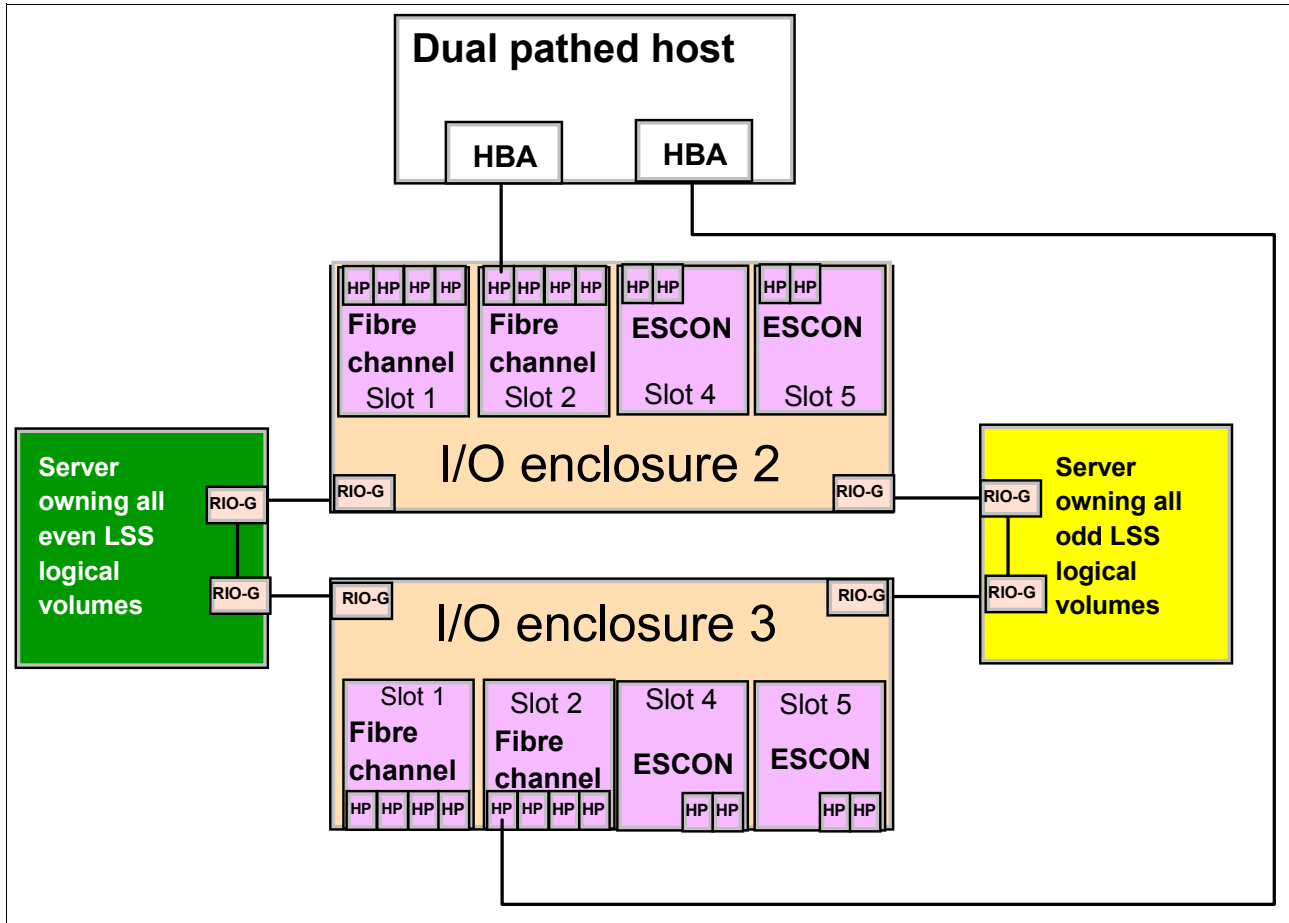


Figure 4-6 Dual pathed host

### SAN/FICON/ESCON switches

Because a large number of hosts may be connected to the DS8000, each using multiple paths, the number of host adapter ports that are available in the DS8000 may not be sufficient to accommodate all the connections. The solution to this problem is the use of SAN switches or directors to switch logical connections from multiple hosts. In a zSeries environment you will need to select a SAN switch or director that also supports FICON. ESCON-attached hosts may need an ESCON director.

A logic or power failure in a switch or director can interrupt communication between hosts and the DS8000. We recommend that more than one switch or director be provided to ensure continued availability. Ports from two different host adapters in two different I/O enclosures should be configured to go through each of two directors. The complete failure of either director leaves half the paths still operating.

### Multi-pathing software

Each attached host operating system now requires a mechanism to allow it to manage multiple paths to the same device, and to preferably load balance these requests. Also, when a failure occurs on one redundant path, then the attached host must have a mechanism to allow it to detect that one path is gone and route all I/O requests for those logical devices to an alternative path. Finally, it should be able to detect when the path has been restored so that the I/O can again be load balanced. The mechanism that will be used varies by attached host operating system and environment as detailed in the next two sections.

## 4.5.1 Open systems host connection

In the majority of open systems environments, IBM strongly recommends the use of the Subsystem Device Driver (SDD) to manage both path failover and preferred path determination. SDD is a software product that IBM supplies free of charge to all customers who use ESS 2105, SAN Volume Controller (SVC), DS6000, or DS8000. There will be a new version of SDD that will also allow SDD to manage pathing to the DS6000 and DS8000 (Version 1.6).

SDD provides availability through automatic I/O path failover. If a failure occurs in the data path between the host and the DS8000, SDD automatically switches the I/O to another path. SDD will also automatically set the failed path back online after a repair is made. SDD also improves performance by sharing I/O operations to a common disk over multiple active paths to distribute and balance the I/O workload. SDD also supports the concept of preferred path for the DS6000 and SVC.

SDD is not available for every supported operating system. Refer to the *IBM TotalStorage DS8000 Host Systems Attachment Guide*, SC26-7628, and the interoperability Web site for direction as to which multi-pathing software may be required. Some devices, such as the IBM SAN Volume Controller (SVC), do not require any multi-pathing software because the internal software in the device already supports multi-pathing. The interoperability Web site is:

<http://www.ibm.com/servers/storage/disk/ds8000/interop.html>

## 4.5.2 zSeries host connection

In the zSeries environment, the normal practice is to provide multiple paths from each host to a disk subsystem. Typically, four paths are installed. The channels in each host that can access each Logical Control Unit (LCU) in the DS8000 are defined in the HCD (hardware configuration definition) or IOCDs (I/O configuration data set) for that host. Dynamic Path Selection (DPS) allows the channel subsystem to select any available (non-busy) path to initiate an operation to the disk subsystem. Dynamic Path Reconnect (DPR) allows the DS8000 to select any available path to a host to reconnect and resume a disconnected operation; for example, to transfer data after disconnection due to a cache miss.

These functions are part of the zSeries architecture and are managed by the channel subsystem in the host and the DS8000.

A physical FICON/ESCON path is established when the DS8000 port sees light on the fiber (for example, a cable is plugged in to a DS8000 host adapter, a processor or the DS8000 is powered on, or a path is configured online by OS/390). At this time, logical paths are established through the port between the host and some or all of the LCUs in the DS8000, controlled by the HCD definition for that host. This happens for each physical path between a zSeries CPU and the DS8000. There may be multiple system images in a CPU. Logical paths are established for each system image. The DS8000 then knows which paths can be used to communicate between each LCU and each host.

### **CUIR**

Control Unit Initiated Reconfiguration (CUIR) prevents loss of access to volumes in zSeries environments due to wrong path handling. This function automates channel path management in zSeries environments, in support of selected DS8000 service actions.

Control Unit Initiated Reconfiguration is available for the DS8000 when operated in the z/OS and z/VM® environments. The CUIR function automates channel path vary on and vary off actions to minimize manual operator intervention during selected DS8000 service actions.



CUIR allows the DS8000 to request that all attached system images set all paths required for a particular service action to the offline state. System images with the appropriate level of software support will respond to such requests by varying off the affected paths, and either notifying the DS8000 subsystem that the paths are offline, or that it cannot take the paths offline. CUIR reduces manual operator intervention and the possibility of human error during maintenance actions, at the same time reducing the time required for the maintenance. This is particularly useful in environments where there are many systems attached to a DS8000.

## 4.6 Disk subsystem

The DS8000 currently supports only RAID-5 and RAID-10. It does not support the non-RAID configuration of disks better known as JBOD (just a bunch of disks).

### 4.6.1 Disk path redundancy

Each DDM in the DS8000 is attached to two 20-port SAN switches. These switches are built into the disk enclosure controller cards. Figure 4-7 illustrates the redundancy features of the DS8000 switched disk architecture. Each disk has two separate connections to the backplane. This allows it to be simultaneously attached to both switches. If either disk enclosure controller card is removed from the enclosure, the switch that is included in that card is also removed. However, the switch in the remaining controller card retains the ability to communicate with all the disks and both device adapters (DAs) in a pair. Equally, each DA has a path to each switch, so it also can tolerate the loss of a single path. If both paths from one DA fail, then it cannot access the switches; however, the other DA retains connection.

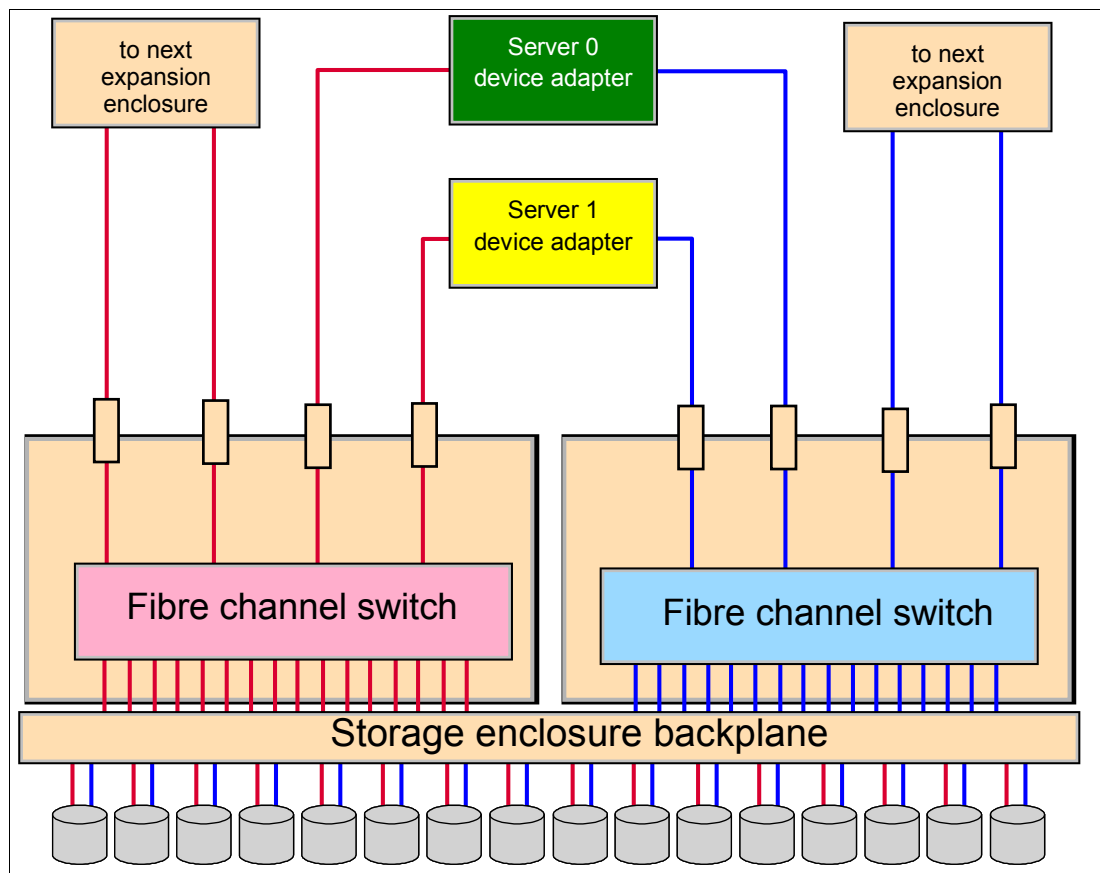


Figure 4-7 Switched disk connections

Figure 4-7 also shows the connection paths for expansion on the far left and far right. The paths from the switches travel to the switches in the next disk enclosure. Because expansion is done in this linear fashion, the addition of more enclosures is completely non-disruptive.

## 4.6.2 RAID-5 overview

RAID-5 is one of the most commonly used forms of RAID protection.

### RAID-5 theory

The DS8000 series supports RAID-5 arrays. RAID-5 is a method of spreading volume data plus parity data across multiple disk drives. RAID-5 provides faster performance by striping data across a defined set of DDMs. Data protection is provided by the generation of parity information for every stripe of data. If an array member fails, then its contents can be regenerated by using the parity data.

### RAID-5 implementation in the DS8000

In a DS8000, a RAID-5 array built on one array site will contain either seven or eight disks depending on whether the array site is supplying a spare. A seven-disk array effectively uses one disk for parity, so it is referred to as a 6+P array (where the P stands for parity). The reason only 7 disks are available to a 6+P array is that the eighth disk in the array site used to build the array was used as a spare. This we then refer to as a 6+P+S array site (where the S stands for spare). An 8-disk array also effectively uses 1 disk for parity, so it is referred to as a 7+P array.

### Drive failure

When a disk drive module fails in a RAID-5 array, the device adapter starts an operation to reconstruct the data that was on the failed drive onto one of the spare drives. The spare that is used will be chosen based on a smart algorithm that looks at the location of the spares and the size and location of the failed DDM. The rebuild is performed by reading the corresponding data and parity in each stripe from the remaining drives in the array, performing an exclusive-OR operation to recreate the data, then writing this data to the spare drive.

While this data reconstruction is going on, the device adapter can still service read and write requests to the array from the hosts. There may be some degradation in performance while the sparing operation is in progress because some DA and switched network resources are being used to do the reconstruction. Due to the switch-based architecture, this effect will be minimal. Additionally, any read requests for data on the failed drive requires data to be read from the other drives in the array and then the DA performs an operation to reconstruct the data.

Performance of the RAID-5 array returns to normal when the data reconstruction onto the spare device completes. The time taken for sparing can vary, depending on the size of the failed DDM and the workload on the array, the switched network, and the DA. The use of arrays across loops (AAL) both speeds up rebuild time and decreases the impact of a rebuild.

## 4.6.3 RAID-10 overview

RAID-10 is not as commonly used as RAID-5, mainly because more raw disk capacity is needed for every GB of effective capacity.

## **RAID-10 theory**

RAID-10 provides high availability by combining features of RAID-0 and RAID-1. RAID-0 optimizes performance by striping volume data across multiple disk drives at a time. RAID-1 provides disk mirroring, which duplicates data between two disk drives. By combining the features of RAID-0 and RAID-1, RAID-10 provides a second optimization for fault tolerance. Data is striped across half of the disk drives in the RAID-1 array. The same data is also striped across the other half of the array, creating a mirror. Access to data is preserved if one disk in each mirrored pair remains available. RAID-10 offers faster data reads and writes than RAID-5 because it does not need to manage parity. However, with half of the DDMs in the group used for data and the other half to mirror that data, RAID-10 disk groups have less capacity than RAID-5 disk groups.

## **RAID-10 implementation in the DS8000**

In the DS8000 the RAID-10 implementation is achieved using either six or eight DDMs. If spares exist on the array site, then six DDMs are used to make a three-disk RAID-0 array which is then mirrored. If spares do not exist on the array site then eight DDMs are used to make a four-disk RAID-0 array which is then mirrored.

## **Drive failure**

When a disk drive module (DDM) fails in a RAID-10 array, the controller starts an operation to reconstruct the data from the failed drive onto one of the hot spare drives. The spare that is used will be chosen based on a smart algorithm that looks at the location of the spares and the size and location of the failed DDM. Remember a RAID-10 array is effectively a RAID-0 array that is mirrored. Thus when a drive fails in one of the RAID-0 arrays, we can rebuild the failed drive by reading the data from the equivalent drive in the other RAID-0 array.

While this data reconstruction is going on, the DA can still service read and write requests to the array from the hosts. There may be some degradation in performance while the sparing operation is in progress because some DA and switched network resources are being used to do the reconstruction. Due to the switch-based architecture of the DS8000, this effect will be minimal. Read requests for data on the failed drive should not be affected because they can all be directed to the good RAID-1 array.

Write operations will not be affected. Performance of the RAID-10 array returns to normal when the data reconstruction onto the spare device completes. The time taken for sparing can vary, depending on the size of the failed DDM and the workload on the array and the DA.

## **Arrays across loops**

The DS8000 implements the concept of arrays across loops (AAL). With AAL, an array site is actually split into two halves. Half of the site is located on the first disk loop of a DA pair and the other half is located on the second disk loop of that DA pair. It is implemented primarily to maximize performance. However, in RAID-10 we are able to take advantage of AAL to provide a higher level of redundancy. The DS8000 RAS code will deliberately ensure that one RAID-0 array is maintained on each of the two loops created by a DA pair. This means that in the extremely unlikely event of a complete loop outage, the DS8000 would not lose access to the RAID-10 array. This is because while one RAID-0 array is offline, the other remains available to service disk I/O.

### **4.6.4 Spare creation**

When the array sites are created on a DS8000, the DS8000 microcode determines which sites will contain spares. The first four array sites will normally each contribute one spare to the DA pair, with two spares being placed on each loop. In general, each device adapter pair will thus have access to four spares.

On the ESS 800 the spare creation policy was to have four DDMs on each SSA loop for each DDM type. This meant that on a specific SSA loop it was possible to have 12 spare DDMs if you chose to populate a loop with three different DDM sizes. With the DS8000 the intention is to not do this. A minimum of one spare is created for each array site defined until the following conditions are met:

- ▶ A minimum of 4 spares per DA pair
- ▶ A minimum of 4 spares of the largest capacity array site on the DA pair
- ▶ A minimum of 2 spares of capacity and RPM greater than or equal to the fastest array site of any given capacity on the DA pair

### **Floating spares**

The DS8000 implements a smart floating technique for spare DDMs. On an ESS 800, the spare *floats*. This means that when a DDM fails and the data it contained is rebuilt onto a spare, then when the disk is replaced, the replacement disk becomes the spare. The data is not migrated to another DDM, such as the DDM in the original position the failed DDM occupied. So in other words, on an ESS 800 there is no post repair processing.

The DS8000 microcode may choose to allow the hot spare to remain where it has been *moved*, but it may instead choose to *migrate* the spare to a more optimum position. This will be done to better balance the spares across the DA pairs, the loops, and the enclosures. It may be preferable that a DDM that is currently in use as an array member be converted to a spare. In this case the data on that DDM will be migrated in the background onto an existing spare. This process does not *fail* the disk that is being migrated, though it does reduce the number of available spares in the DS8000 until the migration process is complete.

A smart process will be used to ensure that the larger or higher RPM DDMs always act as spares. This is preferable because if we were to rebuild the contents of a 146 GB DDM onto a 300 GB DDM, then approximately half of the 300 GB DDM will be wasted since that space is not needed. The problem here is that the failed 146 GB DDM will be replaced with a new 146 GB DDM. So the DS8000 microcode will most likely migrate the data back onto the recently replaced 146 GB DDM. When this process completes, the 146 GB DDM will rejoin the array and the 300 GB DDM will become the spare again. Another example would be if we fail a 73 GB 15k RPM DDM onto a 146 GB 10k RPM DDM. This means that the data has now moved to a slower DDM, but the replacement DDM will be the same as the failed DDM. This means the array will have a mix of RPMs. This is not desirable. Again, a smart migrate of the data will be performed once suitable spares have become available.

### **Hot pluggable DDMs**

Replacement of a failed drive does not affect the operation of the DS8000 because the drives are fully hot pluggable. Due to the fact that each disk plugs into a switch, there is no loop break associated with the removal or replacement of a disk. In addition there is no potentially disruptive loop initialization process.

## **4.6.5 Predictive Failure Analysis® (PFA)**

The drives used in the DS8000 incorporate Predictive Failure Analysis (PFA) and can anticipate certain forms of failures by keeping internal statistics of read and write errors. If the error rates exceed predetermined threshold values, the drive will be nominated for replacement. Because the drive has not yet failed, data can be copied directly to a spare drive. This avoids using RAID recovery to reconstruct all of the data onto the spare drive.

## 4.6.6 Disk scrubbing

The DS8000 will periodically read all sectors on a disk. This is designed to occur without any interference with application performance. If ECC-correctable bad bits are identified, the bits are corrected immediately by the DS8000. This reduces the possibility of multiple bad bits accumulating in a sector beyond the ability of ECC to correct them. If a sector contains data that is beyond ECC's ability to correct, then RAID is used to regenerate the data and write a new copy onto a spare sector of the disk. This scrubbing process applies to both array members and spare DDMs.

## 4.7 Power and cooling

The DS8000 has completely redundant power and cooling. Every power supply and cooling fan in the DS8000 operates in what is known as N+1 mode. This means that there is always at least one more power supply, cooling fan, or battery than is required for normal operation. In most cases this simply means duplication.

### Primary power supplies

Each frame has two primary power supplies (PPS). Each PPS produces voltages for two different areas of the machine:

- ▶ 208V is produced to be supplied to each I/O enclosure and each processor complex. This voltage is placed by each supply onto two redundant power buses.
- ▶ 12V and 5V is produced to be supplied to the disk enclosures.

If either PPS fails, the other can continue to supply all required voltage to all power buses in that frame. The PPS can be replaced concurrently.

**Important:** It should be noted that if you install the DS8000 such that both primary power supplies are attached to the same circuit breaker or the same switch board, then the DS8000 will not be well protected from external power failures. This is a very common cause of unplanned outages.

### Battery backup units

Each frame with I/O enclosures, or every frame if the power line disturbance feature is installed, will have battery backup units (BBU). Each BBU can be replaced concurrently, provided no more than one BBU is unavailable at any one time. The DS8000 BBUs have a planned working life of at least four years.

### Rack cooling fans

Each frame has a cooling fan plenum located above the disk enclosures. The fans in this plenum draw air from the front of the DDMs and then move it out through the top of the frame. There are multiple redundant fans in each enclosure. Each fan can be replaced concurrently.

### Rack power control card (RPC)

The rack power control cards are part of the power management infrastructure of the DS8000. There are two RPC cards for redundancy. Each card can independently control power for the entire DS8000.

## 4.7.1 Building power loss

The DS8000 uses an area of server memory as non-volatile storage (NVS). This area of memory is used to hold data that has not been written to the disk subsystem. If building power were to fail, where both primary power supplies (PPSs) in the base frame were to report a loss of AC input power, then the DS8000 must take action to protect that data.

## 4.7.2 Power fluctuation protection

The DS8000 base frame contains battery backup units that are intended to protect modified data in the event of a complete power loss. If a power fluctuation occurs that causes a momentary interruption to power (often called a brownout) then the DS8000 will tolerate this for approximately 30ms. If the power line disturbance feature is not present on the DS8000, then after that time, the DDMs will stop spinning and the servers will begin copying the contents of NVS to the internal SCSI disks in the processor complexes. For many customers who use UPS (uninterruptible power supply) technology, this is not an issue. UPS-regulated power is in general very reliable, so additional redundancy in the attached devices is often completely unnecessary.

If building power is not considered reliable then the addition of the extended power line disturbance feature should be considered. This feature adds two separate pieces of hardware to the DS8000:

1. For each primary power supply in each frame of the DS8000, a booster module is added that converts 208V battery power into 12V and 5V. This is to supply the DDMs with power directly from the batteries. The PPSs do not normally receive power from the BBUs.
2. Batteries will be added to expansion racks that did not already have them. Base racks and expansion racks with I/O enclosures get batteries by default. Expansion racks that do not have I/O enclosures normally do not get batteries.

With the addition of this hardware, the DS8000 will be able to run for up to 50 seconds on battery power, before the servers begin to copy NVS to SCSI disk and then shutdown. This would allow for a 50 second interruption to building power with no outage to the DS8000.

## 4.7.3 Power control of the DS8000

Unlike the ESS 800, the DS8000 does not possess a white power switch to turn the DS8000 storage unit off and on. All power sequencing is done via the Service Processor Control Network (SPCN) and RPCs. If the user wishes to power the DS8000 off, they must do so using the management tools provided by the Storage Hardware Management Console (S-HMC). If the S-HMC is not functional, then it will not be possible to control the power sequencing of the DS8000 until the S-HMC function is restored. This is one of the benefits that is gained by purchasing a redundant S-HMC.

## 4.7.4 Emergency power off (EPO)

Each DS8000 frame has an emergency power off switch. This button is intended purely to remove power from the DS8000 in the following extreme cases:

- ▶ The DS8000 has developed a fault which is placing the environment at risk, such as a fire.
- ▶ The DS8000 is placing human life at risk, such as the electrocution of a service representative.

Apart from these two contingencies (which are highly unlikely), the EPO switch should never be used. The reason for this is that the DS8000 NVS storage area is not directly protected by batteries. If building power is lost, the DS8000 can use its internal batteries to destage the

data from NVS memory to a variably sized disk area to preserve that data until power is restored. However, the EPO switch does not allow this destage process to happen and all NVS data is lost. This will most likely result in data loss.

If you need to power the DS8000 off for building maintenance, or to relocate it, you should always use the S-HMC to achieve this.

## 4.8 Microcode updates

The DS8000 contains many discrete redundant components. Most of these components have firmware that can be updated. This includes the processor complexes, device adapters, and host adapters. Each DS8000 server also has an operating system (AIX) and Licensed Internal Code (LIC) that can be updated. As IBM continues to develop and improve the DS8000, new releases of firmware and LIC will become available to offer improvements in both function and reliability.

The architecture of the DS8000 allows for concurrent code updates. This is achieved by using the redundant design of the DS8000. In general, redundancy is lost for a short period as each component in a redundant pair is updated.

The S-HMC can hold up to six different versions of code. Each server can hold three different versions of code (the previous version, the active version, and the next version).

### Installation process

The installation process involves several stages.

1. The S-HMC code will be updated. The new code version will be supplied on CD or downloaded via FTP. This may potentially involve updates to the internal Linux version of the S-HMC, updates to the S-HMC LIC, and updates to the firmware of the S-HMC hardware.
2. New DS8000 LIC will be loaded onto the S-HMC and from there to the internal storage of each server.
3. Occasionally, new PPS and RPC firmware may be released. New firmware can be loaded into each Rack Power Control (RPC) card and Primary Power Supply (PPS) directly from the S-HMC. Each RPC and PPS would be quiesced, updated, and resumed one at a time until all have been updated.
4. Occasionally, new firmware for the hypervisor, service processor, system planar, and I/O enclosure planars may be released. This firmware can be loaded into each device directly from the S-HMC. Activation of this firmware may require each processor complex to be shut down and rebooted, one at a time. This would cause each server on each processor complex to fail over its logical subsystems to the server on the other processor complex. Certain updates may not require this step, or it may occur without processor reboots.
5. Updates to the server operating system (currently AIX 5.2) plus updates to the internal LIC will be performed. Every server in a storage image would be updated one at a time. Each update would cause each server to fail over its logical subsystems to its partner server on the other processor complex. This process would also update the firmware running in each device adapter owned by that server.
6. Updates to the host adapters will be performed. For FICON/FCP adapters, these updates will impact each adapter for less than 2.5 seconds, and should not affect connectivity. If an update were to take longer than this, multi-pathing software on the host, or CUIR (for ESCON and FICON), will be used to direct I/O to a different host adapter.

While the installation process described above may seem complex, it will not require a great deal of user intervention. The code installer will normally just start the process and then monitor its progress using the S-HMC.

### **S-HMC considerations**

Before updating the DS8000 code, the Storage Hardware Management Consoles should be updated to the latest code version. This could usually be done at any time prior to the change window, since the DS8000 can continue to operate without an S-HMC, provided no Copy Services or configuration changes are planned while it is unavailable. If you have two S-HMCs you could perform Copy Services operations or configuration changes using the second S-HMC.

### **Different code versions across storage images**

If the DS8000 is partitioned into multiple storage images, it is possible to run each storage image on a different LIC version. This could be beneficial if one image was being used as a production image, while the other was being used for testing.

If this was desired, steps one to four would still affect both storage images since they affect common hardware. However, it is possible to then perform steps 5 and 6 on only one storage image and leave the other storage image downlevel. This situation could then be left until a testing regimen has been performed. At some point the downlevel image could then be updated by performing just steps 5 and 6 on that image.

## **4.9 Management console**

The DS8000 management network consists of redundant Ethernet switches and redundant Storage Hardware Management (S-HMC) consoles.

### **S-HMC**

The S-HMC is used to perform configuration, management, and maintenance activities on the DS8000. It can be ordered to be located either physically inside the base frame or external for mounting in a customer-supplied rack.

If the S-HMC is not operational then it is not possible to perform maintenance, power the DS8000 up or down, or perform Copy Services tasks such as the establishment of FlashCopies. It is thus recommended to order two management consoles to act as a redundant pair.

### **Ethernet switches**

Each DS8000 base frame contains two 16-port Ethernet switches. Two switches are supplied to allow the creation of a fully redundant management network. Each server in the DS8000 has a connection to each switch. Each S-HMC also has a connection to each switch. This means that should a single Ethernet switch fail, all traffic can successfully travel from either S-HMC to any server in the storage unit using the alternate switch.

## **4.10 Summary**

This chapter has described the RAS characteristics of the DS8000. These characteristics combine to make the DS8000 a world leader in reliability, availability, and serviceability.





## Virtualization concepts

This chapter describes virtualization concepts as they apply to the DS8000. In particular, it covers the following topics:

- ▶ Storage system virtualization
- ▶ Abstraction layers for disk virtualization
  - Array sites
  - Arrays
  - Ranks
  - Extent pools
  - Logical volumes
  - Logical storage subsystems
  - Address groups
  - Volume groups
  - Host attachments

## 5.1 Virtualization definition

In our fast-paced world, where you have to react quickly to changing business conditions, your infrastructure must allow for on demand changes. *Virtualization* is key to an on demand infrastructure. However, when talking about virtualization many vendors are talking about different things.

One important feature of the DS8000 is the virtualization of a whole storage subsystem. If you have to run different workloads, for example a service provider might run workloads for different banks, then it might be desirable to completely separate the workloads. This could be done on the processor side with IBM's LPAR technology; the same technology is now also available for a storage subsystem, the IBM TotalStorage DS8000 system.

Another definition of virtualization is the abstraction process going from the physical disk drives to a logical volume that the hosts and servers *see as if it were* a physical disk.

## 5.2 Storage system virtualization

IBM has a long history and experience in virtualization. This goes back several decades, when virtual memory was introduced in the mid 1960s in the operating system. At that time IBM also developed a system that could virtualize a whole processor complex – including the processor, memory, and devices. This operating system was called Virtual Machine (VM). Later on this functionality also became available as a hardware function on S/390 processors. When the processor complex was run in Logical Partition Mode (LPAR), several operating systems could run isolated and independently on the same hardware base.

This LPAR capability heritage from S/390 and zSeries has now become available on the POWER5 pSeries. The DS8000 is based on POWER5 technology, so we can take advantage of its functions, including the LPAR functionality.

The DS8300 Model 9A2 supports LPAR mode. In the current implementation, you can run up to two logical partitions on a physical storage system unit. In each partition you can run a storage facility image. A storage facility image is a virtual storage subsystem with its own copy of Licensed Internal Code (LIC), which consists of the AIX kernel and the functional code. Both storage facility images share the physical hardware and the LPAR hypervisor manages this sharing of the hardware. Currently, however, there are some limitations on the granularity of how the physical resources like processors, memory, cache, and I/O can be split between the LPARs. See Chapter 3, “Storage system LPARs (Logical partitions)” on page 43 for details.

Like in non-LPAR mode, where there are two SMPs running an AIX kernel and forming a storage complex with two servers, server0 and server1, a storage facility image is a storage complex of its own, but since it does not own the physical hardware (the storage unit), you can think of it as a virtual storage system. Each storage facility image has a server 0 and a server 1. Each storage facility image can run its own version of Licensed Internal Code. The storage facility images are totally separated by the LPAR hypervisor. Disk drives and arrays are owned by one or the other storage facility, they cannot be shared.

Figure 5-1 on page 85 illustrates the LPAR concept.

In the following section, when we talk about server 0 or server 1 we could also mean server 0 or server 1 of a storage facility image running in an LPAR.

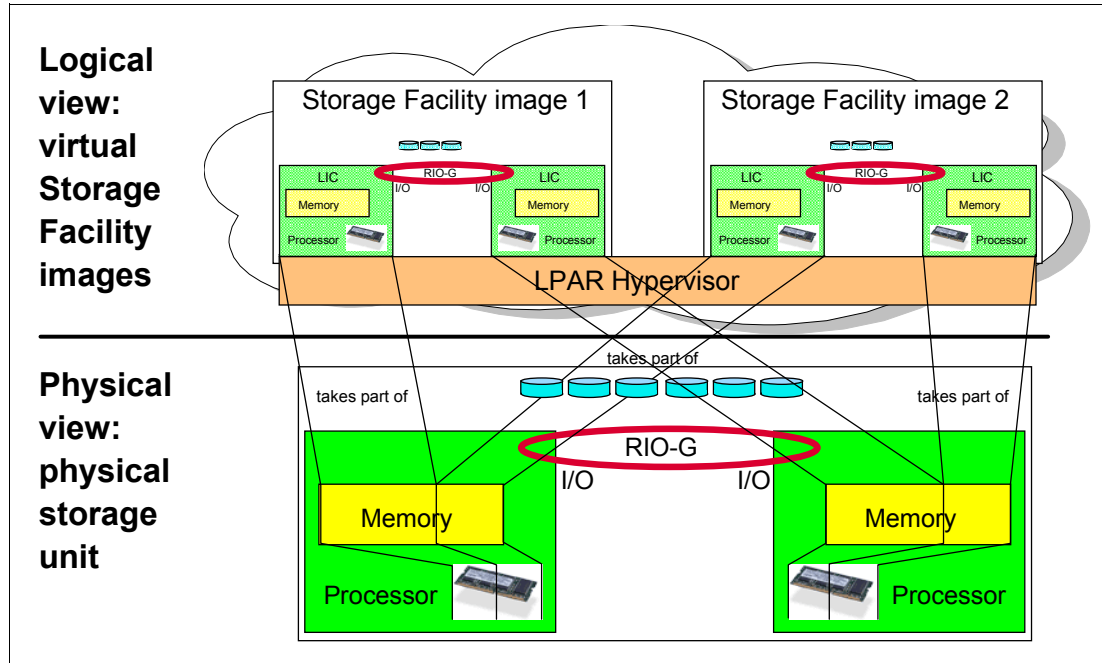


Figure 5-1 Storage Facility virtualization

### 5.3 The abstraction layers for disk virtualization

In this chapter, when talking about virtualization, we are talking about the process of preparing a bunch of physical disk drives (DDMs) to be something that can be used from an operating system, which means we are talking about the creation of LUNs.

The DS8000 is populated with switched FC-AL disk drives that are mounted in disk enclosures. You order disk drives in groups of 16 drives of the same capacity and RPM. The disk drives can be accessed by a pair of device adapters. Each device adapter has four paths to the disk drives. The four paths provide two FC-AL device interfaces, each with two paths, such that either path can be used to communicate with any disk drive on that device interface (in other words, the paths are redundant). One device interface from each device adapter is connected to a set of FC-AL devices such that either device adapter has access to any disk drive through two independent switched fabrics (in other words, the device adapters and switches are redundant).

Each device adapter has four ports and since device adapters operate in pairs, there are eight ports or paths to the disk drives. All eight paths can operate concurrently and could access all disk drives on the attached fabric. In normal operation, however, disk drives are typically accessed by one device adapter. Which device adapter owns the disk is defined during the logical configuration process. This avoids any contention between the two device adapters for access to the disks.

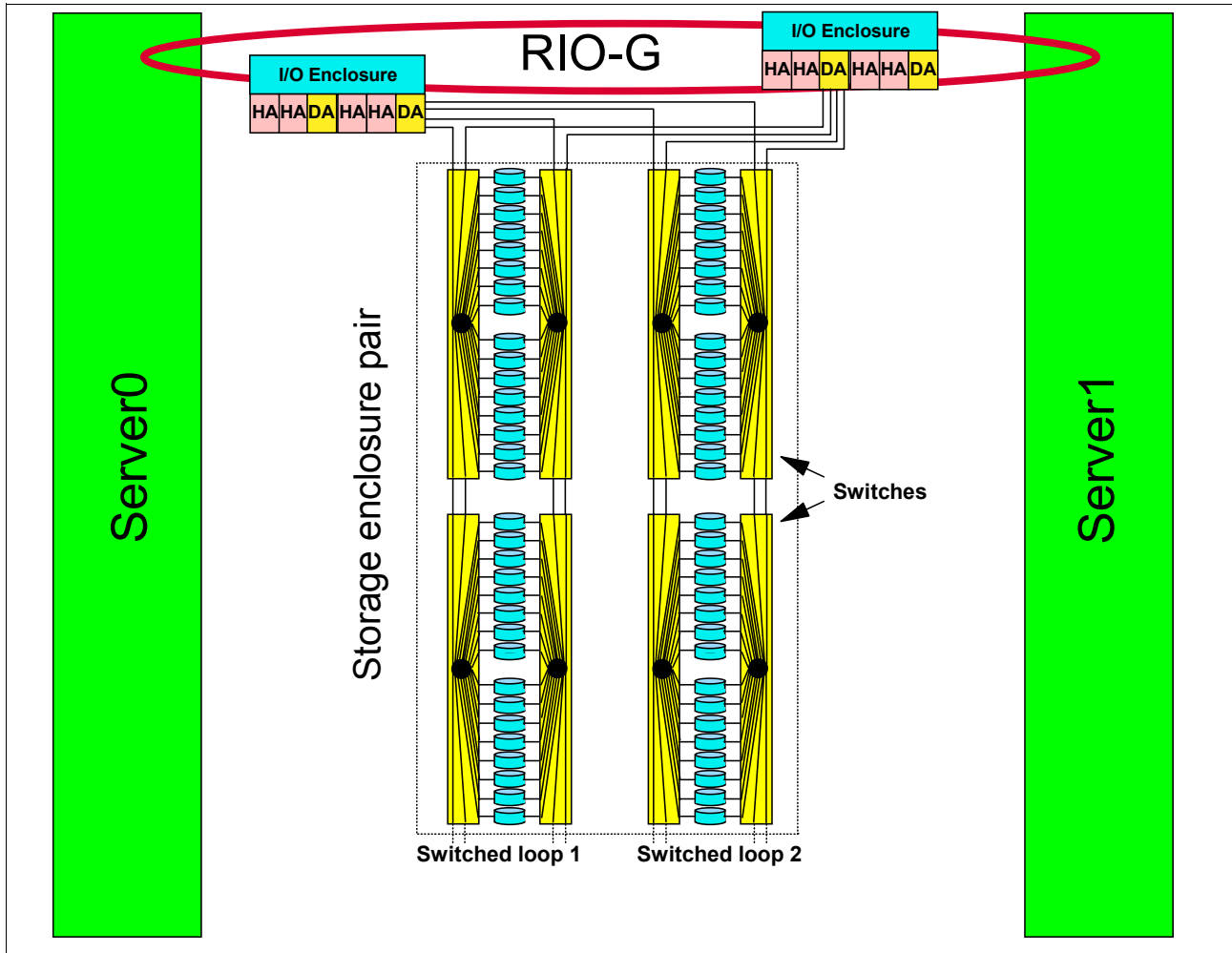


Figure 5-2 Physical layer as the base for virtualization

Figure 5-2 shows the physical layer on which virtualization is based.

Compare this with the ESS design, where there was a real loop and having an 8-pack close to a device adapter was an advantage. This is no longer relevant for the DS8000. Because of the switching design, each drive is in close reach of the device adapter, apart from a few more hops through the Fibre Channel switches for some drives. So, it is not really a loop, but a switched FC-AL loop with the FC-AL addressing schema: Arbitrated Loop Physical Addressing (AL-PA).

### 5.3.1 Array sites

An array site is a group of eight DDMs. What DDMs make up an array site is pre-determined by the DS8000, but note, that there is no pre-determined server affinity for array sites. The DDMs selected for an array site are chosen from two disk enclosures on different loops (see Figure 5-3 on page 87).

The DDMs in the array site are of the same DDM type, which means the same capacity and the same speed (RPM).

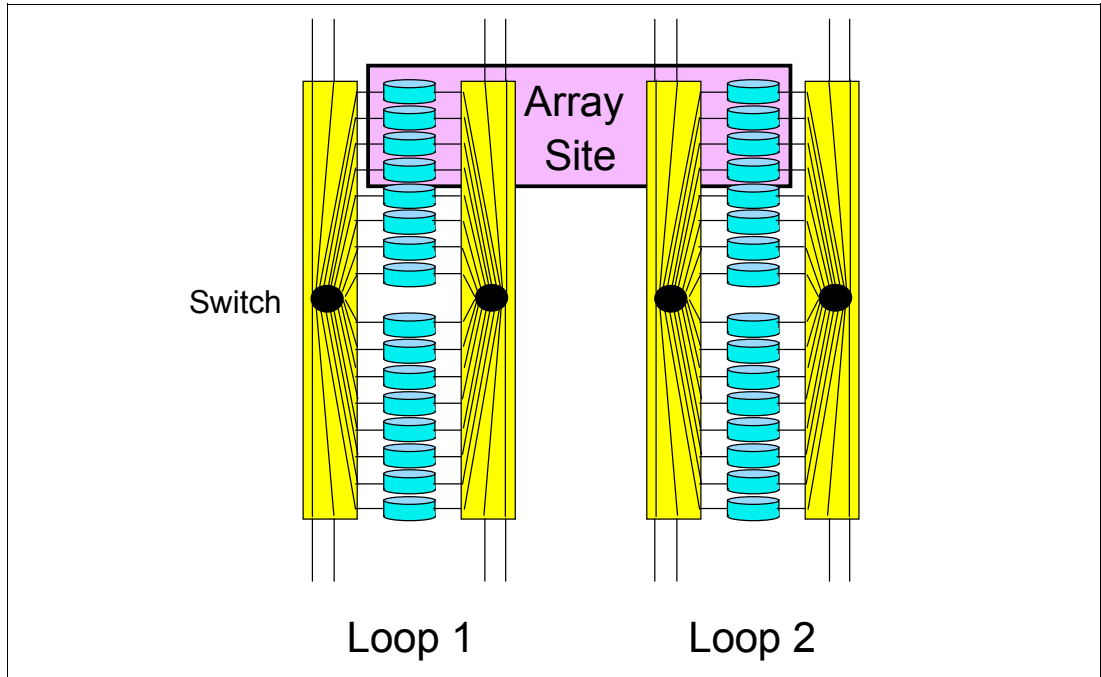


Figure 5-3 Array site

As you can see from Figure 5-3, array sites span loops. Four DDMs are taken from loop 1 and another four DDMs are taken from loop 2.

Array sites are the building blocks used to define arrays.

### 5.3.2 Arrays

An *array* is created from one *array site*. Forming an array means defining it for a specific RAID type. The supported RAID types are RAID-5 and RAID-10 (see 4.6.2, “RAID-5 overview” on page 76 and 4.6.3, “RAID-10 overview” on page 76). For each array site you can select a RAID type. The process of selecting the RAID type for an array is also called *defining* an array.

**Note:** In the DS8000 current implementation, one array is defined using one array site.

According to the DS8000 sparing algorithm, from zero to two spares may be taken from the array site. This is discussed further in Chapter 6, “IBM TotalStorage DS8000 model overview and scalability” on page 103.

Figure 5-4 on page 88 shows the creation of a RAID-5 array with one spare, also called a 6+P+S array (capacity of 6 DDMs for data, capacity of one DDM for parity, and a spare drive). According to the RAID-5 rules, parity is distributed across all seven drives in this example.

On the right-hand side in Figure 5-4 the terms D1, D2, D3, and so on, stand for the set of data contained on one disk within a stripe on the array. If, for example, 1 GB of data is written, it is distributed across all the disks of the array.

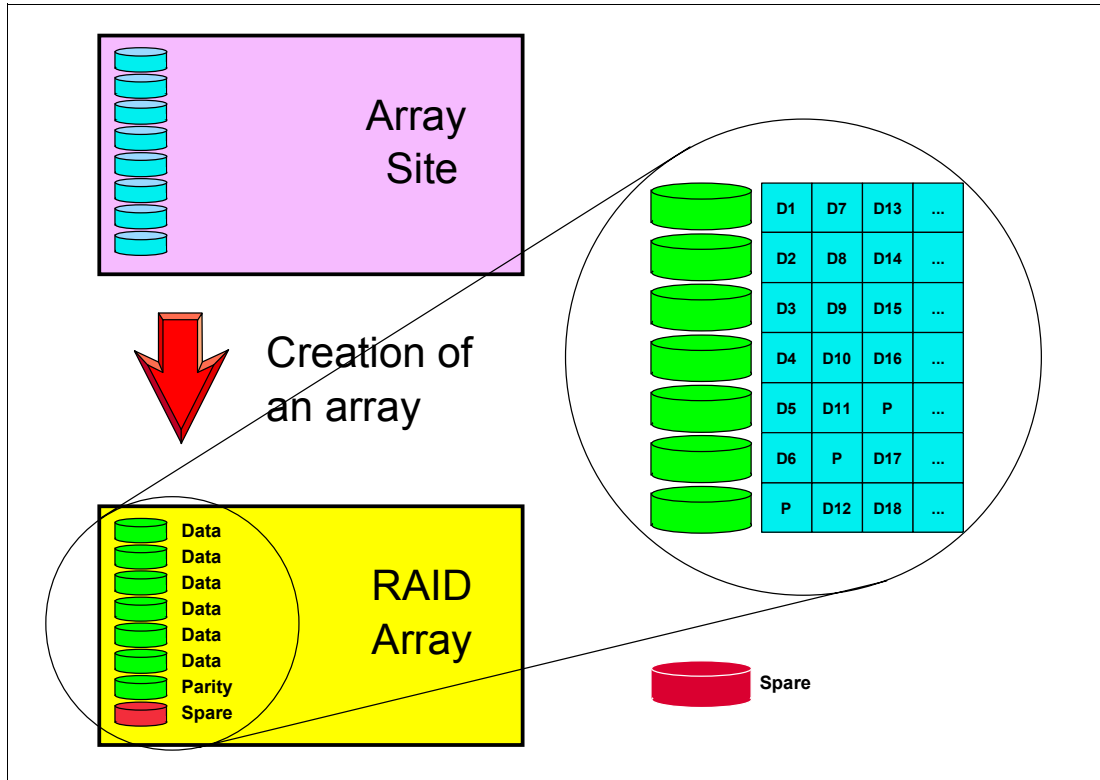


Figure 5-4 Creation of an array

So, an array is formed using one array site, and while the array could be accessed by each adapter of the device adapter pair, it is managed by one device adapter. Which adapter and which server manages this array is defined later in the configuration path.

### 5.3.3 Ranks

In the DS8000 virtualization hierarchy there is another logical construct, a *rank*.

When you define a new rank, its name is chosen by the DS Storage Manager, for example, R1, R2, or R3, and so on. You have to add an array to a rank.

**Note:** In the current DS8000 implementation, a rank is built using just one array.

The available space on each rank will be divided into *extents*. The extents are the building blocks of the logical volumes. An extent is striped across all disks of an array as shown in Figure 5-5 on page 89 and indicated by the small squares in Figure 5-6 on page 90.

The process of forming a rank does two things:

- ▶ The array is formatted for either FB (open systems) or CKD (zSeries) data. This determines the size of the set of data contained on one disk within a stripe on the array.
- ▶ The capacity of the array is subdivided into equal sized partitions, called *extents*. The extent size depends on the *extent type*, FB or CKD.

An FB rank has an extent size of 1 GB (where 1 GB equals  $2^{30}$  bytes).

People who work in the zSeries environment do not deal with gigabytes, instead they think of storage in metrics of the original 3390 volume sizes. A 3390 Model 3 is three times the size of

a Model 1, and a Model 1 has 1113 cylinders which is about 0.94 GB. The extent size of a CKD rank therefore was chosen to be one 3390 Model 1 or 1113 cylinders.

One extent is the minimum physical allocation unit when a LUN or CKD volume is created, as we discuss later. It is still possible to define a CKD volume with a capacity that is an integral multiple of one cylinder or a fixed block LUN with a capacity that is an integral multiple of 128 logical blocks (64K bytes). However, if the defined capacity is not an integral multiple of the capacity of one extent, the unused capacity in the last extent is wasted. For instance, you could define a 1 cylinder CKD volume, but 1113 cylinders (1 extent) is allocated and 1112 cylinders would be wasted.

Figure 5-5 shows an example of an array that is formatted for FB data with 1 GB extents (the squares in the rank just indicate that the extent is composed of several blocks from different DDMs).

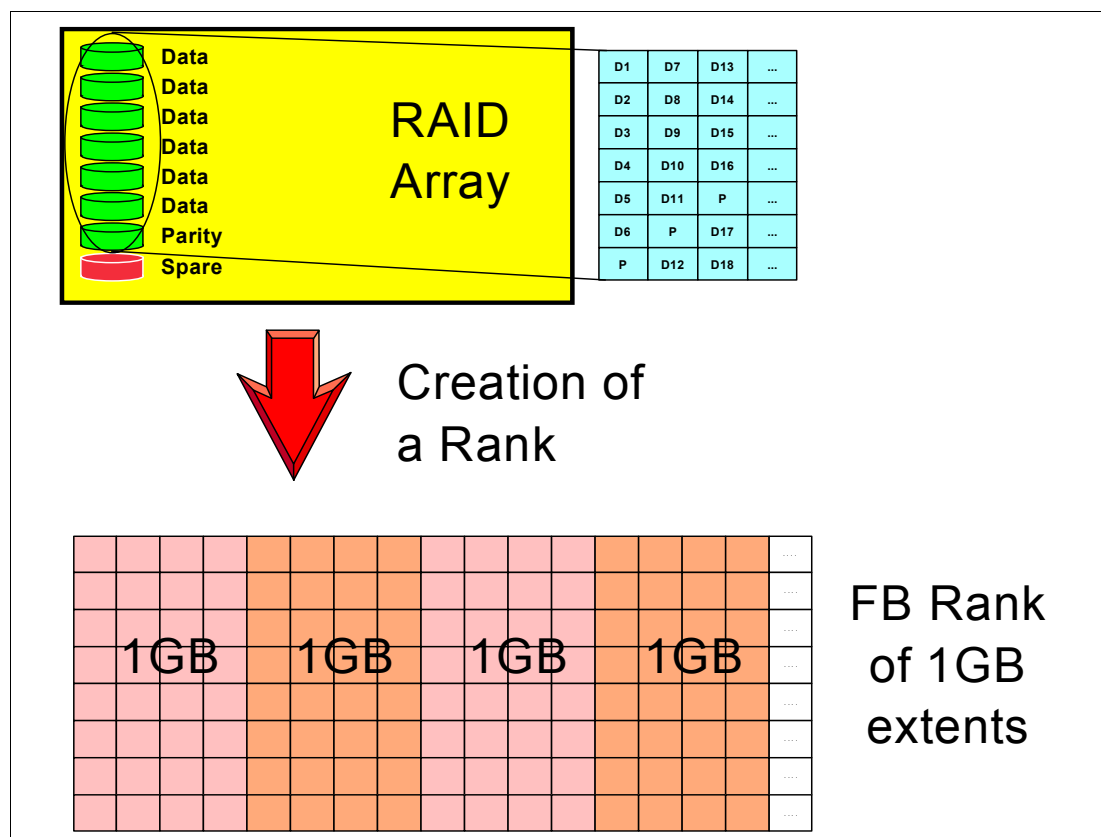


Figure 5-5 Forming an FB rank with 1 GB extents

### 5.3.4 Extent pools

An *extent pool* is a logical construct to aggregate the extents from a set of ranks to form a domain for extent allocation to a logical volume. Typically the set of ranks in the extent pool would have the same RAID type and the same disk RPM characteristics so that the extents in the extent pool have homogeneous characteristics. There is no predefined affinity of ranks or arrays to a storage server. The affinity of the rank (and it's associated array) to a given server is determined at the point it is assigned to an extent pool.

One or more ranks *with the same extent type* can be assigned to an extent pool. One rank can be assigned to only one extent pool. There can be as many extent pools as there are ranks.

The DS Storage Manager GUI guides the user to use the same RAID types in an extent pool. As such, when an extent pool is defined, it must be assigned with the following attributes:

- Server affinity
- Extent type
- RAID type

The minimum number of extent pools is one; however, you would normally want at least two, one assigned to server 0 and the other assigned to server 1 so that both servers are active. In an environment where FB and CKD are to go onto the DS8000 storage server, you might want to define four extent pools, one FB pool for each server, and one CKD pool for each server, to balance the capacity between the two servers. Of course you could also define just one FB extent pool and assign it to one server, and define a CKD extent pool and assign it to the other server. Additional extent pools may be desirable to segregate ranks with different DDM types.

Ranks are organized in two *rank groups*:

- Rank group 0 is controlled by server 0.
- Rank group 1 is controlled by server 1.

**Important:** You should balance your capacity between the two servers for optimal performance.

Figure 5-6 is an example of a mixed environment with CKD and FB extent pools.

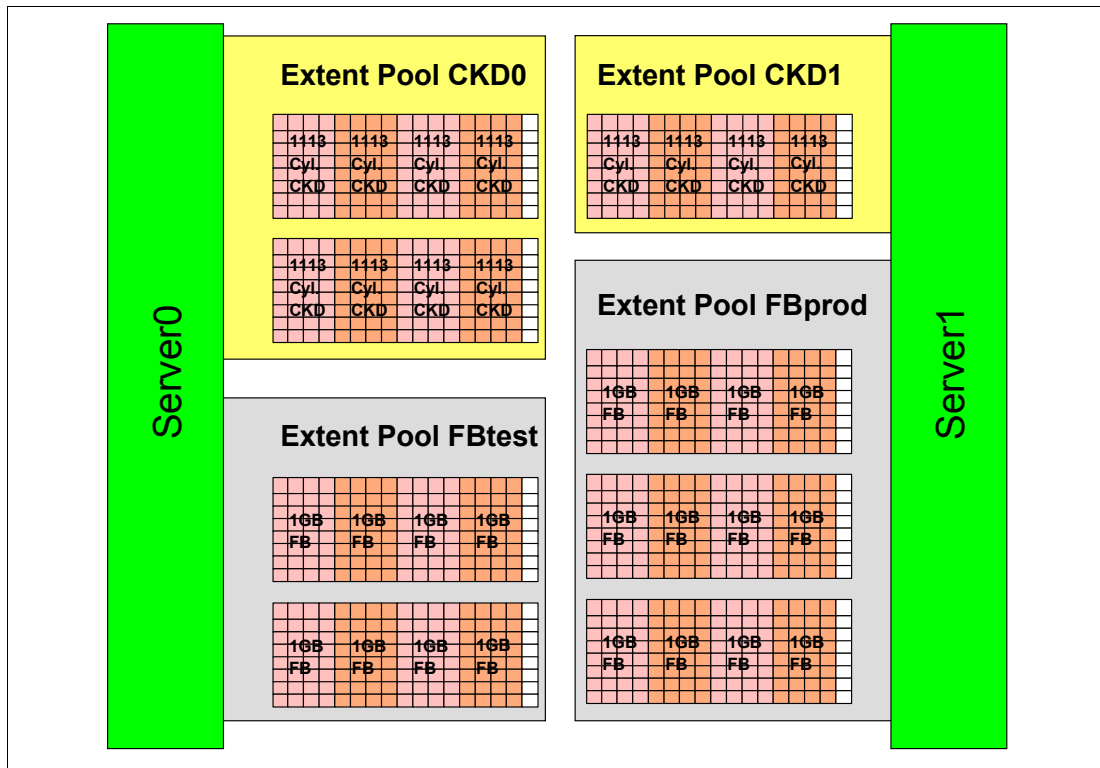


Figure 5-6 Extent pools

You can expand extent pools by adding more ranks to an extent pool.



### 5.3.5 Logical volumes

A logical volume is composed of a set of extents from one extent pool.

On a DS8000 up to 65280 (we use the abbreviation 64K in this discussion, even though it is actually 65536 - 256, which is not quite 64K in binary) volumes can be created (64K CKD, or 64K FB volumes, or a mix of both types, but the sum cannot exceed 64K).

#### Fixed Block LUNs

A logical volume composed of fixed block extents is called a LUN. A fixed block LUN is composed of one or more 1 GB ( $2^{30}$ ) extents from one FB extent pool. A LUN cannot span multiple extent pools, but a LUN can have extents from different ranks within the same extent pool. You can construct LUNs up to a size of 2 TB ( $2^{40}$ ).

LUNs can be allocated in binary GB ( $2^{30}$  bytes), decimal GB ( $10^9$  bytes), or 512 or 520 byte blocks. However, the physical capacity that is allocated for a LUN is always a multiple of 1 GB, so it is a good idea to have LUN sizes that are a multiple of a gigabyte. If you define a LUN with a LUN size that is not a multiple of 1 GB, for example, 25.5 GB, the LUN size is 25.5 GB, but 26 GB are physically allocated and 0.5 GB of the physical storage is unusable.

#### CKD volumes

A zSeries CKD volume is composed of one or more extents from one CKD extent pool. CKD extents are of the size of 3390 Model 1, which has 1113 cylinders. However, when you define a zSeries CKD volume, you do not specify the number of 3390 Model 1 extents but the number of cylinders you want for the volume.

You can define CKD volumes with up to 65520 cylinders, which is about 55.6 GB.

If the number of cylinders specified is not an integral multiple of 1113 cylinders, then some space in the last allocated extent is wasted. For example, if you define 1114 or 3340 cylinders, 1112 cylinders are wasted. For maximum storage efficiency, you should consider allocating volumes that are exact multiples of 1113 cylinders. In fact, integral multiples of 3339 cylinders should be considered for future compatibility.

If you want to use the maximum number of cylinders (65520), you should consider that this is *not* a multiple of 1113. You could go with 65520 cylinders and waste 147 cylinders for each volume (the difference to the next multiple of 1113) or you might be better off with a volume size of 64554 cylinders which is a multiple of 1113 (factor of 58), or even better, with 63441 cylinders which is a multiple of 3339, a model 3 size.

A CKD volume cannot span multiple extent pools, but a volume can have extents from different ranks in the same extent pool.

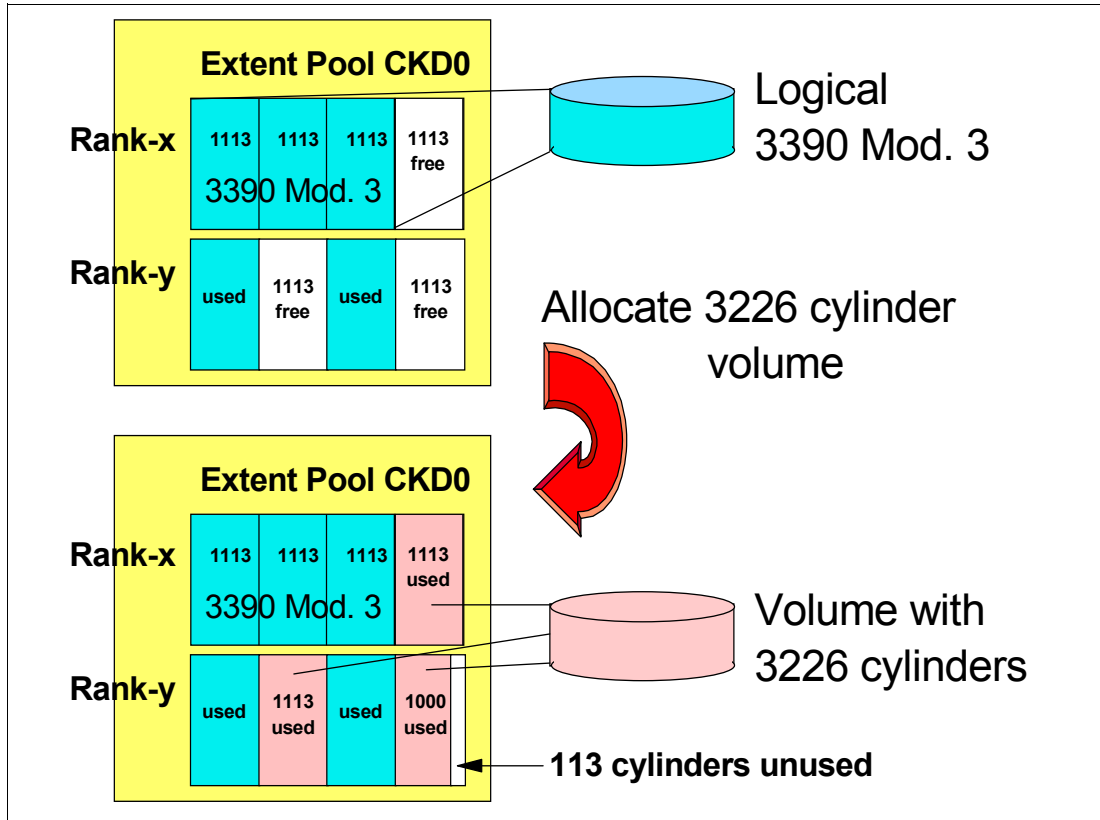


Figure 5-7 Allocation of a CKD logical volume

Figure 5-7 shows how a logical volume is allocated with a CKD volume as an example. The allocation process for FB volumes is very similar and is shown in Figure 5-8 on page 93.

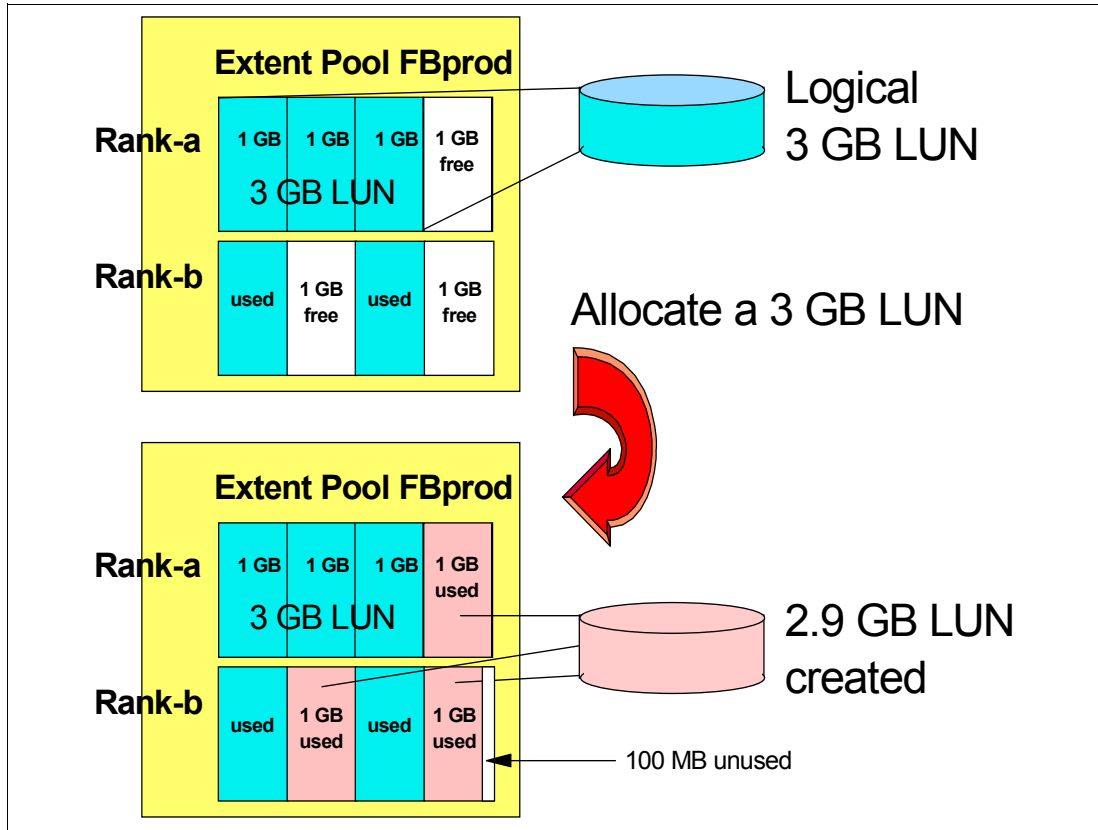


Figure 5-8 Creation of an FB LUN

### iSeries LUNs

iSeries LUNs are also composed of fixed block 1 GB extents. There are, however, some special aspects with iSeries LUNs. LUNs created on a DS8000 are always RAID protected. LUNs are based on RAID-5 or RAID-10 arrays. However, you might want to deceive OS/400® and tell it that the LUN is *not* RAID protected. This causes OS/400 to do its own mirroring. iSeries LUNs can have the attribute *unprotected*, in which case the DS8000 will lie to an iSeries host and tell it that the LUN is not RAID protected.

OS/400 only supports certain fixed volume sizes, for example model sizes of 8.5 GB, 17.5 GB, and 35.1 GB. These sizes are not multiples of 1 GB and hence, depending on the model chosen, some space is wasted. iSeries LUNs expose a 520 byte block to the host. The operating system uses 8 of these bytes so the usable space is still 512 bytes like other SCSI LUNs. The capacities quoted for the iSeries LUNs are in terms of the 512 byte block capacity and are expressed in GB ( $10^9$ ). These capacities should be converted to GB ( $2^{30}$ ) when considering effective utilization of extents that are 1 GB ( $2^{30}$ ). For more information on this topic see Appendix B, “Using DS8000 with iSeries” on page 373.

### Allocation and deletion of LUNs/CKD volumes

All extents of the ranks assigned to an extent pool are independently available for allocation to logical volumes. The extents for a LUN/volume are logically ordered, but they do not have to come from one rank and the extents do not have to be contiguous on a rank. The current extent allocation algorithm of the DS8000 will not distribute the extents across ranks. The algorithm will use available extents within one rank, unless there are not enough free extents available in that rank, but free extents in another rank of the same extent pool. While this algorithm exists, the user may want to consider putting one rank per extent pool to control the

allocation of logical volumes across ranks to improve performance, except for the case in which the logical volume needed is larger than the total capacity of the single rank.

This construction method of using fixed extents to form a logical volume in the DS8000 allows flexibility in the management of the logical volumes. We can now delete LUNs and reuse the extents of those LUNs to create other LUNs, maybe of different sizes. One logical volume can be removed without affecting the other logical volumes defined on the same extent pool. Compared to the ESS, where it was not possible to delete a LUN unless the whole array was reformatted, this DS8000 implementation gives you much more flexibility and allows for on demand changes according to your needs.

Since the extents are *cleaned* after you have deleted a LUN or CKD volume, it may take some time until these extents are available for reallocation. The reformatting of the extents is a background process.

IBM plans to further increase the flexibility of LUN/volume management. We cite from the DS8000 announcement letter the following Statement of General Direction:

*Extension of IBM's dynamic provisioning technology within the DS8000 series is planned to provide LUN/volume: dynamic expansion, online data relocation, virtual capacity over provisioning, and space efficient FlashCopy requiring minimal reserved target capacity.*

### 5.3.6 Logical subsystems (LSS)

A logical subsystem (LSS) is another logical construct. It groups logical volumes, LUNs, in groups of up to 256 logical volumes.

On an ESS there was a fixed association between logical subsystems (and their associated logical volumes) and device adapters (and their associated ranks). The association of an 8-pack to a device adapter determined what LSS numbers could be chosen for a volume. On an ESS up to 16 LSSs could be defined depending on the physical configuration of device adapters and arrays.

On the DS8000, there is no fixed binding between any rank and any logical subsystem. The capacity of one or more ranks can be aggregated into an extent pool and logical volumes configured in that extent pool are not bound to any specific rank. Different logical volumes on the same logical subsystem can be configured in different extent pools. As such, the available capacity of the storage facility can be flexibly allocated across the set of defined logical subsystems and logical volumes.

This predetermined association between array and LSS is gone on the DS8000. Also the number of LSSs has changed. You can now define up to 255 LSSs for the DS8000 (it might be limited at GA to 128 LSSs). You can even have more LSSs than arrays.

For each LUN or CKD volume you can now choose an LSS. You can put up to 256 volumes into one LSS. There is, however, one restriction. We already have seen that volumes are formed from a bunch of extents from an extent pool. Extent pools, however, belong to one server, server 0 or server 1, respectively. LSSs also have an affinity to the servers. All even numbered LSSs (X'00', X'02', X'04', up to X'FE') belong to server 0 and all odd numbered LSSs (X'01', X'03', X'05', up to X'FD') belong to server 1. LSS X'FF' is reserved.

zSeries users are familiar with a logical control unit (LCU). zSeries operating systems configure LCUs to create device addresses. There is a one to one relationship between an LCU and a CKD LSS (LSS X'ab' maps to LCU X'ab'). Logical volumes have a logical volume number X'abcd' where X'ab' identifies the LSS and X'cd' is one of the 256 logical volumes on the LSS. This logical volume number is assigned to a logical volume when a logical volume is

created and determines the LSS that it is associated with. The 256 possible logical volumes associated with an LSS are mapped to the 256 possible device addresses on an LCU (logical volume X'abcd' maps to device address X'cd' on LCU X'ab'). When creating CKD logical volumes and assigning their logical volume numbers, users should consider whether Parallel Access Volumes (PAV) are required on the LCU and reserve some of the addresses on the LCU for alias addresses. For more information on PAV see Chapter 10, "The DS Storage Manager - logical configuration" on page 189.

For open systems, LSSs do not play an important role except in determining which server the LUN is managed by (and which extent pools it must be allocated in) and in certain aspects related to Metro Mirror, Global Mirror, or any of the other remote copy implementations.

Some management actions in Metro Mirror, Global Mirror, or Global Copy operate at the LSS level. For example the freezing of pairs to preserve data consistency across all pairs, in case you have a problem with one of the pairs, is done at the LSS level. With the option now to put all or most of the volumes of a certain application in just one LSS, this makes the management of remote copy operations easier (see Figure 5-9). Of course you could have put all volumes for one application in one LSS on an ESS, too, but then all volumes of that application would also be in one or a few arrays, and from a performance standpoint this was not desirable. Now on the DS8000 you can group your volumes in one or a few LSSs but still have the volumes in many arrays or ranks.

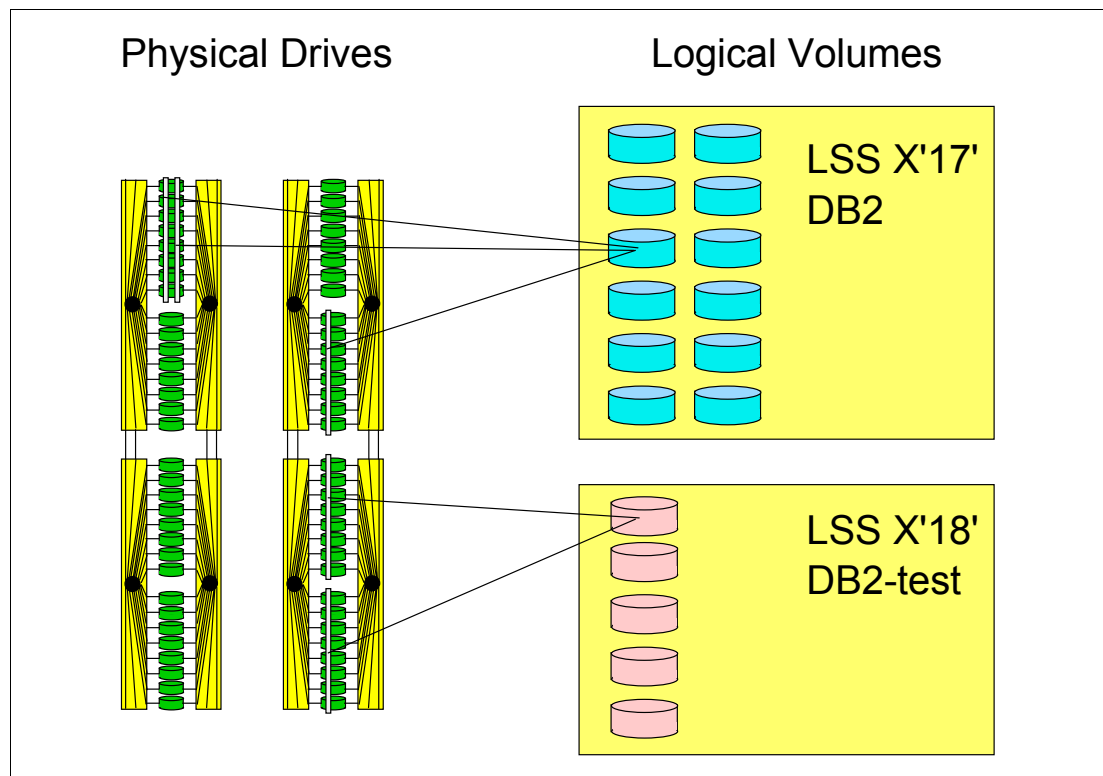


Figure 5-9 Grouping of volumes in LSSs

Fixed block LSSs are created automatically when the first fixed block logical volume on the LSS is created and deleted automatically when the last fixed block logical volume on the LSS is deleted. CKD LSSs require user parameters to be specified and must be created before the first CKD logical volume can be created on the LSS; they must be deleted manually after the last CKD logical volume on the LSS is deleted.

## Address groups

Address groups are created automatically when the first LSS associated with the address group is created, and deleted automatically when the last LSS in the address group is deleted.

LSSs are either CKD LSSs or FB LSSs. All devices in an LSS must be either CKD *or* FB. This restriction goes even further. LSSs are grouped into address groups of 16 LSSs. LSSs are numbered X'ab', where a is the address group and b denotes an LSS within the address group. So, for example X'10' to X'1F' are LSSs in address group 1.

All LSSs within one address group have to be of the same type, CKD or FB. The first LSS defined in an address group fixes the type of that address group.

zSeries clients that still want to use ESCON to attach hosts to the DS8000 should be aware of the fact that ESCON supports only the 16 LSSs of address group 0 (LSS X'00' to X'0F'). Therefore this address group should be reserved for ESCON-attached CKD devices, in this case, and not used as FB LSSs.

Figure 5-10 shows the concept of LSSs and address groups.

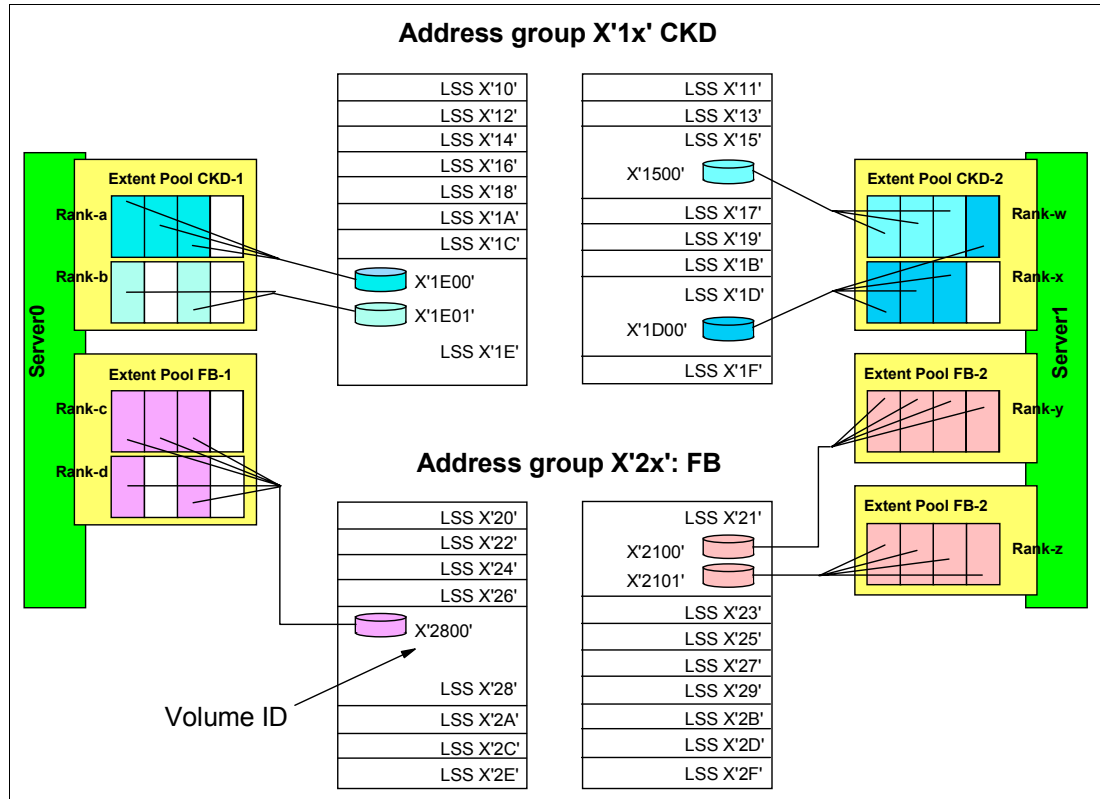


Figure 5-10 Logical storage subsystems

The LUN identifications X'gabb' are composed of the address group X'g', and the LSS number within the address group X'a', and the position of the LUN within the LSS X'bb'. For example, LUN X'2101' denotes the second (X'01') LUN in LSS X'21' of address group 2.

### 5.3.7 Volume access

A DS8000 provides mechanisms to control host access to LUNs. In most cases a server has two or more HBAs and the server needs access to a group of LUNs. For easy management of

server access to logical volumes, the DS8000 introduced the concept of host attachments and volume groups.

## Host attachment

HBAs are identified to the DS8000 in a host attachment construct that specifies the HBAs' World Wide Port Names (WWPNs). A set of host ports can be associated through a port group attribute that allows a set of HBAs to be managed collectively. This port group is referred to as host attachment within the GUI.

A given host attachment can be associated with only one volume group. Each host attachment can be associated with a volume group to define which LUNs that HBA is allowed to access. Multiple host attachments can share the same volume group. The host attachment may also specify a port mask that controls which DS8000 I/O ports the HBA is allowed to log in to. Whichever ports the HBA logs in on, it sees the same volume group that is defined in the host attachment associated with this HBA.

The maximum number of host attachments on a DS8000 is 8192.

## Volume group

A volume group is a named construct that defines a set of logical volumes. When used in conjunction with CKD hosts, there is a default volume group that contains all CKD volumes and any CKD host that logs into a FICON I/O port has access to the volumes in this volume group. CKD logical volumes are automatically added to this volume group when they are created and automatically removed from this volume group when they are deleted.

When used in conjunction with Open Systems hosts, a host attachment object that identifies the HBA is linked to a specific volume group. The user must define the volume group by indicating which fixed block logical volumes are to be placed in the volume group. Logical volumes may be added to or removed from any volume group dynamically.

There are two types of volume groups used with Open Systems hosts and the type determines how the logical volume number is converted to a host addressable LUN\_ID on the Fibre Channel SCSI interface. A *map volume group* type is used in conjunction with FC SCSI host types that poll for LUNs by walking the address range on the SCSI interface. This type of volume group can map any FB logical volume numbers to 256 LUN\_IDs that have zeroes in the last six bytes and the first two bytes in the range of X'0000' to X'00FF'.

A *mask volume group* type is used in conjunction with FC SCSI host types that use the Report LUNs command to determine the LUN\_IDs that are accessible. This type of volume group can allow any and all FB logical volume numbers to be accessed by the host where the mask is a bitmap that specifies which LUNs are accessible. For this volume group type, the logical volume number X'abcd' is mapped to LUN\_ID X'40ab40cd00000000'. The volume group type also controls whether 512 byte block LUNs or 520 byte block LUNs can be configured in the volume group.

When associating a host attachment with a volume group, the host attachment contains attributes that define the logical block size and the Address Discovery Method (LUN Polling or Report LUNs) that are used by the host HBA. These attributes must be consistent with the volume group type of the volume group that is assigned to the host attachment so that HBAs that share a volume group have a consistent interpretation of the volume group definition and have access to a consistent set of logical volume types. The GUI typically sets these values appropriately for the HBA based on the user specification of a host type. The user must consider what volume group type to create when setting up a volume group for a particular HBA.

FB logical volumes may be defined in one or more volume groups. This allows a LUN to be shared by host HBAs configured to different volume groups. An FB logical volume is automatically removed from all volume groups when it is deleted.

The maximum number of volume groups is 8320 for the DS8000.

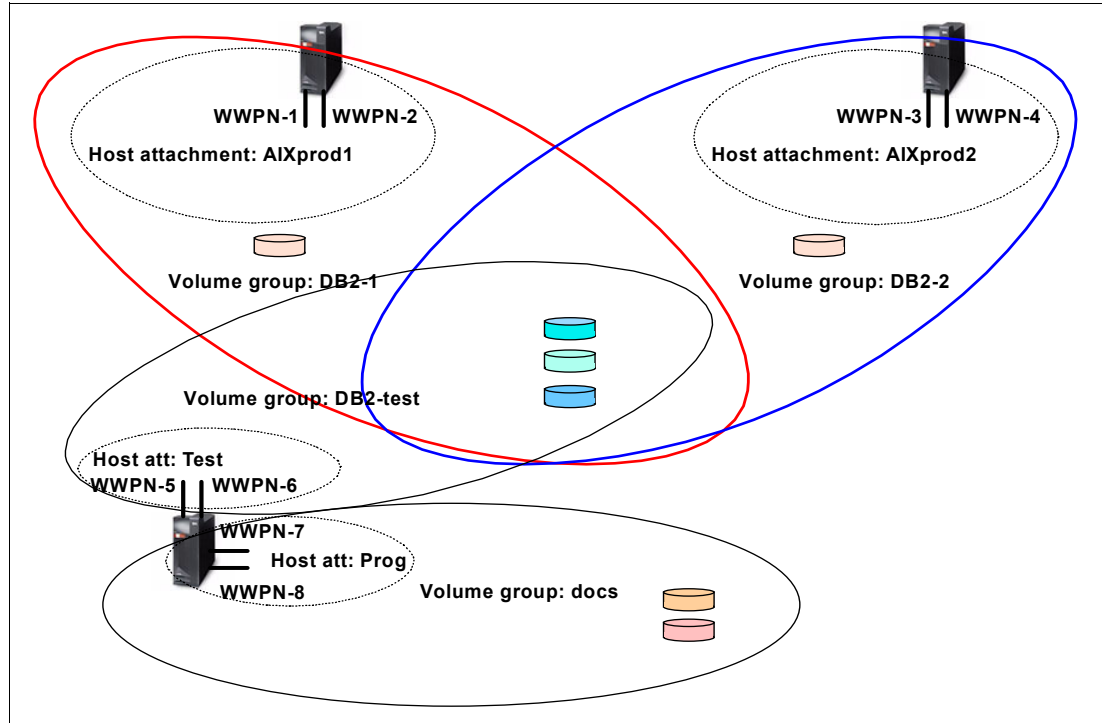


Figure 5-11 Host attachments and volume groups

Figure 5-11 shows the relationships between host attachments and volume groups. Host AIXprod1 has two HBAs, which are grouped together in one host attachment, and both are granted access to volume group DB2-1. Most of the volumes in volume group DB2-1 are also in volume group DB2-2, accessed by server AIXprod2. In our example, there is, however, one volume in each group that is not shared. The server in the lower left part has four HBAs and they are divided into two distinct host attachments. One can access some volumes shared with AIXprod1 and AIXprod2. The other HBAs have access to a volume group called “docs.”

### 5.3.8 Summary of the virtualization hierarchy

Going through the virtualization hierarchy, we started with just a bunch of disks that were grouped in array sites. An array site was transformed into an array, eventually with spare disks. The array was further transformed into a rank with extents formatted for FB or CKD data. Next, the extents were added to an extent pool that determined which storage server would serve the ranks and aggregated the extents of all ranks in the extent pool for subsequent allocation to one or more logical volumes.

Next we created logical volumes within the extent pools, assigning them a logical volume number that determined which logical subsystem they would be associated with and which server would manage them. Then the LUNs could be assigned to one or more volume groups. Finally the host HBAs were configured into a host attachment that is associated with a given volume group.



This new virtualization concept provides for much more flexibility. Logical volumes can dynamically be created and deleted. They can be grouped logically to simplify storage management. Large LUNs and CKD volumes reduce the total number of volumes and this also contributes to a reduction of the management effort.

Figure 5-12 summarizes the virtualization hierarchy.

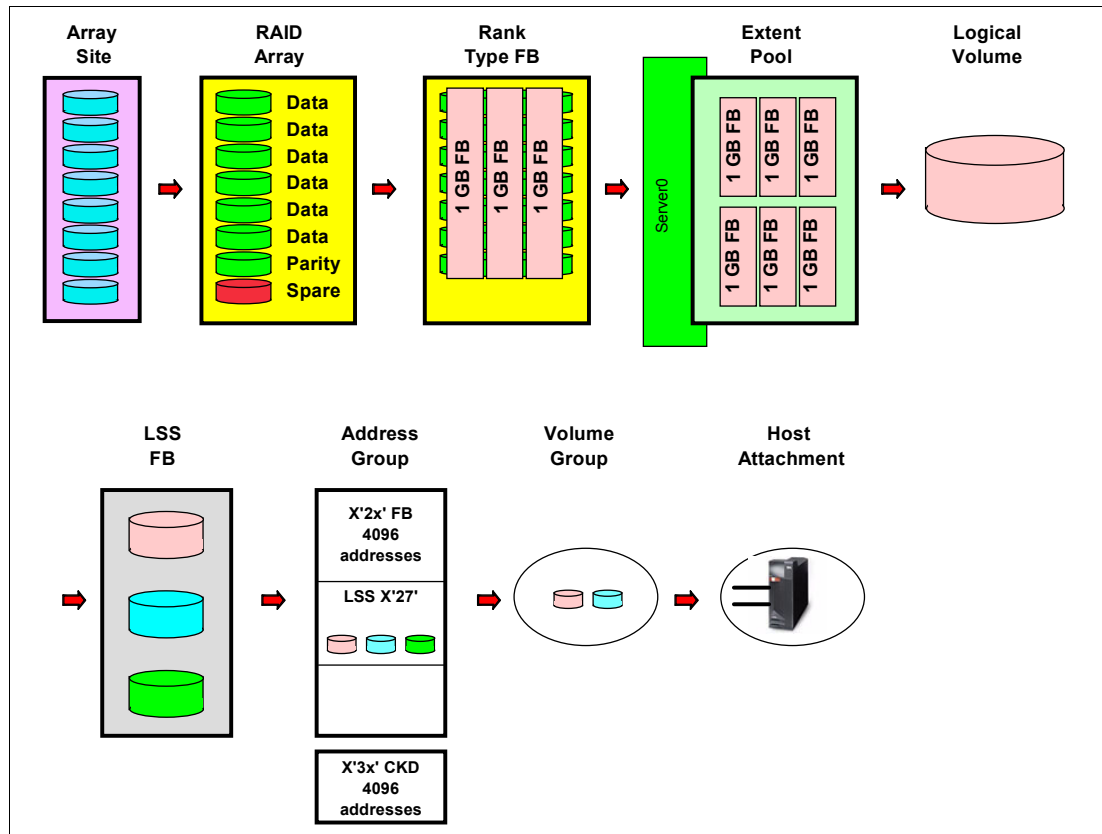


Figure 5-12 Virtualization hierarchy

### 5.3.9 Placement of data

As explained in the previous chapters, there are several options on how to create logical volumes. You can select an extent pool that is owned by one server. There could be just one extent pool per server or you could have several. The ranks of extent pools could come from arrays on different device adapter pairs and different loops or from the same loop. Figure 5-13 on page 100 shows an optimal distribution of eight logical volumes within a DS8000. Of course you could have more extent pools and ranks, but when you want to distribute your data for optimal performance, you should make sure that you spread it across the two servers, across different device adapter pairs, across the loops, and across several ranks.

If you use some kind of a logical volume manager (like LVM on AIX) on your host, you can create a host logical volume from several DS8000 logical volumes (LUNs). You can select LUNs from different DS8000 servers, device adapter pairs, and loops as shown in Figure 5-13. By striping your host logical volume across the LUNs, you will get the best performance for this LVM volume.

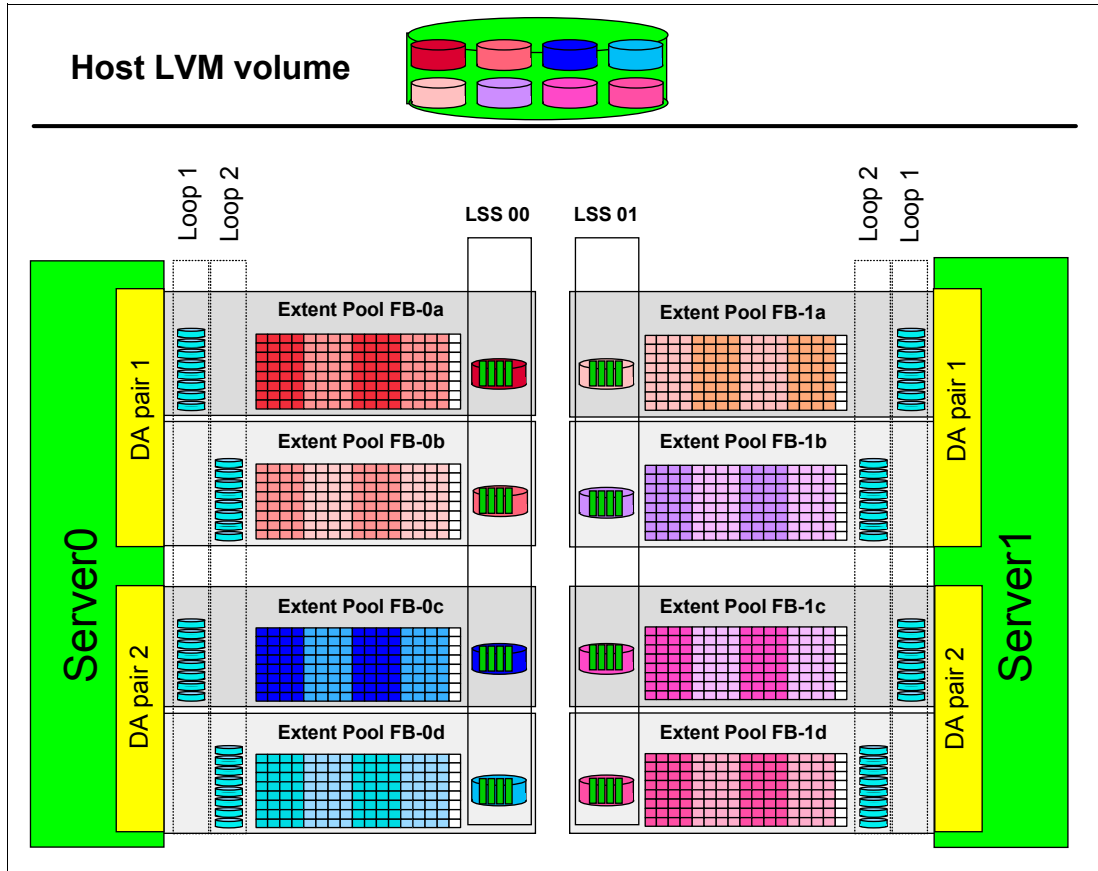


Figure 5-13 Optimal placement of data

## 5.4 Benefits of virtualization

The DS8000 physical and logical architecture defines new standards for enterprise storage virtualization. The main benefits of the virtualization layers are:

- ▶ Flexible LSS definition allows maximization/optimization of the number of devices per LSS.
- ▶ No strict relationship between RAID ranks and LSSs.
- ▶ No connection of LSS performance to underlying storage.
- ▶ Number of LSSs can be defined based upon device number requirements:
  - With larger devices significantly fewer LSSs might be used.
  - Volumes for a particular application can be kept in a single LSS.
  - Smaller LSSs can be defined if required (for systems/applications requiring less storage).
  - Test systems can have their own LSSs with fewer volumes than production systems.
- ▶ Increased number of logical volumes:
  - Up to 65280 (CKD)
  - Up to 65280 (FB)
  - 65280 total for CKD + FB

- ▶ Any mix of CKD or FB addresses in 4096 address groups.
- ▶ Increased logical volume size:
  - CKD: 55.6 GB (65520 cylinders), architected for 219 TB
  - FB: 2 TB, architected for 1 PB
- ▶ Flexible logical volume configuration:
  - Multiple RAID types (RAID-5, RAID-10)
  - Storage types (CKD and FB) aggregated into extent pools
  - Volumes allocated from extents of extent pool
  - Dynamically add/remove volumes
- ▶ Virtualization reduces storage management requirements.





## **IBM TotalStorage DS8000 model overview and scalability**

This chapter provides an overview of the IBM TotalStorage DS8000 storage server which is from here on referred to as a DS8000.

We include information on the two models and how well they scale regarding capacity and performance.

## 6.1 DS8000 highlights

The DS8000 is a member of the DS product family. It offers disk storage servers with high performance and has the capability to scale very well to the highest disk storage capacities. The scalability is designed to support continuous operations.

The DS8000 series models follow the 2105 (ESS 800) as device type 2107 and are based on IBM POWER5 server technology.

The DS8000 series models consist of a storage unit and one or two management consoles. A graphical user interface (GUI) or the command-line interface (CLI) allows a storage administrator to configure and logically partition storage.

The DS8000 series currently offers two models with base and expansion units to configure storage servers that meet the client's performance and capacity requirements.

### 6.1.1 Model naming conventions

Before we get into the details about each model, we start with an explanation of the model naming conventions in the DS8000.

The DS8000 has three models at general availability (GA), model 921, 922, and 9A2. The difference in models is the number of processors and the capability of storage system LPARs. You can also order expansion frames with the base frame. The expansion frame is described as a model 92E or 9AE.

The last position of the three characters means the number of 2-way processors on each processor complex (xx1 means 2-way and xx2 means 4-way, xxE means expansion frame (no processors)). The middle position of the three characters means LPAR or Non-LPAR model (x2x means non-LPAR model and xAx means LPAR model).

We summarize these rules in Figure 6-1.

<h2>Model Naming Conventions</h2>			
9xy	Y=1	Y=2	Y=E
X=2	Non LPAR/2-way	Non LPAR/4-way	Non LPAR Expansion
X=a	N/A	LPAR 4-way	LPAR Expansion

For example,  
921 : Non-LPAR / 2-way base frame  
9AE : LPAR / Expansion frame

Figure 6-1 Model naming conventions

In the following sections, we describe these models further:

1. DS8100 Model 921

This model features a dual two-way processor complex and it includes a base frame and an optional expansion frame.

2. DS8300 Models 922 and 9A2

The DS8300 is a dual four-way processor complex with a Model 922 base frame and up to two optional Model 92E expansion frames. Model 9A2 is also a dual four-way processor complex, but it offers two IBM TotalStorage storage system Logical Partitions (LPARs) in one machine. The Model 9A2 can also be expanded with up to two Model 9AE expansion units.

## 6.1.2 DS8100 Model 921

The DS8100 Model 921 has the following features:

- ▶ Two processor complexes with pSeries POWER5 1.5 GHz two-way CEC each.
- ▶ Up to 128 DDMs for a maximum disk storage capacity of 38.4 TB with 300 GB DDMs.
- ▶ Up to 128 GB of processor memory, which was referred as *cache* in the ESS 800.
- ▶ Up to 16 host adapters (HAs) with four ports per HA and 2 Gbps per port. Each port can be independently configured as either a Fibre Channel protocol (FCP) port for open systems host connection or PPRC FCP links, but also as a FICON port to connect to zSeries hosts. This totals up to 64 ports with any mix of FCP and FICON ports. ESCON host connection is also supported. But with ESCON, an HA contains only two ESCON ports. A mix of ESCON ports and FCP ports on the same HA is not possible. A DS8000 can have both ESCON adapters and FCP/FICON adapters at the same time.

The DS8100 Model 921 connects to a wide variety of host server platforms. For an up-to-date connectivity list, see:

<http://www-1.ibm.com/servers/storage/disk/ds8000/interop.html>



Figure 6-2 DS8100 Model 921 base frame open (left) and with Model 92E for max config

The DS8100 Model 921 can connect to one expansion frame. This expansion frame is called a Model 92E. Figure 6-2 on page 105 displays the front view of a DS8100 Model 921 with the cover off, and the Model 921 with an expansion Model 92E with covers. The base and expansion frame together allow for a maximum capacity for a DS8100 with 384 DDMs. There are 128 DDMs in the base frame and 256 DDMs in the expansion frame. With all DDMs being 300 GB, this results in a maximum disk storage capacity of 115.2 TB.

Figure 6-3 shows the maximum configuration of a Model 921 with the 921 base frame plus a 92E expansion frame and provides the front view of the basic structure and placement of the hardware components within both frames.

**Note:** A Model 921 can be upgraded to a Model 922 or to a Model 9A2.

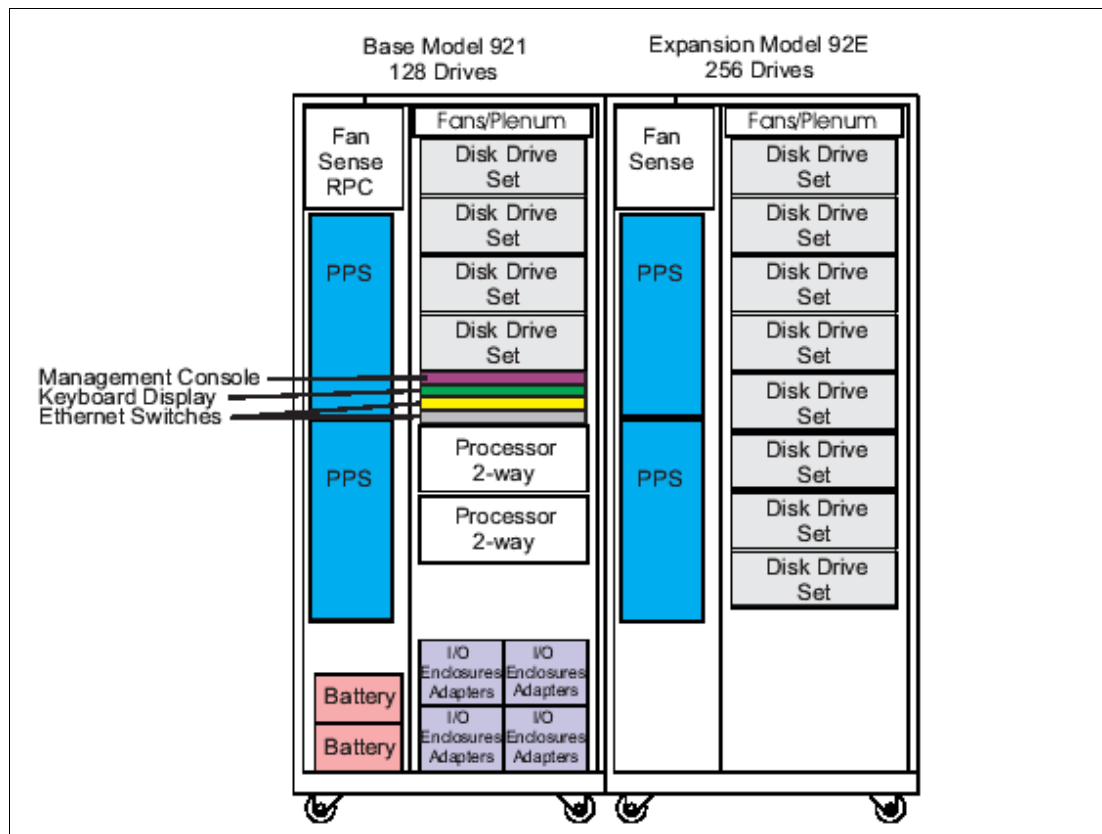


Figure 6-3 Maximum configuration for the Model 921

### 6.1.3 DS8300 Models 922 and 9A2

The DS8300 Model 922 and 9A2 offer higher capacity and performance than the DS8100.

Model 9A2 provides two storage images through two storage system LPARs within the same physical storage unit. Both storage images split the processor, cache, and adapter resources. The split ratio in the beginning is a 50:50 split with the potential to subdivide these resources later at a different split ratio between the storage images. This future flexibility is going to also allow a finer granularity to split the processor resources than on an entire processor level, which is the case for the first editions of the Model 9A2.



Both models provide the following features:

- ▶ Two processor complexes with pSeries POWER5 1.9 GHz four-way CEC each.
- ▶ Up to 128 DDMs for a maximum of 38.4 TB with 300 GB DDMs. This is the same as for the DS8100 Model 921.
- ▶ Up to 256 GB of processor memory, which was referred to as *cache* in the ESS 800.
- ▶ Up to 16 host adapters (HAs). Each HA provides four 2 Gbps Fibre Channel ports which can freely be configured as:
  - FCP ports to open system hosts
  - PPRC FCP links
  - FICON ports to connect to zSeries hosts
- ▶ A HA can also contain two ESCON ports instead of four Fibre Channel ports. You cannot mix ESCON ports with Fibre Channel ports on the same HA.

DS8300 models connect to a wide variety of host server platforms. Refer to the following Web site for an up-to-date interoperability list for the DS8300 models:

<http://www-1.ibm.com/servers/storage/disk/ds8000/interop.html>



Figure 6-4 DS8300 Model 922/9A2 base frame rear view with and without covers

Figure 6-4 displays a DS8300 Model 922 from the rear. On the left is the rear view with closed covers. The right shows the rear view of the Model 922 with no covers. The middle view is another rear view but only with one cover off. This view shows the standard 19 inch rack mounted hardware including disk drives, processor complexes, and I/O enclosures.

DS8300 models can connect to one or two expansion frames. This provides the following configuration alternatives:

- ▶ With an expansion frame 92E or 9AE the disk storage capacity and number of adapters can expand to:
  - Up to 384 DDMs in total - as for the DS8100. This is a maximum disk capacity of 115.2 TB with 300 GB DDMs.

- Up to 32 host adapters (HAs) with four 2 Gbps Fibre Channel ports on each HA. Each HA can be freely configured to hold:
  - FCP ports to connect to FCP-based host servers
  - FCP ports for PPRC links, which in turn can be shared between PPRC and FCP-based hosts through a SAN fabric
  - FICON ports to connect to one or more zSeries servers
  - Any combination within a HA
- A HA can also have two ESCON ports, but you cannot mix ESCON ports with Fibre Channel ports on the very same HA.
- ▶ With two 92E expansion frames or two 9AE expansion frames for the storage system LPAR model the capacity expands to:
  - Up to 640 DDMs in total for a maximum of 192 TB disk storage when utilizing 300 GB DDMs. Figure 6-5 shows such a maximum configuration for Model 922 with two expansion frames, Model 92E, or a Model 9A2 with two expansion frames Model 9AE. This figure also shows the basic hardware components and how they are distributed across all three frames.

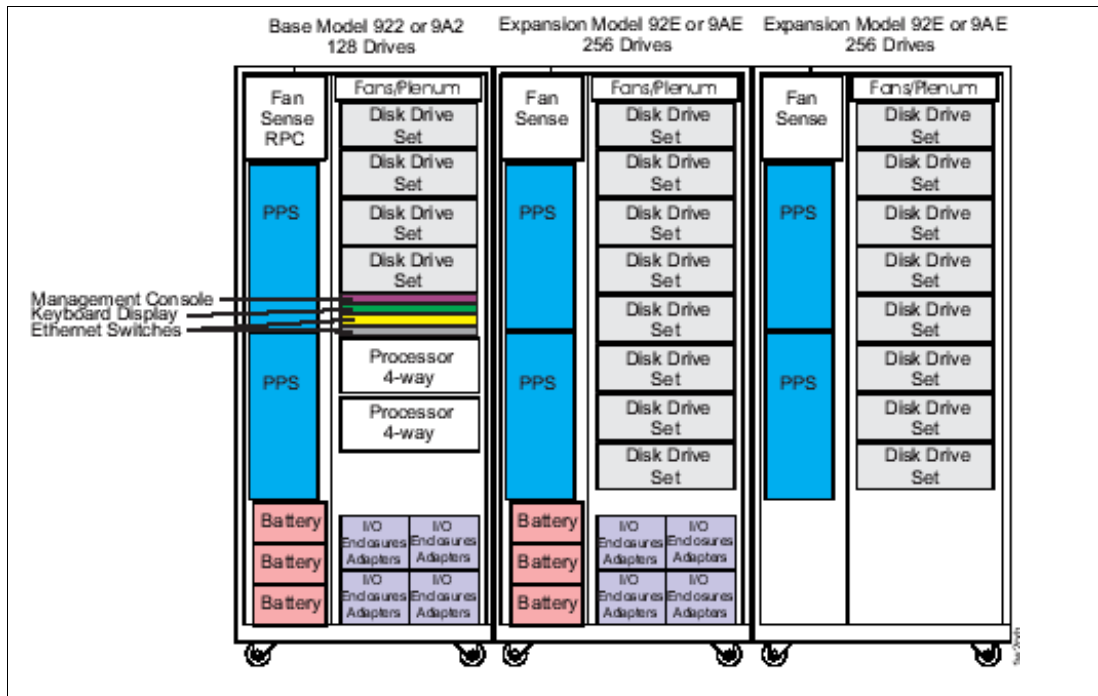


Figure 6-5 DS8300 max config for Model 922/9A2 with 2 expansion frames (either 92E or 9AE)

Note the additional I/O enclosures in the first expansion frame, which is the middle frame in Figure 6-5. Each expansion frame has twice as many DDMs as the base frame, so with 128 DDMs in the base frame and 2 x 256 DDMs in the expansion frames, a total of 640 DDMs is possible.

## 6.2 Model comparison

DS8000 models vary in processor type, disk capacity, processor memory, and host adapters. Table 6-1 on page 109 provides an overview of the different features and numbers for the currently available DS8000 models.

Table 6-1 DS8000 model comparison

Base Mod	Images	Exp Mod	Proc type	DDMs	Proc Mem	HAs
921	1	None	2-way 1.5 GHz	<= 128	<= 128 GB	<= 16
		1 x 92E		<= 384		
922	1	None	4-way 1.9 GHz	<= 128	<= 256 GB	<= 16
		1 x 92E		<= 384		<= 32
		2 x 92E		<= 640		
9A2	2	None		<= 128		<= 16
		1 x 9AE		<= 384		<= 32
		2 x 9AE		<= 640		

Depending on the DDM sizes, which can be different within a 921, 922, or 9A2, and the number of DDMs, the total capacity is calculated accordingly.

Each FCP/FICON adapter has four Fibre Channel ports, which provide up to 128 Fibre Channel ports. Each ESCON adapter has two ports, therefore, the maximum number of ports is 64.

## 6.3 Designed for scalability

One of the advantages of the DS8000 series is its linear scalability for capacity and performance. If your business (or your customer's business) grows rapidly, you may need much more storage capacity, faster storage performance, or both. The DS8000 series can meet these demands within a single storage unit. We explain the scalability in this section.

### 6.3.1 Scalability for capacity

The DS8000 series has a linear capacity growth up to 192 TB and can add additional DDMs without a system disruption. IBM plans to offer larger models in the future.

#### Large and scalable capacity

You can have from 16 DDMs up to 384 DDMs (Model 921) or 640 DDMs (Model 922 or 9A2) in a DS8000.

- ▶ Each base frame can have up to 128 DDMs and an expansion frame can have up to 256 DDMs.
- ▶ The DS8100 can have one expansion frame and the DS8300 can have two expansion frames
- ▶ The DS8000 can contain three types of DDMs, 73 GB (15,000 RPM), 146 GB (10,000 RPM), and 300 GB (10,000 RPM).

Therefore, you can select a physical capacity from 1.1 TB (73 GB x 16 DDMs) up to 192 TB (300 GB x 640 DDMs).

We summarize these characteristics in Table 6-2 on page 110.

Table 6-2 Comparison of models for capacity

	2-way (Base frame only)	2-way + Expansion frame	4-way or LPAR (Base frame only)	4-way or LPAR + Expansion frame	4-way or LPAR + 2 Expansion frames
Device adapters	2 to 8 (1 - 4 Fibre Loops)	2 to 8	2 to 8	2 to 16	2 to 16
Drives 73 GB (15k RPM) 146 GB (10k RPM) 300 GB (10k RPM)	16 to 128 (increments of 16)	16 to 384 (increments of 16)	16 to 128 (increments of 16)	16 to 384 (increments of 16)	16 to 640 (increments of 16)
Physical capacity	1.1 to 37.2 TB	1.1 to 111.6 TB	1.1 to 37.2 TB	1.1 to 111.6 TB	1.1 to 186.3 TB

### Adding DDMs

A significant benefit of the DS8000 series is the ability to add DDMs without disruption for maintenance. Even if it is difficult to take time for system maintenance for your system (for example, if you are operating a 24x7 online system), you don't need to order larger capacity in advance to prepare for future growth. You can begin with a small configuration at first, and grow the capacity on demand without disruption.

**Note:** According to the announcement letter, the following activities are disruptive in the current status:

- ▶ Field attachment of a Model 92E with the 922
- ▶ Field attachment of a Model 9AE with the 9A2

In both cases, *only when the first expansion frame is attached to the base frame*, you do need a disruptive maintenance. The first expansion enclosure for the Model 922 and 9A2 has I/O enclosures and these I/O enclosures must be connected into the current RIO-G loops. Currently, this operation is not supported non-disruptively.

If you install the base frame and the first expansion frame for the Model 922 or 9A2 at the beginning, you don't need a disruptive upgrade to add DDMs. The expansion enclosure for the DS8100 and the second expansion enclosure for the DS8300 has no I/O enclosure, therefore you can attach them to the existing frame without disruption.

### Future plan

In addition, the architecture is designed to scale with 300 GB disk technology up to 1 PB. According to the announcement letter, IBM has issued the following Statement of General Direction:

*Future models in the series are planned to provide customers additional flexibility and scalability. IBM next plans to provide 8-way processors with corresponding scalability from the 2-way and 4-way models currently offered. In addition, physical capacity will be scalable through increased disk drives within a configuration.*

## 6.3.2 Scalability for performance

The DS8000 series also has linear scalability for performance. This capability is due to the architecture of the DS8000 series.

## Linear-scalable architecture

The following two figures illustrate how you can realize the linear scalability in the DS8000 series.

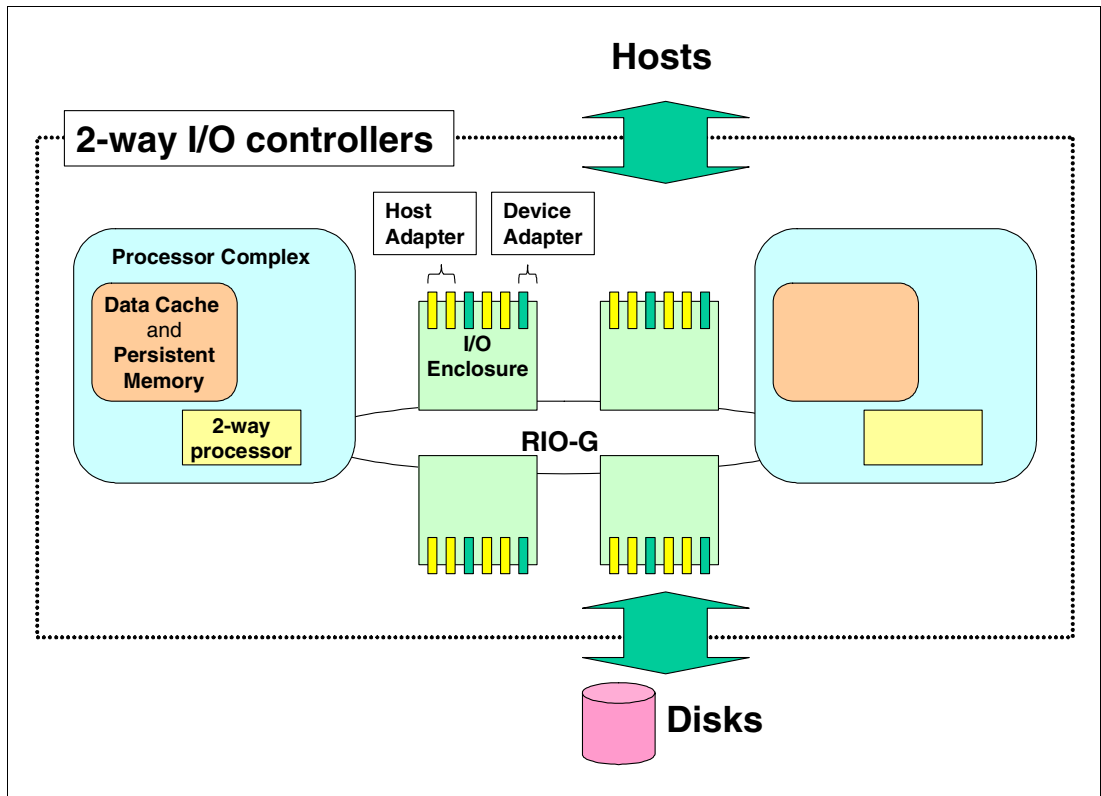


Figure 6-6 2-way model components

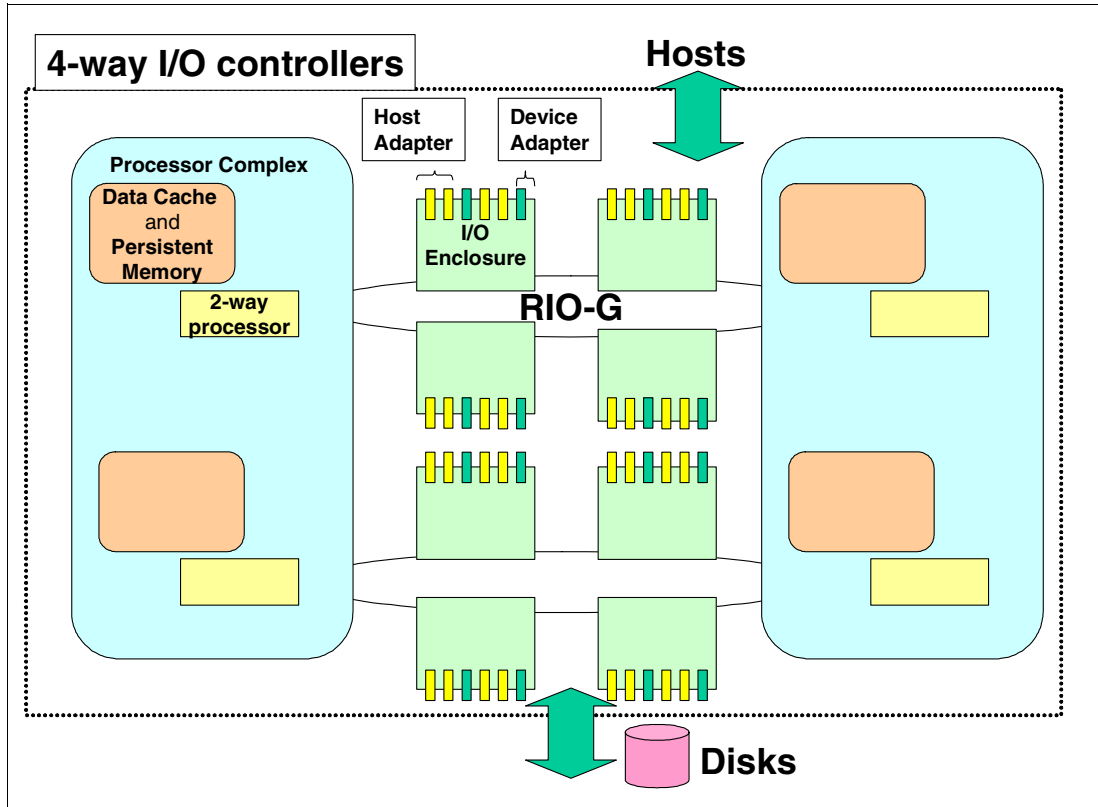


Figure 6-7 4-way model components

These two figures describe the main components of the I/O controller for the DS8000 series. The main components include the I/O processors, data cache, internal I/O bus (RIO-G loop), host adapters, and device adapters. Figure 6-6 on page 111 is a 2-way model and Figure 6-7 is a 4-way model. You can easily see that, if you upgrade from the 2-way model to the 4-way model, the number of main components doubles within a storage unit.

For further information about performance for the DS8000 series, see Chapter 12, “Performance considerations” on page 253.

## Future Plan

According to the announcement letter, IBM has issued the following Statement of General Direction:

*Future models in the series are planned to provide customers additional flexibility and scalability. IBM next plans to provide 8-way processors with corresponding scalability from the 2-way and 4-way models currently offered. In addition, physical capacity will be scalable through increased disk drives within a configuration.*

In the 8-way model, the number of main components is four times that of a 2-way model.

## The benefit of the DS8000 for scalability

Because the DS8000 series adopts this architecture for the scaling of models, the DS8000 series has the following benefits for scalability:

- ▶ The DS8000 series is easily scalable for performance and capacity.
- ▶ The DS8000 series architecture can be easily upgraded.

- ▶ The DS8000 series has a longer life cycle than other storage devices.

### 6.3.3 Model upgrades

The DS8000 series models are modular systems that are designed to be built upon and upgraded from one model to another in the field, helping clients respond swiftly to changing business requirements.

IBM service representatives can upgrade a Model 921 in the field when you order a model conversion to a Model 922 or Model 9A2. You can also change the storage system LPAR configuration, both from non-LPAR to LPAR, from LPAR to non-LPAR (According to the announcement letter, model upgrades will be offered after GA).

Figure 6-8 summarizes the model conversions planned for the DS8000 series.

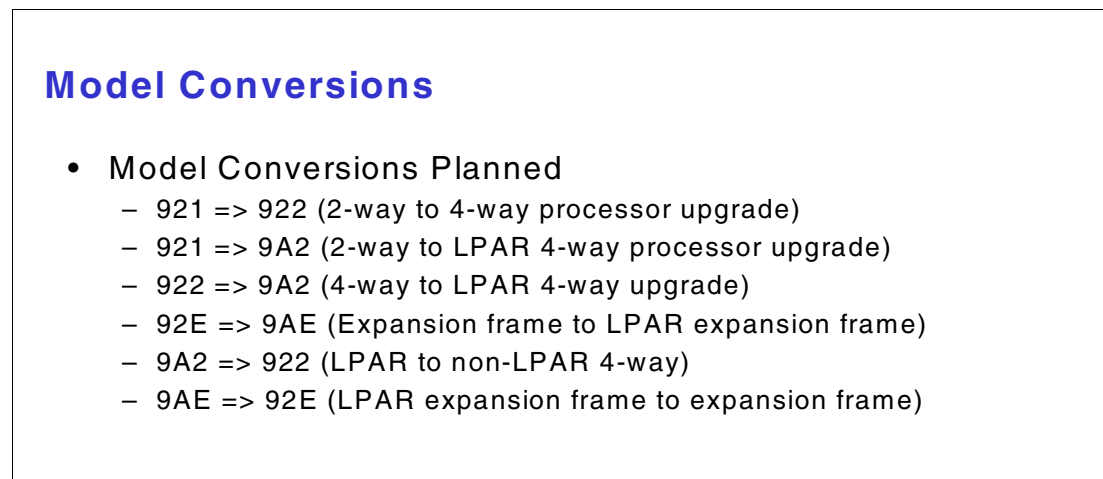


Figure 6-8 Model conversions

**Note:** According to the announcement letter, model conversions are disruptive activities in the current status.







## Copy Services

In this chapter, we describe the architecture and functions of Copy Services for the DS8000 series. Copy Services is a collection of functions that provide disaster recovery, data migration, and data duplication functions. Copy Services run on the DS8000 storage unit and they support open systems and zSeries environments.

Copy Services has four interfaces: a Web-based interface (DS Storage Manager), a command-line interface (DS CLI), an application programming interface (DS API), and host I/O commands from zSeries servers.

This chapter discusses the following topics:

- ▶ Introduction to Copy Services
- ▶ Copy Services functions
- ▶ Interfaces for Copy Services
- ▶ Interoperability with IBM TotalStorage Enterprise Storage Server (ESS)

## 7.1 Introduction to Copy Services

Copy Services is a collection of functions that provide disaster recovery, data migration, and data duplication functions. With the Copy Services functions, for example, you can create backup data with little or no disruption to your application, and you can back up your application data to the remote site for the disaster recovery.

Copy Services run on the DS8000 storage unit and support open systems and zSeries environments. These functions are supported also on the previous generation of storage systems called the IBM TotalStorage Enterprise Storage Server (ESS).

Many design characteristics of the DS8000 and data copying and mirroring capabilities of Copy Services features contribute to the protection of your data, 24 hours a day and seven days a week. The licensed features included in Copy Services are the following:

- ▶ FlashCopy, which is a Point-in-Time Copy function
- ▶ Remote Mirror and Copy functions, previously known as Peer-to-Peer Remote Copy or PPRC, which include:
  - IBM TotalStorage Metro Mirror, previously known as Synchronous PPRC
  - IBM TotalStorage Global Copy, previously known as PPRC Extended Distance
  - IBM TotalStorage Global Mirror, previously known as Asynchronous PPRC
- ▶ z/OS Global Mirror, previously known as Extended Remote Copy (XRC)
- ▶ z/OS Metro/Global Mirror

We explain these functions in detail in the next section.

You can manage the Copy Services functions through a command-line interface (DS CLI) and a new Web-based interface (DS Storage Manager). You also can manage the Copy Services functions through the open application programming interface (DS Open API). When you manage the Copy Services through these interfaces, these interfaces invoke Copy Services functions via the Ethernet network. In zSeries environments, you can invoke the Copy Service functions by TSO commands, ICKDSF, the DFSMSdss™ utility, and so on.

We explain these interfaces in 7.3, “Interfaces for Copy Services” on page 136.

## 7.2 Copy Services functions

We describe each function and the architecture of the Copy Services in this section. There are two primary types of Copy Services functions: *Point-in-Time Copy* and *Remote Mirror and Copy*. Generally, the Point-in-Time Copy function is used for data duplication and the Remote Mirror and Copy function is used for data migration and disaster recovery.

### 7.2.1 Point-in-Time Copy (FlashCopy)

The Point-in-Time Copy feature, which includes FlashCopy, enables you to create full volume copies of data in a storage unit. When you set up a FlashCopy operation, a relationship is established between the source and target volumes, and a bitmap of the source volume is created. Once this relationship and bitmap are created, the target volume can be accessed as though all the data had been physically copied. While a relationship between the source and target volume exists, optionally, a background process copies the tracks from the source to the target volume.

**Note:** In this chapter, *track* means a piece of data in the DS8000; the DS8000 uses the logical tracks to manage the Copy Services functions.

See Figure 7-1 for an illustration of FlashCopy concepts.

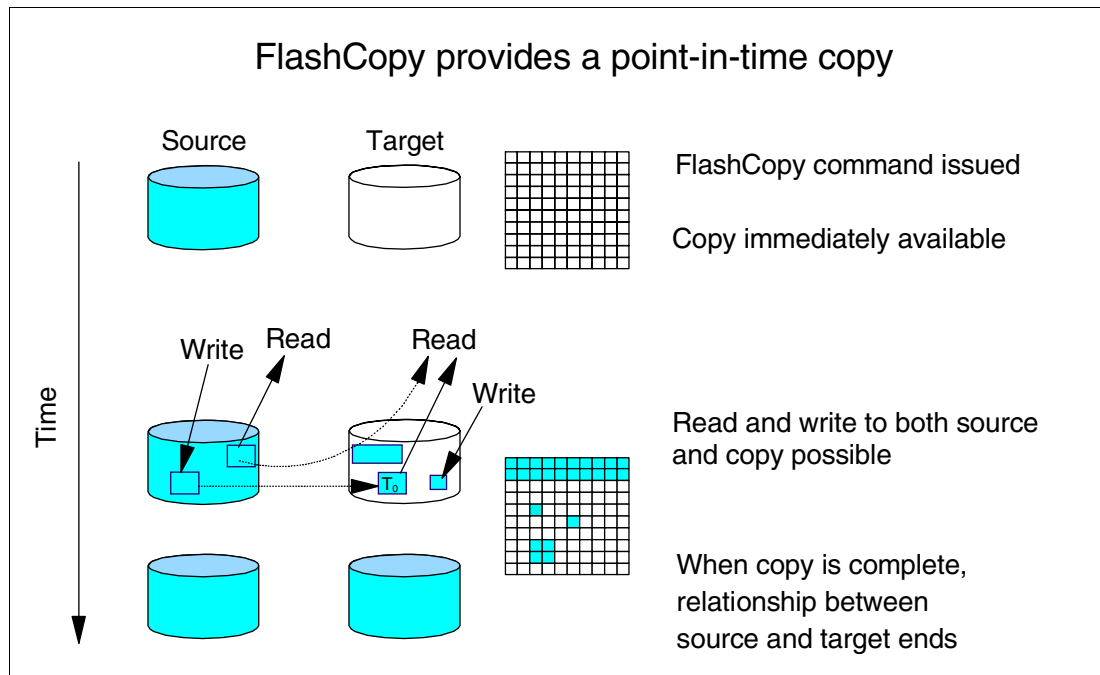


Figure 7-1 FlashCopy concepts

When a FlashCopy operation is invoked, it takes only a few seconds to complete the process of establishing the FlashCopy pair and creating the necessary control bitmaps. Thereafter, you have access to a point-in-time copy of the source volume. As soon as the pair has been established, you can read and write to both the source and target volumes.

After creating the bitmap, a background process begins to copy the real-data from the source to the target volumes. If you access the source or the target volumes during the background copy, FlashCopy manages these I/O requests as follows:

► **Read from the source volume**

When you read some data from the source volume, it is simply read from the source volume.

► **Read from the target volume**

When you read some data from the target volume, FlashCopy checks the bitmap and:

- If the backup data is already copied to the target volume, it is read from the target volume.
- If the backup data is not copied yet, it is read from the source volume.

► **Write to the source volume**

When you write some data to the source volume, at first the updated data is written to the data cache and persistent memory (write cache). And when the updated data is destaged to the source volume, FlashCopy checks the bitmap and:

- If the backup data is already copied, it is simply updated on the source volume.

- If the backup data is not copied yet, first the backup data is copied to the target volume, and after that it is updated on the source volume.

► **Write to the target volume**

When you write some data to the target volume, it is written to the data cache and persistent memory, and FlashCopy manages the bitmaps to not overwrite the latest data. FlashCopy does not overwrite the latest data by the physical copy.

The background copy may have a slight impact to your application because the physical copy needs some storage resources, but the impact is minimal because the host I/O is prior to the background copy. And if you want, you can issue FlashCopy with the *no background copy* option.

### **No background copy option**

If you invoke FlashCopy with the no background copy option, the FlashCopy relationship is established without initiating a background copy. Therefore, you can minimize the impact of the background copy. When the DS8000 receives an update to a source track in a FlashCopy relationship, a copy of the point-in-time data is copied to the target volume so that it is available when the data from the target volume is accessed. This option is useful for customers who don't need to issue FlashCopy in the opposite direction.

### **Benefits of FlashCopy**

The point-in-time copy created by FlashCopy is typically used where you need a copy of the production data to be produced with little or no application downtime (depending on the application). It can be used for online backup, testing of new applications, or for creating a database for data-mining purposes. The copy looks exactly like the original source volume and is an instantly available, binary copy.

### **Point-in-Time Copy function authorization**

FlashCopy is an optional function. To use it, you must purchase the Point-in-Time Copy 2244 function authorization model, which is 2244 Model PTC.

## **7.2.2 FlashCopy options**

FlashCopy has many options and expanded functions to help provide data duplication. We explain these options and functions in this section.

### **Refresh target volume (also known as Incremental FlashCopy)**

Refresh target volume provides the ability to *refresh* a LUN or volume involved in a FlashCopy relationship. When a subsequent FlashCopy operation is initiated, only the tracks changed on both the source and target need to be copied from the source to the target. The direction of the *refresh* can also be reversed.

In many cases, at most 10 to 20 percent of your entire data is changed in a day. In such a situation, if you use this function for daily backup, you can save the time for the physical copy of FlashCopy.

Figure 7-2 on page 119 explains the architecture for Incremental FlashCopy.

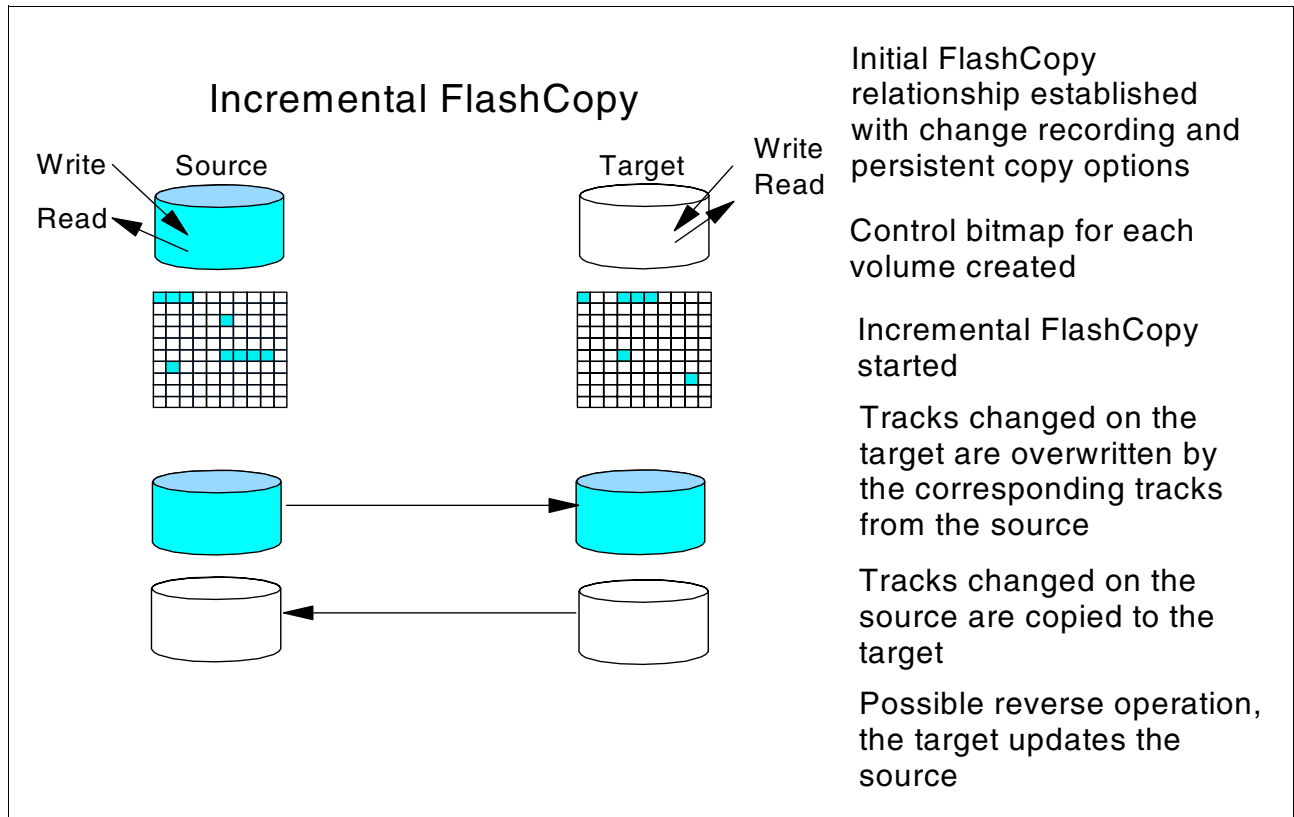


Figure 7-2 Incremental FlashCopy

In the Incremental FlashCopy operations:

1. At first, you issue full FlashCopy with the *change recording* option. This option is for creating change recording bitmaps in the storage unit. The change recording bitmaps are used for recording the tracks which are changed on the source and target volumes after the last FlashCopy.
2. After creating the change recording bitmaps, Copy Services records the information for the updated tracks to the bitmaps. The FlashCopy relationship persists even if all of the tracks have been copied from the source to the target.
3. The next time you issue Incremental FlashCopy, Copy Services checks the change recording bitmaps and copies only the changed tracks to the target volumes. If some tracks on the target volumes are updated, these tracks are overwritten by the corresponding tracks from the source volume.

You can also issue incremental FlashCopy from the target volume to the source volumes with the *reverse restore* option. The reverse restore operation cannot be done unless the background copy in the original direction has finished.

### Data Set FlashCopy

Data Set FlashCopy allows a FlashCopy of a data set in a zSeries environment.

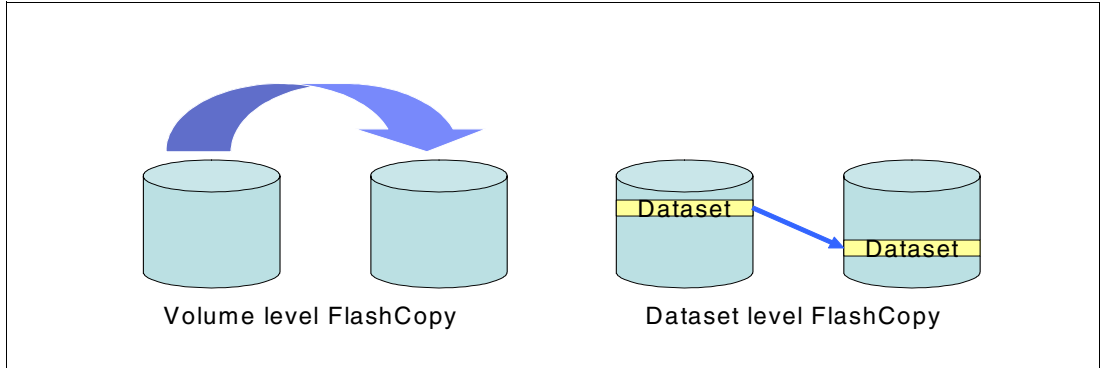


Figure 7-3 Data Set FlashCopy

**Multiple Relationship FlashCopy**

Multiple Relationship FlashCopy allows a source to have FlashCopy relationships with multiple targets simultaneously. A source volume or extent can be FlashCopied to up to 12 target volumes or target extents, as illustrated in Figure 7-4.

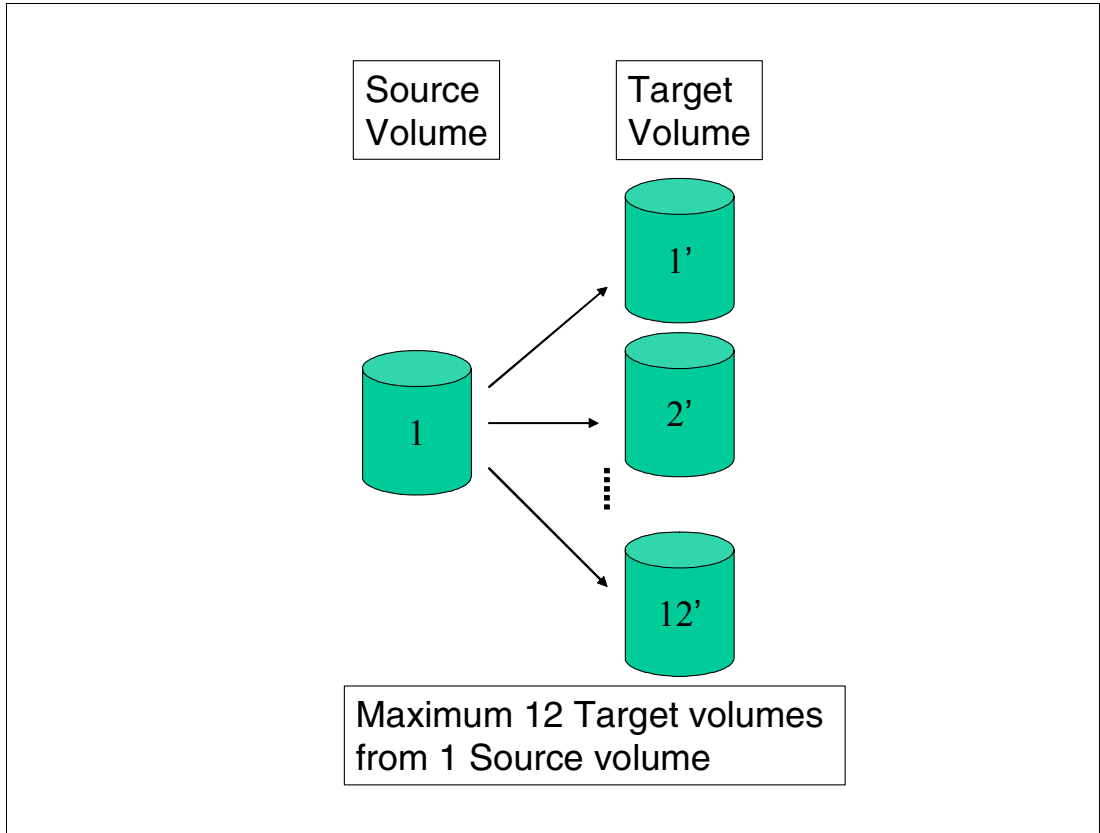


Figure 7-4 Multiple Relationship FlashCopy

**Note:** If a FlashCopy source volume has more than one target, that source volume can be involved only in a single incremental FlashCopy relationship.

## Consistency Group FlashCopy

Consistency Group FlashCopy allows you to freeze (temporarily queue) I/O activity to a LUN or volume. Consistency Group FlashCopy helps you to create a consistent point-in-time copy across multiple LUNs or volumes, and even across multiple storage units.

### What is Consistency Group FlashCopy?

If a consistent point-in-time copy across many logical volumes is required, and the user does not wish to quiesce host I/O or database operations, then the user may use Consistency Group FlashCopy to create a consistent copy across multiple logical volumes in multiple storage units.

In order to create this consistent copy, the user would issue a set of **Establish FlashCopy** commands with a **freeze** option, which will hold off host I/O to the source volumes. In other words, Consistency Group FlashCopy provides the capability to temporarily queue (at the host I/O level, not the application level) subsequent write operations to the source volumes that are part of the Consistency Group. During the temporary queueing, Establish FlashCopy is completed. The temporary queueing continues until this condition is reset by the **Consistency Group Created** command or the time-out value expires (the default is two minutes).

Once all of the Establish FlashCopy requests have completed, a set of **Consistency Group Created** commands must be issued via the same set of DS network interface servers. The Consistency Group Created commands are directed to each logical subsystem (LSS) involved in the consistency group. The Consistency Group Created command allows the write operations to resume to the source volumes.

This operation is illustrated in Figure 7-5.

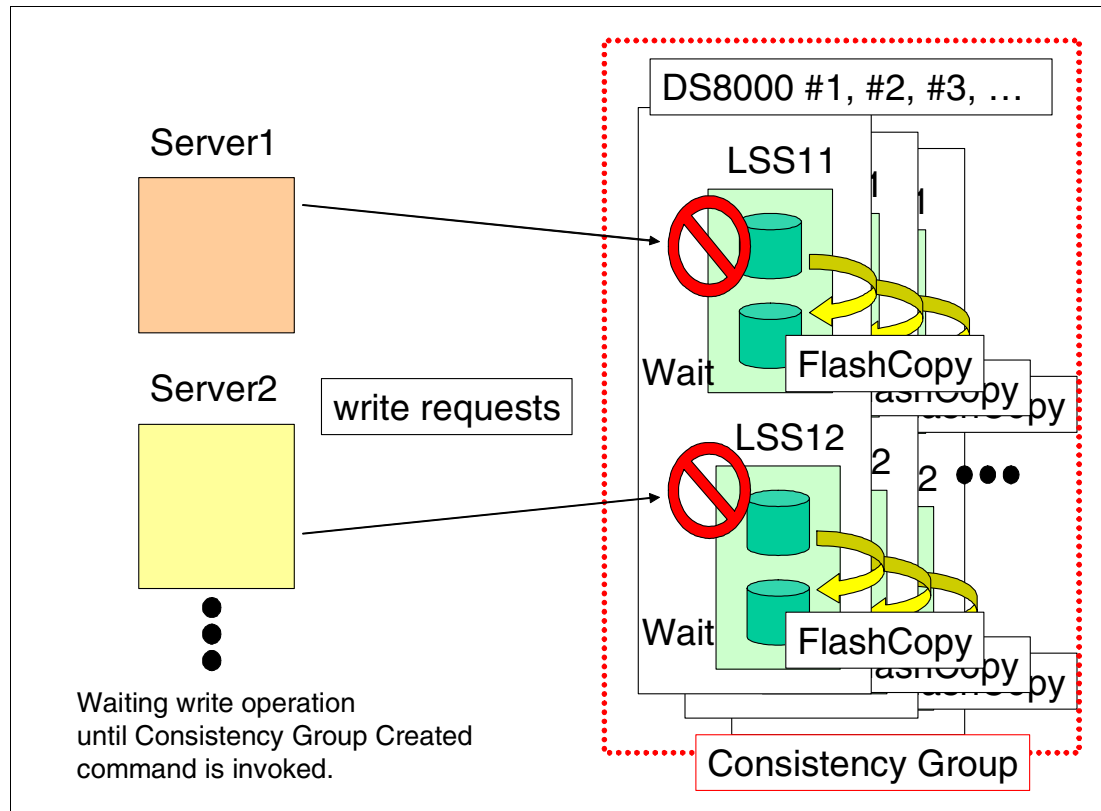


Figure 7-5 Consistency Group FlashCopy

A more detailed discussion of the concept of *data consistency* and how to manage the Consistency Group operation is in 7.2.5, “What is a Consistency Group?” on page 132.

**Important:** Consistency Group FlashCopy can create host-based consistent copies, they are not application-based consistent copies. The copies have *power-fail* or *crash* level consistency. This means that if you suddenly power off your server without stopping your applications and without destaging the data in the file cache, the data in the file cache may be lost and you may need recovery procedures to restart your applications. To start your system with Consistency Group FlashCopy target volumes, you may need the same operations as the crash recovery.

For example, If the Consistency Group source volumes are used with a journaled file system (like AIX JFS) and the source LUNs are not unmounted before running FlashCopy, it is likely that **fsck** will have to be run on the target volumes.

**Note:** Consistency Group FlashCopy is not available through the use of the DS Storage Manager GUI at the current time.

### Establish FlashCopy on existing Remote Mirror and Copy primary

This option allows you to establish a FlashCopy relationship where the target is also a remote mirror primary volume. This enables you to create full or incremental point-in-time copies at a local site and then use remote mirroring commands to copy the data to the remote site. We explain the functions of Remote Mirror and Copy in the 7.2.3, “Remote Mirror and Copy (Peer-to-Peer Remote Copy)” on page 123.

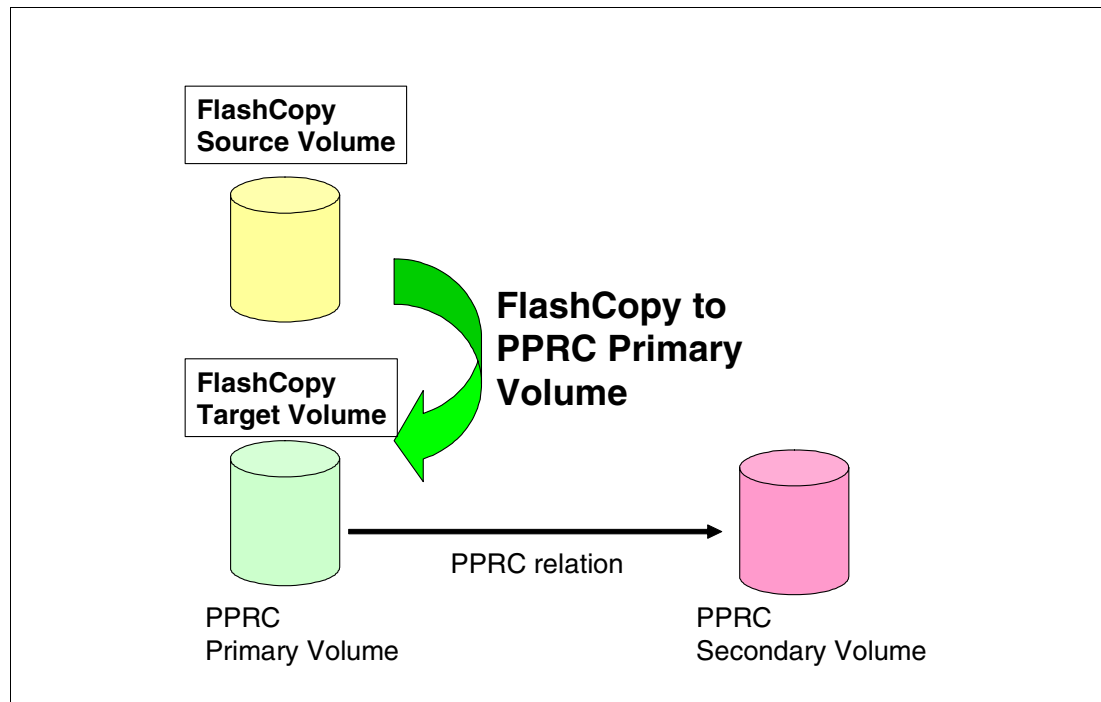


Figure 7-6 Establish FlashCopy on existing Remote Mirror and Copy Primary

**Note:** You cannot FlashCopy from a source to a target, where the target is also a Global Mirror primary volume.



## Persistent FlashCopy

Persistent FlashCopy allows the FlashCopy relationship to remain even after the copy operation completes. You must explicitly delete the relationship.

## Inband Commands over Remote Mirror link

In a remote mirror environment, commands to manage FlashCopy at the remote site can be issued from the local or intermediate site and transmitted over the remote mirror Fibre Channel links. This eliminates the need for a network connection to the remote site solely for the management of FlashCopy.

**Note:** This function is not available through the use of the DS Storage Manager GUI at the current time.

## 7.2.3 Remote Mirror and Copy (Peer-to-Peer Remote Copy)

The Remote Mirror and Copy feature (formally called Peer-to-Peer Remote Copy, or PPRC) is a flexible data mirroring technology that allows replication between volumes on two or more disk storage systems. You can also use this feature for data backup and disaster recovery. Remote Mirror and Copy is an optional function. To use it, you must purchase the Remote Mirror and Copy 2244 function authorization model, which is 2244 Model RMC.

DS8000 storage units can participate in Remote Mirror and Copy solutions with the ESS Model 750, ESS Model 800, and DS6000 storage units. To establish a PPRC relationship between the DS8000 and the ESS, the ESS needs to have licensed internal code (LIC) version 2.4.2 or later.

The Remote Mirror and Copy feature can operate in the following modes:

### Metro Mirror (Synchronous PPRC)

Metro Mirror provides real-time mirroring of logical volumes between two DS8000s that can be located up to 300 km from each other. It is a synchronous copy solution where write operations are completed on both copies (local and remote site) before they are considered to be complete.

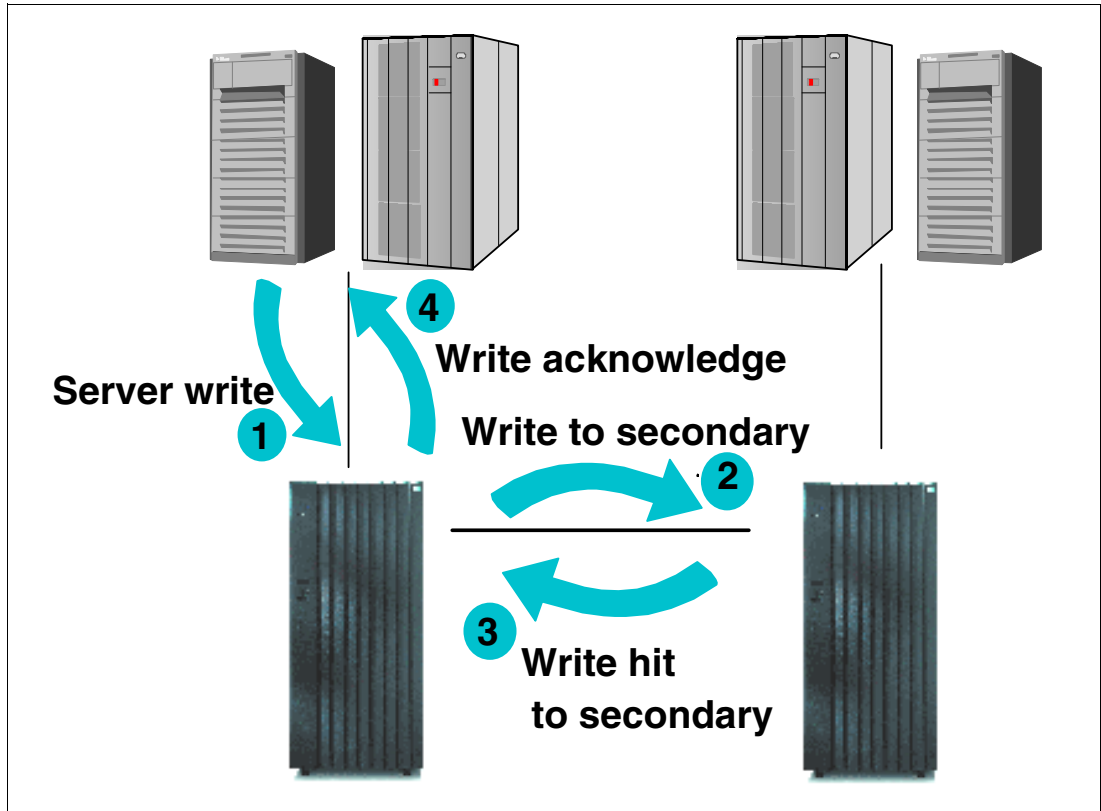


Figure 7-7 Metro Mirror

### Global Copy (PPRC-XD)

Global Copy copies data non-synchronously and over longer distances than is possible with Metro Mirror. When operating in Global Copy mode, the source volume sends a periodic, incremental copy of updated tracks to the target volume, instead of sending a constant stream of updates. This causes less impact to application writes for source volumes and less demand for bandwidth resources, while allowing a more flexible use of the available bandwidth.

Global Copy does not keep the sequence of write operations. Therefore, the copy is normally fuzzy, but you can make a consistent copy through synchronization (called a go-to-sync operation). After the synchronization, you can issue FlashCopy at the secondary site to make the backup copy with data consistency. After the establish of the FlashCopy, you can change the PPRC mode back to the non-synchronous mode.

**Note:** When you change PPRC mode from synchronous to non-synchronous mode, you change the PPRC mode from synchronous to suspend mode at first, and then you change PPRC mode from suspend to non-synchronous mode.

If you want make a consistent copy with FlashCopy, you must purchase a Point-in-Time Copy function authorization (2244 Model PTC) for the secondary storage unit.

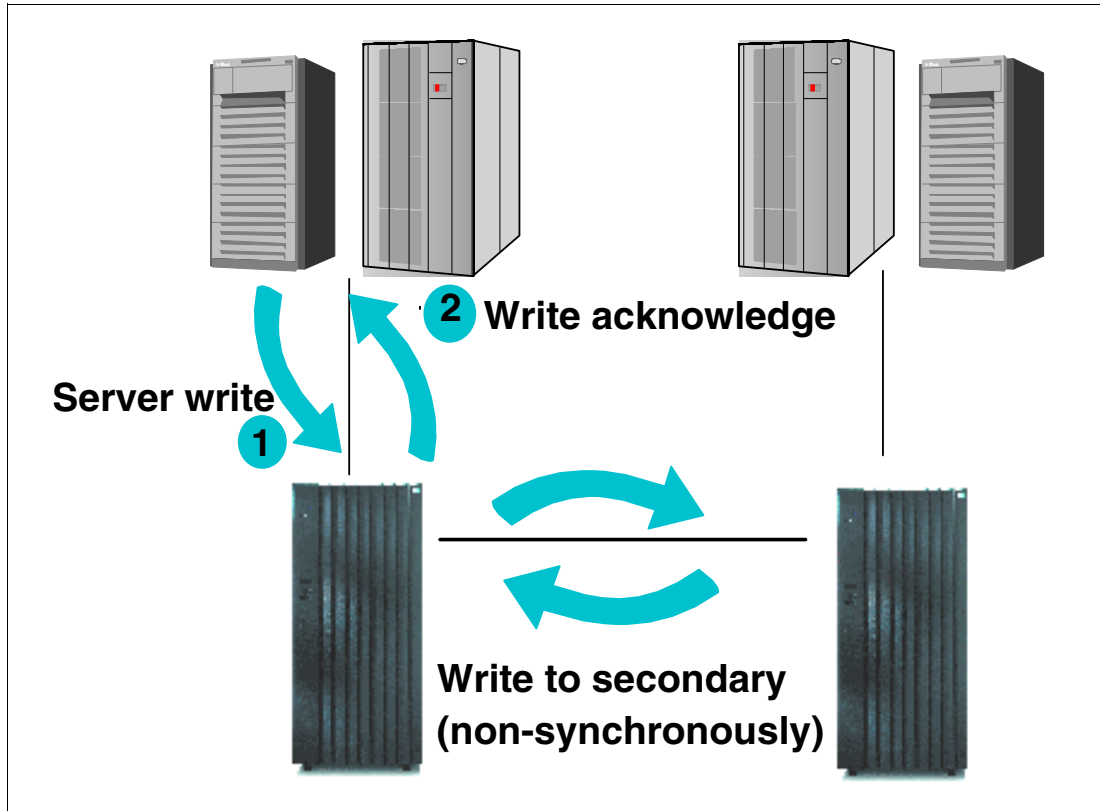


Figure 7-8 Global Copy

### Global Mirror (Asynchronous PPRC)

Global Mirror provides a long-distance remote copy feature across two sites using asynchronous technology. This solution is based on the existing Global Copy and FlashCopy. With Global Mirror, the data that the host writes to the storage unit at the local site is asynchronously shadowed to the storage unit at the remote site. A consistent copy of the data is automatically maintained on the storage unit at the remote site.

Global Mirror operations provide the following benefits:

- ▶ Support for virtually unlimited distances between the local and remote sites, with the distance typically limited only by the capabilities of the network and the channel extension technology. This *unlimited* distance enables you to choose your remote site location based on business needs and enables site separation to add protection from localized disasters.
- ▶ A consistent and restartable copy of the data at the remote site, created with minimal impact to applications at the local site.
- ▶ Data currency where, for many environments, the remote site lags behind the local site typically 3 to 5 seconds, minimizing the amount of data exposure in the event of an unplanned outage. The actual lag in data currency that you experience can depend upon a number of factors, including specific workload characteristics and bandwidth between the local and remote sites.
- ▶ Dynamic selection of the desired recovery point objective, based upon business requirements and optimization of available bandwidth.
- ▶ Session support whereby data consistency at the remote site is internally managed across up to eight storage units that are located across the local and remote sites.

- ▶ Efficient synchronization of the local and remote sites with support for failover and failback modes, helping to reduce the time that is required to switch back to the local site after a planned or unplanned outage.

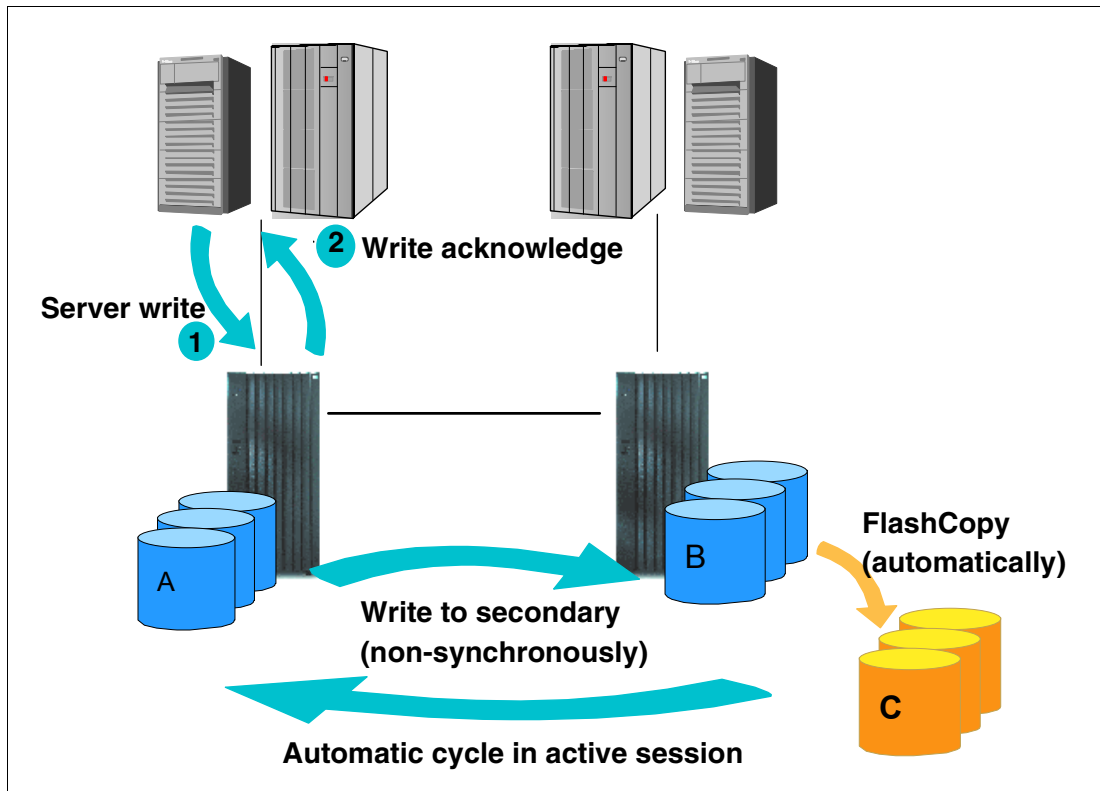


Figure 7-9 Global Mirror

### **How Global Mirror works**

We explain how Global Mirror works in Figure 7-10 on page 127.

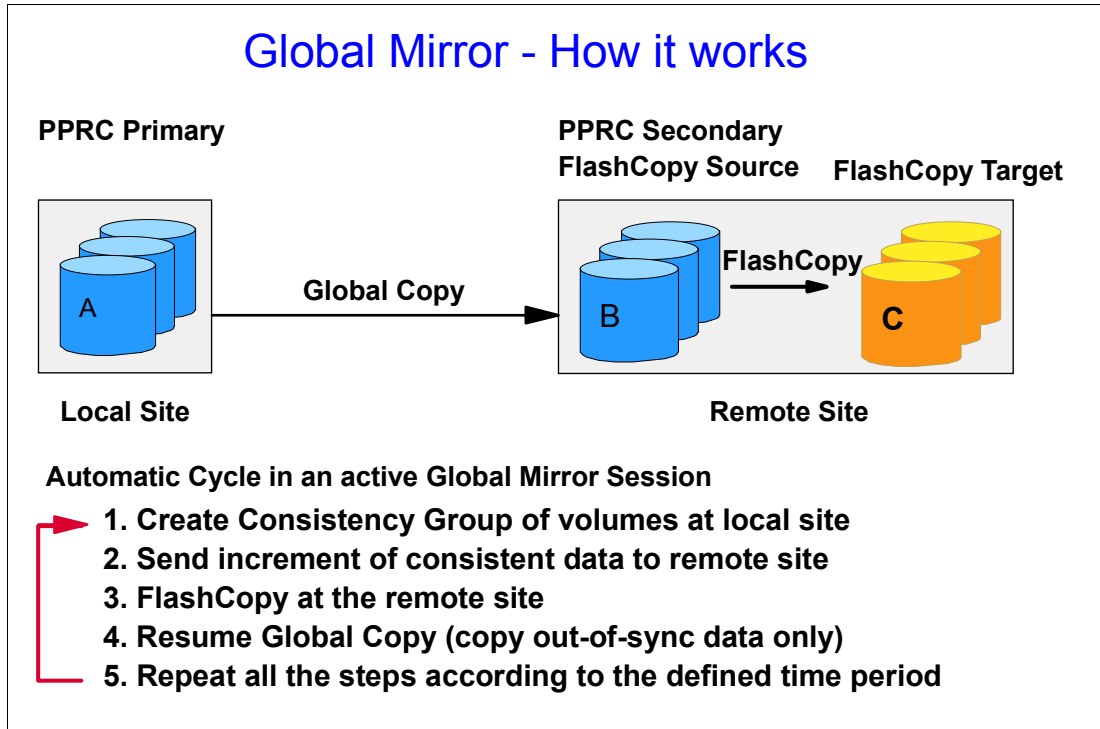


Figure 7-10 How Global Mirror works

The A volumes at the local site are the production volumes and are used as Global Copy primary volumes. The data from the A volumes is replicated to the B volumes, which are Global Copy secondary volumes. At a certain point in time, a Consistency Group is created using all of the A volumes, even if they are located in different ESS boxes. This has no application impact because the creation of the Consistency Group is very quick (on the order of milliseconds).

**Note:** The copy created with Consistency Group is a power-fail consistent copy, not an application-based consistent copy. When you recover with this copy, you may need recovery operations, such as the `fsck` command in an AIX filesystem.

Once the Consistency Group is created, the application writes can continue updating the A volumes. The increment of the consistent data is sent to the B volumes using the existing Global Copy relationship. Once the data reaches the B volumes, it is FlashCopied to the C volumes.

The C volumes now contain the *consistent* copy of data. Because the B volumes usually contain a *fuzzy* copy of the data from the local site (not when doing the FlashCopy), the C volumes are used to hold the last point-in-time consistent data while the B volumes are being updated by the Global Copy relationship.

**Note:** When you implement Global Mirror, you set up the FlashCopy between the B and C volumes with *No Background copy* and *Start Change Recording* options. It means that before the latest data is updated to the B volumes, the last consistent data in the B volume is moved to the C volumes. Therefore, at some time, a part of consistent data is in the B volume, and the other part of consistent data is in the C volume.

If a disaster occurs during the FlashCopy of the data, special procedures are needed to finalize the FlashCopy.

In the recovery phase, the consistent copy is created in the B volumes. You need some operations to check and create the consistent copy.

You need to check the status of the B volumes for the recovery operations. Generally, these check and recovery operations are complicated and difficult with the GUI or CLI in a disaster situation. Therefore, you may want to use some management tools, (for example, Global Mirror Utilities), or management software, (for example, Multiple Device Manager Replication Manager), for Global Mirror to automate this recovery procedure.

The data at the remote site is current within 3 to 5 seconds, but this recovery point (RPO) depends on the workload and bandwidth available to the remote site.

In contrast to the previously mentioned Global Copy solution, Global Mirror overcomes its disadvantages and automates all of the steps that have to be done manually when using Global Copy.

If you use Global Mirror, you must adhere to the following additional rules:

- ▶ You must purchase a Point-in-Time Copy function authorization (2244 Model PTC) for the secondary storage unit.
- ▶ If Global Mirror will be used during failback on the secondary storage unit, you must also purchase a Point-in-Time Copy function authorization for the primary system.

**Note:** PPRC can do failover and failback operations. A failover operation is the process of temporarily switching production to a backup facility (normally your recovery site) following a planned outage, such as a scheduled maintenance period or an unplanned outage, such as a disaster. A failback operation is the process of returning production to its original location. These operations use Remote Mirror and Copy functions to help reduce the time that is required to synchronize volumes after the sites are switched during a planned or unplanned outage.

## **z/OS Global Mirror (XRC)**

DS8000 storage complexes support z/OS Global Mirror only on zSeries hosts. The z/OS Global Mirror function mirrors data on the storage unit to a remote location for disaster recovery. It protects data consistency across all volumes that you have defined for mirroring. The volumes can reside on several different storage units. The z/OS Global Mirror function can mirror the volumes over several thousand kilometers from the source site to the target recovery site. With z/OS Global Mirror, you can suspend or resume service during an outage. You do not have to terminate your current data-copy session. You can suspend the session, then restart it. Only data that changed during the outage needs to be re-synchronized between the copies. The z/OS Global Mirror function is an optional function. To use it, you must purchase the remote mirror for z/OS 2244 function authorization model, which is 2244 Model RMZ.

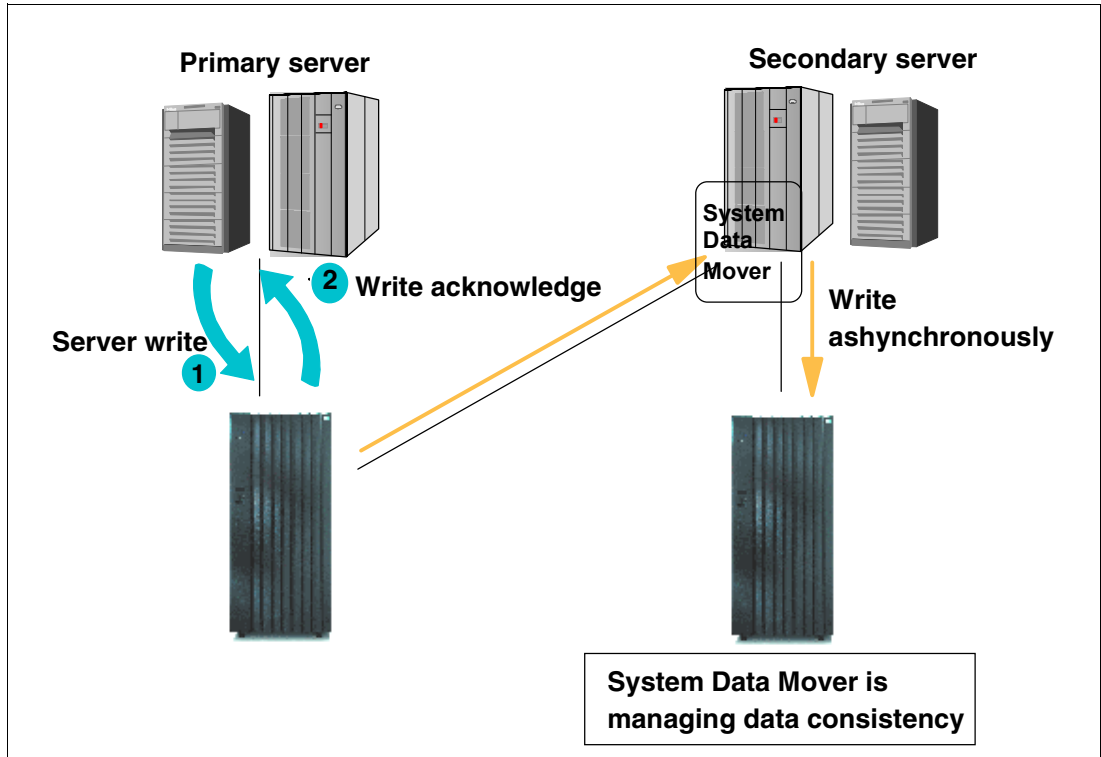


Figure 7-11 z/OS Global Mirror

### **z/OS Metro/Global Mirror (3-site z/OS Global Mirror and Metro Mirror)**

This mirroring capability uses z/OS Global Mirror to mirror primary site data to a location that is a long distance away and also uses Metro Mirror to mirror primary site data to a location within the metropolitan area. This enables a z/OS 3-site high availability and disaster recovery solution for even greater protection from unplanned outages. The z/OS Metro/Global Mirror function is an optional function. To use it, you must purchase both of the following functions:

- ▶ Remote Mirror for z/OS (2244 Model RMZ)
- ▶ Remote Mirror and Copy function (2244 Model RMC) for both the primary and secondary storage units

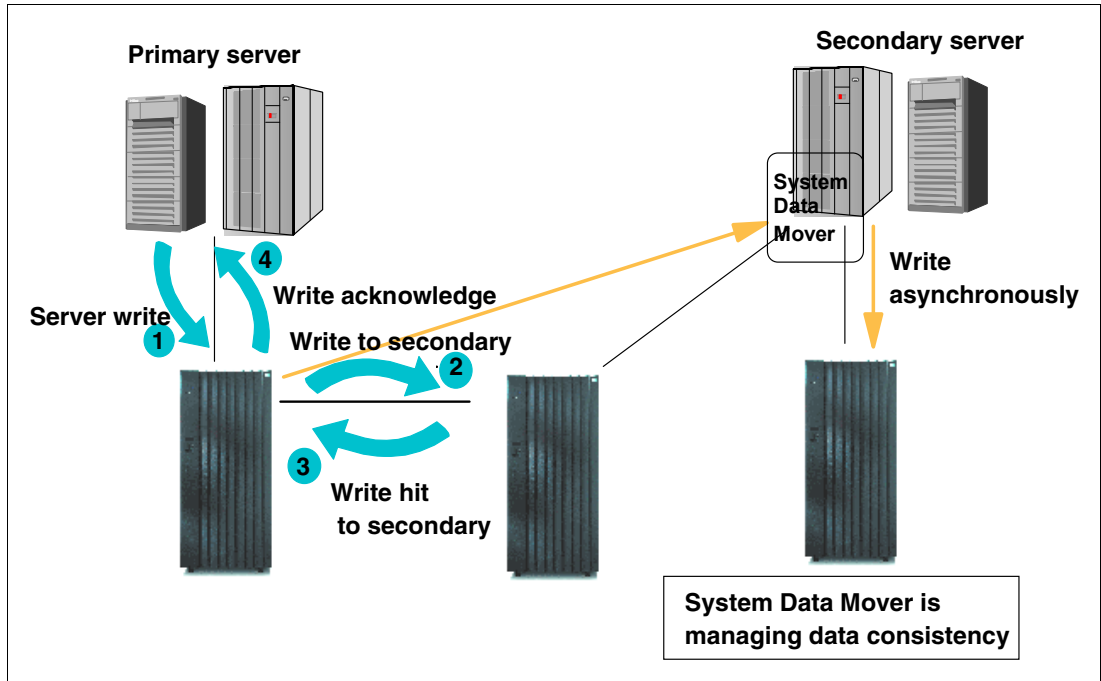


Figure 7-12 z/OS Metro/Global Mirror

## 7.2.4 Comparison of the Remote Mirror and Copy functions

In this section we summarize the use of and considerations for Remote Mirror and Copy functions.

### Metro Mirror (Synchronous PPRC)

- ▶ **Description**  
Metro Mirror is a function for synchronous data copy at a distance.
- ▶ **Advantages**  
There is no data loss and it allows for rapid recovery for distances up to 300 km.
- ▶ **Considerations**  
There may be slight performance impact for write operations.



**Note:** If you want to use a PPRC copy from the application server which has the PPRC primary volume, you need to compare its function with OS mirroring.

You will have some disruption to recover your system with PPRC secondary volumes in an open system environment because PPRC secondary volumes are not online to the application servers during the PPRC relationship.

You may also need some operations before assigning PPRC secondary volumes. For example, in an AIX environment, AIX assigns specific IDs to each volume (PVID). PPRC secondary volumes have the same PVID as PPRC primary volumes. AIX cannot manage the volumes with the same PVID as different volumes. Therefore, before using the PPRC secondary volumes, you need to clear the definition of the PPRC primary volumes or reassign PVIDs to the PPRC secondary volumes.

Some operating systems (OS) or file systems (for example, AIX LVM) have a function for disk mirroring. OS mirroring needs some server resources, but usually can keep operating with the failure of one volume of the pair and recover from the failure non-disruptively. If you use a PPRC copy from the application server for recovery, you need to consider which solution (PPRC or OS mirroring) is better for your system.

## Global Copy (PPRC-XD)

- ▶ Description

Global Copy is a function for continuous copy without data consistency.

- ▶ Advantages

It can copy your data at nearly an unlimited distance, even if you are limited by the network and channel extender capabilities. It is suitable for data migration and daily backup to the remote site.

- ▶ Considerations

The copy is normally *fuzzy* but can be made consistent through synchronization.

**Note:** When you create a consistent copy for Global Copy, you need the go-to-sync (synchronize the secondary volumes to the primary volumes) operation. During the go-to-sync operation, PPRC changes from a non-synchronous copy to a synchronous copy. Therefore, the go-to-sync operation may cause performance impact to your application system. If the data is heavily updated and the network bandwidth for PPRC is limited, the time for the go-to-sync operation becomes longer.

## Global Mirror (Asynchronous PPRC)

- ▶ Description:

Global Mirror is an asynchronous copy; you can create a consistent copy in the secondary site with an adaptable Recovery Point Objective (RPO).

**Note:** Recovery Point Objective (RPO) specifies how much data you can afford to re-create should the system need to be recovered.

- ▶ Advantages:

Global Mirror can copy with nearly an unlimited distance. It is scalable across the storage units. It can realize a low RPO with enough link bandwidth. Global Mirror causes only a slight impact to your application system.

► Considerations:

When the link bandwidth capability is exceeded with a heavy workload, the RPO might grow.

**Note:** To manage Global Mirror, you need many complicated operations. Therefore, we recommend management utilities (for example, Global Mirror Utilities) or management software (for example, IBM Multiple Device Manager) for Global Mirror.

### **z/OS Global Mirror (XRC)**

► Description

z/OS Global Mirror is an asynchronous copy controlled by z/OS host software, called *System Data Mover*.

► Advantages

It can copy with nearly unlimited distance. It is highly scalable, and it has very low RPO.

► Considerations

Additional host server hardware and software is required. The RPO might grow if bandwidth capability is exceeded, or host performance might be impacted.

## **7.2.5 What is a Consistency Group?**

With Copy Services, you can create *Consistency Groups* for FlashCopy and PPRC. Consistency Group is a function to keep *data consistency* in the backup copy. Data consistency means that the order of dependent writes is kept in the copy.

In this section we define *data consistency* and *dependent writes*, and then we explain how Consistency Group operations keep data consistency.

### ***What is data consistency?***

Many applications, such as databases, process a repository of data that has been generated over a period of time. Many of these applications require that the repository is in a consistent state in order to begin or continue processing. In general, consistency implies that the order of dependent writes is preserved in the data copy. For example, the following sequence might occur for a database operation involving a log volume and a data volume:

1. Write to log volume: Data Record #2 is being updated.
2. Update Data Record #2 on data volume.
3. Write to log volume: Data Record #2 update complete.

If the copy of the data contains any of these combinations then the data is consistent:

- Operation 1, 2, and 3
- Operation 1 and 2
- Operation 1

If the copy of data contains any of those combinations then the data is *inconsistent* (the order of dependent writes was *not* preserved):

- Operation 2 and 3
- Operation 1 and 3
- Operation 2

► Operation 3

In the Consistency Group operation, data consistency means this sequence is always kept in the backup data.

And, the order of non-dependent writes does not necessarily need to be preserved. For example, consider the following two sequences:

1. Deposit paycheck in checking account A
2. Withdraw cash from checking account A
3. Deposit paycheck in checking account B
4. Withdraw cash from checking account B

In order for the data to be consistent, the deposit of the paycheck must be applied *before* the withdraw of cash for each of the checking accounts. However, it does not matter whether the deposit to checking account A or checking account B occurred first, as long as the associated withdrawals are in the correct order. So for example, the data copy would be consistent if the following sequence occurred at the copy. In other words, the order of updates is not the same as it was for the source data, but the order of *dependent* writes is still preserved.

1. Deposit paycheck in checking account B
2. Deposit paycheck in checking account A
3. Withdraw cash from checking account B
4. Withdraw cash from checking account A

***How does Consistency Group keep data consistency?***

Consistency Group operations cause the storage units to hold I/O activity to a volume for a time period by putting the source volume into an *extended long busy* state. This operation can be done across multiple LUNs or volumes, and even across multiple storage units.

In the storage subsystem itself, each command is managed with each logical subsystem (LSS). This means that there are slight time lags until each volume in the different LSS is changed to the extended long busy state. Some people are concerned that the time lag causes you to lose data consistency, but, it is not true. We explain how to keep data consistency in the Consistency Group environments in the following section.

See Figure 7-13 on page 134. In this case, three write operations (1st, 2nd, and 3rd) are dependent writes. It means that these operations must be completed sequentially.

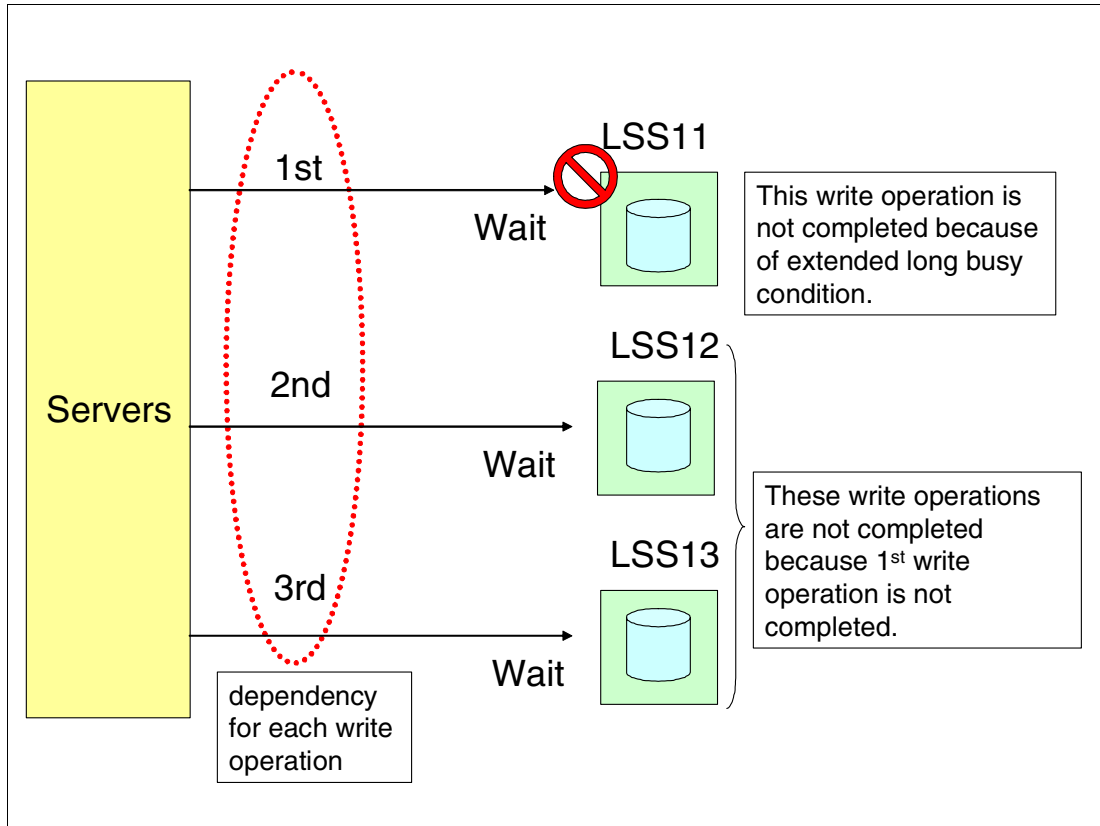


Figure 7-13 Consistency Group: Example1

Because of the time lag for Consistency Group operations, some volumes in some LSSs are in an extended long busy state and other volumes in the other LSSs are not.

In Figure 7-13, the volumes in LSS11 are in an extended long busy state, and the volumes in LSS12 and 13 are not. The 1st operation is not completed because of this extended long busy state, and the 2nd and 3rd operations are not completed, because the 1st operation has not been completed. In this case, 1st, 2nd, and 3rd updates are not included in the backup copy. Therefore, this case is consistent.

See Figure 7-14 on page 135. In this case, the volumes in LSS12 are in an extended long busy state and the other volumes in LSS11 and 13 are not. The 1st write operation is completed because the volumes in LSS11 are not in an extended long busy state. The 2nd write operation is not completed because of the extended long busy state. The 3rd write operation is also not completed because the 2nd operation is not completed. In this case, the 1st update is included in the backup copy, and the 2nd and 3rd updates are not included. Therefore, this case is consistent.

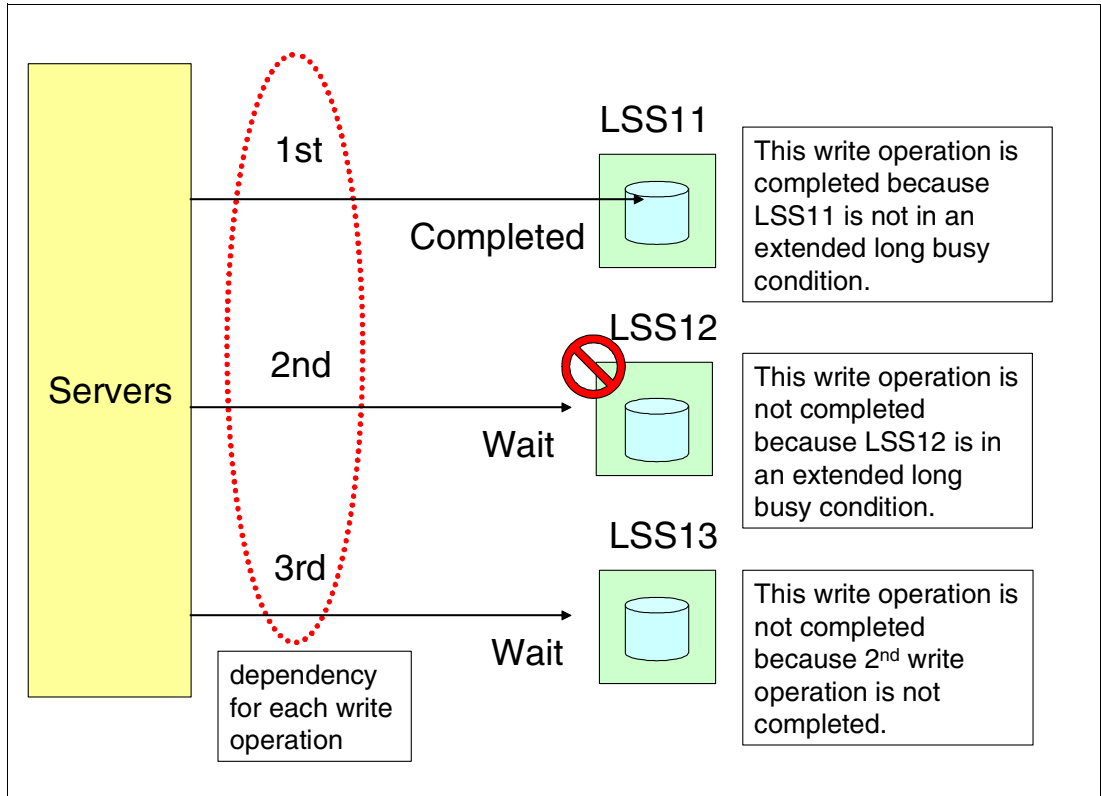


Figure 7-14 Consistency Group: Example 2

In all cases, if each write operation is dependent, Consistency Group can keep the data consistent in the backup copy.

If each write operation is not dependent, the I/O sequence is not kept in the copy that is created by the Consistency Group operation. See Figure 7-15 on page 136. In this case, the three write operations are independent. If the volumes in LSS12 are in an extended long busy state and the other volumes in LSS11 and 13 are not, the 1st and 3rd operations are completed and the 2nd operation is not completed.

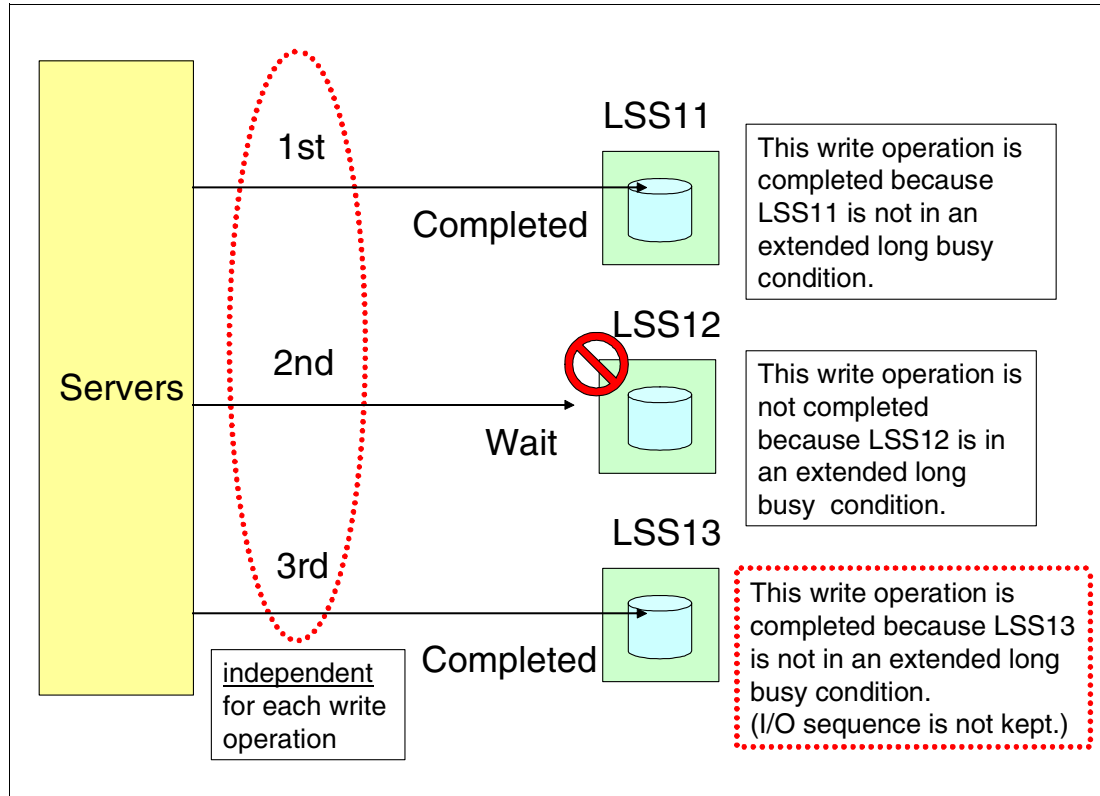


Figure 7-15 Consistency Group: Example.3

In this case, the copy created by the Consistency Group operation reflects only the 1st and 3rd write operation, not including the 2nd operation.

If you accept this result, you can use the Consistency Group operation with your applications. But, if you cannot accept it, you should consider other procedures without Consistency Group operation. For example, you could stop your applications for a slight interval for the backup operations.

## 7.3 Interfaces for Copy Services

There are multiple interfaces for invoking Copy Services. We describe them in this section.

### 7.3.1 Storage Hardware Management Console (S-HMC)

Copy Services functions can be initiated over the following interfaces:

- ▶ zSeries Host I/O Interface
- ▶ DS Storage Manager web-based Interface
- ▶ DS Command-Line Interface (DS CLI)
- ▶ DS open application programming interface (DS Open API)

DS Storage Manager, DS CLI, and DS Open API commands are issued via the Ethernet network, and these commands are invoked by the Storage Hardware Management Console (S-HMC). When the S-HMC has the command requests, including those for Copy Services,

from these interfaces, S-HMC communicates with each server in the storage units via the Ethernet network. Therefore, the S-HMC is a key component to configure and manage the DS8000.

The network components for Copy Services are illustrated in Figure 7-16.

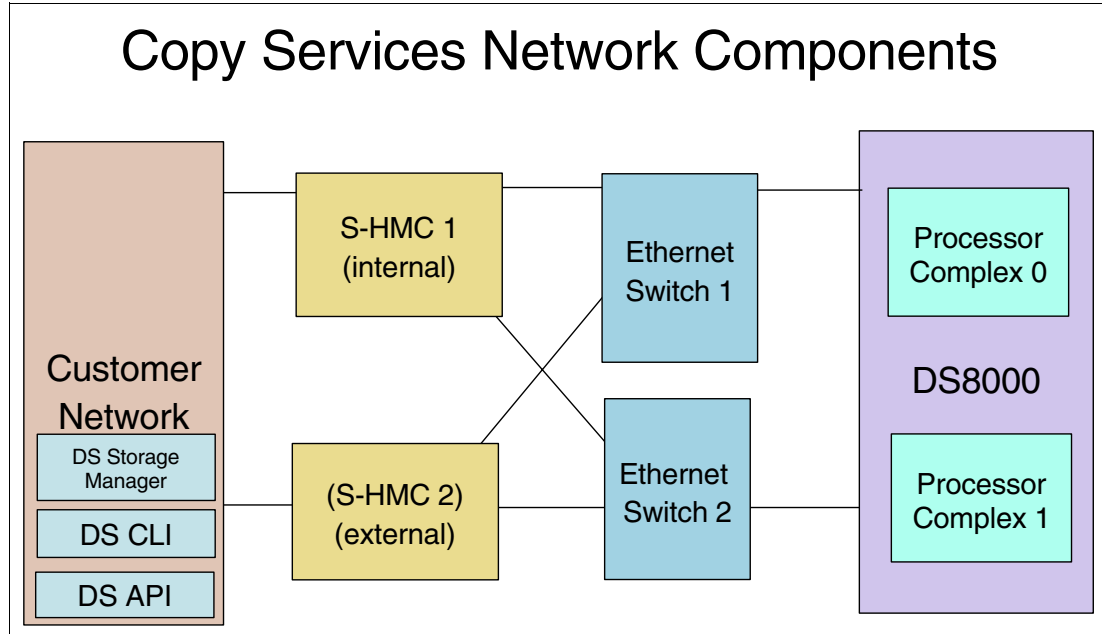


Figure 7-16 DS8000 Copy Services network components

Each DS8000 will have an internal S-HMC in the base frame, and you can have an external S-HMC for redundancy.

For further information about the S-HMC, see Chapter 9, "Configuration planning" on page 157.

### 7.3.2 DS Storage Manager Web-based interface

DS Storage Manager is a Web-based management interface. It is used for managing the logical configurations and invoking the Copy Services functions. The DS Storage Manager has an online mode and an offline mode; only the online mode is supported for Copy Services.

DS Storage Manager is already installed in the S-HMC. You can also install it in other computers that you prepare. When you manage the Copy Services functions with DS Storage Manager in your computers, DS Storage Manager issues its command to the S-HMC via the Ethernet network.

The DS Storage Manager can be used for almost all functions for Copy Services. The following functions cannot be issued from the DS Storage Manager in the current implementation:

- ▶ Consistency Group operation (FlashCopy and PPRC)
- ▶ Inband commands over Remote Mirror link

### 7.3.3 DS Command-Line Interface (DS CLI)

The IBM TotalStorage DS Command-Line Interface (CLI) helps enable open systems hosts to invoke and manage the Point-in-Time Copy and Remote Mirror and Copy functions through batch processes and scripts. The CLI provides a full-function command set that allows you to check your storage unit configuration and perform specific application functions when necessary. The following list highlights a few of the specific types of functions that you can perform with the DS CLI:

- ▶ Check and verify your storage unit configuration.
- ▶ Check the current Copy Services configuration that is used by the storage unit.
- ▶ Create new logical storage and Copy Services configuration settings.
- ▶ Modify or delete logical storage and Copy Services configuration settings.

**Tip: *What is changed from the ESS CLI?***

ESS CLI is a prior version of the command-line interface. It is used for managing the ESS. There are differences between the ESS CLI and DS CLI.

The ESS CLI needs two steps to issue Copy Service functions:

1. Register a Copy Services task from the Web user interface.
2. Issue the registered task from CLI.

The DS CLI has no need to register a Copy Services task before you issue the CLI. You can easily implement and dynamically change your Copy Services operation without the GUI.

For further information about the DS CLI, see Chapter 11, “DS CLI” on page 231.

### 7.3.4 DS Open application programming Interface (API)

The DS Open application programming interface (API) is a non-proprietary storage management client application that supports routine LUN management activities, such as LUN creation, mapping and masking, and the creation or deletion of RAID-5 and RAID-10 volume spaces. The DS Open API also enables Copy Services functions such as FlashCopy and Remote Mirror and Copy. It supports these activities through the use of the Storage Management Initiative Specification (SMIS), as defined by the Storage Networking Industry Association (SNIA)

The DS Open API helps integrate DS configuration management support into storage resource management (SRM) applications, which allow customers to benefit from existing SRM applications and infrastructures. The DS Open API also enables the automation of configuration management through customer-written applications. Either way, the DS Open API presents another option for managing storage units by complementing the use of the IBM TotalStorage DS Storage Manager web-based interface and the DS Command-Line Interface.

You must implement the DS Open API through the IBM TotalStorage Common Information Model (CIM) agent, a middleware application that provides a CIM-compliant interface. The DS Open API uses the CIM technology to manage proprietary devices such as open system devices through storage management applications. The DS Open API allows these storage management applications to communicate with a storage unit.



IBM will support IBM TotalStorage Multiple Device Manager (MDM) for the DS8000 under the IBM TotalStorage Productivity Center in the future. MDM consists of software components which enable storage administrators to monitor, configure, and manage storage devices and subsystems within a SAN environment. MDM also has a function to manage the Copy Services functions, called the *MDM Replication Manager*.

For further information about MDM, see 15.5, “IBM TotalStorage Productivity Center” on page 326.

## 7.4 Interoperability with ESS

Copy Services for DS8000 also supports the IBM Enterprise Storage Server Model 800 (ESS 800) and Model 750. To manage the ESS 800 from the Copy Services for DS8000, you need to install licensed internal code version 2.4.2 or later on the ESS 800.

The DS CLI supports the DS8000, DS6000, and ESS 800 at the same time. DS Storage Manager does not support ESS 800.

**Note:** DS8000 does not support PPRC with an ESCON link. If you want to configure a PPRC relationship between a DS8000 and ESS 800, you have to use a FCP link.

## 7.5 Future Plans

According to the announcement letter, IBM has issued the following Statement of General Direction:

*IBM intends to offer a long-distance business continuance solution across three sites allowing for recovery from the secondary or tertiary site with full data consistency.*



# Planning and configuration

In this part we present an overview of the planning and configuration necessary before installing your DS8000. The topics include:

- ▶ Installation planning
- ▶ Configuration planning
- ▶ Logical configuration
- ▶ CLI - Command-Line Interface
- ▶ Performance





## Installation planning

This chapter discusses various physical considerations and preparation involved in planning the physical installation of a new DS8000 in your environment. The following topics are covered:

- ▶ Installation site preparation
- ▶ Host attachments
- ▶ Network and SAN requirements

## 8.1 General considerations

Successful installation of a DS8000 requires careful planning. The main considerations when planning for the physical installation of a new DS8000 include the following:

- ▶ Delivery
- ▶ Floor loading
- ▶ Space occupation
- ▶ Electrical power
- ▶ Operating environment
- ▶ FAN and cooling
- ▶ Host attachment
- ▶ Network and SAN requirements

Always refer to the most recent information for physical planning in the *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495. You can download this at:

<http://www.ibm.com/servers/storage/disk/ds8000>

## 8.2 Delivery requirements

Before you receive your DS8000 shipment, ensure that you meet all delivery requirements. You need to consider the size and weight of the DS8000 rack in its shipping container, as well as how it will be moved from the loading dock to the final installation location.

Use the following steps to ensure that your receiving area and loading ramp can safely accommodate the delivery of your storage unit:

1. Using Table 8-1, review the packaged weight and dimensions of the DS8000 container and other containers that you will receive.

To calculate the weight of your total shipment, add the weight of each model container that you will receive and the weight of one ship group container for each model. If you ordered any external Storage Hardware Management Console (S-HMC), add the weight of those containers, as well.

Table 8-1 shows the final packaged dimensions and maximum packaged weight of the storage unit shipments.

Table 8-1 Packaged dimensions and weight for DS8000 models

Shipping container	Packaged Dimensions (in centimeters and inches)	Maximum Packaged Weight (in kilograms and pounds)
Model 921 pallet or crate	Height 207.5 cm (81.7 in.) Width 101.5 cm (40 in.) Depth 137.5 cm (54.2 in.)	1309 kg (2886 lb)
Model 922 pallet or crate	Height 207.5 cm (81.7 in.) Width 101.5 cm (40 in.) Depth 137.5 cm (54.2 in.)	1368 kg (3016 lb)
Model 9A2 pallet or crate	Height 207.5 cm (81.7 in.) Width 101.5 cm (40 in.) Depth 137.5 cm (54.2 in.)	1368 kg (3016 lb)
Model 92E (expansion unit) pallet or crate	Height 207.5 cm (81.7 in.) Width 101.5 cm (40 in.) Depth 137.5 cm (54.2 in.)	1209 kg (2665 lb)

Shipping container	Packaged Dimensions (in centimeters and inches)	Maximum Packaged Weight (in kilograms and pounds)
Model 9AE (expansion unit) pallet or crate	Height 207.5 cm (81.7 in.) Width 101.5 cm (40 in.) Depth 137.5 cm (54.2 in.)	1209 kg (2665 lb)
(If ordered) External S-HMC container	Height 69.0 cm (27.2 in.) Width 80.0 cm (31.5 in.) Depth 120.0 cm (47.3 in.)	75 kg (165 lb)

**Attention:** A fully configured model in the packaging can weigh over 1406 kg (3100 lbs). Use of fewer than three persons to move it can result in injury.

2. Ensure that your loading dock, receiving area, and elevators can safely support the packaged weight and dimensions of the shipping containers.
3. To compensate for the weight of the DS8000 shipment, ensure that the loading ramp at your site does not exceed an angle of 10°.

## 8.3 Installation site preparation

Before you begin to install a new DS8000, you must ensure that the location where you plan to install your DS8000 storage units meets all requirements.

The topics in this section discuss how to prepare the installation site to meet all of these requirements.

### 8.3.1 Floor and space requirements

When you are planning the location of your storage units, you need to use the following steps to ensure that your planned installation location meets the space and floor load requirements:

1. Identify the base models and expansion models that are included in your storage units. If your storage units use an external Storage Hardware Management Console (S-HMC), include the racks containing the external S-HMCs.
2. Decide whether the storage units will be installed on a raised or nonraised floor.
  - a. If the location has a raised floor, plan where the floor tiles must be cut to accommodate the cables.
  - b. If the location has a nonraised floor, resolve any safety problems caused by the location of cable exits and routing.
3. Determine whether the floor of the location meets the floor load requirements for the storage units. This floor load rating is the concrete sub-floor rating, not the raised floor rating.
4. Calculate the amount of space that the storage units will use.
  - a. Identify the total amount of space that is needed for the storage units using the dimensions of the models and the weight distribution areas calculated in step 3. Service clearances between the DS8000 and an adjacent piece of equipment can overlap, but weight distribution areas cannot overlap.
  - b. Ensure that the area around each stand-alone model and each storage unit meets the service clearance requirements.

**Important:** Any expansion units within the storage unit must be attached to the base model on the right side (as you face the front of the units).

## Installing on raised or nonraised floors

You can install your DS8000 storage units on a raised or nonraised floor.

However, installing the models on a raised floor provides the following benefits:

- ▶ Improves operational efficiency and allows greater flexibility in the arrangement of equipment.
- ▶ Increases air circulation for better cooling.
- ▶ Protects the interconnecting cables and power receptacles.
- ▶ Prevents tripping hazards because cables can be routed underneath the raised floor.
- ▶ Aesthetically more appealing.

## Meeting floor load requirements

Use the following steps to ensure that your location meets the floor load requirements and to determine the weight distribution area required for the floor load:

1. Find the concrete sub-floor load rating in the location where you plan to install the storage units. Refer to Chapter 4, “Meeting DS8000 delivery and installation requirements” in the *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

**Important:** If you do not know or are not certain about the floor load rating of the installation site, be sure to check with the building engineer or another appropriate person. A structural engineer may be needed to assist in this evaluation.

2. Determine whether the floor load rating of the location meets the following requirements:
  - The design target used by IBM is 342 kg per m<sup>2</sup> (70 lb per ft<sup>2</sup>).
  - When you install a storage unit, which includes both base models and expansion models, the minimum floor load rating is 361 kg per m<sup>2</sup> (74 lb per ft<sup>2</sup>). At 342 kg per m<sup>2</sup> (70 lb per ft<sup>2</sup>), the side dimension for the weight distribution area exceeds the 76.2 cm (30 in.) allowed maximum, as defined in IBM Corporate Standards.
  - The per caster transferred weight to a raised floor panel is 450 kg (1000 lb).

## Calculating space requirements

When you are planning the installation location, you must first calculate the total amount of space that is needed for the storage units.

Use the following steps to calculate enough space for your storage units:

1. Determine the dimensions of each model configuration in your storage units. Table 8-2 on page 147 provides the dimensions of the DS8000 models.



Table 8-2 The DS8000 dimensions

Configuration/ Attribute	2107-921/922/9A2 (Base frame only)	2107-921/922/9A2 2107-92E/9AE (Expansion frame)	2107-922/9A2 2107-92E/9AE (2 Expansion frames)
Dimensions with covers Height x Width x Depth (Std/Metric)	76 x 33.3 x 46.7 inch 193 x 84.7 x 118.3 cm	76 x 69.7 x 46.7 inch 193 x 172.7 x 118.3 cm	76 x 102.6 x 46.7 inch 193 x 260.9 x 118.3 cm
Footprint® (Std/Metric)	10.77ft <sup>2</sup> 1.002 m <sup>2</sup>	22.0 ft <sup>2</sup> 2.05 m <sup>2</sup>	33.23 ft <sup>2</sup> 3.095 m <sup>2</sup>

2. Calculate the total area that is needed for the model configuration by adding the weight distribution area to the dimensions.
3. Determine the total space that is needed for the storage units by planning where you will place each model configuration in the storage units and how much area each configuration will need based on step 2.
4. Verify that the planned space and layout also meets the service clearance requirements for each unit and system.

**Note:** The DS8000 requires service clearances of 121.9 cm (48 inches) at the front and 76.2 cm (30 inches) at the rear. Racks can be placed side by side if floor loading restrictions allow this.

### 8.3.2 Power requirements

When you consider the DS8000 storage complex location, consider the following issues:

- ▶ Power control selections
- ▶ Power outlet requirements
- ▶ Input voltage requirements
- ▶ Power connector requirements
- ▶ Power consumption and environment

#### Power control

The DS8000 provides power controls on the model racks and through the Management Console.

The DS8000 models have the following manual power controls in the form of physical switches on the racks:

- ▶ Remote force power off (Remote FPO) switch (available on base models)
  - Planning requirements:** To use this feature, you must supply a remote force power off circuit. See the *DS8000 Introduction and Planning Guide*, GC35-0495, for more detailed information.
- ▶ Local/remote switch (available on base models)
- ▶ Local power on/local force power off switch (available on base models)
- ▶ Unit emergency power off (UEPO) switch (available on all models)

**Note:** Use this switch only in extreme emergencies. Using this switch may result in data loss.

You can use the following power controls through the DS Storage Manager (running on the Management Console):

- ▶ Local power control mode (visible in the DS Storage Manager)
- ▶ Remote power control mode (visible in the DS Storage Manager)

If you select the Remote power control mode, you choose one of the following remote mode options:

- **Remote Management Console, Manual:** Your use of the DS Storage Manager power on/off page controls when the unit powers on and off.
- **Remote Management Console, Scheduled:** A schedule, which you set up, controls when the unit powers on and off.
- **Remote Management Console, Auto:** This setting applies only in situations in which input power is lost. In those situations, the unit powers on as soon as external power becomes available again.
- **Remote Auto/Scheduled:** A schedule, which you set up, controls when the unit powers on and off. A power on sequence is also initiated if the unit was powered off due to an external power loss during the time that the units are scheduled to be on and external power becomes available again.
- **Remote zSeries Power Control:** One or more attached zSeries units control the power on and power off sequences.

**Planning requirements:** If you choose the **Remote zSeries Power Control** options, you must have the remote zSeries power control feature. DS8000 feature code 1000 is needed in order to provide the necessary power sequence cables to connect to the zSeries processor.

## Power outlet requirements

You must supply the following power outlets for the installation of your storage units:

- ▶ Two independent power outlets for the two DS8000 power line cords are needed by each base model and expansion model.

**Note:** To eliminate a single point of failure, the outlets must be independent. This means that each outlet must use a separate power source and each power source must have its own wall circuit breaker.

- ▶ Two outlets that are within 3.1 m (10 ft) of the external Storage Hardware Management Console (S-HMC). These outlets may be provided from Power Distribution Units in the rack or from an external power source.

## Input voltage requirements

All power inputs are balanced three phase.

Due to different power requirements across the world, you must specify the nominal AC voltage (phase to phase) that is supported by the power supply that is installed on the model.

Use the following feature codes when you specify the input voltage for your base or expansion model:

- ▶ **9090** AC input voltage: 200 V to 240 V
- ▶ **9091** AC input voltage: 380 V to 480 V

### Power connector requirements

The cable connectors supplied with various line cords, and the required receptacles are covered in Chapter 4, “Meeting DS8000 delivery and installation requirements” in the *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

### Power consumption and environmental information

When you are planning the power requirements for the DS8000, consider the power consumption and other environmental points of the storage unit. Table 8-3 provides the power consumption for various DS8000 models.

Table 8-3 Power consumption for the DS8000 Models

Measurement	Units	Model 921	Model 922	Model 9A2	Expansion Model 92E	Expansion Model 9AE
Peak electric power	kilovolt amperes (kVA)	5.5	7.0	7.0	6.0	6.0

For more information, refer to the Chapter 4, “Meeting DS8000 delivery and installation requirements” in the *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

## 8.3.3 Environmental requirements

To properly maintain your DS8000 storage unit, you must install your storage unit in a location that meets the operating environment requirements.

Take the following steps to ensure that you meet these requirements:

1. Note where the air intake locations are on the models that compose your storage unit.
2. Verify that you can meet the environmental operating requirements at the air intake locations.
3. Consider optimizing the air circulation and cooling for the storage unit by using a raised floor, adjusting the floor layout, and adding perforated tiles around the air intake areas.

### Fans and air intake areas

The DS8000 models provide air circulation through various fans throughout the frame. You must maintain the correct operating environment requirements for your models at each air intake location.

### Operating environment requirements

The DS8000 should be maintained within an operating temperature range of 20 to 25 degrees Celsius (68 to 77 degrees Fahrenheit). The recommended operating temperature with the power on is 22 degrees Celsius (72 degrees Fahrenheit). We strongly recommend that you avoid running the DS8000, or any disk storage equipment, at temperatures outside this temperature range.

The humidity range should be maintained between 40% and 50%. The recommended operating point with the power on is 45%.

## Cooling the storage complex

Adequate airflow needs to be maintained to ensure effective cooling. You can take steps to optimize the air circulation and cooling for your storage units.

To optimize the cooling around your storage units, prepare the location of your storage units as recommended in the following steps:

1. Install the storage unit on a raised floor. Although you can install the storage unit on a non-raised floor, installing the storage unit on a raised floor provides increased air circulation for better cooling.
2. Install perforated tiles in the front and back of each base model and expansion model as follows:
  - a. For a stand-alone base model, install two fully perforated tiles in front of each base model and one partially perforated tile at the back of each base model.
  - b. For a row of machines, install a row of perforated tiles in front of the machines and one or two fully perforated tiles at the back of each two machines.
  - c. For groupings of machines, where a hot aisle/cold aisle layout is used, use a cold aisle row of perforated tiles in front of all machines. For hot aisles, install a perforated tile per pair of machines.

## 8.4 Host attachment

The DS8000 storage unit provides a variety of host attachments so that you can consolidate storage capacity and workloads for open systems hosts, S/390 hosts, and eServer zSeries hosts.

**Note:** There is no SCSI attachment support for the DS8000 storage unit.

The DS8100 Model 921 supports a maximum of 16 host adapters and 4 device adapter pairs and the DS8300 Models 922 and 9A2 support a maximum of 32 host adapters and 8 device adapter pairs.

You can configure the storage unit for any of the following system adapter types and protocols:

- ▶ Fibre channel adapters, for support of Fibre Channel protocol (FCP) and fibre connection (FICON) protocol
- ▶ Enterprise Systems Connection Architecture® (ESCON) adapters

### 8.4.1 Attaching to open systems hosts

You can attach a DS8000 storage unit to an open systems host with Fibre Channel adapters.

Fibre Channel is a 1 Gbps or 2 Gbps, full-duplex, serial communications technology to interconnect I/O devices and host systems that are separated by tens of kilometers.

The IBM TotalStorage DS8000 series supports 1 Gbps and 2 Gbps connections. The DS8000 series negotiates automatically and determines whether it is best to run at 1 Gbps link or 2 Gbps link.

Fibre channel connections are established between Fibre Channel ports that reside in I/O devices, host systems, and the network that interconnects them. Each storage unit Fibre

Channel adapter has four ports. Each port has a unique worldwide port name (WWPN). You can configure the port to operate with the SCSI-FCP upper-layer protocol. Shortwave adapters and longwave adapters are available on the storage unit. Fibre channel adapters for SCSI-FCP support provide the following configurations:

- ▶ A maximum of 64 host ports for DS8100 Model 921 and a maximum of 128 host ports for DS8300 Models 922 and 9A2
- ▶ A maximum of 8K host logins per Fibre Channel port
- ▶ A maximum of 2000 N-port logins per storage image
- ▶ Access to all 65,280 LUNs per target (one target per host adapter), depending on host type
- ▶ Either arbitrated loop, switched fabric, or point-to-point topologies

For more support information about the open system platforms, refer to 15.1.1, “Supported operating systems and servers” on page 320.

### 8.4.2 ESCON-attached S/390 and zSeries hosts

You can attach the DS8000 storage unit to the ESCON-attached S/390 and zSeries hosts. With ESCON adapters, the storage unit provides the following configurations:

- ▶ A maximum of 32 host ports for DS8100 Model 921 and a maximum of 64 host ports for DS8300 Models 922 and 9A2.
- ▶ A maximum of 64 logical paths per port.
- ▶ Access to 16 control-unit images (4096 CKD devices) over a single ESCON port on the storage unit.
- ▶ Zero to 64 ESCON channels; two per ESCON host adapter.
- ▶ Two ESCON links with each link supporting up to 64 logical paths.
- ▶ DS8100 storage unit supports up to 16 host adapters that provide a maximum of 32 ESCON links per machine. A DS8300 storage unit supports up to 32 host adapters that provide a maximum of 64 ESCON links per machine.

The FICON bridge card in ESCON director 9032 Model 5 enables a FICON bridge channel to connect to ESCON host adapters in the storage unit. The FICON bridge architecture supports up to 16384 devices per channel. The storage unit supports the following operating systems for S/390 and zSeries hosts:

- ▶ Transaction Processing Facility (TPF)
- ▶ Virtual Storage Extended/Enterprise Storage Architecture (VSE/ESA™)
- ▶ z/OS
- ▶ z/VM
- ▶ Linux

### 8.4.3 FICON-attached S/390 and zSeries hosts

You can attach the DS8000 storage unit to FICON-attached S/390 and zSeries hosts.

Each storage unit Fibre Channel adapter has four ports. Each port has a unique world wide port name (WWPN). You can configure the port to operate with the FICON upper-layer protocol. When configured for FICON, the Fibre Channel port supports connections to a maximum of 128 FICON hosts. On FICON, the Fibre Channel adapter can operate with fabric

or point-to-point topologies. With Fibre Channel adapters that are configured for FICON, the storage unit provides the following configurations:

- ▶ Either fabric or point-to-point topologies
- ▶ A maximum of 64 host ports for DS8100 Model 921 and a maximum of 128 host ports for DS8300 Models 922 and 9A2
- ▶ A maximum of 2048 logical paths on each Fibre Channel port
- ▶ Access to all 64 control-unit images (16,384 CKD devices) over each FICON port

#### 8.4.4 Where to get the updated information for host attachment

Due to frequent updates and changes of support information, always refer to the latest *IBM DS8000 Interoperability Matrix* at the following Web site for the details about types, models, adapters, and the operating systems that the storage unit supports:

<http://www.ibm.com/servers/storage/disk/ds8000/interop.htm>

#### Host systems attachment

Refer to the *IBM TotalStorage DS8000 Host Systems Attachment Guide*, SC26-7628, for detailed information on attaching servers to the DS8000 series.

#### Host adapters

For information about supported Fibre Channel HBAs and the recommended or required firmware and device driver levels for all IBM storage systems, you can visit the *IBM HBA Search Tool* site, sometimes also referred to as the *Fibre Channel host bus adapter firmware and driver level matrix*:

<http://knowledge.storage.ibm.com/HBA/HBASearchTool>

Additionally, review host adapter vendor documentation and Web pages to obtain information regarding host adapter configuration planning, hardware and software requirements, driver levels, and release notes.

ATTO:

<http://www.attotech.com/>

Emulex:

<http://www.emulex.com/ts/dds.html>

JNI:

<http://www.jni.com/OEM/oem.cfm?ID=4>

QLogic:

[http://www.qlogic.com/support/oem\\_detail\\_all.asp?oemid=22](http://www.qlogic.com/support/oem_detail_all.asp?oemid=22)

#### SAN Fabric products

Fabric product vendor documentation and Web pages should be reviewed to obtain information regarding configuration planning, hardware and software requirements, firmware and driver levels, and release notes.

IBM:

<http://www.ibm.com/storage/ibmsan/products/sanfabric.html>

CNT (INRANGE):

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

McDATA:

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

Cisco:

<http://www.cisco.com/go/ibm/storage>

### **Channel extension technology products**

Channel extension technology product vendor documentation and Web pages should be reviewed to obtain information regarding configuration planning, hardware and software requirements, firmware and driver levels, and release notes.

Cisco:

<http://www.cisco.com/go/ibm/storage>

CIENA:

<http://www.ciena.com/products/transport/shorthaul/cn2000/index.asp>

CNT:

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

Nortel:

<http://www.nortelnetworks.com/>

ADVA:

<http://www.advaoptical.com/>

## **8.5 Network and SAN requirements**

Your DS8000 storage units must be placed in a location that meets the network and communications requirements.

You should keep in mind the following network and communications issues when you plan the location and interoperability of your storage units:

- ▶ Storage Hardware Management Console network requirements
- ▶ Remote support connection requirements
- ▶ Remote power control requirements
- ▶ SAN considerations

### **8.5.1 S-HMC network requirements**

Generally, each S-HMC requires a dedicated connection to the network. See 9.2, “Storage Hardware Management Console (S-HMC)” on page 158.

## 8.5.2 Remote support connection requirements

You must meet the requirements for the modem and for an outside connection if you will use remote support.

The DS8000 S-HMC contains a modem to take advantage of remote support, which can include outbound support (call home) or inbound support (remote service performed by an IBM next level support representative). For each S-HMC, you must provide the following equipment close enough to the S-HMC to support the modem connection:

- ▶ One analog telephone line for initial setup
- ▶ A telephone cable to connect the modem to a telephone jack

To enable remote support you must allow an external connection, such as one of the following:

- ▶ A telephone line
- ▶ An internet connection through your firewall that allows IBM to use a VPN connection to your S-HMC.

## 8.5.3 Remote power control requirements

Remote power control allows you to control the power of your storage complex through the DS Storage Manager (running on the S-HMC).

There are several settings for remote power control. Only the remote zSeries power control setting requires planning on your part.

The remote zSeries power control setting allows you to power on and off a room from one zSeries interface. If you use the remote zSeries power control setting, you must meet the following requirements:

- ▶ You must order the remote zSeries power control feature.
- ▶ You can allow up to four zSeries remote power-control interfaces.

## 8.5.4 SAN requirements

A Fibre Channel storage area network (SAN) is a specialized, high-speed network that attaches servers and storage devices. With a SAN, you can perform an any-to-any connection across the network using interconnecting elements such as routers, gateways, hubs, and switches.

For a DS8000 configuration, you can use SANs to attach storage units and to attach hosts to the storage unit.

When you connect your DS8000 storage units to a SAN, you must meet the following requirements:

- ▶ When a SAN is used to attach both disks and hosts to the storage unit, any storage device that is managed by the storage unit must be visible to the host systems.
- ▶ When concurrent device adapters and host adapter operations are supported through the same I/O port, the SAN attached to the port must provide both host and device access.
- ▶ Fibre channel host adapters must be configured to operate in a point-to-point mode fabric topology. See the *IBM TotalStorage DS8000 Host Systems Attachment Guide*, SC26-7628, for more information.



You should also keep the following considerations in mind:

- ▶ Fibre channel SANs can provide the capability to interconnect open systems and storage in the same network as S/390 and zSeries host systems and storage.
- ▶ A single Fibre Channel host adapter can have physical access to multiple Fibre Channel ports on the storage unit.

For some Fibre Channel attachments, you can establish zones to limit the access of host adapters to storage system adapters. By establishing zones, you reduce the possibility of interactions between system adapters in switched configurations.

You can configure switch ports that are attached to the storage unit in more than one zone. This enables multiple system adapters to share access to the storage unit Fibre Channel ports. Shared access to a storage unit Fibre Channel port might come from host platforms that support a combination of bus adapter types and operating systems.





# Configuration planning

This chapter discusses planning considerations related to implementing the DS8000 series in your environment. The topics covered are:

- ▶ Configuration planning overview
- ▶ Storage Hardware Management Console (S-HMC)
- ▶ DS8000 licensed functions
- ▶ Capacity planning
- ▶ Data migration planning
- ▶ Planning for performance

For a complete discussion of these topics, see the *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

## 9.1 Configuration planning overview

When installing a DS8000 disk system, various physical requirements need to be addressed to successfully integrate the DS8000 into your existing environment. These requirements include:

- ▶ The DS Management Console (S-HMC), which is the focal point for configuration, Copy Services management, and maintenance for a DS8000 storage unit.
- ▶ Storage features, which include disk enclosures, disk drives sets, standby capacity on demand, disk enclosure fillers, and disk drives cables.
- ▶ I/O adapter features, which are separated into I/O enclosures, device adapters and cables, and host adapters and cables.
- ▶ Processor memory, referring to the amount of memory you want for the processors on your model, which can vary from 16 GB to 256 GB.
- ▶ Other configuration features such as power, power line cords, input voltage requirements, battery assemblies, extended power line disturbance, remote zSeries power control, and shipping weight reduction.
- ▶ Licensed functions include both required and optional features such as the operating environment, Point-in-Time Copy, Remote Mirror and Copy, Remote Mirror for z/OS and Parallel Access Volumes. See 9.3, “DS8000 licensed functions” on page 167 for an in-depth discussion of the licensed functions.
- ▶ Delivery requirements to receive delivery of the DS8000 at your location.
- ▶ Installation site requirements for adequate floor space, power, environmentals, external S-HMC installation, network, and communication.

For a complete discussion of these issues, see *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

## 9.2 Storage Hardware Management Console (S-HMC)

The S-HMC feature looks similar to a laptop. It consists of a workstation processor, keyboard, monitor, modem, and Ethernet cables. The S-HMC is a closed system appliance. Each DS8000 will have an internal S-HMC (feature code 1100) in the base frame, together with a pair of Ethernet switches installed and cabled to the processor complex or external S-HMC, or both. It is a focal point with multiple functions such as:

- ▶ Storage configuration
- ▶ LPAR management
- ▶ Advanced Copy Services invocations
- ▶ Interface for local service personnel
- ▶ Remote service and support

The S-HMC is connected to the storage facility by way of redundant private Ethernet networks. Figure 9-1 on page 159 shows the back of a single S-HMC and a pair of Ethernet switches. The S-HMC has 2 built-in Ethernet ports, one dual-port Ethernet PCI adapter and one PCI modem for asynchronous call home support. The S-HMC's private Ethernet ports shown are configured into port 1 of each Ethernet switch to form the private DS8000 network. The customer Ethernet port indicated is the primary port to be used to connect to the customer network. The empty Ethernet port is normally not used. Corresponding private Ethernet ports of the external S-HMC (FC1110) would be plugged into port 2 of the switches

as shown. To interconnect two DS8000 base frames, FC1190 would provide a pair of 31m Ethernet cables to connect from port 16 of each switch in the second base frame into port 15 of the first frame. If the second S-HMC is installed in the second DS8000, it would remain plugged into port 1 of *its Ethernet* switches.

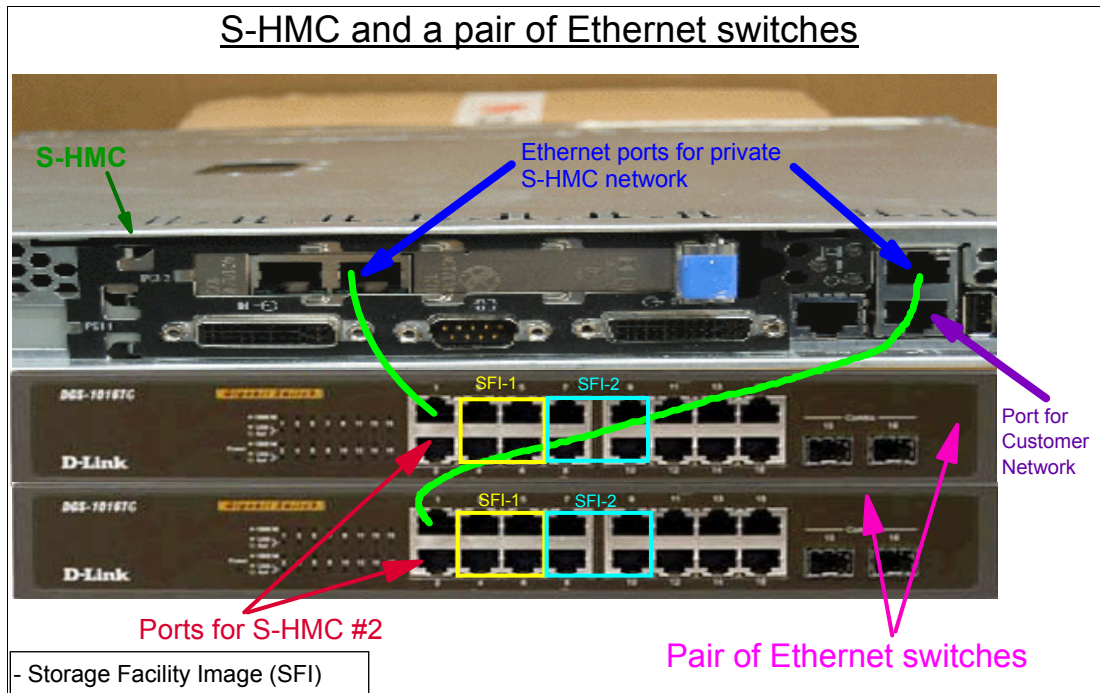


Figure 9-1 S-HMC and Ethernet switches

### 9.2.1 External S-HMC

Additionally, the DS8000 can have an external S-HMC (feature code 1110), which may be ordered today in preparation for its availability later in 2005 when this will be a supported configuration. The external S-HMC is equivalent hardware to the internal S-HMC and needs to be installed in a customer-supplied 19-inch IBM or a non-IBM rack (1U server/1U display). Configuration management is achieved either in real-time or offline configuration mode.

In environments with high availability requirements and Advanced Copy Services, it is recommended that an external S-HMC is installed to eliminate single points of failure. Since the S-HMC is the only service personnel interface available, an external S-HMC will greatly enhance maintenance operational capabilities as a result of internal S-HMC failures. The external S-HMC is an optional priced feature. To help preserve console functionality, the external and the internal S-HMCs are not available to be used as a general purpose computing resource.

**Tip:** To ensure that IBM service representatives can quickly and easily access an external S-HMC, place the external S-HMC rack within 15.2 m (50 ft.) of the storage units connected to it.

## 9.2.2 S-HMC software components

The S-HMC consists of the following software functions:

- ▶ Remote services
- ▶ DS Storage Manager
- ▶ System and partition management
- ▶ Service
- ▶ Storage facility RAS
- ▶ Storage Management Initiative Specification Common Information Mode (SMI-S CIM) server

Figure 9-2 shows the different software components that were available in the ESS 2105 master console, the ESS operating system functions and the pSeries hardware management console, which are now integrated into the S-HMC.

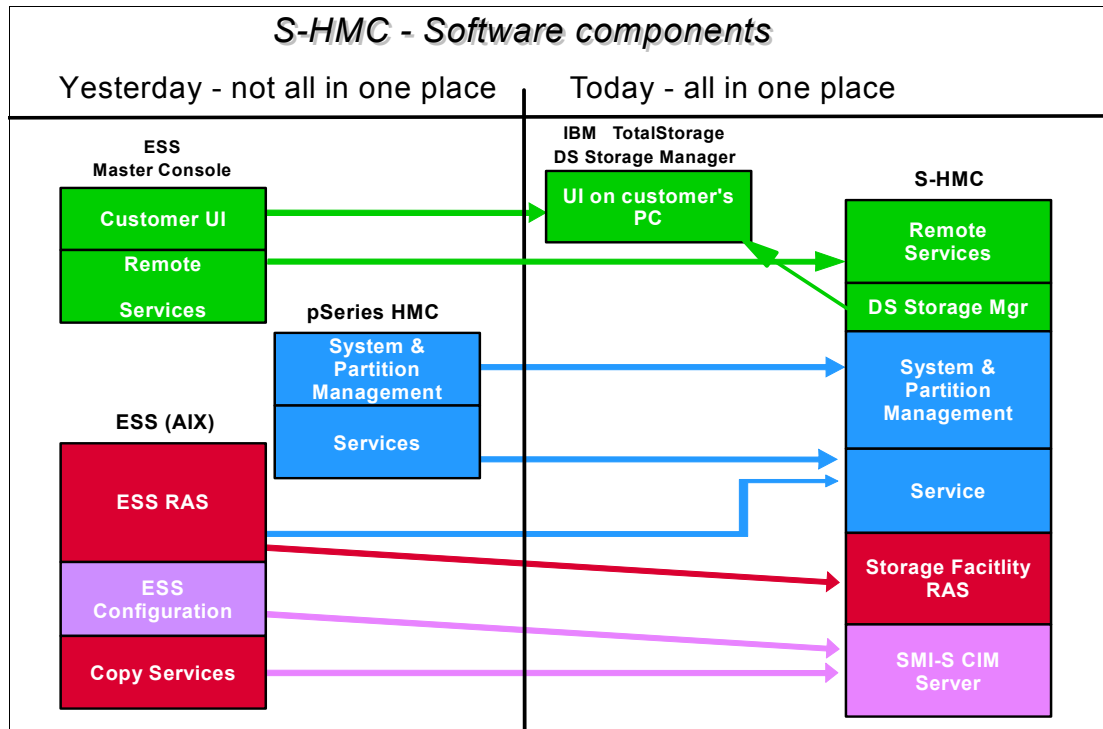


Figure 9-2 Software functions of Management Console

### Remote services

IBM Service personnel located outside of the customer facility log in to the S-HMC to provide service and support. The methods available for IBM to connect to the S-HMC are configured by the IBM SSR at the direction of the customer, and may include dial-up only access or access through the high-speed internet connection.

### DS Storage Manager

The DS Storage Manager is a single integrated Web-based (HTML) graphical user interface that offers:

- ▶ Offline configuration to allow a user to create and save logical configurations and apply them to an online DS8000 series system. This offline configuration tool is installed on a

customer server and can be used for the configuration of a DS8000 series system at initial installation or for reconfiguration activities.

- ▶ Online configuration, which provides real-time configuration management support.
- ▶ Copy Services to allow a user to execute Copy Services functions.

The DS Storage Manager on the internal S-HMC provides only real-time configuration, as opposed to offline configuration. The user may *also* opt to install an additional DS Storage Manager on the user's own workstation. With this DS Management Console, the user will be able to perform both online and offline configurations. In order to perform online configuration the user's workstation must be connected to the S-HMC that is connected to the DS8000. The workstation must meet the following minimum requirements:

- ▶ 1.4 GHz Pentium® 4
- ▶ 256 KB cache
- ▶ 256 MB memory
- ▶ 1 GB disk space for the system management software
- ▶ 1 GB work space per server
- ▶ IP Network connectivity to each port on S-HMC
- ▶ Serial connectivity to your storage unit

You may also have additional IP Network connectivity to an external network to enable call home and remote support.

Operating systems supported on the customer-supplied PC are:

- ▶ Microsoft Windows 2000
- ▶ Microsoft Windows 2000 Server
- ▶ Microsoft Windows XP Professional
- ▶ Microsoft Windows 2003 Server
- ▶ Linux (RedHat AS 2.1)

Table 9-1 shows all the ports needed to connect to and fully operate the DS Storage Manager.

*Table 9-1 Required ports*

Port number from/to	Protocol
1720/1720	tcp
1722/1722	tcp
1750/1750	tcp
8451/8455	tcp

The online configuration and Copy Services are available via a Web browser interface installed on the S-HMC. The DS Storage Manager is provided with the DS8000 series at no additional charge. The S-HMC is required for all customer storage configuration operations. All storage configuration operations are accomplished through invoking SMI-S operations, either through the DS8000 Storage Manager, through the DS8000 Command-Line Interface (CLI), or through any SMI-S compliant storage management agent, such as the IBM TotalStorage Productivity Center (TPC).

## System and partition management

Customer access to the S-HMC is provided to allow the management of optional LPAR environments. This access allows for configuration, activation, and deactivation of logical partitions.

### Service

The service function of the S-HMC is typically used to perform service actions such as a repair action, the installation of a storage facility or an MES, or the installation or upgrade of licensed internal code releases. These service actions can be accessed using the supplied S-HMC or a laptop or equivalent. The laptop will connect to the service port on the Ethernet switch (ESSNet) on the storage facility that is being serviced. This laptop needs to be configured for dynamic host configuration protocol (DHCP) before connecting to the ESSNet. The laptop will use the WebSM client to access the service functions on the S-HMC. The WebSM client is available for downloading to the laptop from the S-HMC.

### Storage facility RAS

This component, together with the hardware, provides the reliability, availability, and serviceability functions of the DS8000. See Chapter 4, “RAS” on page 61 for a more complete description.

### SMI-S CIM server

The SMI-S server is an implementation of the industry standard of the Storage Networking Industry Association (SNIA). The ESS API is implemented through the IBM TotalStorage Common Information Model Agent (CIM Agent) for the DS8000, a middleware application designed to provide a CIM-compliant interface. The CIM-compliant interface allows Tivoli® and third-party software management tools to discover, monitor, and control the DS8000. The DS8000 API and CIM Agent are provided with the DS8000 at no additional charge. The CIM Agent is available for the AIX, Linux, and Windows 2000 operating system environments.

### Advanced Copy Services invocations

The same CIM server that is required to support storage configuration is also used to initiate Advanced Copy Services operations in a storage complex. The S-HMC is required only to initiate changes to the Copy Services environment. Existing Copy Services relationships (such as FlashCopy pairs or PPRC pairs) continue to operate as designed even in the absence of the S-HMC. Copy Services operations against Count Key Data (CKD) volumes can also be initiated through inband channel commands issued across ESCON/FICON channels from z/Series hosts. These commands are issued directly to the storage facility and can be executed without support from the S-HMC.

**Note:** All service interfaces require an authenticated login procedure to access service functions.

## 9.2.3 S-HMC network topology

In order to connect the S-HMC to your network, you will need the following network information:

- ▶ One IP address per S-HMC. With an external S-HMC, you will need two IP addresses.
- ▶ Host names for the S-HMCs (for example, MC1 and MC2).
- ▶ Domain name for the S-HMC (for example, us.ibm.com®).
- ▶ Whether you will use local time or a different time zone.



- ▶ TCP/IP interface network mask (for example, 255.255.254.0).
- ▶ If you plan to use a Domain Name Server (DNS) to resolve network names, you will need the IP address of your DNS server and the name of your DNS. You may have more than one DNS.
- ▶ You can specify a default gateway in dotted decimal form or just the name in case you are using a DNS.

The S-HMC can also be connected to the IBM organization's support services using:

- ▶ A dial-up connection
- ▶ A secure high-speed connection

### **Dial-up connection**

This is a low-speed asynchronous modem connection to a telephone line. This connection typically favors small amounts of data transfers. When configuring for a dial-up connection, have the following information available:

- ▶ Which dialing mode will be used, either tone or pulse.
- ▶ Whether a dialing prefix is required when dialing an outside line.

### **Secure high-speed connection**

This connection is through a high-speed Ethernet connection that can be configured through a secure virtual private network (VPN) internet connection to ensure authentication and data encryption. IBM has chosen to use a graphical interface (WebSM) for servicing the storage facility, and for the problem determination activity logs, error logs, and diagnostic dumps that may be required for effective problem resolution. These logs can be significant in size. For this reason, a high-speed connection would be the ideal infrastructure for effective support services.

A remote connection can be configured to meet the following customer requirements:

- ▶ Allow call on error (machine-detected)
- ▶ Allow connection for a few days (customer-initiated)
- ▶ Allow remote error investigation (Service-initiated)

Figure 9-3 on page 164 shows a typical redundant S-HMC configuration. In this configuration, we have MC1 (internal S-HMC) and MC2 (external S-HMC) connected to the 172.16-BLACK network and the 172.16-GRAY network. These two networks are the private Ethernet networks of the DS8000. MC1 and MC2 are also connected to the customer's network by way of the customer Ethernet ports on the pair of Ethernet switches that are installed at installation time.

## S-HMC - Network Topology

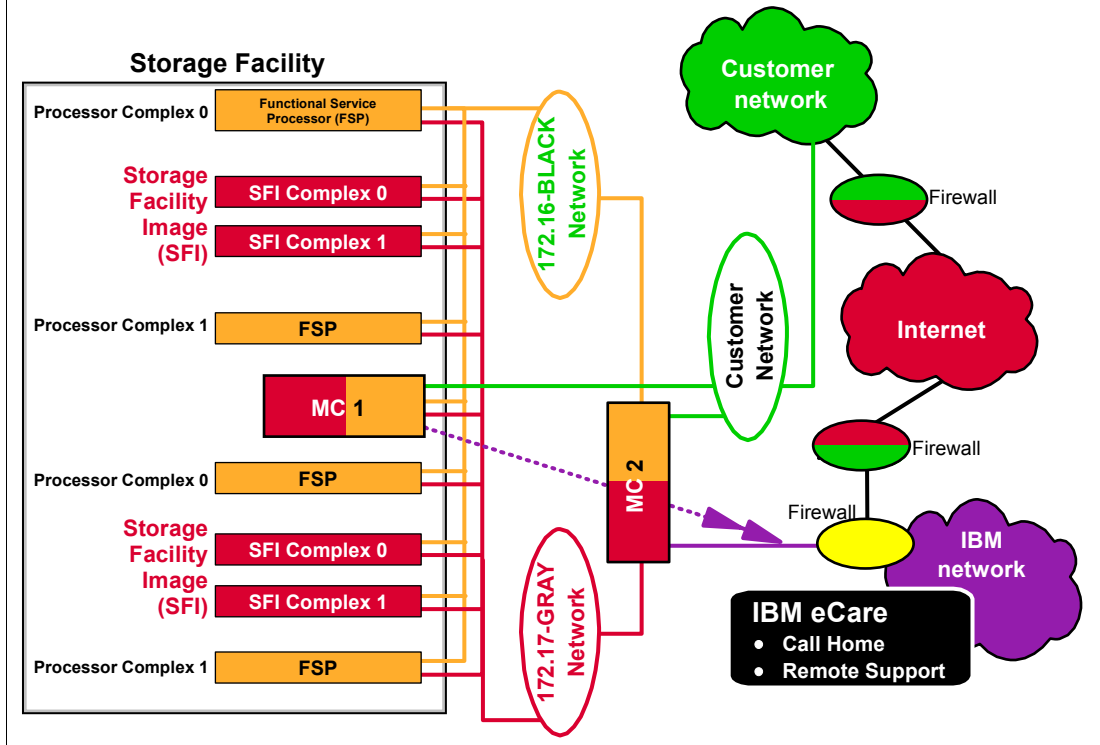


Figure 9-3 S-HMC network topology

In this configuration, capabilities exist to execute remote services such as:

- ▶ Call home
- ▶ Remote access

### Call home

Call home is the capability of the S-HMC to contact IBM support services to report a problem. This is referred to as *call home for service*. The S-HMC will also provide machine-reported product data (MRPD) information to IBM by way of the call home facility. The MRPD information includes installed hardware, configurations, and features. The storage plex will use the call home method to send heartbeat information to IBM and by doing this, will ensure that the S-HMC will be able to initiate a call home to IBM in the case of an error. Should the heartbeat information not reach IBM, a service call to the customer will be initiated by IBM to investigate the status of the DS8000. The call home can either be configured for modem or internet setup. The call home service can only be initiated by the S-HMC. This facility cannot be initiated from the outside. In our example, MC1 will initiate a call home for service when detecting an error with the DS8000. This action will pass through the customer's network across the internet to the IBM service center. Depending on the complexity of the error, IBM will request remote access.

### Remote access

Remote access is required when the local service personnel cannot correct problems with the storage plex and IBM product engineer assistance is required to resolve the problems. In this case the S-HMC will initiate a call home for service and IBM support service will initiate a remote access session, using the basic ASCII terminal to either dial the S-HMC modem to

request an IP connection, or if the modem is unavailable, will use voice phone or e-mail to request an IP connection.

**Note:** The IP connection initiated by the S-HMC will always be to a specific telephone number for modem access or to a specific IP address for internet access.

The communication will be by way of VPN and will use data encryption. This network configuration will position IBM to provide faster support assistance and problem resolution since IBM will be able to connect to the S-HMC relatively quickly and error logs can be transmitted to IBM over a high-speed network. This will eliminate any delays associated with transmitting huge error logs over a dial-up connection.

There can also be a direct connection to the IBM network by way of VPN. This is depicted in the diagram with the arrow from MC1 to the IBM network.

**Note:** IBM recommends that the S-HMC be connected to the customer's public network over a secure VPN connection, instead of a dial-up connection.

## S-HMC security considerations

Allowing access between the Internet and computers in a customer network brings valid security concerns which need to be addressed. IBM has taken the steps necessary to provide secure network access for the S-HMC. Even after securing access to the S-HMC, there are additional levels of security built into the Service applications available on the S-HMC. In the following sections we discuss the security protection securing access to the S-HMC from the Internet, and then we describe the internal security of the S-HMC itself.

### Security mechanism 1 - Console must initiate session

The first security measure that is employed to protect the console is to only allow network sessions or conversations to be initiated from the console itself. This means that there are no applications running on the console that are listening on TCPIP ports to establish a session. If a session is needed from the console to enable a service action, an IBM Service representative may initiate this session by dialing into the console using the modem, and requesting that the console establish the session. This session will only be initiated to one of the defined TCPIP addresses which represent the IBM Service centers.

At installation time, the customer may decide to only allow a service session when manually requested, by the customer, through the console interface. These installation options are briefly described here, and explained in detail in *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

### Security mechanism 2 - Public key encryption

The S-HMC uses a public key encryption mechanism to maintain the security of data exchanged between the console and the IBM Service organization. Each S-HMC, during manufacturing or during the installation process, generates a public encryption key based on the private key that the console will use for encryption and decryption.

During the installation process at the customer site, the IBM SSR will connect to the IBM Service organization, by way of modem, internet connection, or the SSRs MOST portable console, and will transmit the public key for the installed console to a database maintained within the IBM secure network. Whenever IBM Service requires access to the console located at the customer site, the IBM personnel will have to retrieve the console-specific public key from the database and use this key to establish the communication session needed for service.

### **Security mechanism 3 - Login security**

When the network connection and session are established, the IBM Service personnel will be able to log into the S-HMC, without a secure password, for the purpose of collecting problem determination and sending the problem determination to the IBM data collection site.

If problem analysis shows that additional actions are needed to further refine the problem definition, the next level of IBM Service may have a requirement for a higher level of access to the storage facility. If this is the case, all of the previous security measures will also apply, but to obtain a higher level of authorization, the service organization will be required to log in to secure userids on the S-HMC. These userids are protected by S-HMC user management.

### **S-HMC user management**

The S-HMC's user management is governed by the following rules:

- ▶ The allowed number of users is pre-defined.
- ▶ No additional users can be defined to the S-HMC.
- ▶ The password to these predefined users can be changed by IBM support personnel.
- ▶ Users with higher privileges are protected by a challenge/authentication scheme.
- ▶ The root user ID is locked.
- ▶ User login is only allowed from the private network, or from an IBM remote support service connection.
- ▶ User activity is logged.
- ▶ The S-HMC has auditing capabilities.

The functionality of standard software components is restricted to provide added security advantages when integrating the S-HMC into your private network. These restrictions are as follows:

- ▶ The S-HMC acts as a firewall or proxy for incoming traffic.
- ▶ The S-HMC does not have any IP forwarding or gateway capabilities.
- ▶ Many standard services such FTP and TELNET do not exist on the S-HMC.
- ▶ Only Secure Shell (SSH) is permitted over the remote connection.
- ▶ No TCP/IP connection from the outside is allowed into the S-HMC, except the VPN connection.
- ▶ In an Internet configuration, the following firewall ports need to be open: 500 udp, 500 esp (VPN), and 4500 udp.
- ▶ The allowed destination IP addresses will only be the IBM services support centers: 207.25.252.196, 129.42.160.16, and 207.25.252.198.

## **9.2.4 FTP Offload option**

As an alternative to a VPN connection via the Internet, the S-HMC can be set up to use the file transfer protocol (ftp) for sending error data to IBM. It is the customer's responsibility to provide a secure path from the S-HMC to the destination server (testcase.software.ibm.com) on the Internet. Usually this involves some kind of ftp proxy or relay firewall. The S-HMC supports seven different types of ftp firewalls.

Connectivity via ftp is also required for downloading DS8000 code packages from an IBM remote code repository on the Internet.

**Note:** Data sent using the ftp protocol is neither encrypted nor authenticated.

Regardless of the type of connectivity (VPN or ftp), no customer data is transmitted to IBM.

## 9.3 DS8000 licensed functions

*Licensed functions* are the storage unit's operating system and functions. Each licensed function indicator feature selected on a DS8000 base unit enables that function at the system level. These features enable a licensed function subject to you applying a feature activation code made available by IBM. They are also used for maintenance billing purposes. These include both required features and optional features.

Table 9-2 provides the appropriate licensed function indicators for each licensed function.

Table 9-2 *Licensed function indicators*

Licensed function	IBM 2107 indicator feature number	IBM 2244 function authorization models and features
Operating environment	0700	Model OEL 70xx
Point-in-Time Copy	0720	Model PTC 72xx
Remote Mirror and Copy	0740	Model RMC 74xx
Remote Mirror for z/OS	0760	Model RMZ 76xx
Parallel Access Volumes	0780	Model PAV 78xx

### 9.3.1 Operating environment license (OEL) - required feature

You must order an operating environment license (OEL) feature, the IBM TotalStorage DS Operating Environment, for every storage unit. The operating environment model and features establish the extent of IBM authorization for the use of the IBM TotalStorage DS Operating Environment. This operating environment license support function is called the 2244 Model OEL. The OEL licenses the operating environment and is based on the total physical capacity of the storage unit (base model plus any expansion models). It authorizes you to use the model configuration at a given capacity level.

Once the OEL has been technically activated for the storage unit, you can configure the storage unit. Activating the OEL means that you have obtained the feature activation key from the IBM disk storage feature activation (DSFA) Web site and entered it into the DS8000 Storage Manager. The feature activation process is discussed in more detail in 9.3.7, "Disk storage feature activation" on page 173.

**Note:** Standby CoD disk drive features do not count toward the physical capacity.

Table 9-3 on page 168 provides the feature codes for the operating environment license. (The codes apply to models 921,922, and 9A2.)

Table 9-3 Operating environment license feature codes

Feature code	Description
7000	OEL-inactive
7001	OEL-1TB
7002	OEL-5TB
7003	OEL-10TB
7004	OEL-25TB
7005	OEL-50TB
7010	OEL-100TB

Licensed functions are activated and enforced within a defined license scope. *License scope* refers to the following types of storage and servers with which the function can be used:

- ▶ Fixed block (FB) - The function can be used only with data from Fibre Channel-attached servers.
- ▶ Count key data (CKD) - The function can be used only with data from FICON- or ESCON-attached servers.
- ▶ Both FB and CKD (ALL) - The function can be used with data from all attached servers.

Some licensed functions have multiple license scope options, while other functions have only the scope of a single license.

Table 9-4 provides the license scope options for each licensed function.

Table 9-4 License scope for each DS8000 licensed function

Licensed function	License scope options
Operating environment	ALL
Point-in-Time Copy	FB, CKD, or ALL
Remote Mirror and Copy	FB, CKD, or ALL
Remote Mirror for z/OS	CKD
PAV	CKD

### Optional features

The following optional features are available for the DS8000:

- ▶ Point-in-Time Copy, which includes IBM TotalStorage FlashCopy.
- ▶ Remote Mirror and Copy, which includes IBM TotalStorage Metro Mirror, IBM TotalStorage Global Mirror and IBM TotalStorage Global Copy.
- ▶ Remote Mirror for z/OS, which is the IBM TotalStorage z/OS Global Mirror.
- ▶ Parallel Access Volumes (PAV).

## 9.3.2 Point-in-Time Copy function (2244 Model PTC)

The Point-in-Time Copy licensed function model and features establish the extent of IBM authorization for the use of the IBM TotalStorage FlashCopy. When you order Point-in-Time

Copy functions, you specify the feature code that represents the physical capacity to authorize for the function.

Table 9-5 provides the feature codes for the Point-in-Time Copy function. (The codes apply to models 921,922, and 9A2.)

*Table 9-5 Point-in-Time Copy (PTC) feature codes*

Feature code	Description
7200	PTC-inactive
7201	PTC-1TB unit
7202	PTC-5TB unit
7203	PTC-10TB unit
7204	PTC-25TB unit
7205	PTC-50TB unit
7210	PTC-100TB unit

You can combine feature codes to order the exact capacity you need. For example, if you determine that you need 23 TB of Point-in-Time capacity, you can order two 7203 features (this will give you 20 TB) and three 7201 features (this will give you 3 TB), giving you a total of 23 TB.

### 9.3.3 Remote Mirror and Copy functions (2244 Model RMC)

The Remote Mirror and Copy licensed function model and features establish the extent of IBM authorization for the use of the Metro Mirror (Synchronous PPRC), Global Mirror (Asynchronous PPRC) and Global Copy (PPRC Extended Distance).

Table 9-6 provides the feature codes for the Remote Mirror and Copy functions. (The codes apply to models 921,922, and 9A2.)

*Table 9-6 Remote Mirror and Copy (RMC) feature codes*

Feature code	Description
7400	RMC-inactive
7401	RMC-1TB unit
7402	RMC-5TB unit
7403	RMC-10TB unit
7404	RMC-25TB unit
7405	RMC-50TB unit
7410	RMC-100TB unit

### 9.3.4 Remote Mirror for z/OS (2244 Model RMZ)

The Remote Mirror for z/OS licensed function model and features establish the extent of IBM authorization for the use of IBM TotalStorage z/OS Global Mirror.

Table 9-7 on page 170 provides the feature codes for Remote Mirror for zSeries functions. (The codes apply to models 921,922, and 9A2.)

Table 9-7 Remote Mirror for zSeries (RMZ) feature codes

Feature code	Description
7600	RMZ-inactive
7601	RMZ-1TB unit
7602	RMZ-5TB unit
7603	RMZ-10TB unit
7604	RMZ-25TB unit
7605	RMZ-50TB unit
7610	RMZ-100TB unit

### 9.3.5 Parallel Access Volumes (2244 Model PAV)

The Parallel Access Volumes model and features establish the extent of IBM authorization for the use of the Parallel Access Volumes licensed function.

Table 9-8 provides the feature codes for the PAV function. (The codes apply to models 921,922, and 9A2.)

Table 9-8 Parallel Access Volumes (PAV) feature codes

Feature code	Description
7800	PAV-Disable
7801	PAV-1TB unit
7802	PAV-5TB unit
7803	PAV-10TB unit
7804	PAV-25TB unit
7805	PAV-50TB unit
7810	PAV-100TB unit

### 9.3.6 Ordering licensed functions

An OEL is required for every DS8000 base unit. This license is for the total physical capacity of the entire storage unit, including the base and any expansion models, and for both FB and CKD data, but excludes Standby CoD disk drives. For example, if the total capacity of the storage unit is 30 TB (15 TB in the base frame and 15 TB in the expansion frame), then you will purchase an OEL feature for 30 TB.

Should you increase your capacity in the future, for example, from the 30 TB to 40 TB, whether you are using Standby CoD disk drives or additional disk drives, you then must purchase an additional 10 TB of 2244 Model OEL features.

The other optional features such as PTC, RMC, RMZ, and PAV can use any combination of the features to cover the required capacity of the storage unit. For example, if the storage unit has a capacity of 20 TB of data and only 5 TB of the data will be PTC, then you only need to purchase a PTC license for 5 TB of data – not the full 20 TB.

Figure 9-4 shows an example of a FlashCopy feature code authorization. In this case, the user is authorized up to 25 TB of CKD data. The user cannot FlashCopy any FB data.



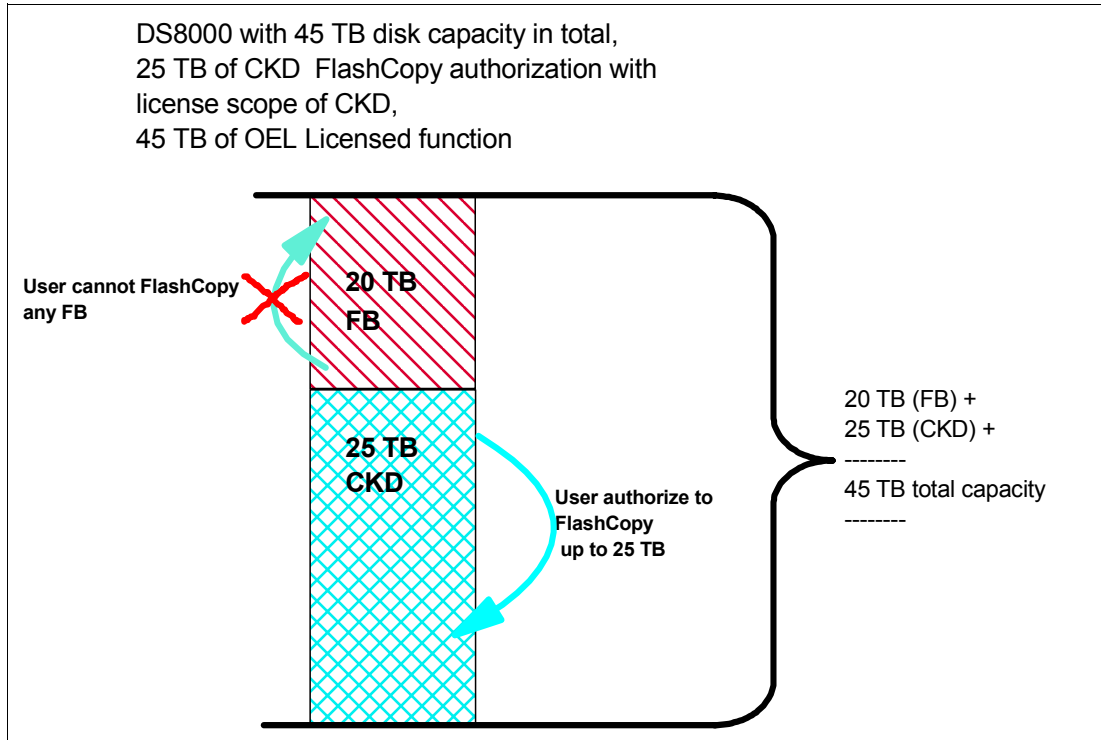


Figure 9-4 User authorize to FlashCopy 25 TB of CKD data

As illustrated in Figure 9-5 on page 172, the user decides to change the license scope from CKD to ALL and FlashCopy the FB data as well. This increase of licensed scope from CKD to ALL is non-disruptive. Changing the license scope from FB to CKD, or CKD to FB (lateral change), or reduction in license scope from ALL to FB or CKD, will be disruptive and will require an IML. In this example, the user decides to FlashCopy 10 TB of FB and 12 TB of CKD data. The user has disk capacity of 20 TB of FB and 25 TB of CKD. In order to implement the changes, the user now has to purchase a Point-in-Time Copy function authorization of 45 TB, the total FB and CKD capacity.

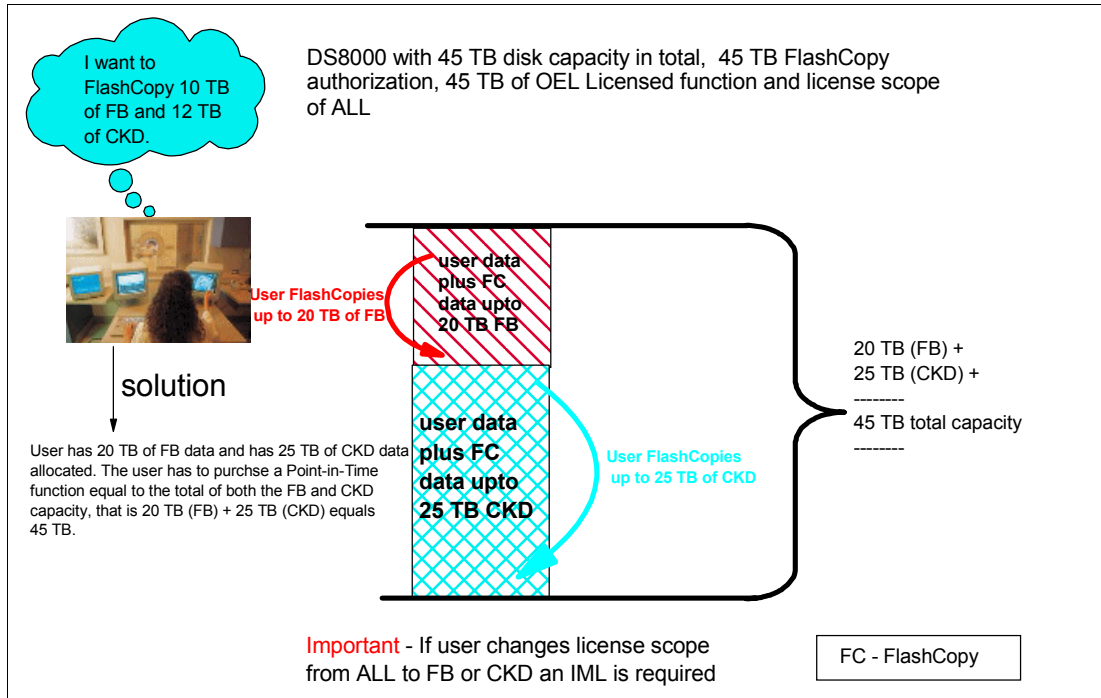


Figure 9-5 User authorize to FlashCopy 45 TB of data with a license scope of ALL

For PTC and RMC, you may have FB data for open systems only, or you may have CKD for zSeries only, or you may have both FB and CKD data. For RMC, you will need the license feature for both the primary storage unit and the secondary storage unit.

Figure 9-6 on page 173 shows an example of a Metro Mirror configuration. In this case, the user has to purchase Remote Mirror and Copy function authorization for 45 TB with Metro Mirror feature for both the primary DS8000 and secondary DS8000.

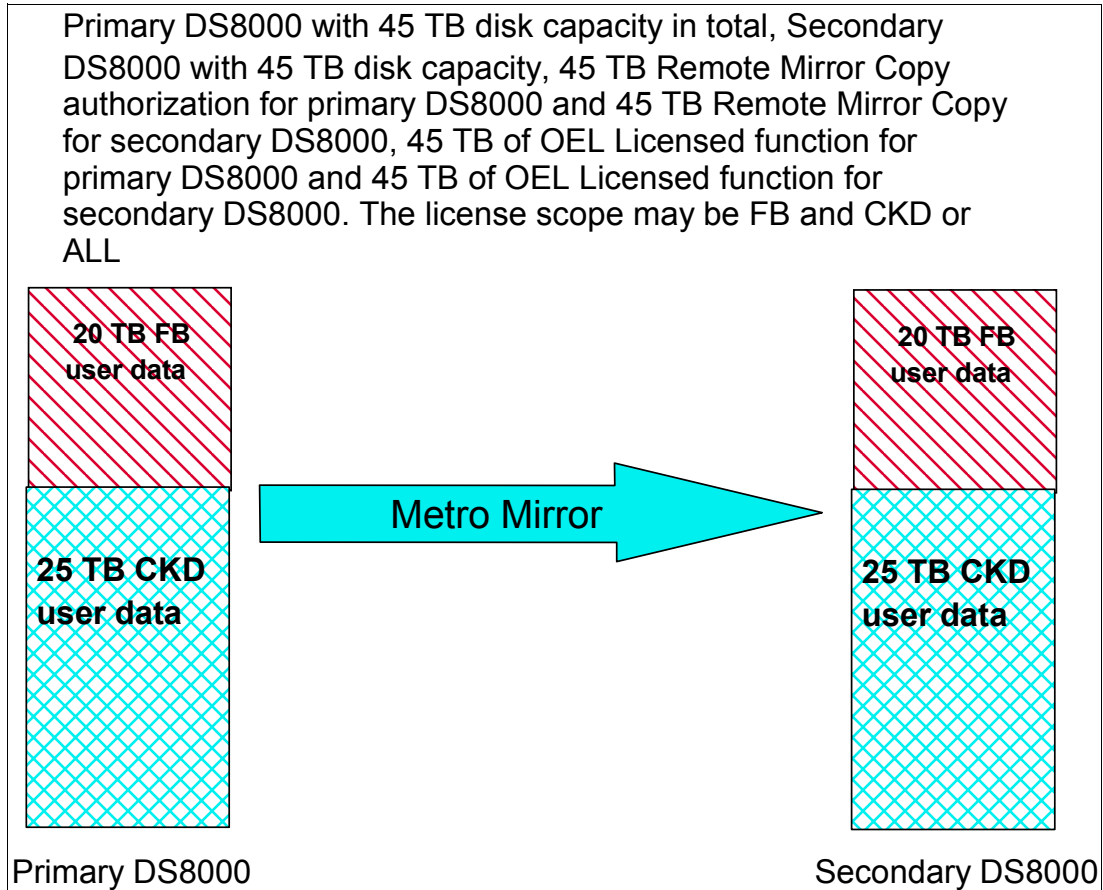


Figure 9-6 Remote Mirror and Copy

Global Mirror requires both RMC and PTC functions. Both primary and secondary storage units require RMC functions and the secondary requires PTC functions as well. If Global Mirror will be used during failback on the secondary storage unit, a Point-in-Time Copy function authorization must also be purchased for the primary system.

RMZ and PAV are only applicable to zSeries, so you will only have CKD data. In RMZ configurations that exploit the failback technique, you will have to purchase a license feature for the secondary storage unit too.

The initial enablement of any optional DS8000 licensed function is a concurrent activity (assuming the appropriate level of microcode is installed on the machine for the given function). The removal of a DS8000 licensed function to deactivate the function is a disruptive activity and requires a machine IML.

### 9.3.7 Disk storage feature activation

Managing and activating licensed functions is your responsibility. You manage and activate functions through the IBM Disk Storage Feature Activation (DSFA) Web site at:

<http://www.ibm.com/storage/dsfa>

Management refers to the use of the IBM Disk Storage Feature Activation (DSFA) Web site to select a license scope and to assign a license value. You perform these activities and then activate the function.

*Activation* refers to the retrieval and installation of the feature activation code into the DS8000 system. The feature activation code is obtained using the DSFA Web site and is based on the license scope and license value.

The high-level steps for storage feature activation are:

- ▶ Have machine-related information available (model, serial number, and machine signature). This information is obtained from the DS8000 Storage Manager.
- ▶ Log onto the DSFA Web site.
- ▶ Obtain feature activation keys by completing the machine-related information. The feature activation keys can either be directly retrieved, saved onto a diskette, or they can be written down.
- ▶ Complete feature key activation by logging onto your DS8000 Storage Manager application and apply the keys to the storage unit.

### 9.3.8 Scenarios for managing licensing

There are many ways to manage the license of a DS8000. Some scenarios may include adding storage capacity to an existing licensed function or reallocating a license between storage images.

#### **Adding storage capacity to an existing licensed function**

Assume you initially purchased a 2244 Point-in-Time Copy feature (2244-PTC) for 25 terabytes. After several months, you need an additional 20 TB for your Point-in-Time Copy operations. To increase storage for the feature requires that you obtain a new license key, and not install a larger license. However, this is a non-disruptive activity and does not require that you reboot your machine. You have to complete the following steps to activate the keys:

- ▶ Order two of feature 7203 (10 TB each of 2244-PTC) for example. You can order a different combination of the features to get to your required amount.
- ▶ You will receive a 2244 function authorization serial number from an IBM representative.
- ▶ Retrieve the license key from the DSFA Web site.
- ▶ This new license key represents the total capacity that you now have licensed (or 45 TB). It licenses the original 25 TB plus the additional 20 TB, just ordered.
- ▶ Enter the license key into the DS Storage Manager. This will replace the existing license key with the new license key.
- ▶ After successful installation of the license key, you now have 45 TB of 2244-PTC capacity.

## 9.4 Capacity planning

The DS8000 offers high scalability while maintaining excellent performance. We have already explained the scalability in 6.3, “Designed for scalability” on page 109. You know that the physical storage capacity is not equal to the logical storage capacity that you can use for your system. We explain how to estimate the logical capacity for the DS8000 in this section.

### 9.4.1 Logical configurations

The total capacity of physical disk drives in the DS8000 is not equal to the logical capacity you can use because you configure RAID protection to protect your important data. The

DS8000 adopts a virtualization concept, which was introduced in Chapter 5, “Virtualization concepts” on page 83.

The logical capacity depends on the RAID type and the number of spare disks in a rank. We explain the capacity with the following figures.

**Attention:**

- ▶ The figures used in this section are still subject to change. Specifically, the exact number of extents per array may be slightly different.
- ▶ Generally speaking, you might ensure there is at least 5% of additional capacity if you are providing an exact number of devices.
  - There will also potentially be unusable space for devices which are not an integer number of extents in size and for the last extent in an extent pool.
  - To prepare for the unexpected situation, a certain amount of margin is required for important systems.

**Note:** IBM will offer Capacity Magic for the DS8000 in the future. Capacity Magic calculates the physical and effective storage capacity of a DS8000. IBM engineers and IBM Business Partners can download this tool from an IBM Web site. The following figures are intended to be used only until Capacity Magic is available.

## CKD RAID rank capacity

Rank	Spare / No spare	Extents	Binary GB	Decimal GB	3390-3 devices	3390-9 devices	32760 cyl devices	65520 cyl devices
RAID10 73GB	Spare	216	190.30	204.34	72	24	7	3
RAID10 73GB	No spare	288	253.74	272.45	96	32	9	4
RAID10 146GB	Spare	435	383.25	411.51	145	48	15	7
RAID10 146GB	No spare	581	511.88	549.63	193	64	20	10
RAID10 300GB	Spare	883	777.96	835.32	294	98	30	15
RAID10 300GB	No spare	1,178	1,037.86	1,114.39	392	130	40	20
RAID5 73GB	Spare	434	382.37	410.57	144	48	14	7
RAID5 73GB	No spare	507	446.69	479.62	169	56	17	8
RAID5 146GB	Spare	873	769.14	825.86	291	97	30	15
RAID5 146GB	No spare	1,018	896.89	963.03	339	113	35	17
RAID5 300GB	Spare	1,771	1,560.32	1,675.38	590	196	61	30
RAID5 300GB	No spare	2,066	1,820.22	1,954.45	688	229	71	35

**Notes**

There may be space left over when defining the devices shown in the table.

Devices that are not a multiple of 1113 cylinders will use additional space on the disk subsystem as the allocation unit is an extent of this size.

Figure 9-7 CKD RAID rank capacity

## FB RAID rank capacity

Rank	Spare / No spare	Extents	Binary GB	Decimal GB
RAID10 73GB	Spare	192	192	206.16
RAID10 73GB	No spare	256	256	274.88
RAID10 146GB	Spare	386	386	414.46
RAID10 146GB	No spare	519	519	557.27
RAID10 300GB	Spare	785	785	842.89
RAID10 300GB	No spare	1,048	1,048	1,125.28
RAID5 73GB	Spare	386	386	414.46
RAID5 73GB	No spare	450	450	483.18
RAID5 146GB	Spare	779	779	836.44
RAID5 146GB	No spare	909	909	976.03
RAID5 300GB	Spare	1,582	1,582	1,698.66
RAID5 300GB	No spare	1,844	1,844	1,979.98

### Notes

Device sizes that are not a multiple of 1 binary GB will use additional space on the disk subsystem as the allocation unit is an extent of this size

Figure 9-8 FB RAID rank capacity

For example, if you configure a RAID-5 rank with 146 GB DDMs with a spare disk in open system environments, the capacity of the rank totals 779 extents (779 GB).

**Note:** In the DS8000, extent size is specified with binary size, not decimal size. For example, 1 GB in binary is described as 1024 x 1024 x 1024, and 1GB in decimal is described as 1000 x 1000 x 1000. Operating systems adopt the binary format for calculating their storage.

### 9.4.2 Sparring rules

To estimate your usable storage capacity, it is helpful to understand the rules of sparring since the capacity of the rank is different with or without spare disks in the rank. The sparring rules for the DS8000 are as follows:

- ▶ A minimum of one spare is required for each array site defined until the following considerations are met:
  - Minimum of 4 spares per DA pair.  
The spares are balanced between the two device interfaces.
  - Minimum of 4 spares per the largest capacity array site on the DA pair.  
The spares are balanced between the two device interfaces.
  - Minimum 2 spares of capacity and RPM greater than or equal to the fastest array site of any given capacity on the DA pair.  
The spares are balanced between the two device interfaces.
- ▶ Spares are not necessarily allocated on the first arrays to be formatted.
- ▶ Intermix configuration complicates the sparring rules. You should use Capacity Magic for these calculations.

### 9.4.3 Sparing examples

This section provides some examples for configuring each rank according to the rule of sparing disks.

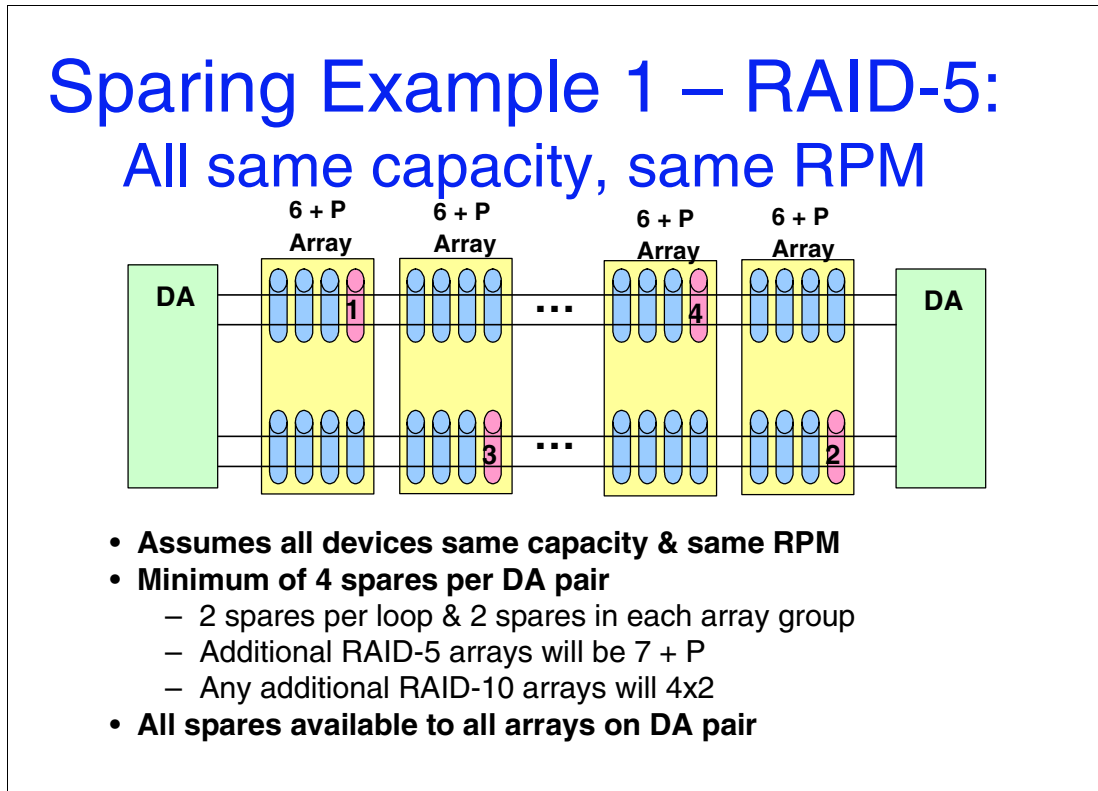
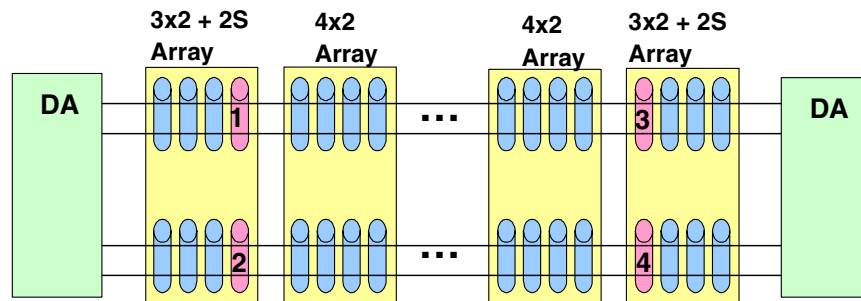


Figure 9-9 Sparing example 1: RAID-5 - All same capacity, same RPM

In Figure 9-9, four RAID-5 arrays are installed in a DA pair. Each array needs a spare disk drive (6+P+S). Two spare disks are in one loop and the other two spare disks are in the other loop in the DA pairs.

If you add other arrays configured with the same capacity and RPM as in this DA pair, additional arrays will not need spare disks (the arrays will be configured as RAID-5:7+P, RAID-10:4x2).

## Sparring Example 2 – RAID-10: All same capacity, same RPM



- Assumes all devices same capacity & same RPM
- Minimum of 4 spares per DA pair
  - 2 spares per loop & 2 spares in each array group
  - Additional RAID-10 arrays will be 4x2
  - Any additional RAID-5 arrays will 7 + P
- All spares available to all arrays on DA pair

Figure 9-10 Sparring example 2: RAID-10 - All same capacity, same RPM

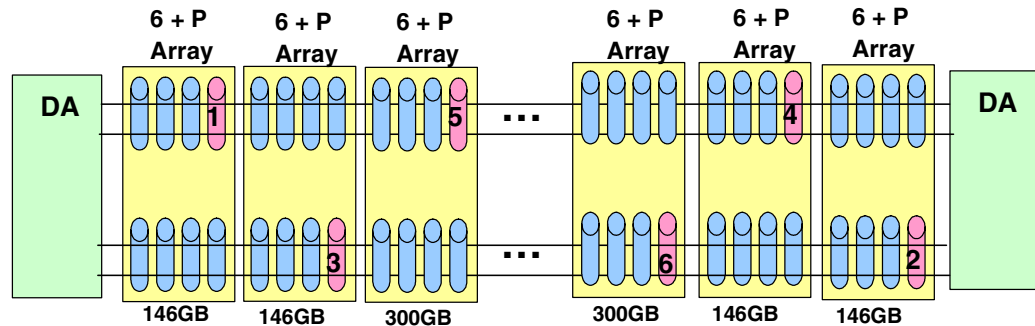
In Figure 9-10, four RAID-10 arrays are installed in a DA pair. Two arrays need two spare disk drives ( $3 \times 2 + 2S$ ). One spare disk is in one loop and other spare disk is in the other loop in an array. The other two arrays don't need spare disks ( $4 \times 2$ ).

If you add other arrays configured with the same capacity and RPM as in this DA pair, additional arrays don't need spare disks (the arrays will be configured as RAID-5:7+P, RAID-10:4x2).



## Sparing Example 3 – RAID-5:

1<sup>st</sup> 4 arrays 146GB & next 2 arrays 300GB (same RPM)



- Assumes all devices same RPM
- Minimum of 4 spares per DA pair
  - 2 spares per loop & 2 spares in each array group
  - Additional 146GB RAID arrays will be 7 + P (RAID-5) or 4x2 (RAID-10)
- Minimum 4 spares of the largest capacity array site on the DA pair
  - Next 2 300GB arrays will also be 6 + P (if RAID-5)
    - If next 300GB array configured is RAID-10, then that array will be 3x2 and any additional 300GB arrays on this DA pair will not have spares
- All spares available to all arrays on the DA pair

Figure 9-11 Sparing example 3: RAID-5 - Different capacity, same RPM

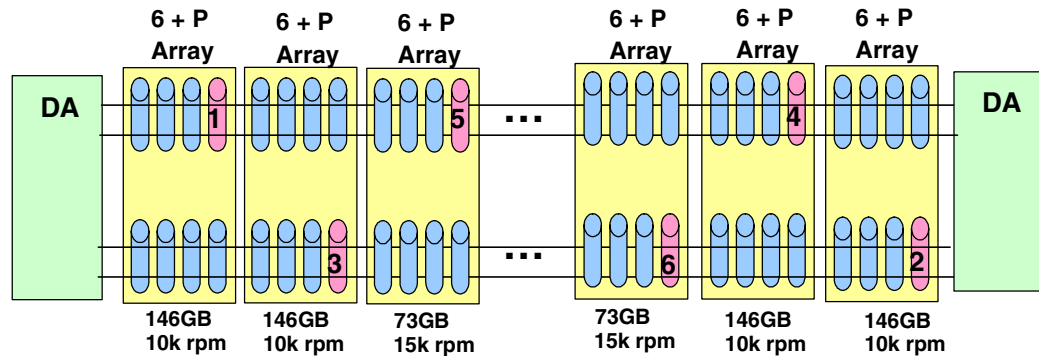
Figure 9-11 illustrates an intermix configuration in a DA pair.

1. At first, four RAID-5 arrays with 146 GB DDMs are installed in a DA pair. Each array needs a spare disk drive (6+P+S). Two spare disks are in one loop and other two spare disks are in the other loop in the DA pairs.
2. Next, two RAID-5 arrays with 300 GB DDMs are added in the DA pair. According to the rule of *Minimum 4 spare disks of the largest capacity array site on the DA pair*, an additional 2 arrays need a spare disk (6+P+S).
3. If you add other arrays with 300 GB DDMs, you need spare disks in additional arrays. You need a total of four 300 GB spare disks in the DA pair; then, two 6+P+S arrays are needed for RAID-5, and one 3x2+2S array is needed for RAID-10.

**Note:** Intermix of DDM size and RPM configuration is planned for the future.

## Sparing Example 4 – RAID-5:

1<sup>st</sup> 4 arrays 146GB 10k RPM & next 4 arrays 73GB 15k RPM



- Assumes mixture of different RPM devices
- Minimum of 4 spares per DA pair
  - 2 spares per loop & 2 spares in each array group
  - Additional 146GB / 10k RPM RAID arrays will be 7 + P (RAID-5) or 4x2 (RAID-10)
- Minimum 2 spares of capacity and RPM greater than or equal to the fastest array site of any given capacity on the DA pair
  - Next 2 73GB / 15k RPM arrays will also be 6 + P (if RAID-5)
    - Any additional 73GB / 15k RPM arrays on this DA pair will not have spares
- All spares available to all arrays on the DA pair

Figure 9-12 Sparing example 4: RAID-5 - Different capacity, different RPM

Figure 9-12 illustrates another intermix configuration in a DA pair.

1. At first, four RAID-5 arrays with 146 GB/10k RPM DDMs are installed in a DA pair. Each array needs a spare disk drive (6+P+S). Two spare disks are in one loop and the other two spare disks are in the other loop in the DA pairs.
2. Two RAID-5 arrays with 73 GB / 15k RPM DDMs are added in the DA pair. According to the rule of *Minimum 2 spares of capacity and RPM greater than or equal to the fastest array site of any given capacity on the DA pair*, the additional two arrays need a spare disk (6+P+S).
3. If you add other arrays with 73 GB//15k RPM DDMs, you do not need spare disks in additional arrays. You already have four spare disks of the largest capacity (146 GB) and two spare disks of the fastest (15k RPM) in the DA pair.

**Note:** Intermix of DDM size and RPM configuration is planned for the future.

### 9.4.4 IBM Standby Capacity on Demand (Standby CoD)

To help further meet the changing storage needs of growing businesses, the DS8000 series can use the IBM Standby Capacity on Demand option, which is designed to allow clients to access extra capacity quickly whenever the need arises. With all these capabilities, the DS8000 series can quickly respond to changing business needs.

The IBM Standby Capacity on Demand (Standby CoD) offering allows the installation of inactive disk drives that can be easily activated as business needs dictate.

A Standby CoD disk set contains 16 disk drives of the same capacity and RPM (10000 or 15000 RPM). With this offering, up to four Standby CoD disk drive sets (64 disk drives) can be factory or field installed into your system.

To activate, you logically configure the disk drives for use. This is a non-disruptive activity that does not require intervention from IBM. Upon activation of any portion of the Standby CoD disk drive set, you must place an order with IBM to initiate billing for the activated set. At that time, you can also order replacement Standby CoD disk drive sets.

For more details, refer to the *IBM TotalStorage DS8000 Introduction and Planning Guide*, GC35-0495.

### 9.4.5 Capacity and well-balanced configuration

The DDMs are installed in the DS8000 according to a rule of installation sequence of disk enclosures. Sometimes, the installation rule is related to the arrangement of your data. The installation rule is defined as follows:

- ▶ Each disk enclosure can contain 16 DDMs or dummy carriers, and a pair of disk enclosures (32 DDMs) is installed in the DS8000 at the same time.
- ▶ Each disk enclosure is connected to a specific DA pair, and two disk enclosure pairs (64 DDMs) are connected to a DA pair sequentially.

For example, if you install 96 DDMs in a DS8000, the first two disk enclosure pairs (64 DDMs) are connected to the DA pair 0, and next one disk enclosure pair (32 DDM) is connected to the DA pair 1.

The following figures illustrate the installation rule of the disk enclosures.

Figure 9-13 on page 182 shows a DS8100 Model 921 (2-way model).

# DDM to DA Mapping -- 2-way

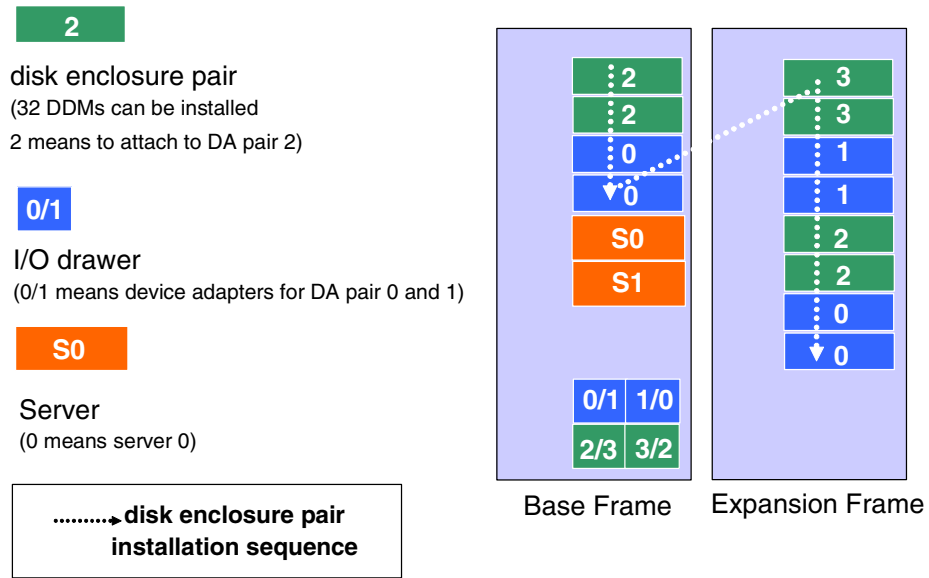


Figure 9-13 DDM to DA mapping (2-way model)

Model 921 can have four DA pairs and 12 disk enclosure pairs (1 disk enclosure pair can have  $16 \times 2 = 32$  DDMs). And each two disk enclosure pairs (64 DDMs) is connected to a DA pair in order.

For example, the first 64 DDMs are connected to the DA pair 2, the next 64 DDMs are connected to the DA pair 0, and so on. Therefore, a 256 DDMs configuration is a well-balanced configuration (all four DA pairs have 64 DDMs).

If you install more than 256 DDMs in the DS8100, the next disk enclosures are connected to the DA pair 2 again. When you have the maximum DDMs (384 DDMs), DA pair 0 and 1 have 128 DDMs and DA pair 2 and 3 have 64 DDMs.

Figure 9-14 on page 183 shows the DS8300 Model 922 (4-way model).

# DDM to DA Mapping -- 4-way

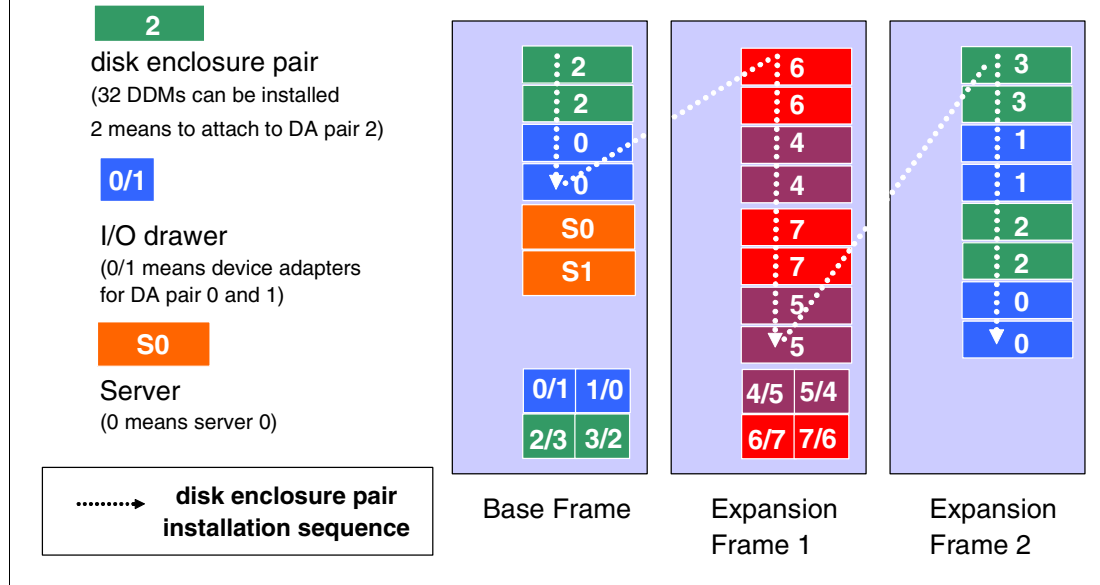


Figure 9-14 DDM to DA Mapping (4-way model)

Model 922 can have eight DA pairs and 20 disk enclosure pairs.

Therefore, a 512 DDMs configuration is a well-balanced configuration (all eight DA pairs have 64 DDMs).

When you have the maximum number of DDMs (640 DDMs), DA pair 0 and 1 have 128 DDMs, and DA pairs 2 through 7 have 64 DDMs.

## 9.5 Data migration planning

When migrating data, the migration objectives should be clearly defined. Use the following key questions to define your generic migration environment:

- ▶ Why is the data migrating?
- ▶ How much data is migrating?
- ▶ How quickly must the migration be performed?
- ▶ What duration of service outage can be tolerated?
- ▶ Is the data migration to or from the same type storage, for example from an ESS Model 800 to a DS8000?
- ▶ What resources are available for the migration?

After answering these questions, you will be in a position to choose the appropriate tools and utilities, such as standard operating system mirroring, basic commands, software packages, remote copy technologies, and migration appliances to achieve your data migration objectives.

## 9.5.1 Operating system mirroring

Logical volume mirroring (LVM) and Veritas Volume Manager have little or no application service disruption and the original copy will stay intact while the second copy is being made. The disadvantages of this approach include:

- ▶ Host cycles are utilized.
- ▶ It is labor intensive to set up and test.
- ▶ There is a potential for application delays due to dual writes occurring.
- ▶ It does not allow for *point-in-time* copy or easy backout once the first copy is removed.

## 9.5.2 Basic commands

Basic commands such as **cpio** (in a UNIX environment) or **copy** (in a Windows environment) are easy and common to use. The disadvantages of basic commands are:

- ▶ Length of service disruption varies by commands used.
- ▶ It is tedious to handle large numbers of LUNs.
- ▶ Command scripting is prone to human error.
- ▶ Security file level access issues can arise, depending on which commands used.

## 9.5.3 Software packages

You can employ data migration, backup, and restore packages, as well as database tools, to migrate data. An example of a data migration package is BRMS. Tivoli Storage Manager (TSM) may also be used in a backup and restore fashion to achieve data migrations. DB2 utilities and tools such as unload, load, and copy can also be used to migrate data. Other third-party data migration, backup and restore, and database packages are available. Contact the respective vendor for further assistance.

The advantages of using software packages are:

- ▶ Small application impact when using data migration package.
- ▶ Little disruption for *non-database* files.
- ▶ Backup or restore cycles are offloaded to another server.
- ▶ They are often standard database utilities.

The disadvantages of using a software package are:

- ▶ Cost of data migration package.
- ▶ Bigger impact or disruption with large databases due to lack of checkpoint-restart capabilities.
- ▶ Possibly lengthy application outage to back up or restore environment.
- ▶ Application service interruptions are possible.

## 9.5.4 Remote copy technologies

Remote copy technologies include synchronous and asynchronous mirroring. They include Metro Mirror, Metro/Global Copy and Global Copy.

The advantages of remote technologies are:

- ▶ Other than Global Copy for zSeries, they are operating system independent.

- ▶ Minimal host application outages.

The disadvantages of remote copy technologies are:

- ▶ The same storage device types are required. For example, in a Metro Mirror configuration you need ESS 800 mirroring to a DS8000 (or an IBM approved configuration), but cannot have a non-IBM disk system mirroring to a DS8000.
- ▶ Physical volume ID (PVID) and device name are not maintained if not under LVM.

### 9.5.5 Migration services and appliances

The following IBM migration services and appliances are available:

- ▶ IBM Piper Services Offering
- ▶ IBM SAN Volume Controller

These data migration services and appliances provide smooth, risk-averse data migration to your DS8000. The advantages of these migration methods are:

- ▶ Facilitated by custom migration tools.
- ▶ Minimal customer involvement by IT staff, except planning.
- ▶ Minimal disruption or outages to IT operations.
- ▶ Online migrations can occur while continuing IT operations.
- ▶ Tunable migration rate to eliminate impact to applications.
- ▶ Transparent to application servers.
- ▶ High throughput tool minimizes migration duration.
- ▶ Delivered by team experienced with many migrations.
- ▶ Maintains data integrity during migration.
- ▶ Can fall back to the original data and storage device.
- ▶ Large number of operating systems and storage systems supported.

The disadvantages of migration appliances are:

- ▶ Cost of migration appliance or service.
- ▶ Application disruption to install and remove appliance.

See Appendix C, “Service and support offerings” on page 407 for storage services offerings.

### 9.5.6 z/OS data migration methods

Figure 9-15 on page 186 lists a number of data migration methods available to migrate data from existing disk systems to the DS8000 in a zSeries environment.

Data migration methods	
Environment	Data migration method
S/390	IBM TotalStorage Global Mirror, Remote Mirror and Copy (when available)
zSeries	IBM TotalStorage Global Mirror, Remote Mirror and Copy (when available)
Linux environment	IBM TotalStorage Global Mirror, Remote Mirror and Copy (when available)
z/OS operating system	DFSMSdss (simplest method)
	DFSMSHsm
	IDCAMS Export/Import (VSAM)
	IDCAMS Repro (VSAM, SAM, BDAM)
	IEBCOPY
	ICEGENER, IEBGENER (SAM)
	Specialized database utilities for CICS, DB2 or IMS
	Softtek Replicator (previously known as TDMF)
	INNOVATION FDR Plug and Swap (FDRPAS)
VM operating system	DASD Dump Restore
	CMDISK
	COPYFILE
	PTAPE
VSE operating system	VSE fastcopy
	VSE ditto
	VSE power
	VSE REPRO or EXPORT/IMPORT

Figure 9-15 Different data migration methods

See Chapter 14, “Data migration in zSeries environments” on page 293 for a complete discussion of the different methods available for data migration from any other disk system to the DS8000 family.

## 9.6 Planning for performance

IBM TotalStorage DS8000 is a high-performance, high-capacity series of disk storage that is designed to support continuous operations and allows your workload to be easily consolidated into a single storage subsystem. To have a well-balanced disk system the following components that affect performance need to be considered:

- ▶ Disk Magic
- ▶ Size of cache storage
- ▶ Number of channels
- ▶ Remote Copy
- ▶ Parallel Access Volume
- ▶ I/O priority queuing
- ▶ Monitoring performance
- ▶ Hotspot avoidance
- ▶ Balance I/O activity



## 9.6.1 Disk Magic

An IBM representative or an IBM Business Partner can model your workload using Disk Magic before migrating to the DS8000. Modelling should be based on performance data covering several time intervals, and should include peak I/O rate, peak R/T and peak (read and write) MB/second throughput. Disk Magic will provide insight when you are considering deploying remote technologies such as Metro Mirror. Consult your sales representative for assistance with Disk Magic.

**Note:** Disk Magic is available to IBM sales representatives and IBM Business Partners only.

## 9.6.2 Size of cache storage

Having adequate cache storage to sustain your peak workload is imperative for consistent disk system performance. The cache storage is divided into write cache and persistent cache. The DS8100 Model 921 offers up to 128 GB of processor memory and the DS8300 Models 922 and 9A2 offer up to 256 GB of processor memory. In addition, the Non-Volatile Storage (NVS) scales to the processor memory size selected, which can also help optimize performance.

## 9.6.3 Number of host ports/channels

You should plan to have an adequate number of host ports or channels to provide the required bandwidth to support your workload. The ports must also be balanced across the entire DS8000. FICON ports are faster and more efficient than ESCON ports.

## 9.6.4 Remote copy

If the DS8000 is a primary disk system in a remote copy configuration, it will consume more resources, such as cache and channels, compared to a standalone disk system, and planning should be done accordingly.

## 9.6.5 Parallel Access Volumes (z/OS only)

Configuring the DS8000 with PAV will minimize or eliminate IOSQ delays and improve disk performance. PAV can either be static or dynamic. Dynamic PAV should be implemented when possible, as this provides more flexibility. z/OS Workload Manager (WLM) will manage the PAV devices as a group, instead of having a static relationship with the base device. In a static relationship the base device cannot borrow a PAV device from another base device that is not in use. In a dynamic environment, idle PAV devices are assigned to base devices that need more PAV devices to manage the workload. WLM manages the PAV devices on an LSS group level.

## 9.6.6 I/O priority queuing (z/OS only)

I/O priority queuing allows the DS8000 series to use I/O priority information provided by the z/OS Workload Manager to manage the processing sequence of I/O operations.

## 9.6.7 Monitoring performance

A number of monitoring tools are available to measure the performance of your DS8000 once it is installed into your configuration.

For example, in the zSeries environment, the RMF™ RAID rank report can be used to investigate RAID rank saturation when the DS8000 is already installed in your environment. New counters will be reported by RMF that will provide statistics on a volume level for channel and disk analysis.

In an open system environment the **iostat** command is useful to determine whether a system's I/O load is balanced or whether a single volume is becoming a performance bottleneck. The tool reports I/O statistics for TTY devices, disks, and CD-ROMs. It monitors I/O device throughput and utilization by observing the time the disks are active in relation to their average transfer rates. The **vmstat** utility may also be used to take a quick snapshot or overview of the system performance.

### 9.6.8 Hot spot avoidance

Workload activity concentrated on a limited number of RAID ranks will saturate the RAID ranks. This may result in poor response times, so balancing I/O activity across any disk system is important. Spreading your I/O activity evenly across the available DS8000s will enable you to optimally exploit the DS8000 resources, thus providing better performance. I/O activity attributes to consider include spreading I/O activity across:

- ▶ The two servers of the DS8000
- ▶ The loops of the DS8000
- ▶ All available RAID ranks
- ▶ Across multiple DS8000s



## The DS Storage Manager - logical configuration

In this chapter, the following topics are discussed:

- ▶ Configuration hierarchy, terminology, and concepts
- ▶ Summary of the DS Storage Manager logical configuration steps
- ▶ Introducing the GUI and logical configuration panels
- ▶ The logical configuration process

## 10.1 Configuration hierarchy, terminology, and concepts

The DS Storage Manager provides a powerful, flexible, and easy to use application for the logical configuration of the DS8000. It is the client's responsibility to configure the storage server to fit their specific needs. It is not in the scope of this redbook to show detailed steps and scenarios for every possible setup. Help and guidance can be obtained from an IBM FTSS or an IBM Business Partner if the client requires further assistance.

### 10.1.1 Storage configuration terminology

An understanding of the following concepts and terminology may help you use and configure the DS Storage Manager configurator:

#### **Storage unit**

A storage unit, also known as a *storage facility*, is a single physical storage subsystem (DS8000).

#### **Storage complex**

A storage complex consists of one or more physical storage units that can be managed from a central management point. DS8000 units can be placed together to form a complex. Currently one DS8000 is managed by one S-HMC. In 1Q05 you will be able to have one S-HMC (or also a redundant S-HMC) manage two DS8000s.

#### **Storage image**

A storage image, also known as a *storage facility image (SFI)*, is a logical storage subsystem and only applies to the DS8000. It consists of two LPARs, one on each processor complex.

#### **Host attachment**

A host attachment is a group of host ports that you want to manage the same way. The GUI allows grouping of one or more host ports, installed on a given host, into what is called a *host attachment*. The host ports we are referring to here are the HBA ports on the host, not the DS8000. The GUI allows the user to specify the WWPNs of each HBA on the host that you want to group together to form a host attachment. You specify how many HBAs you want to connect to on the host, then click a checkbox to indicate that you want to group these (host HBAs) together.

This makes it easy to set up volume access, via storage image I/O ports and volume groups, in the same way for all the host ports, since the same settings are applied to all the host ports grouped into the same host attachment. Volumes can be assigned to volume groups, and volume groups can then be assigned to host attachments, for presentation to the host operating systems.

Traditionally we think of hosts with one or more Fibre Channel adapters (HBAs), with each adapter having one or more Fibre Channel ports. Each port has a unique Fibre Channel address called the World Wide Port Name (WWPN). The WWPN is the address to which we assign volumes. A host attachment is a grouping of ports and their World Wide Port Names. Definitions are made about hosts and attachments in the GUI. The concepts and limitations are explained in the following list:

- ▶ Multiple Fibre Channel ports' WWPNs on the same host system can be specified in one or more host attachments, in one host definition called *host system* in the GUI. A host attachment does not always mean a single port. For example, refer to Figure 10-1 on page 191.

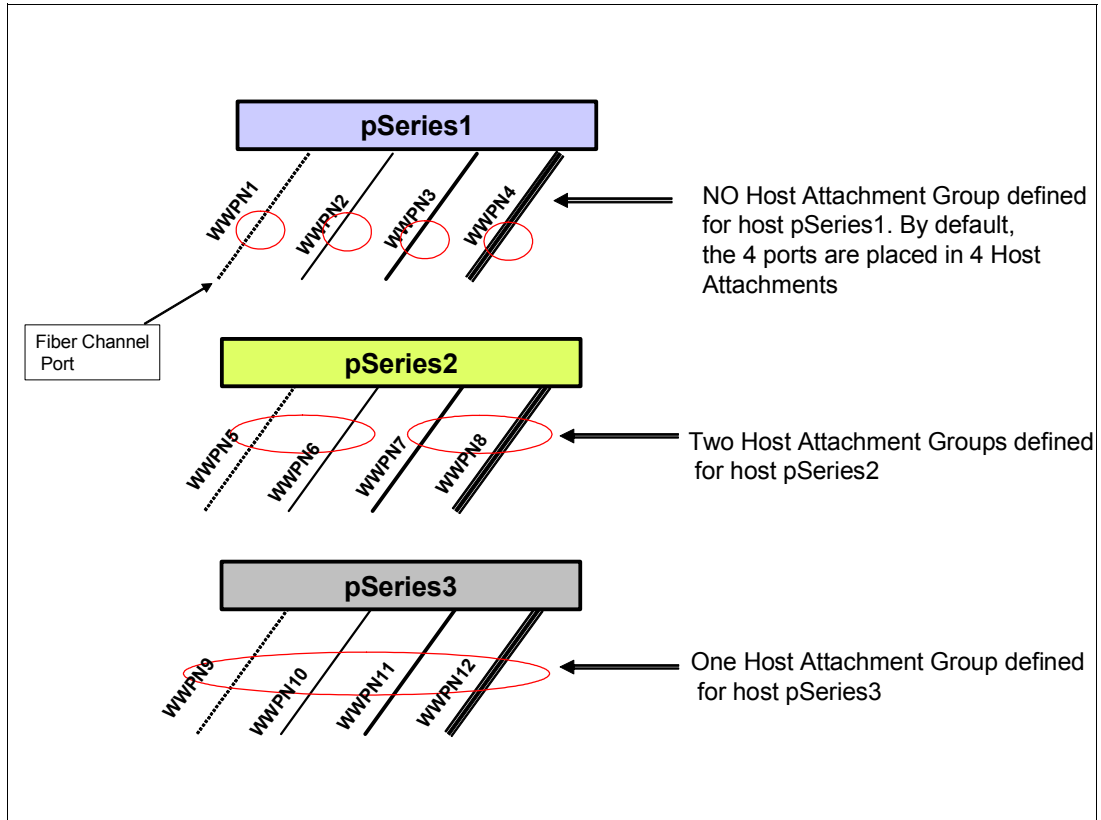


Figure 10-1 Diagram of hosts and host attachment groups

- ▶ In Figure 10-1 we show three pSeries hosts with four host ports each. The first host, pSeries1, has no specific attachment groupings, so each port could be defined as an attachment. Server ports can be grouped in attachments for convenience. For example, pSeries2 shows two host attachments, with two ports in each attachment. Server pSeries3 shows one host attachment with four ports grouped in the one attachment.
- ▶ A host attachment can be configured to access specific disk subsystem I/O ports or all valid disk subsystem I/O ports.
- ▶ Host attachments can be configured to access specific volume groups.
- ▶ A specific host attachment (one port or set of grouped ports) can access only one volume group.
- ▶ Multiple host attachments, even different open system host types, with the same block size and addressing mode, can access the same volume group. The safest approach to this concept is to configure one host per volume group. If shared access to the LUN is required, for example, for dual pathing or clustering, then the shared LUNs may be placed in multiple volume groups as shown in Figure 10-3 on page 195.
- ▶ FICON/ESCON attachment access is controlled by zSeries HCD/IOGEN definitions. Some ESCON/FICON attachment considerations follow.
  - One storage subsystem FICON host definition will be needed in order to configure I/O adapters to use FICON protocol instead of FCP protocol.
  - No storage subsystem host definition will be required for ESCON hosts since ESCON adapters are single purpose and do not require configuration.
  - Default volume groups are automatically created, allowing anonymous access for ESCON/FICON.

One set of host definitions may be used for multiple storage images, storage units, and storage complexes.

### ***DDM***

A Disk Drive Module (DDM) is a field replaceable unit that consists of a single disk drive and its associated packaging. DDMs are ordered in a group of 16 called a drive set. A drive set is installed across two disk enclosures, 8 DDMs in the front disk enclosure and 8 DDMs in the rear disk enclosure. When ordering disk enclosures they come in pairs, each able to hold 16 drives. If any disk enclosure is not full of DDMs, then the empty slots must be filled with dummy carriers called *disk enclosure fillers*. Disk enclosure fillers can be ordered in groups of 16 per order.

### ***Array sites***

An array site is a predetermined grouping of eight individual DDMs of the same speed and capacity. Four disks from a rear disk enclosure and four disks from a front disk enclosure make up the array site.

### ***Arrays***

Arrays consist of DDMs from an array site, used to construct one RAID array. An array is given either a RAID-5 or RAID-10 format.

### ***Ranks***

One array forms one CKD or Fixed Block (FB) rank. When the rank is configured, either CKD or FB characteristics are taken on at this point. Presently only one array can reside in a rank, but in the future one or more arrays will be able to reside in one rank.

**Note:** The ranks have no pre-determined relation to an LSS.

### ***Extent pools***

An extent pool consists of one or several ranks. Ranks in the same extent pool must be of the same data format (CKD or FB). Each extent pool is associated with server 0 or server 1. Although it is possible to create extent pools with ranks of different drive capacities, speeds, and RAID types, we recommend you create them to consist of the same RAID type, speed, and capacity. Also, we recommend that you configure only half of the total number of ranks to reside in one pool (server 0) and the other half in the other pool (server 1). Extent pools contain one or more ranks divided into fixed-size extents as follows:

- ▶ CKD extents are equal to a 3390 Mod1
- ▶ FB extents are 1GB

The storage in an extent pool (the extents from each rank in the extent pool) is used to create logical volumes. We recommend that you create one extent pool out of only one rank to start with, unless the size required for a Fixed Block logical volume (LUN) is larger than the combined free extents residing in one single rank.

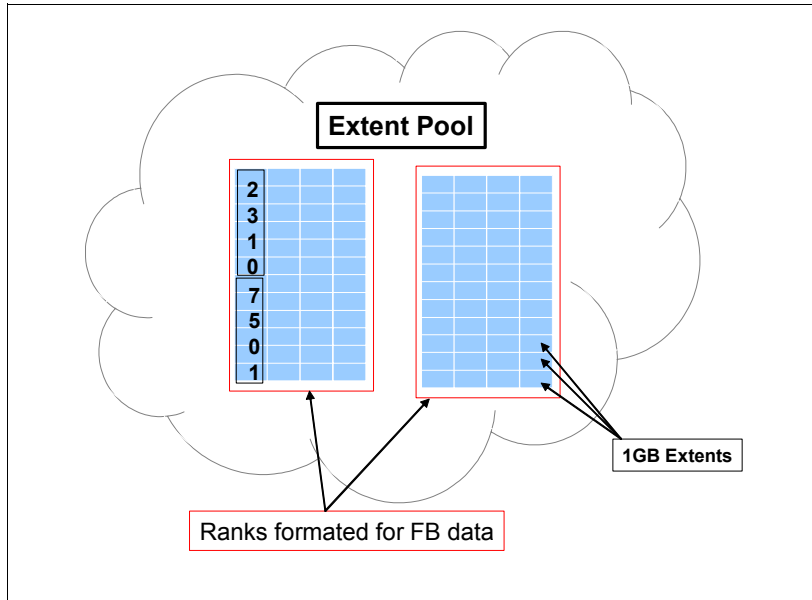


Figure 10-2 An extent pool containing 2 volumes

Figure 10-2 is an example of one extent pool composed of ranks formatted for FB data. Two logical volumes are defined (volumes 2310 and 7501). Each volume is made up of 6 extents at 1 GB each. This makes each volume 6 GBs. The numbering sequence of LUN id 2310, shown in the diagram as the top volume, translates into an address. The volume identification number is built using the following rules:  $xyzz$ ,  $x$  = the LSS address group,  $xy$  = LSS number itself, and  $zz$  = the volume id(void). For example, LUN 2310 has an address group number of 2, is located in LSS 23, and has a void of 10.

Different volumes on a single extent pool can be assigned to the same or different LSSs.

Extent pools are assigned to server 0 and server 1 during configuration and receive their server affinity at this time. If you are using the custom configuration, we recommend, for user manageability reasons, that you associate the rank even numbers to server 0 and the rank odd numbers to server 1 when defining the extent pools during the configuration process.

The following rules apply when creating extent pools:

- ▶ You must configure a minimum of two extent pools to utilize server 0 and server 1.
- ▶ More than one rank can reside in an extent pool, but you cannot make two extent pools out of only one rank. We recommend that you create one extent pool out of one rank, unless the LUN capacity is greater than the capacity of one rank in the extent pool.

Some general considerations are:

- ▶ One rank per pool will not constrain addresses.
- ▶ Ranks can be added to an extent pool at any time.
- ▶ The logical volumes defined in one extent pool can be in different LSSs.
- ▶ The logical volumes in different extent pools can be in the same LSS; they are limited only by the odd and even server affinity.
- ▶ Ranks can be removed from an extent pool if no extents on the rank are currently assigned to the logical volumes.
- ▶ Any extent can be used to make a logical volume.

- ▶ There are thresholds that warn you when you are nearing the end of space utilization on the extent pool.
- ▶ There is a Reserve space option that will prevent the logical volume from being created in reserved space until the space is explicitly released.

**Note:** A user can't control or specify which ranks in an extent pool are used when allocating extents to a volume.

### **Logical volumes**

Logical volumes, also known as *LUNs* when configured for open systems or *CKD volumes* when configured for zSeries, can only be made from one or more extents residing in the same extent pool. This means that a logical volume cannot span multiple extent pools. Ranks must be added to extent pools to make larger LUNs. We use the terms *volume* and *LUN* interchangeably throughout the rest of this chapter when referring to logical volumes.

Some limitations are:

- ▶ A specific volume is in one LSS.
- ▶ Multiple volumes in one extent pool or one rank can be in the same or different LSSs, as shown in Figure 10-6 on page 198.
- ▶ Multiple volumes in different extent pools and on different ranks can be in the same LSS, also as shown in Figure 10-6.
- ▶ The minimum volume/LUN size is one extent.
  - For CKD the minimum size is a 3390 Mod1.
  - For FB the minimum size is 1GB.

**Note:** The user can specify volume sizes in binary, decimal, or block sizes. When you specify binary form, 1 GB is equal to 1073741924 bytes, instead of the decimal size, as in the ESS, where 1GB is 1000000000 bytes.

- ▶ The maximum volume/LUN size is equal to the size of the extent pool, with the following limitations: 56 GB for CKD, with appropriate zSeries software support, and 2 TB for FB. For example, if only one rank was residing in the extent pool, then the maximum LUN size would be equal to the capacity of that one rank.
- ▶ The maximum number of logical volumes at GA for the DS8000 is 64K; of the 64K LUNs, a maximum of 32K can be for FB and 32K can be for CKD.
- ▶ Volumes can be deleted and the extents reused without having to format the ranks or arrays they reside in.

### **Volume groups**

A volume group is a collection of logical volumes. Volume groups are created to provide FB LUN masking by assigning logical volumes and host attachments to the same volume group. For CKD volumes, one volume group for ESCON and FICON attachment with an anonymous host attachment is automatically created.

Volume groups can be thought of as LUN groups. Do not confuse the term volume group here with that of volume groups on pSeries. The DS Storage Manager volume groups have the following properties:

- ▶ Volume groups enable FB LUN masking.



- ▶ They contain one or more host attachments from different hosts and one or more LUNs. This allows sharing of the volumes/LUNs in the group with other host port attachments or other hosts that might be, for example, configured in clustering.
- ▶ A specific host attachment can be in only one volume group. Several host attachments can be associated with one volume group. For example, a port with a WWPN number of 10000000C92EF123 can only reside in one volume group, not two or more.
- ▶ Assigning host attachments from multiple host systems, even running different operating system types, are allowed in the same volume group.

**Note:** We recommend that you configure one volume group per host. Presently the GUI does not allow you to make more than one volume group per host attachment. If you want multiple volume groups on a single host, the GUI does allow you to make more than one host attachment to a host. This is the way you achieve multiple volume groups on a host.

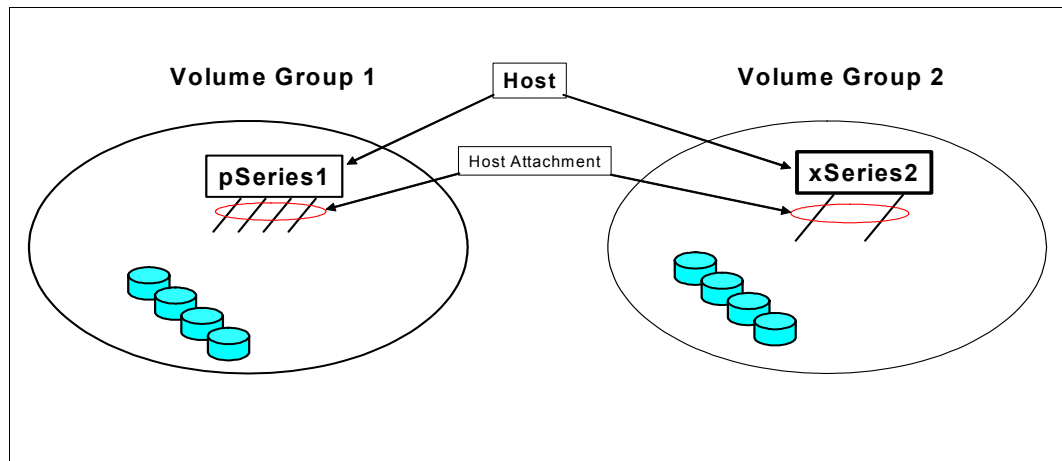


Figure 10-3 Diagram of the relationship of a host attachment to a volume group

In Figure 10-3, we show two volume groups, volume group 1 and volume group 2. A pSeries server with 1 host attachment (four ports grouped in that attachment) resides in volume group 1. The xSeries2 server has 1 host attachment (2 ports grouped into the attachment). The ports are grouped together in one attachment definition. For example, the server, xSeries2 is dual pathed to the LUNs through one attachment group definition.

- ▶ In order to share LUNs across multiple host attachments, LUNs can be in more than one volume group as shown in the example in Figure 10-4 on page 196.

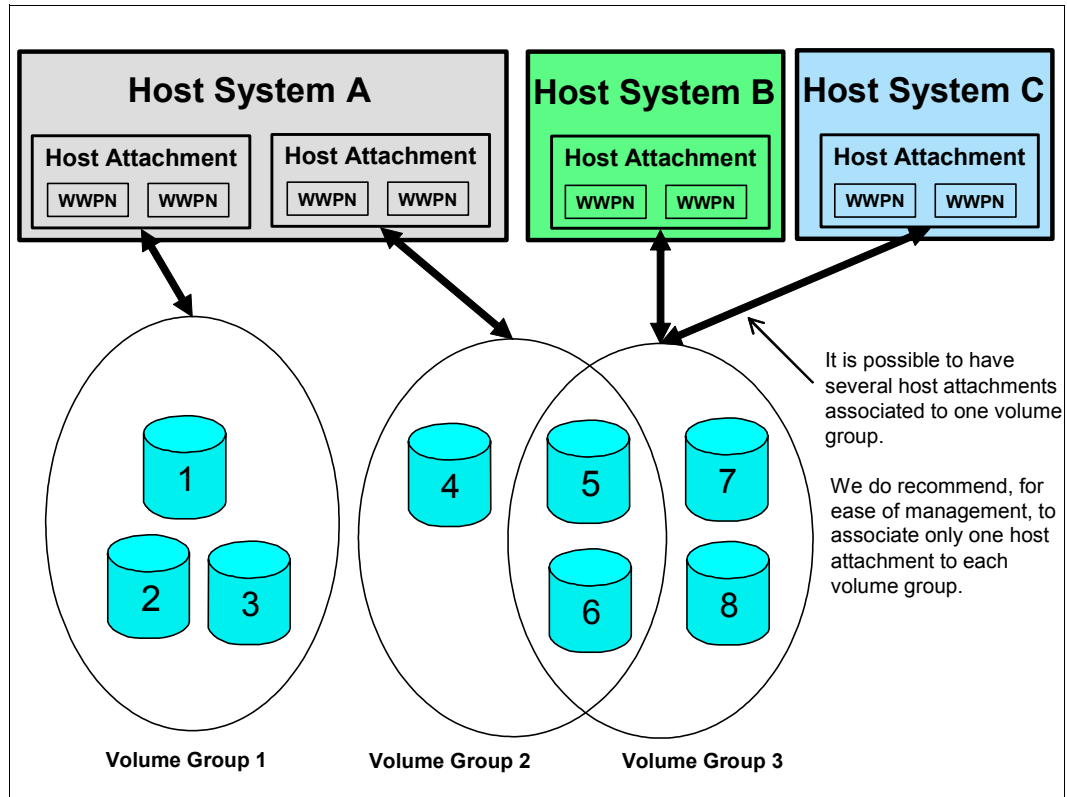


Figure 10-4 Example of volume groups, LUNs and host attachment definitions

In Figure 10-4 we show three hosts (host A, host B, and host C) defined in the logical configuration as three different *host systems*. The user will group the WWPN of each host system in groups called host attachments. As shown in the diagram, each host attachment is assigned to one volume group. Volumes can belong to several volume groups. For example, volumes 5 and 6 are in volume group 2 and volume group 3, so they will be shared by host A, host B, and host C. Several host attachments can be associated to the same volume group. For example, hosts B and C will share volumes 5, 6, 7, and 8 because their host attachments are assigned to the volume group 3. However, for management simplification, we recommend that only one host attachment is assigned to each volume group.

- The maximum number of volume groups for the DS8000 is 8320.

### Address groups

An address group is a group of FB or CKD LSSs. An address group has up to 16 LSSs.

The DS8000 supports up to 16 address groups. Each address group is identified by one hexadecimal digit: address group 0 to address group F.

### LSS/LCU

A logical subsystem (LSS) is a topological construct that consists of a group of up to 256 logical volumes. The DS8000 architecture allows up to 255 LSSs (hex 00 to FE). However, at GA, a DS8000 will allow you to have up to 64 CKD logical subsystems (16384 CKD logical volumes) and up to 64 fixed-block logical subsystems (16384 fixed-block logical devices). There is a one-to-one mapping between a CKD logical subsystem and a zSeries control-unit image. ESCON-attached hosts can only access LSSs 00 to 0F, so these should be reserved for CKD LSSs if you plan to use ESCON host adapters.

Figure 10-5 shows an example of the relationship between LSSs, extent pools, and volume groups: Extent pool 4, consisting of two LUNs, LUN 2210 and LUN 7401; and extent pool 5, consisting of three LUNs, LUN 2313, 7512, and 7515.

Here are some considerations regarding the relationship between LSS, extent pools, and volume groups:

- ▶ Volumes from different LSSs and different extent pools can be in one volume group as shown in Figure 10-5. Volume group 1 consists of extent pool 4, LUN 2210, and extent pool 5, LUN 7512.
- ▶ Volumes from the same LSS or the same extent pool, or both, can be in different volume groups. For example in Figure 10-5, LUN 7512 in extent pool 5 and LUN 7515 in extent pool 5, both of which reside in LSS75, also reside in different volume groups.

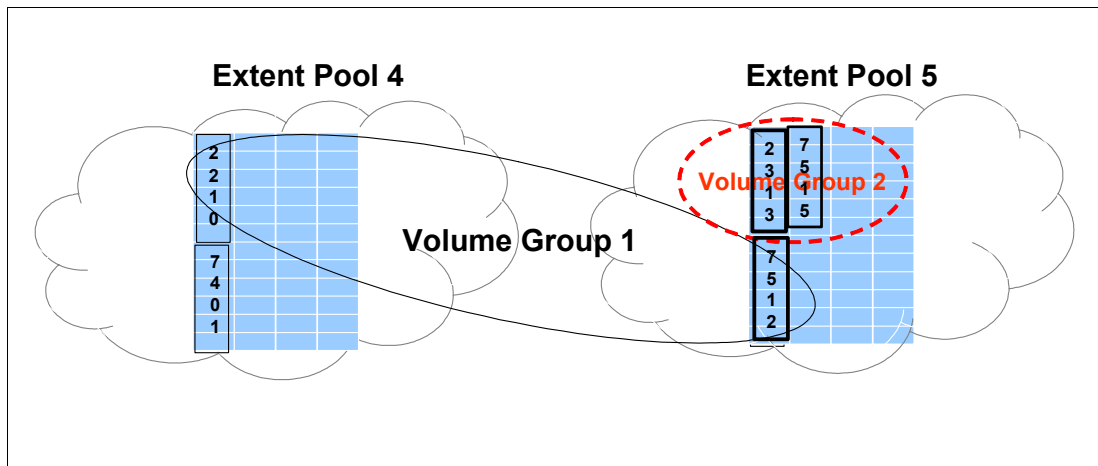


Figure 10-5 Example of relationship between LSSs, extent pools, and volume group

The LSS and LCU provide the logical grouping of volumes/LUNs for Copy Services and various other purposes. Some of these purposes are explained along with general considerations about LSSs in the DS8000, as follows:

- ▶ The LSS or LCU determines the addressing, address groups, and PAVs.
  - Each logical volume has a 4 digit Hexadecimal xyzz identification number built using the following rules: x = the address group, xy = LSS number itself, and zz = the volume id (valid). For example, LUN 2310 has an address group number of 2, is located in LSS 23, and has a valid of 10.
  - There are up to 16 LSSs in an address group. For example, 00 to 0F for address group 0, 10-1F for address group 1, 20-2F for address group 2, and so forth.
  - Any given PAV can only be used within one LCU.
- ▶ LSSs are used for Copy Services, PPRC paths, and consistency group properties or time-outs.
- ▶ There are a maximum number of LSSs. For the DS8000 there are a maximum of 255 LSSs. When speaking of ESCON, the maximum is 16 ESCON CKD LSSs.
- ▶ The LSSs have a pre-determined association with Server0 or Server1.
  - The even LSSs are associated with Server0.
  - The odd LSSs are associated with Server1.
- ▶ The LSSs are configured to be either CKD or FB.
  - CKD LSSs' definitions are configured during the LCU creation.

- FB LSSs' definitions are configured during the volume creation.
- ▶ LSSs have no predetermined relation to physical ranks or extent pools other than their server affinity to either Server0 or Server1.
  - One LSS can contain volumes/LUNs from different extent pools.
  - One extent pool can contain volumes/LUNs that are in different LSSs, as shown in Figure 10-6.
  - One LCU can contain CKD volumes/LUNs of different types, for example, type 3390 Model 3 and Model 9.
  - LSSs can have a many-to-many Copy Services relationship as shown by the arrows in Figure 10-6.

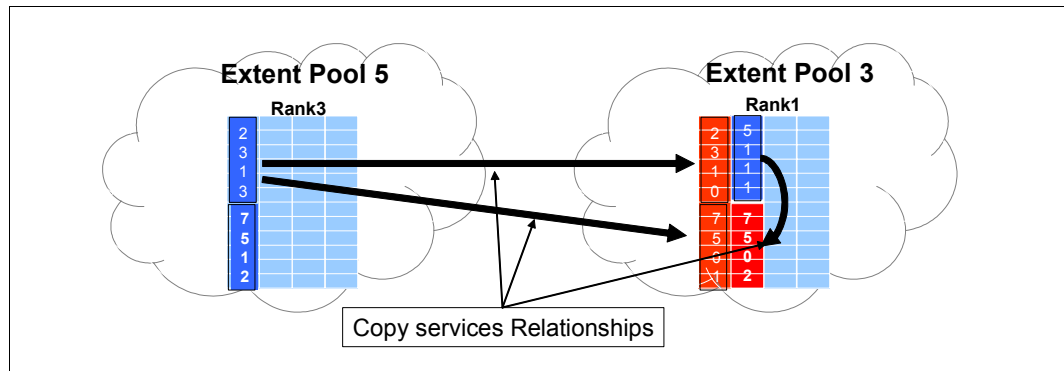


Figure 10-6 Example of Copy Services relationship between LSSs in the same storage image

- ▶ An entire address group will be either FB or CKD; for example, all 16 LSSs in that group would be either FB or CKD type LSSs.
- ▶ Initially there will only be 8 address groups for the DS8000 at GA, however more will be available at a later time.
- ▶ One address group has 4096 addresses (16 LSS x 256 logical volume = 4096 logical volumes or addresses). This means that if you had eight address groups, you would have 32768 (8 X 4096 = 32768) addresses, which translates to 32768 logical volumes or addresses.

**Attention:** The address groups and LSSs are predefined in the DS8000. Initially, they do not have any data format attributes (FB or CKD). Their data format attributes are set when the first LUN is added to one LSS in an unused address group or when you create an LCU using an unused address group.

When creating a logical volume, the user is prompted to add the volume to an LSS. Adding the volume to an unused LSS will set its data format characteristics (FB or CKD) and also will set the address group data format characteristics. It also sets the data format of the remaining 15 LSSs in that address group.

As an example: When creating a logical volume from an extent pool built with FB ranks, the GUI will prompt for which LSS this FB volume will be placed in, and will propose a list of LSSs. If the user chooses LSS 14 and if it is the first LUN to be placed in this LSS, then the attributes of address group 1 will be Fixed Block, reserving LSS 10 to 1F for Fixed Block volumes. The user will not be able to create any LCU using address group 1.

**Tip:** We recommend that you reserve all LSSs in address group 0 (LSSs 00-0F) for CKD ESCON attachments because ESCON attachments must use address group 0.

## 10.1.2 Summary of the DS Storage Manager logical configuration steps

It is our recommendation that you review the following concepts before performing the logical configuration. These recommendations are discussed in this chapter.

### Planning

When configuring available space to be presented to the host, we suggest that you approach the configuration from the host and work up to the DDM (raw physical disk) level. This is just the opposite way that you would configure the raw DDMs into host volumes. Refer to Figure 10-7 on page 200 to understand the hierarchy in the virtualization layers in the DS8000.

1. Determine the number of hosts and type of hosts (zSeries or Open System) in the environment that will use external capacity.
2. Determine the amount of capacity needed for each host and for each data format type.
3. Determine the number and the size of logical volumes needed to fulfill the capacity requirements for zSeries and open system hosts.
4. Determine the number of rank types (FB or CKD) and the number of ranks per extent pool to be able to build the specific logical volumes. The recommendation is to have one rank per extent pool unless the LUN size requires you to spread the LUN on several ranks in an extent pool.
5. Determine the number of address groups that will be assigned for CKD LSSs and the number of address groups that will be assigned for FB LSSs, and reserve the corresponding address group. Note that address group 0 needs to be reserved for ESCON attachment.

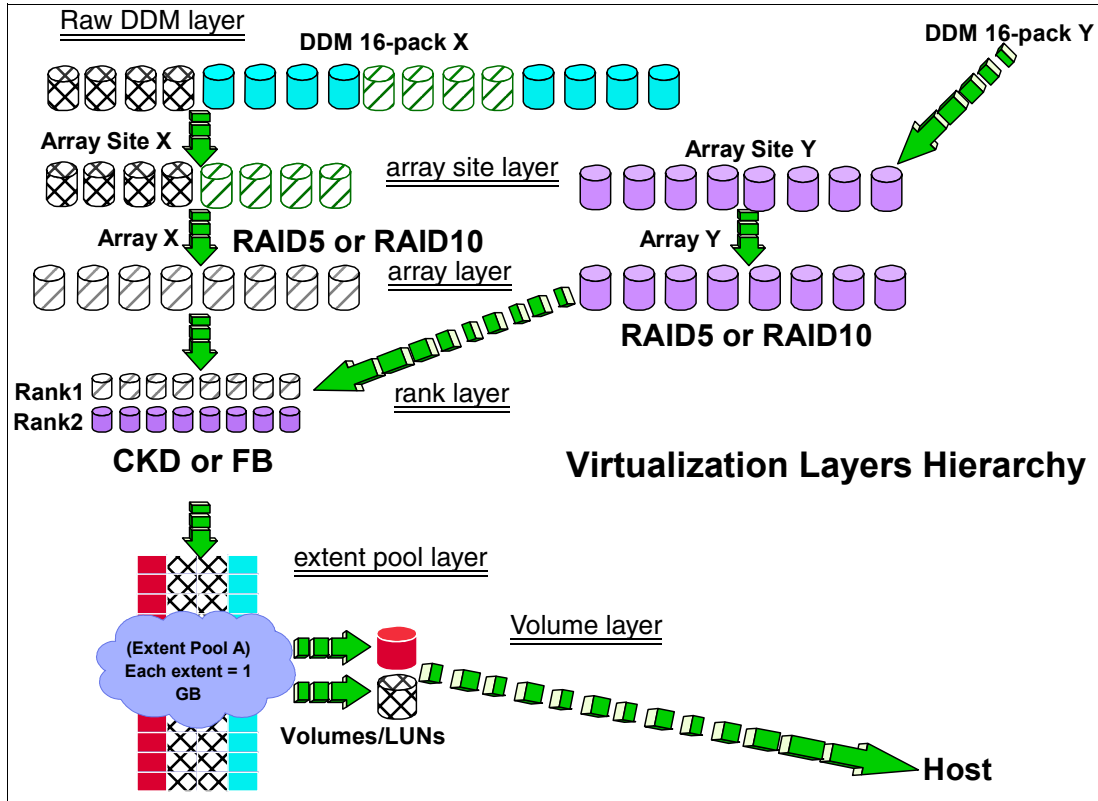


Figure 10-7 View of the raw DDM to LUN relationship

### Raw or physical DDM layer

At the very top of Figure 10-7 you can see the raw DDMs. There are 16 DDMs in a disk drive set. DDM X represents one 16-pack and DDM Y represents another 16-pack. Upon placing the 16-packs into the DS8000, each 16-pack is grouped into array sites, shown as the second layer.

### Array site layer

At the array site level, predetermined groups of eight DDMs of the same speed and capacity are arranged. An *arrays across loop* strategy is used in the predetermined groupings so that an array does not consist of the same, eight physical raw disks in the disk drive set.

### Array layer

This level is where the format is placed on the array. Sparring rules are enforced depending on which RAID format is chosen. If you choose RAID-5, then one spare is created and a RAID-5 format is striped across the remaining 7 drives. You must calculate the equivalent of one disk that is used for parity out of the array. Although parity is not placed on one physical disk, but striped across all the remaining disks, that parity equals one disk's worth of capacity. See Figure 10-8 on page 201. In this figure the RAID format is a 6 + P + S. If you add up the parity chunks it equals one disk's worth of capacity. If you chose RAID-10 then you would have two spares with no parity and a 3 X 3 + 2(spares) configuration. This would continue for each RAID array until the sparring rules are met.

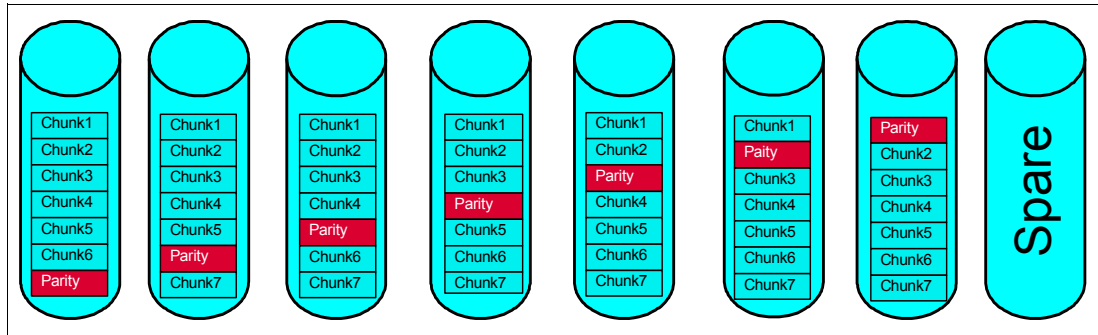


Figure 10-8 Diagram of how parity is striped across physical disks

### **Rank layer**

At this level the ranks are formed. Presently only one array can make up a rank.

### **Extent pool layer**

At this level the extent pools are formed. In Figure 10-7 we show that extent pool A is made up of 2 ranks: rank 1 and rank 2. The extents in the pool are 1GB.

### **Logical volume layer**

Layer 6 is the final level of the LUN formation. As illustrated in Figure 10-7, two LUNs were created out of extent pool A. The top LUN is made up of 12 extents, making it a 12 GB LUN. The bottom LUN is made up of 24 extents, making it a 24 GB LUN.

### **Logical Configuration flow**

Figure 10-9 on page 202 shows the recommended flow for performing the Logical Configuration using the GUI.

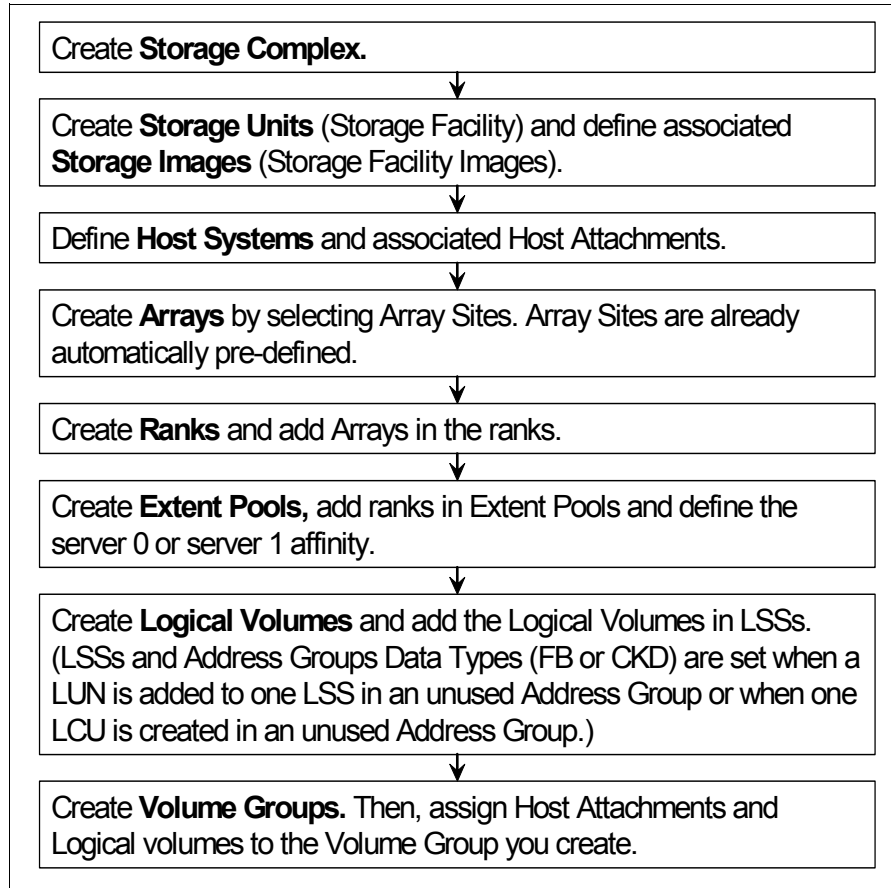


Figure 10-9 Recommended logical configuration steps

## 10.2 Introducing the GUI and logical configuration panels

The IBM TotalStorage DS Storage Manager is a program interface that is used to perform logical configurations and Copy Services management functions. The DS Storage Manager program is installed via a GUI (graphical mode) or as an unattended (silent mode) installation for the supported operating systems. It can be accessed from any location that has network access using a Web browser. This section describes the DS Storage Manager GUI and logical configuration concepts and steps that allow the user a simple and flexible way to successfully configure FB and CKD storage.

### 10.2.1 Connecting to the DS8000

To connect to the DS8000 through the browser, enter the URL of either the default Storage Hardware Management Console (S-HMC) or the optional S-HMC you may have purchased. You can connect through either S-HMC, but we recommend that once you start updating and modifying from one S-HMC, you continue to make your changes through that S-HMC for the duration of the change. The URL consists of the TCP/IP address as shown in Figure 10-10, or a fully qualified name that the DNS server can resolve as shown in Figure 10-11 on page 203.



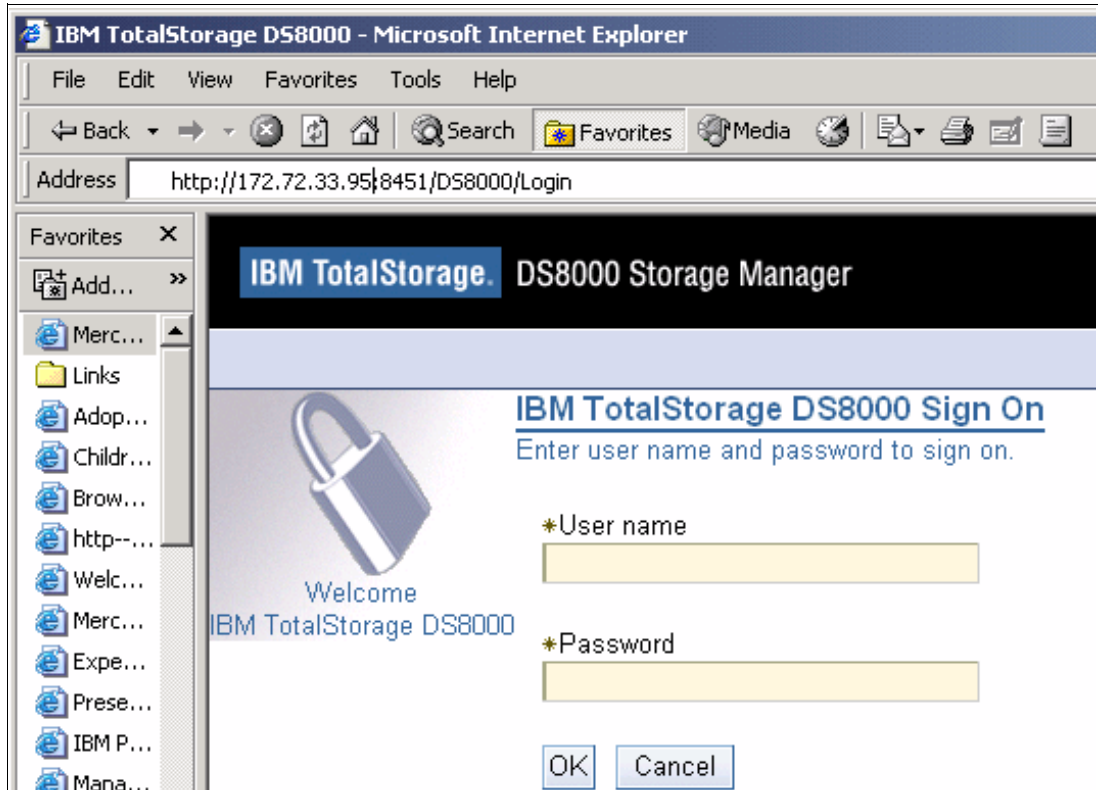


Figure 10-10 Entering the URL using the TCP/IP address for the S-HMC

In Figure 10-10, we show the TCP/IP address and the port number 8451 afterwards.



Figure 10-11 Entering the URL using the fully qualified name of your S-HMC

In Figure 10-11, we show the fully qualified name and the port number 8452 separated by a colon.

For ease of identification, you could add a suffix such as 0 or 1 to the selected fully qualified name, for example, SHMC\_0 for the default S-HMC as shown in Figure 10-11. Then bookmark it for ease of use. When assigning the number, we recommend that you make the last field of the optional S-HMC only one digit higher than the default S-HMC.

## 10.2.2 The Welcome panel

The IBM TotalStorage DS8000 Storage Manager (DS Storage Manager) is a software application that runs on the S-HMC. It is the interface provided for the user to define and maintain the configuration of the DS8000. The DS Storage Manager can be accessed using a

Web browser running directly to the on-board S-HMC, or in a remote machine connected into the user's network.

Once the GUI is started and the user has successfully logged on, the Welcome panel shown in Figure 10-12 is displayed.



Figure 10-12 The Welcome panel

Figure 10-12 shows the Welcome panel's two menu choices. Click the triangle beside either menu item to expand the menu; this displays the options needed to configure the storage.

The DS8000 Storage Manager configurator can be used either in *Real-time* (online) or *Simulated* (offline) mode as shown in Figure 10-12. Either mode can be used to manipulate the storage configuration process for a DS8000, defining CKD and fixed block (FB) storage. Either mode can also be used to modify an existing configuration.

It is important to know that the Simulated Manager is limited in its function, and is used to pre-configure new configurations or modify existing configurations; the modifications are executed at a later time. For example, the Simulated Manager could be used to execute or modify changes at an off-peak hour.

► **Real-time Manager configuration**

You can use the Real-time mode selections of the DS Storage Manager if you chose **Real-time** during the installation of the DS Storage Manager. Part of the Real-time configuration process requires you to input the OEL license activation key. You can obtain this key and all of the license activation keys from the Disk Storage Feature Activation (DSFA) Web site at:

<http://www.ibm.com/storage/dsfa>

This application provides logical configuration and Copy Services functions for a storage unit attached to the network. This feature provides you with real-time (online) configuration

support. A view of the fully expanded Real-time manager menu choices is shown in Figure 10-13.



Figure 10-13 The fully expanded Real-time manager menu choices

#### – Copy Services

You can use the Copy Services selections of the DS Storage Manager if you chose Real-time during the installation of the DS Storage Manager and you purchased these optional features. A further requirement to using the Copy Services features is to apply the license activation keys. You need to obtain the Copy Services license activation keys (including the one for the use of PAVs) from the Disk Storage Feature Activation (DSFA) Web site.

#### ► Simulated Manager configuration

You can begin using the simulated mode immediately after logging on to the DS Storage Manager. However, if you want to make your configurations usable you need to obtain the license activation keys from the Disk Storage Feature Activation (DSFA) Web site.

You need to input these activation keys and save them as part of your configuration input. This application provides the ability to create or modify logical configurations when disconnected from the network. After creating the configuration, you can save it and then apply it to a storage unit attached to the network at a later time.

**Note:** The actual keys are *not* entered via the Simulated manager, only the capacities. The keys must be entered and applied via the Real-time manager.

A view of the fully expanded Simulated Manager menu choices is shown in Figure 10-14 on page 206.



Figure 10-14 The fully expanded Simulated manager menu choices

The following items should be considered as first steps in the use of either of these modes.

► **Log in**

Logging in to the DS Storage Manager requires that you provide your user name and password. This function is generally administered through your system administrator and by your company policies.

The preconfigured userid and password is as follows:

userid = admin

password = admin

► **Creating and defining the users and passwords**

Click **User administration**, as shown in Figure 10-15 on page 207, to add users and set passwords. The names and passwords must conform to the following standards:

- The user name can be up to 16 characters.
- Passwords must contain at least 5 alphabetic characters, and at least one special character, with an alphabetic character in the first and last positions. Passwords are limited to a total of 16 characters. The user name cannot be part of the password. This entry will appear as asterisks.

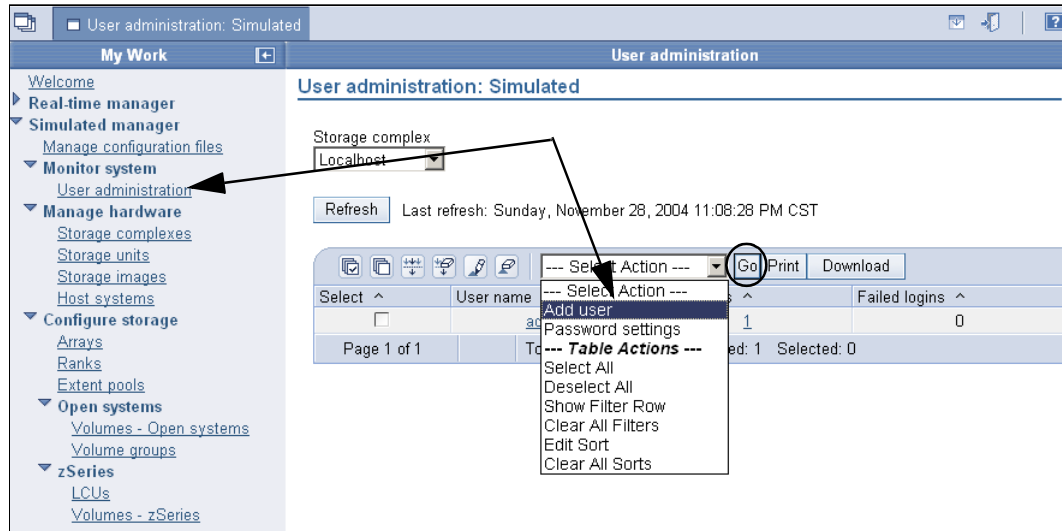


Figure 10-15 User administration panel

Click **Go** to advance to the panel shown in Figure 10-16.

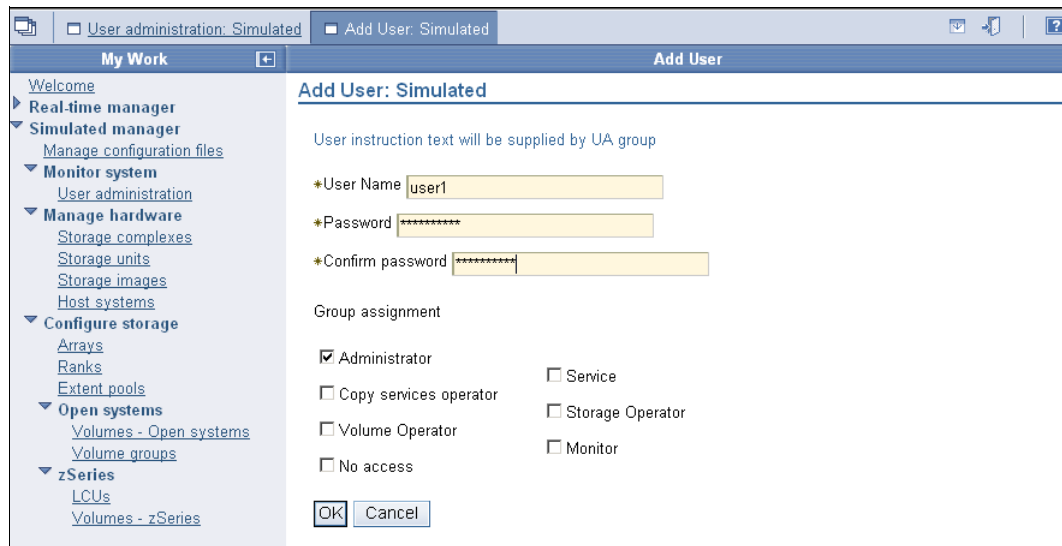


Figure 10-16 Add User panel

You can grant user access privileges from the panel shown in Figure 10-16.

► **Using the help panels (information center)**

The information center displays product and application information. The system provides a graphical user interface for browsing and searching online documentation.

The broad range of topics covered includes accessibility, Copy Services, device storage, host system attachments, concurrent code loads, input/output configuration programs, and volume storage.

To use the information center click the question mark (?) icon that appears in the top right corner of the screen.

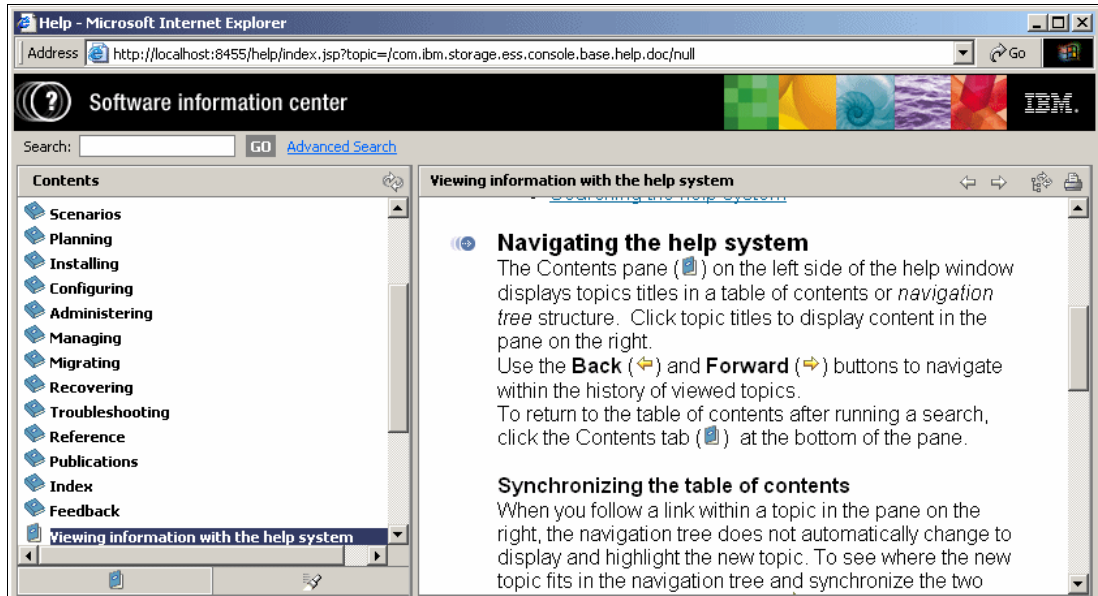


Figure 10-17 View of the information center

### 10.2.3 Navigating the GUI

Knowing what icons, radio buttons, and check boxes to click in the GUI will help you properly navigate your way through the configurator and successfully configure your storage.

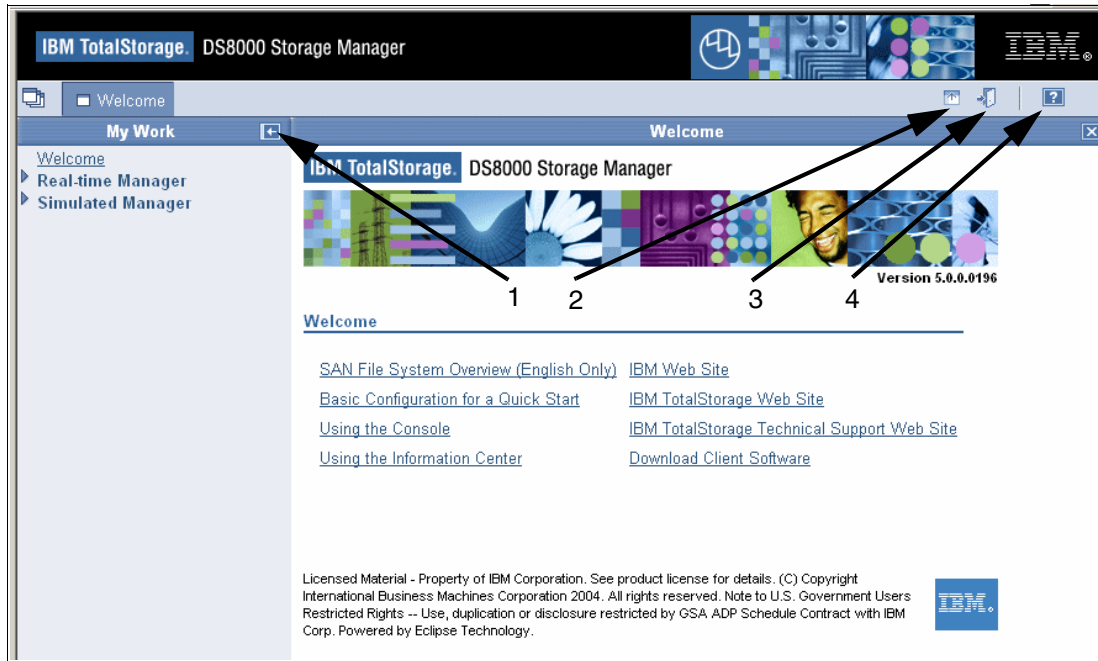


Figure 10-18 The DS Storage Manager Welcome panel

The picture icons that appear near the top of the Welcome screen and that are identified by number in Figure 10-18 have the following meanings:

1. Icon 1 as identified in the figure allows you to hide the My Work menu area to increase the space for displaying the main panel.

2. Icon 2 will hide the black banner across the top of the screen, again to increase the space to display the panel you are working on.
3. Icon 3 allows you to properly log out and exit the DS Storage Manager GUI.
4. Icon 4 accesses the Information Center. You get a help menu screen that prompts you for input on help topics.

Figure 10-19 shows how the screen looks if you expand the work area using icons 1 and 2 from the previous illustration. The icons highlighted in this figure can be used to reduce the work area and restore display of the My Work menu and product banner areas.

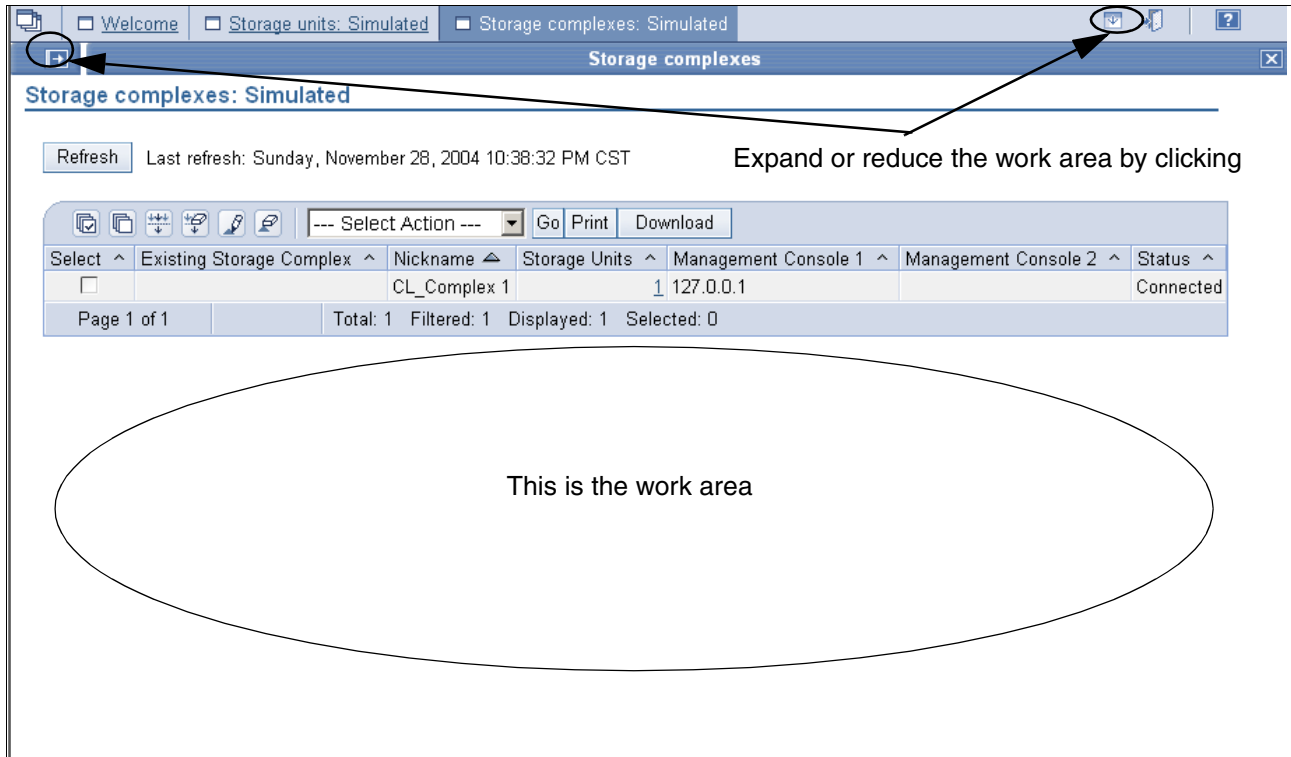


Figure 10-19 View of the storage complexes in the work area

To reduce the work area and work from the Real-time or Simulated Manager menu selection again, simply click the button shown on the left of Figure 10-19.

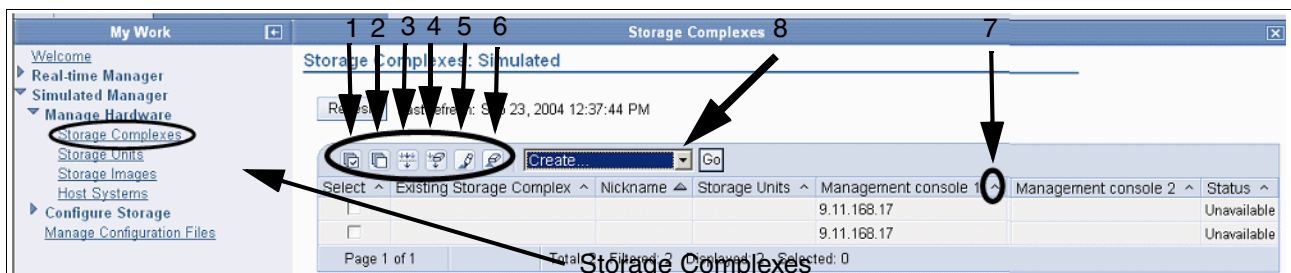


Figure 10-20 View of the Storage Complexes section

The buttons highlighted in Figure 10-20 have the following meanings:

- ▶ Boxes 1 through 6 are for selecting and filtering. Their specific meaning are:
  - 1 Select All**



- 2 Deselect All
- 3 Show Filter Row
- 4 Clear All Filters
- 5 Edit Sort
- 6 Clear All Sorts

- ▶ The caret button (number 7) is for a simple ascending/descending sort on a single column.
- ▶ Clicking the pull-down (number 8) results in the expanded action list shown in Figure 10-21. Near the center of the list you can access the same selection and filtering options mentioned previously.

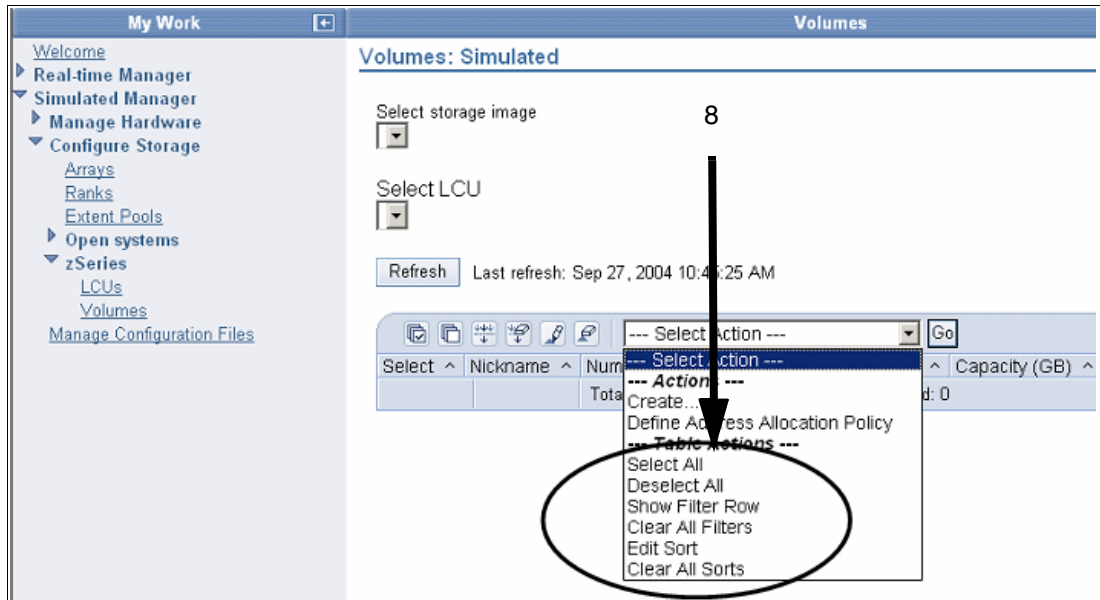


Figure 10-21 Storage unit view of the pull-down

**Radio buttons and check boxes**

Figure 10-22 illustrates the difference between radio buttons and check boxes.

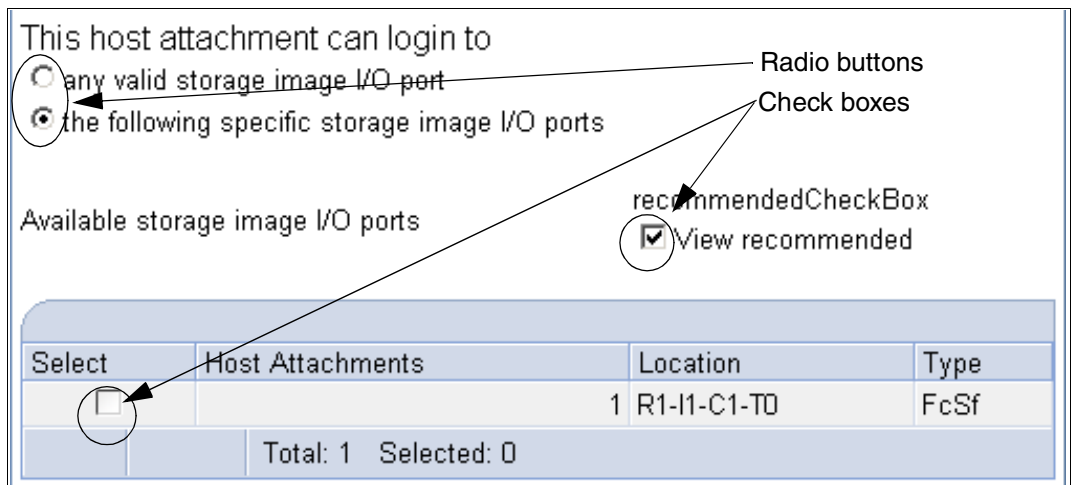


Figure 10-22 View of radio buttons and check boxes in the host attachment panel



In the example shown in Figure 10-22, the radio button is checked to allow specific host attachments for selected storage image I/O ports only. The check box has also been selected to show the recommended location view for the attachment.

## 10.3 The logical configuration process

We recommend that you configure your storage environment in the following order. This does not mean that you have to follow this guide exactly. You can get the same results by following a different order, as long as you define your storage complex, unit, and images first. This is only a suggestion.

### 10.3.1 Configuring a storage complex

To create the storage complex, expand the **Manage Hardware** (1) section, click **Storage Complexes** (2), click **Create** from the Select Action (3) pull-down and click **Go** (4), as shown in Figure 10-23. Follow the panel directions with each advancing panel.

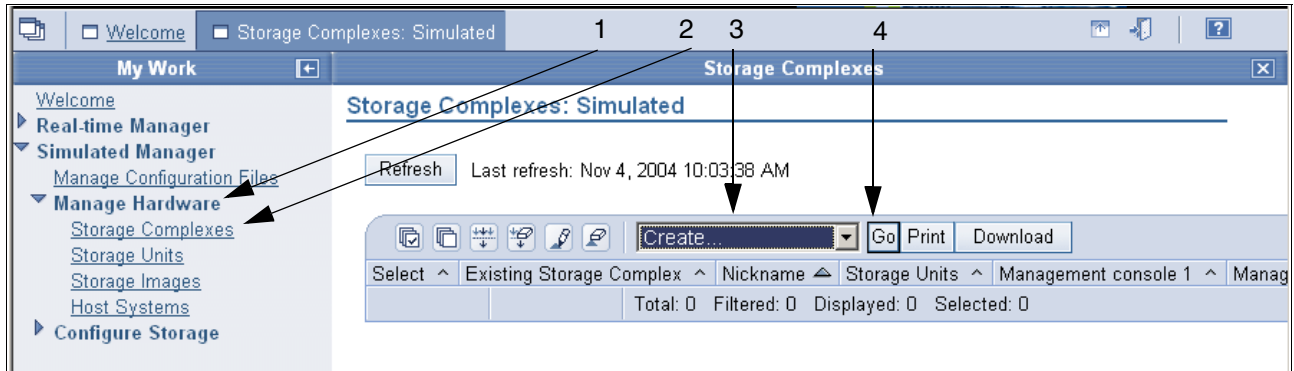


Figure 10-23 View of the Select Action pull-down menu with Create, selected

Under the Define Properties panel, type the storage complex **Nickname** and **Description**.

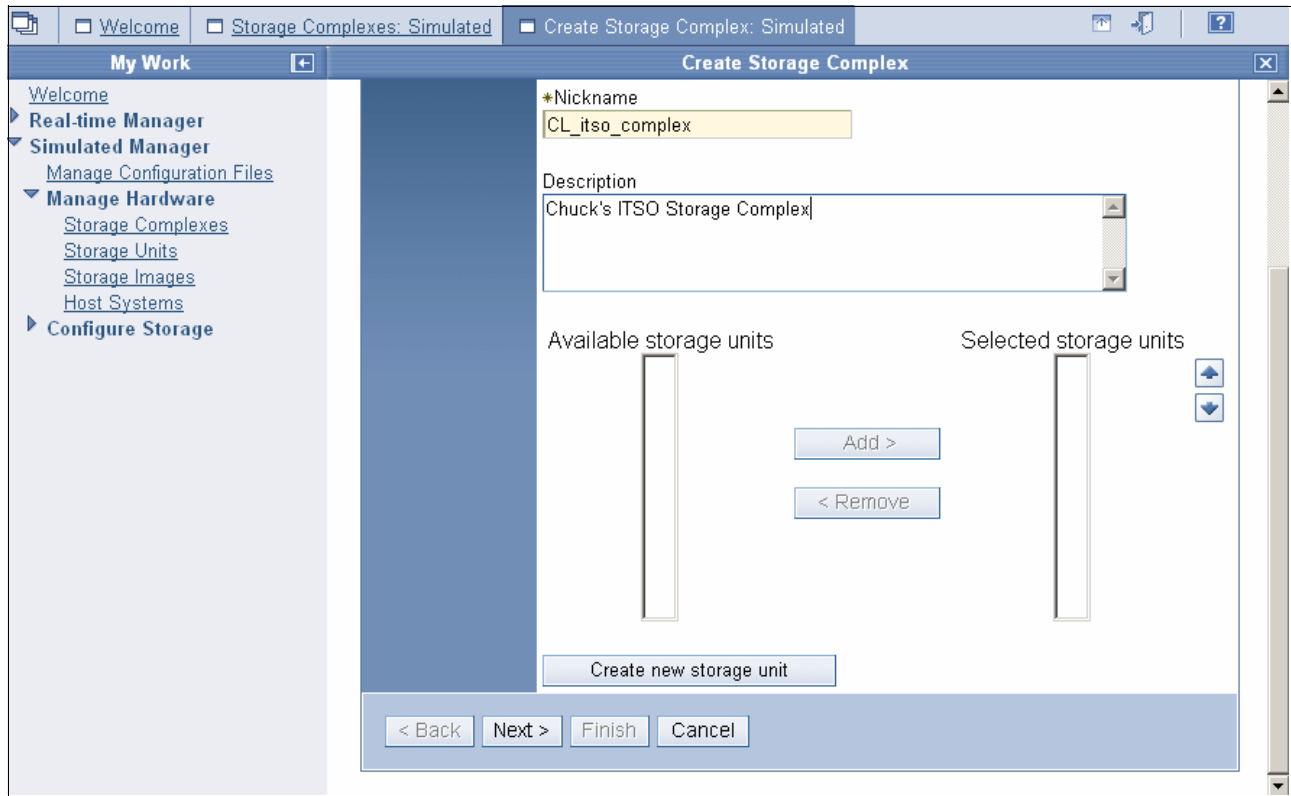


Figure 10-24 The Create Storage Complex panel, with the Nickname and Description defined

Do *not* click **Create new storage unit** at the bottom of the screen as shown in Figure 10-24. Click **Next**, then **Finish** in the verification step.

### 10.3.2 Configuring the storage unit

To create the storage unit, expand the **Manage Hardware** section, click **Storage Units (2)**, click **Create** from the Select Action pull-down, and click **Go**. Follow the panel directions with each advancing panel.

After clicking **Go**, you will see the General storage unit information panel shown in Figure 10-25 on page 213.

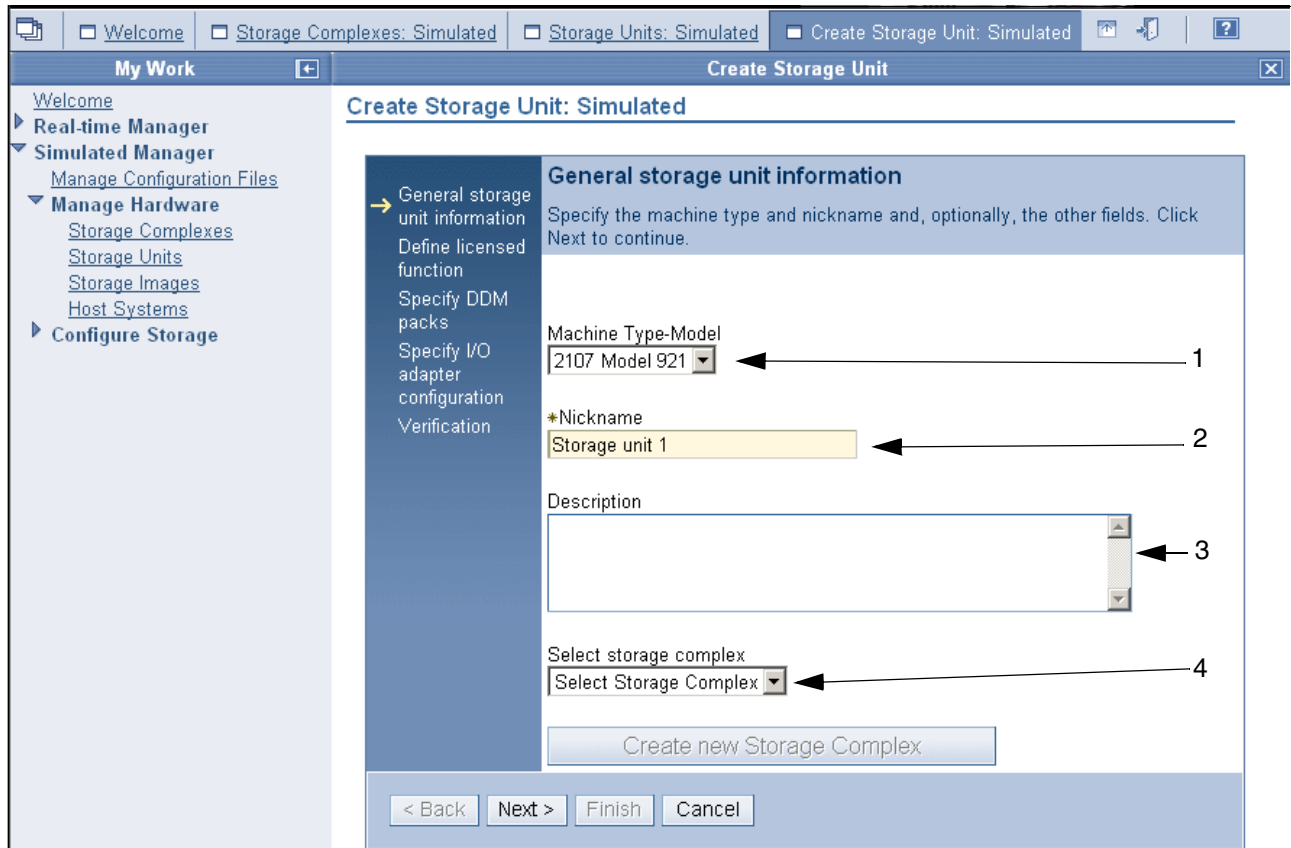


Figure 10-25 The General storage unit information panel

Fill in the required fields as shown in Figure 10-25, and choose the following:

1. Click the **Machine Type-Model** from the pull-down.
2. Fill in the **Nickname**.
3. Type in the **Description**.
4. Click the **Select Storage complex** from the pull-down, and choose the storage complex on which you wish to create the storage image.

Click **Next to** advance you to the Define licensed function panel, under the Create Storage Unit path, as shown in Figure 10-26 on page 214.

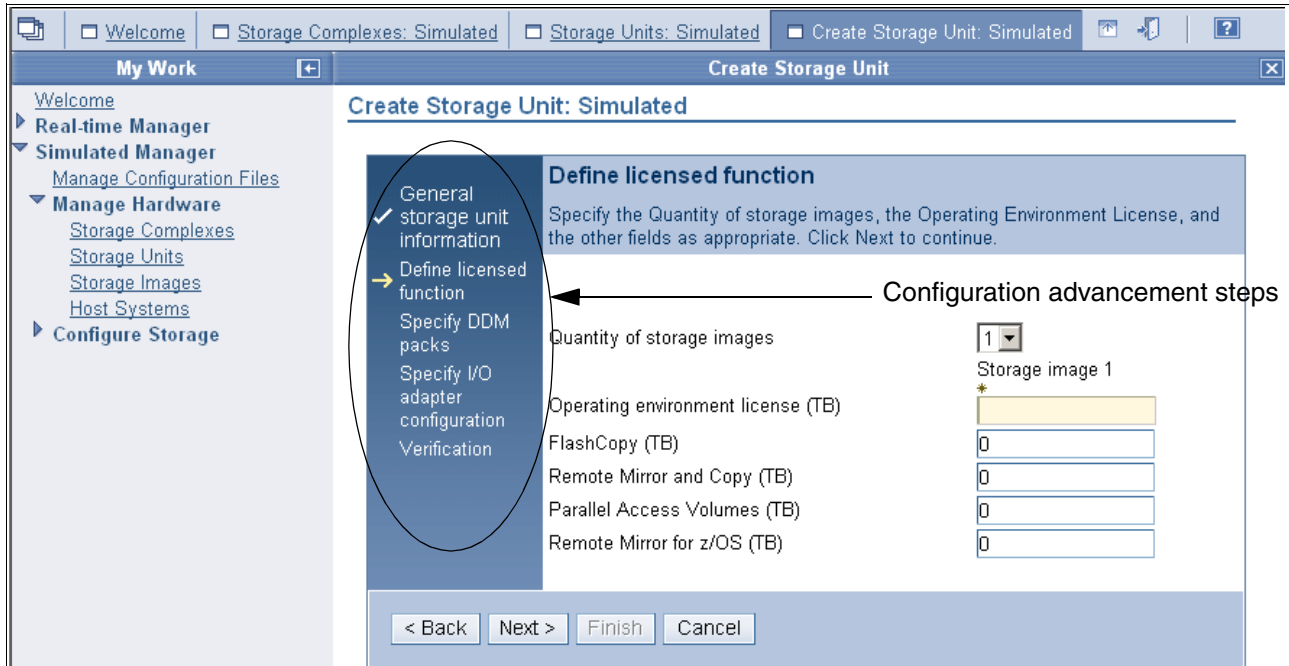


Figure 10-26 View of the Defined licensed function panel

Fill in the fields shown in Figure 10-26 with the following information:

- ▶ The quantity of images
- ▶ The number of licensed TBs for the Operating environment
- ▶ The quantity of storage covered by the FlashCopy License, in TB.
- ▶ The amount of disk for Remote Mirror and Copy in TB
- ▶ The amount of TB for Parallel Access Volumes (PAV)
- ▶ The amount in TB for Remote Mirror for z/OS

Click **Next** to advance to the Specify DDM packs panel shown in Figure 10-27 on page 215. Fill in the proper information for your specific environment.

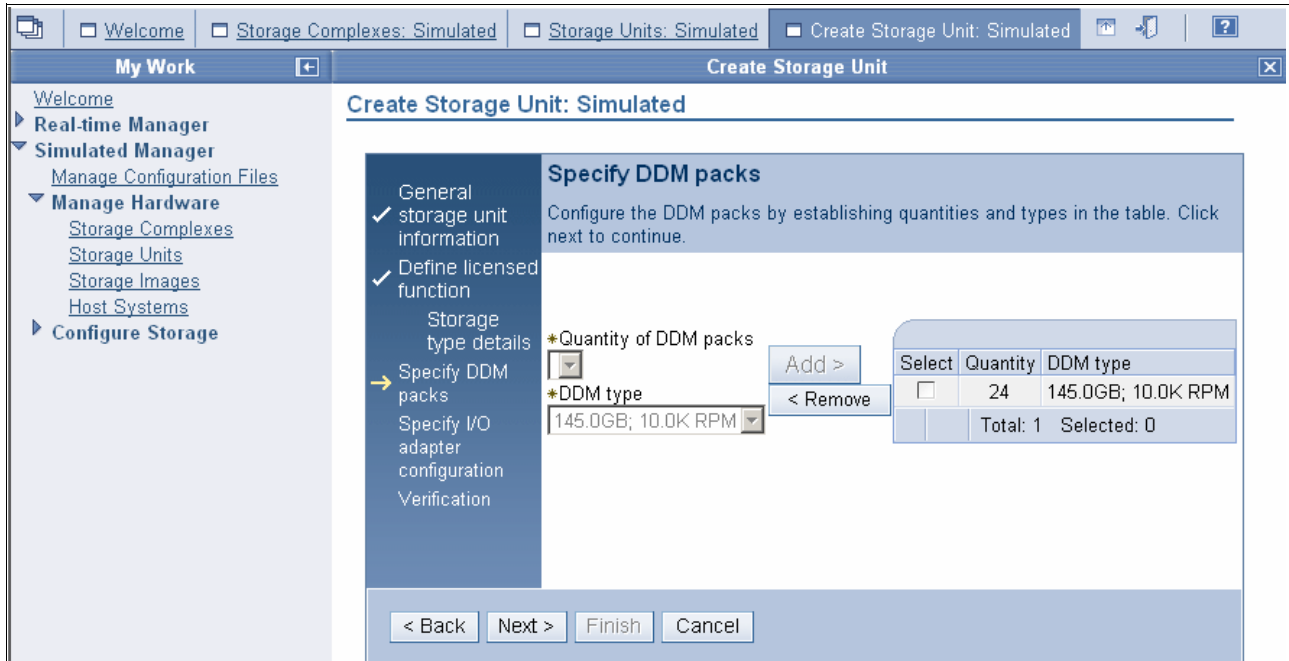


Figure 10-27 View of Specify DDM packs panel, with the Quantity and DDM type added

5. Click **Add** and **Next** to advance to the Specify I/O adapter configuration panel shown in Figure 10-28.

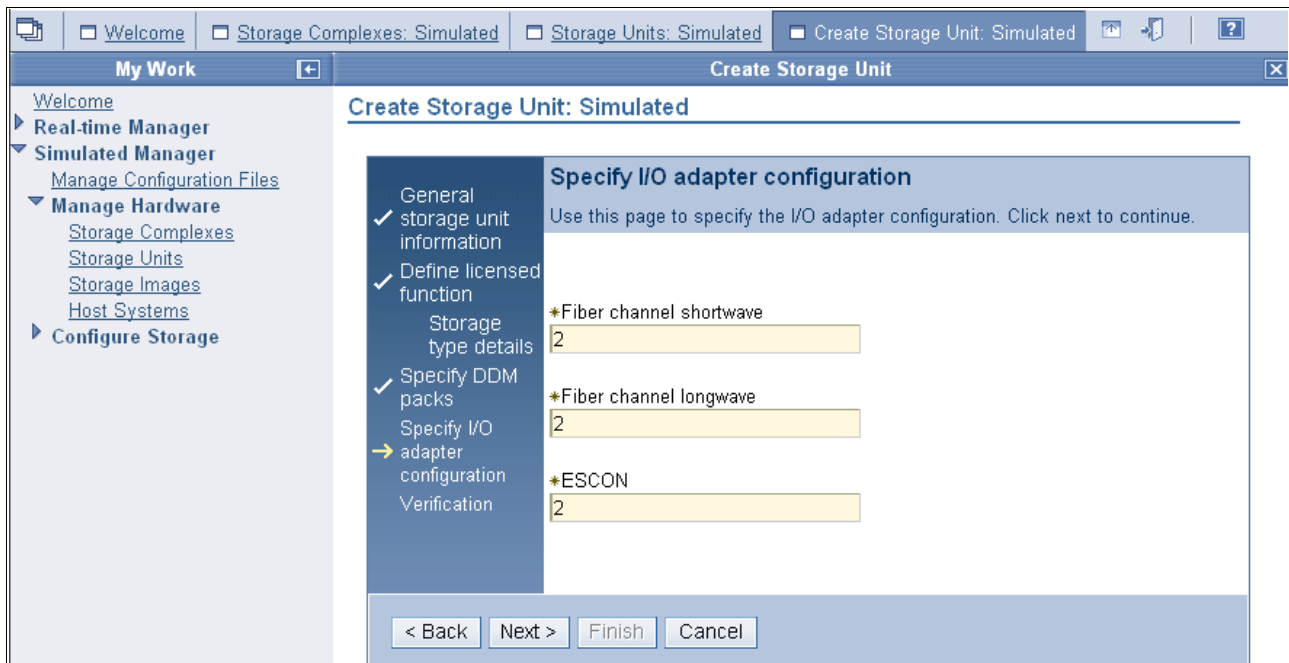


Figure 10-28 Specify I/O adapter configuration panel

Enter the appropriate information and click **Next**.

The storage facility image will be created automatically, by default, as you create the storage unit.

### 10.3.3 Configuring the logical host systems

To create a logical host for the storage unit that you just created, click **Host Systems** as shown in Figure 10-29. You may want to expand the work area.

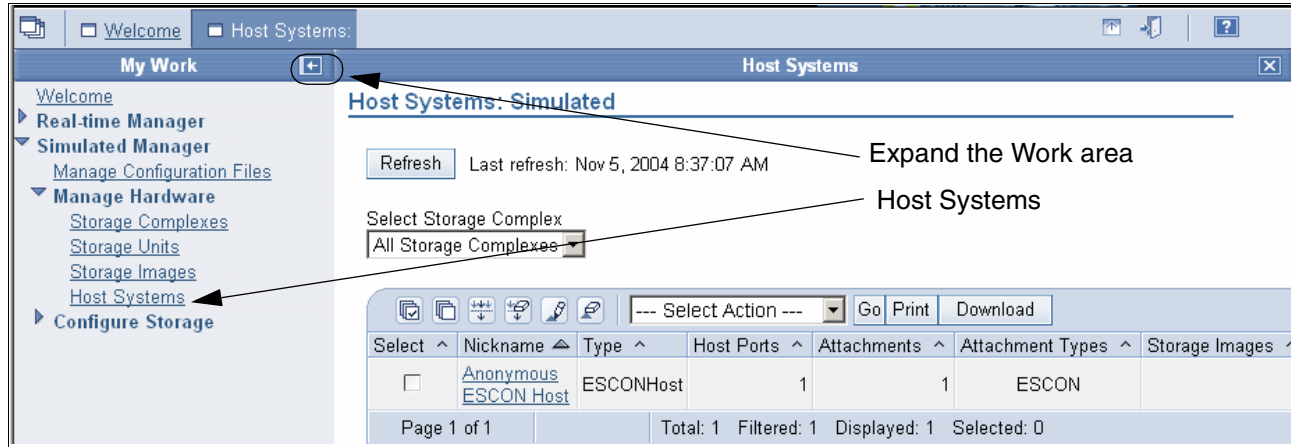


Figure 10-29 Create host systems, screen 1

You can expand the view by clicking the left arrow in the My Work area as shown in Figure 10-29; the expanded view is shown in Figure 10-30.

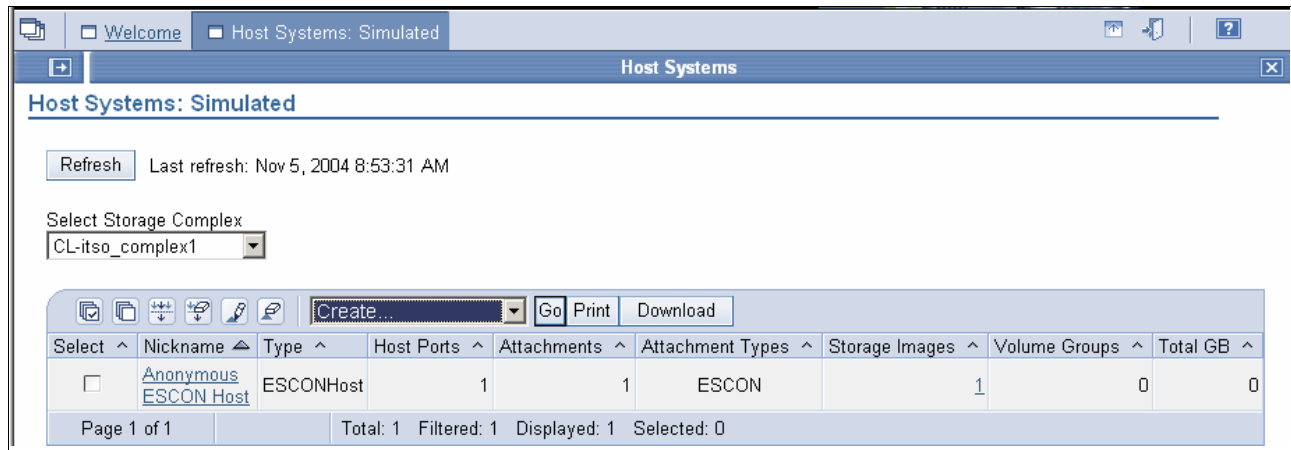


Figure 10-30 View of Host Systems panel, with the **Go** button selected

Click the Select Storage Complex action pull-down, highlight the storage complex you wish to configure, then click **Create** and **Go**. The screen will advance to the General host information panel shown in Figure 10-31 on page 217.

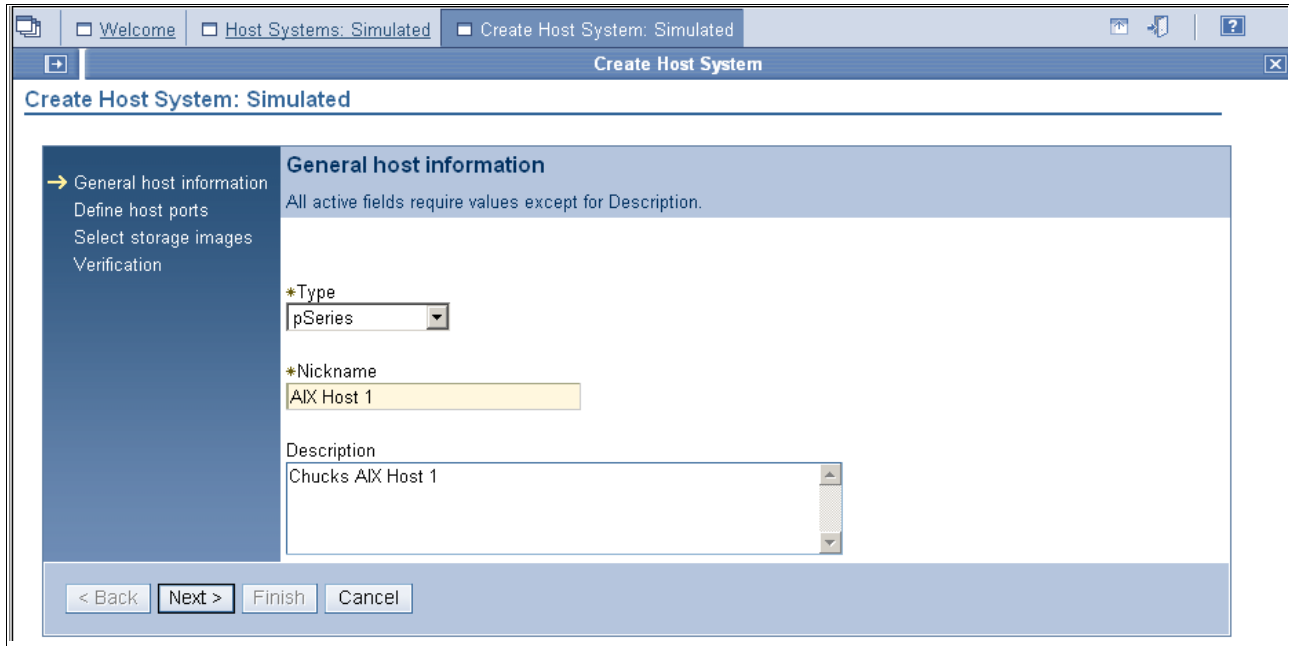


Figure 10-31 View of the General host information panel

Click **Next** to advance to the Define host ports panel shown in Figure 10-32.

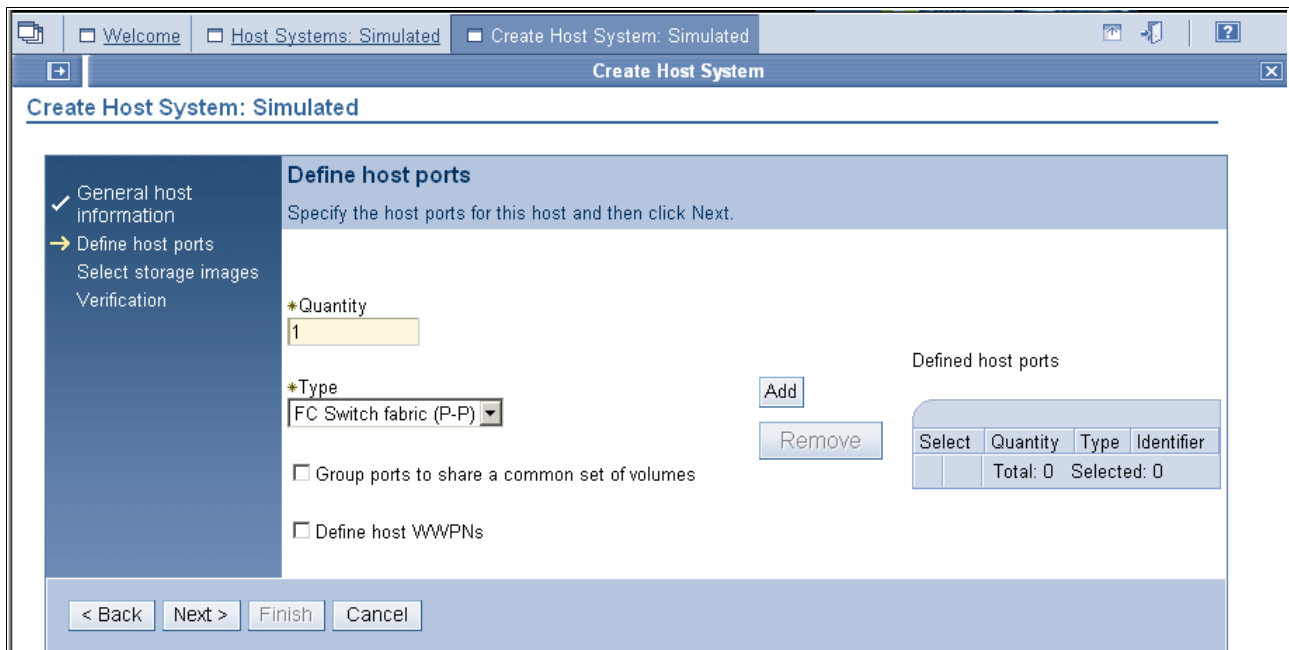


Figure 10-32 View of Define host port panel

Enter the appropriate information on the Define host ports panel shown in Figure 10-32.

**Note:** Selecting “Group ports to share a common set of volumes” will group the host ports together into one attachment. Each host port will require a WWPN to be entered; now, if you are using the Real-time Manager; or later, if you are using the Simulated Manager.

Click **Add**, and the Define host ports panel will be updated with the new information as shown in Figure 10-33.

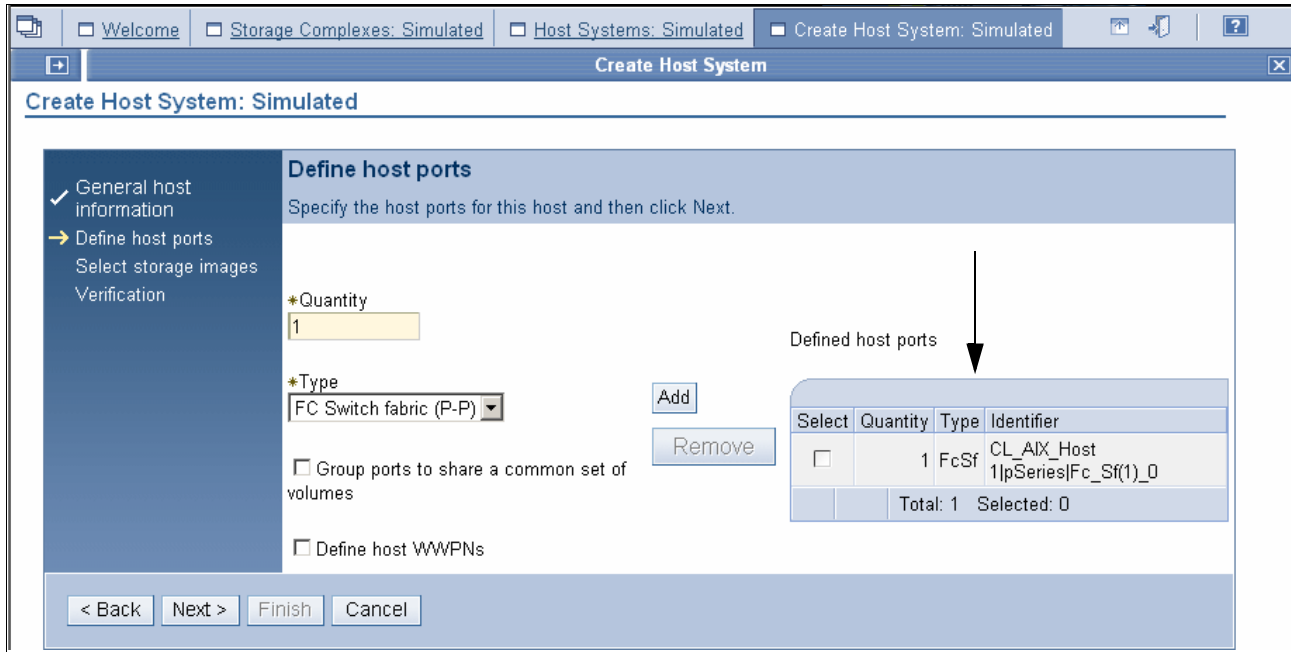


Figure 10-33 Define host ports panel, with updated host information

Click **Next**, and the screen will advance to the Select storage images panel shown in Figure 10-34.

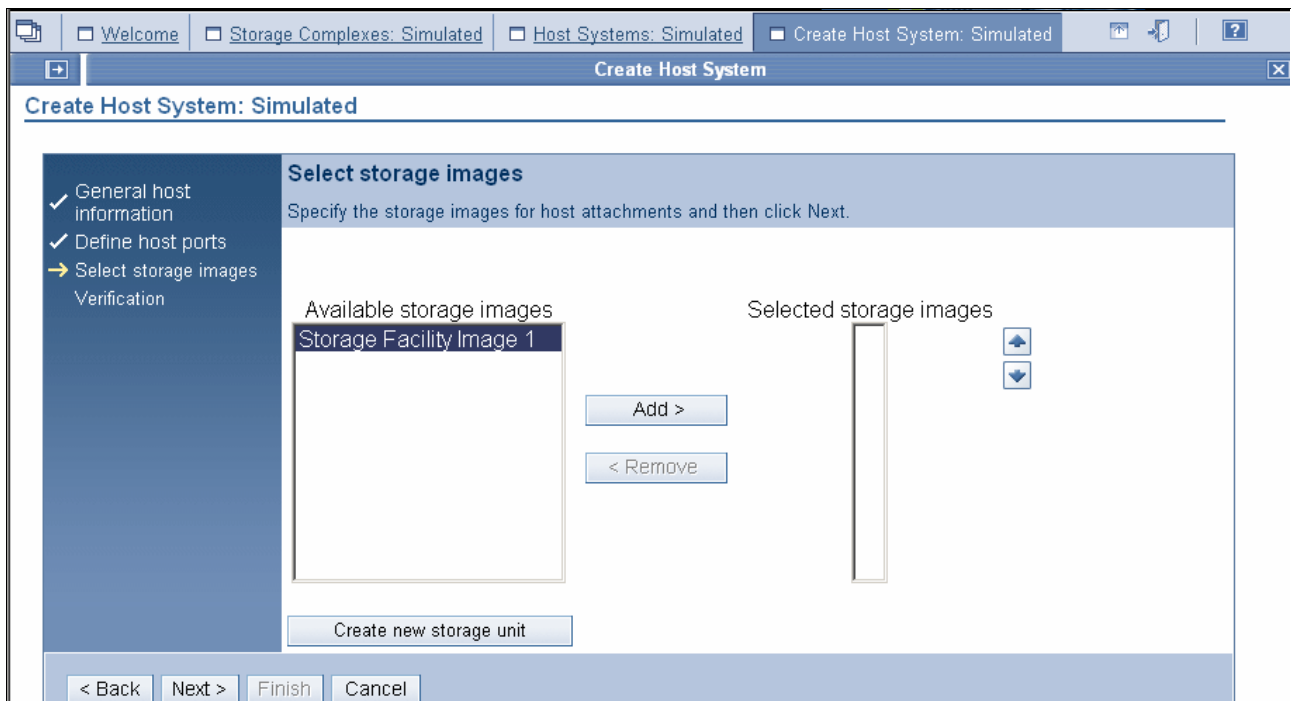


Figure 10-34 Select storage images panel

Highlight the Available storage images that you wish, click **Add** and **Next**.



The screen will advance to the Specify storage image parameters section shown in Figure 10-35.

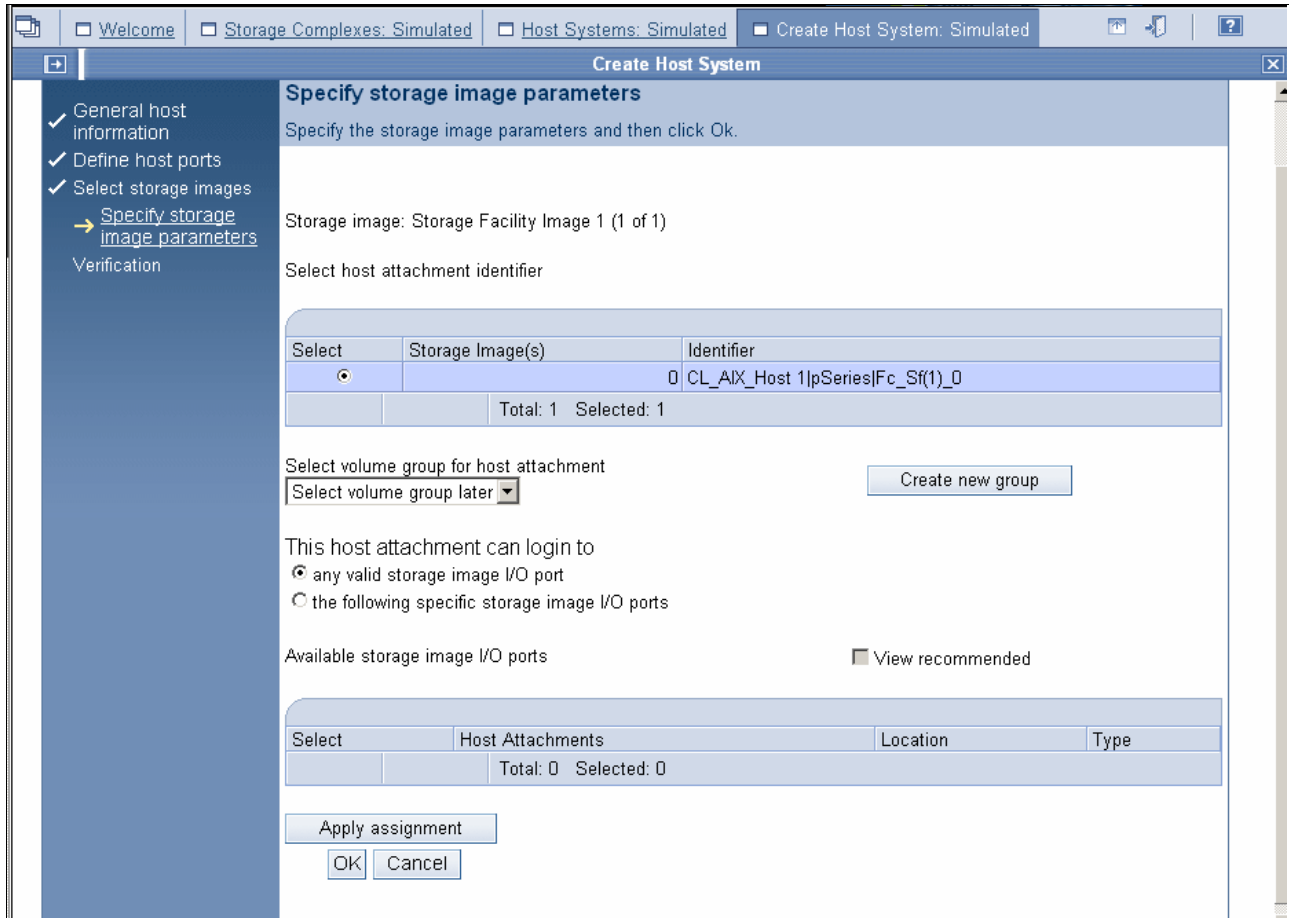


Figure 10-35 Specify storage image parameters panel

Make the following entries and selections on the Specify storage image parameters panel:

1. Click the Select volume group for host attachment pull-down and highlight **Select volume group later**.
2. Click **any valid storage image I/O port** under the “This host attachment can login to” field.
3. Click **Apply assignment** and **OK**.
4. Verify and click **Finish**.

### 10.3.4 Creating arrays from array sites

Under **Configure Storage**, click **Arrays**. The screen will advance to the Create Array: Simulated panel. Click the Storage complex pull-down, highlight the storage complex you wish to configure, click **Create** and **Go** (the panel is not pictured here). The screen will advance to the Definition method panel shown in Figure 10-36 on page 220.

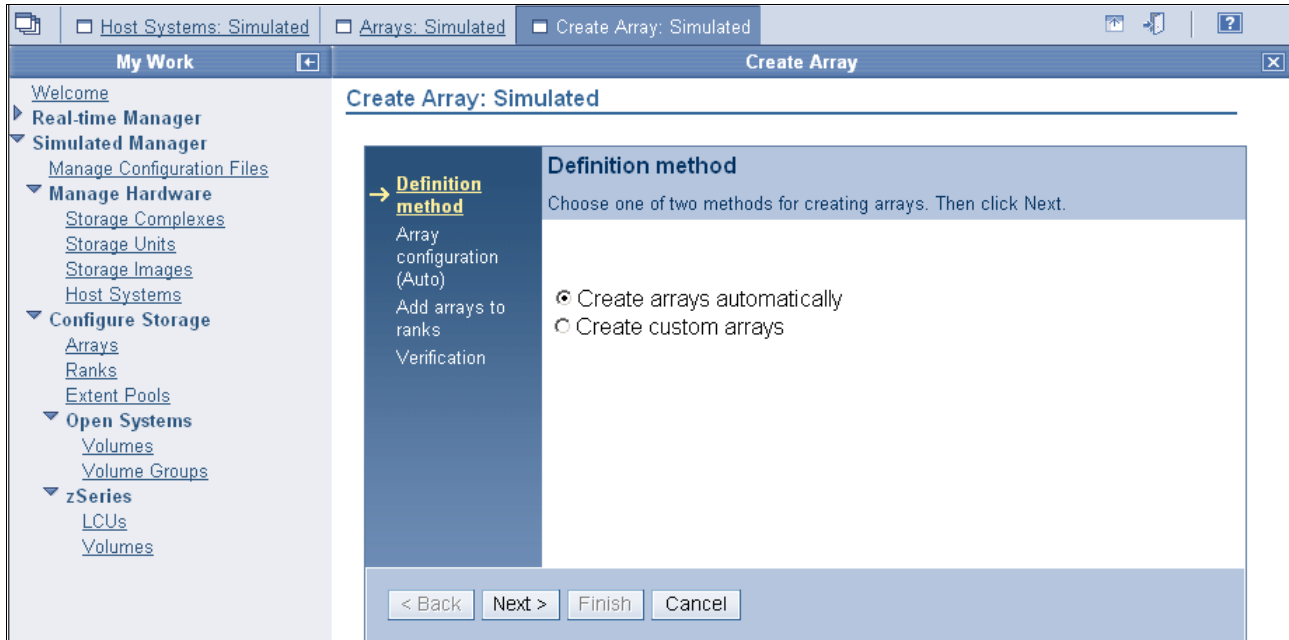


Figure 10-36 The Definition method panel

From the Definition method panel, if you choose **Create arrays automatically**, the system will automatically take all the space from the array site and place it into an array. Physical disks from any array site could be placed, through a predetermined algorithm, into the array. It is at this point that you create the RAID-5 or RAID-10 format and striping in the array being created.

If you choose to create arrays automatically, the screen will advance to the Array configuration (Auto) panel shown in Figure 10-37.

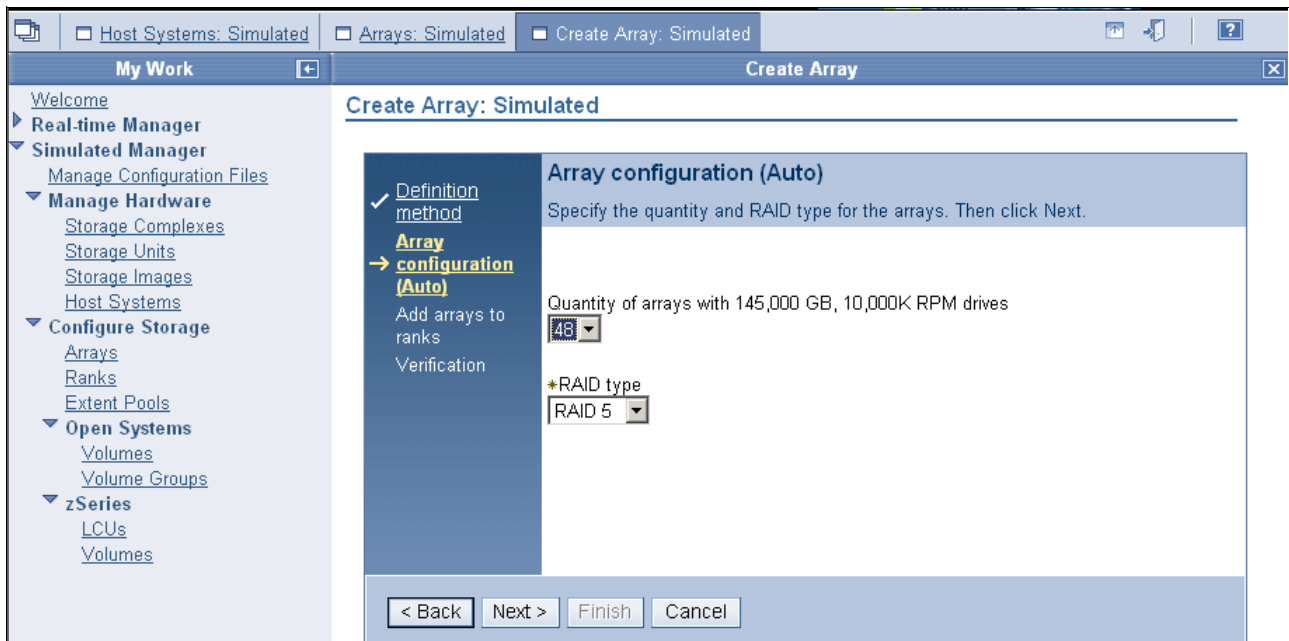


Figure 10-37 The Array configuration (Auto) panel

Enter the appropriate information for the quantity of the arrays and the RAID type.

Click **Next** to advance to the Add arrays to ranks panel shown in Figure 10-38.

If you click the check box next to **Add these arrays to ranks** you will not have to configure the ranks separately at a later time. The ranks can be either **FB** or **CKD**; this is specified in the Storage type pull-down shown in Figure 10-38.

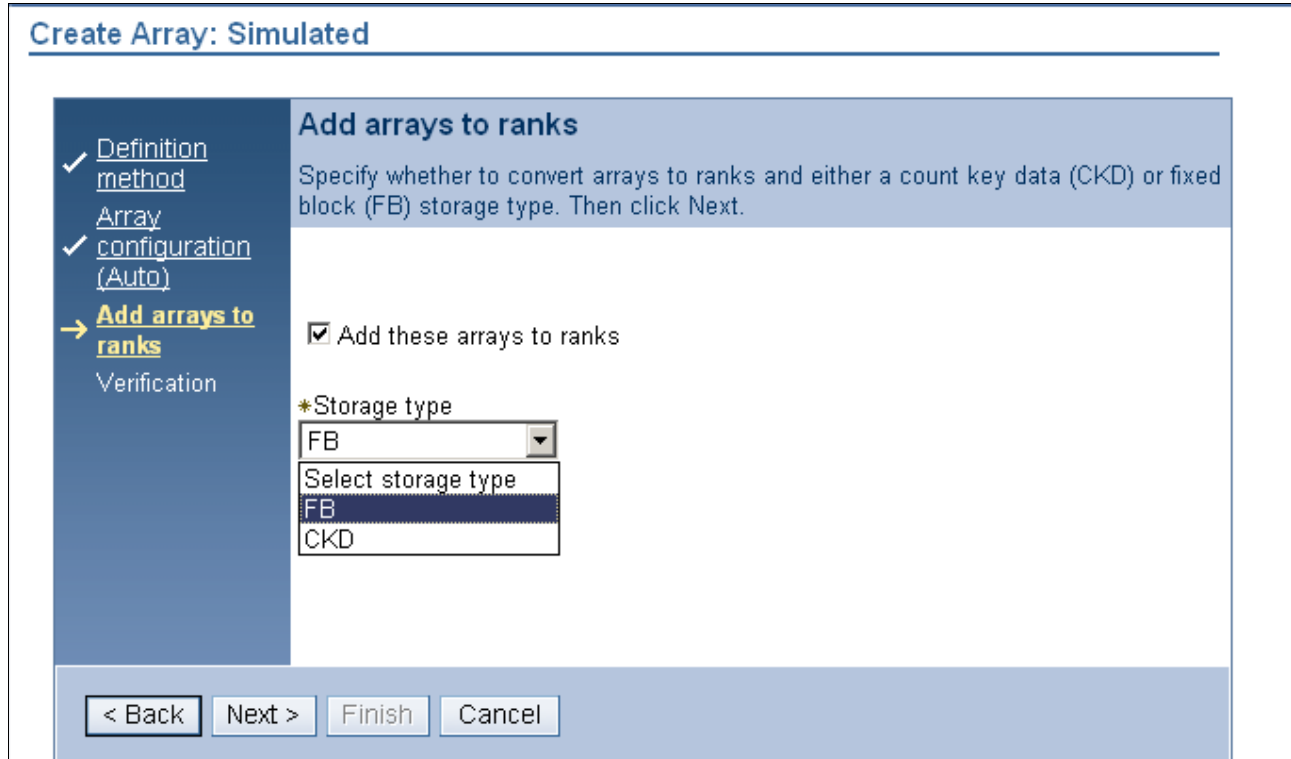


Figure 10-38 The Add arrays to ranks panel with FB selected

Click **Next** and **Finish** to configure the arrays and ranks in one step.

### 10.3.5 Creating extent pools

To create extent pools, expand the **Configure Storage** section, click **Extent pools**, click **Create** from the Select Action pull-down and click **Go**. Follow the panel directions with each advancing window.

You can select either the **Custom extent pool** or the **Create extent pool automatically based on storage requirements** radio button as shown in Figure 10-39 on page 222.

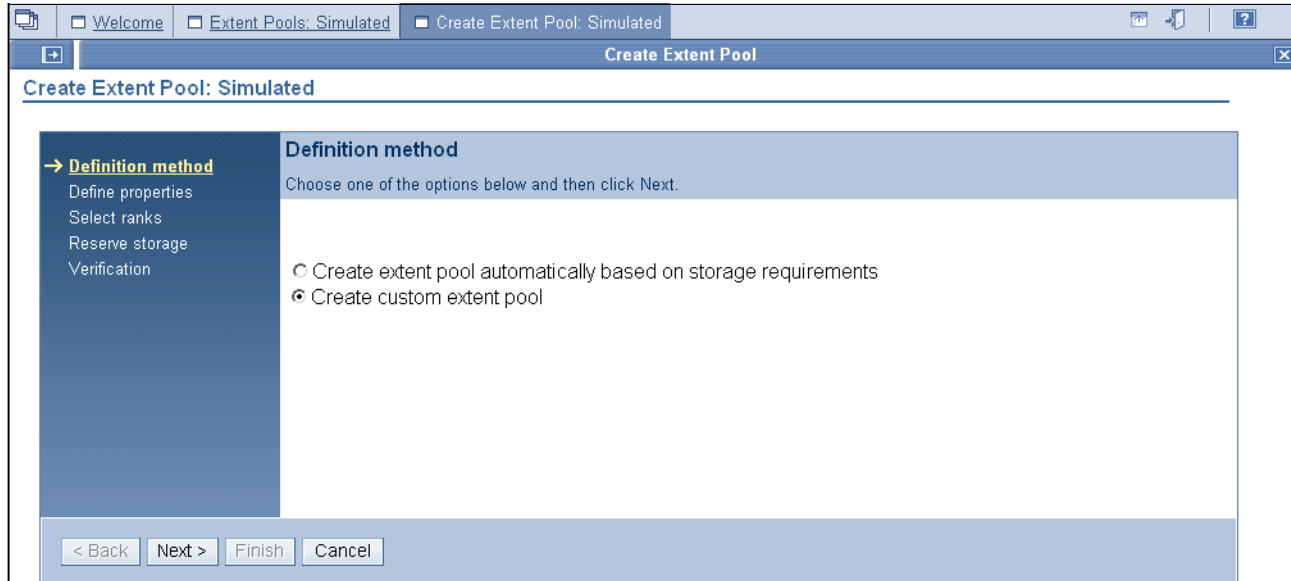


Figure 10-39 The Definition method panel

The extent pools are given either a server 0 or server 1 affinity at this point, as shown in Figure 10-40.

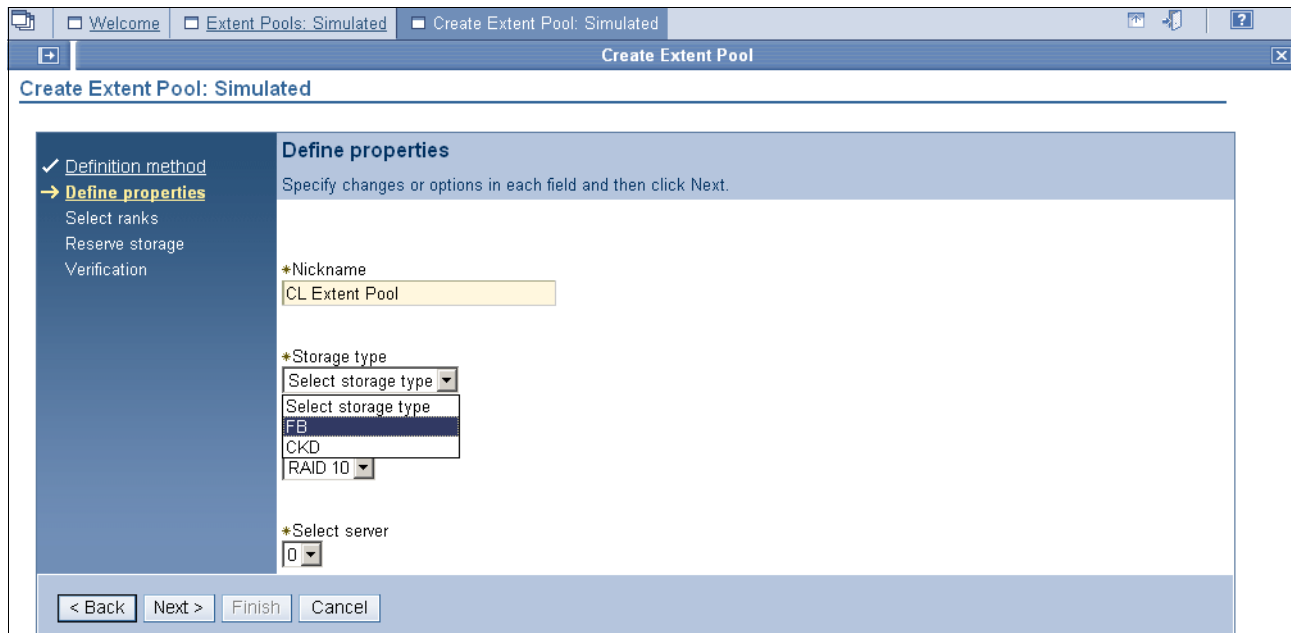


Figure 10-40 The Define properties panel

Click **Next** and **Finish**.

### 10.3.6 Creating FB volumes from extents

Under the **Simulated Manager**, expand the **Open Systems** section and click **Volumes**.

Click **Create** from the Select Action pull-down and click **Go**. Follow the panel directions with each advancing window.

Choose the extent pool from which you wish to configure the volumes, as shown in Figure 10-41.

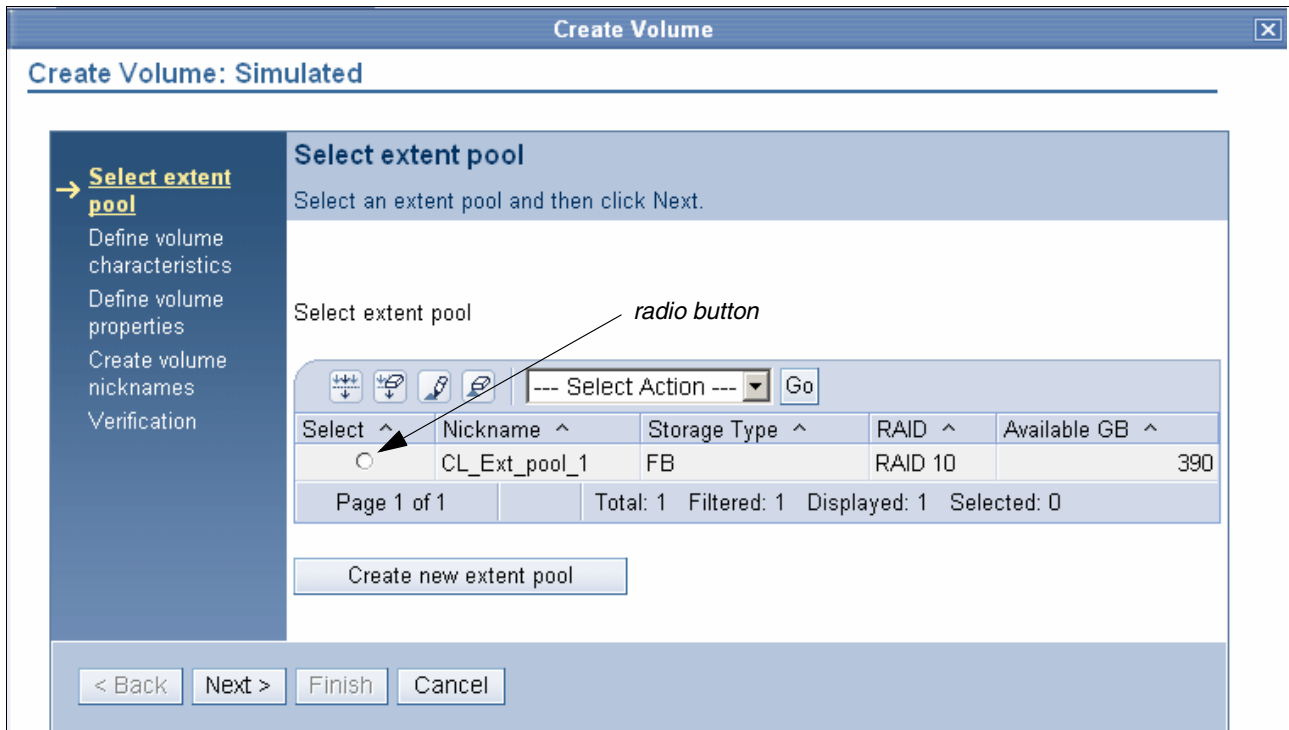


Figure 10-41 The Select extent pool panel

Determine the quantity and size of the volumes. Use the calculators to determine the max size versus quantity, as shown in Figure 10-42.

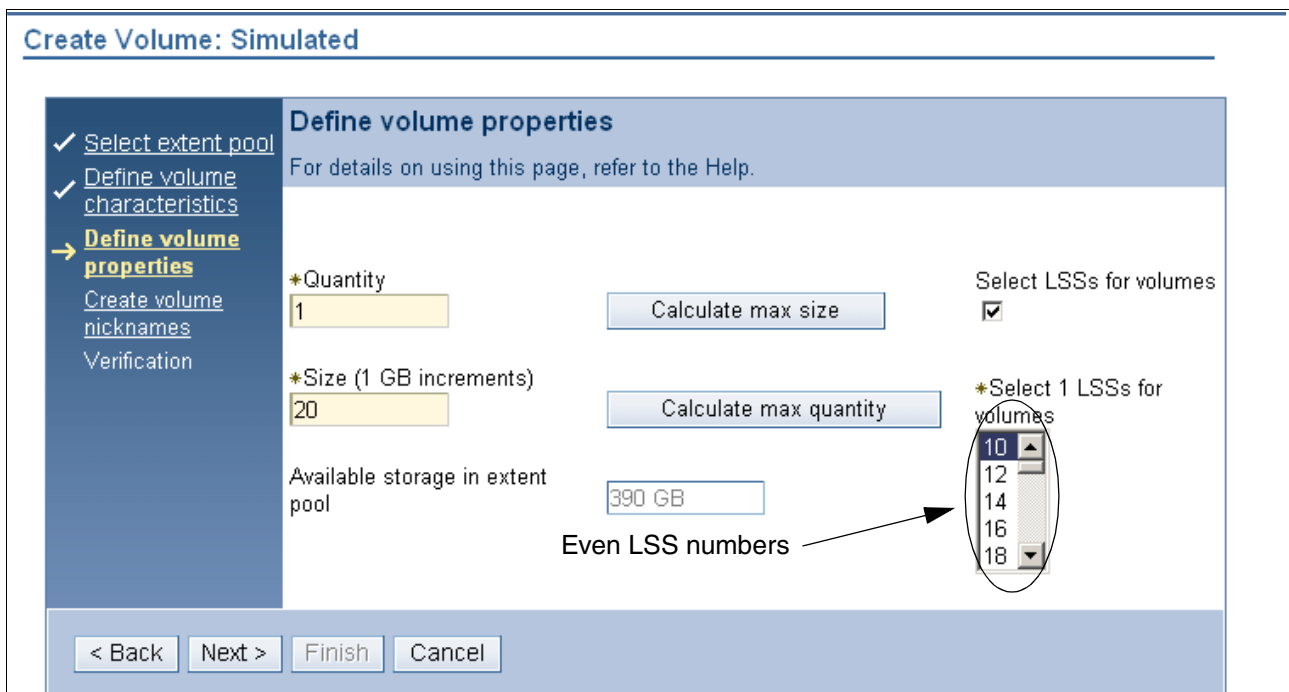


Figure 10-42 The Define volume properties panel

It is here that the volume will take on the LSS numbering affinity.

**Note:** Since server 0 was selected for the extent pool, only even LSS numbers are selectable, as shown in Figure 10-42.

You can give the volume a unique name and number as shown in Figure 10-43. This can be helpful for managing the volumes.

**Create Volume: Simulated**

✓ Select extent pool  
✓ Define volume characteristics  
✓ Define volume properties  
→ **Create volume nicknames**  
Verification

**Create volume nicknames**

Check the box for "Generate a sequence of nicknames based on the following" to enter data in the following fields.

Quantity of volumes

Generate a sequence of nicknames based on the following

Prefix (e.g. Vol)  Suffix (e.g. 0001)

< Back Next > Finish Cancel

Figure 10-43 The Create volume nicknames panel

Click **Next** and **Finish** to end the process of creating the volumes.

### 10.3.7 Creating volume groups

Under **Simulated Manager, Open Systems**, perform the following steps to configure the volume groups:

1. Click **Volume Groups**.
2. Click **Create**.
3. Click **Go**.

Fill in the appropriate information as directed in the panel menu. You will have to specify the host type as shown in Figure 10-44 on page 225.

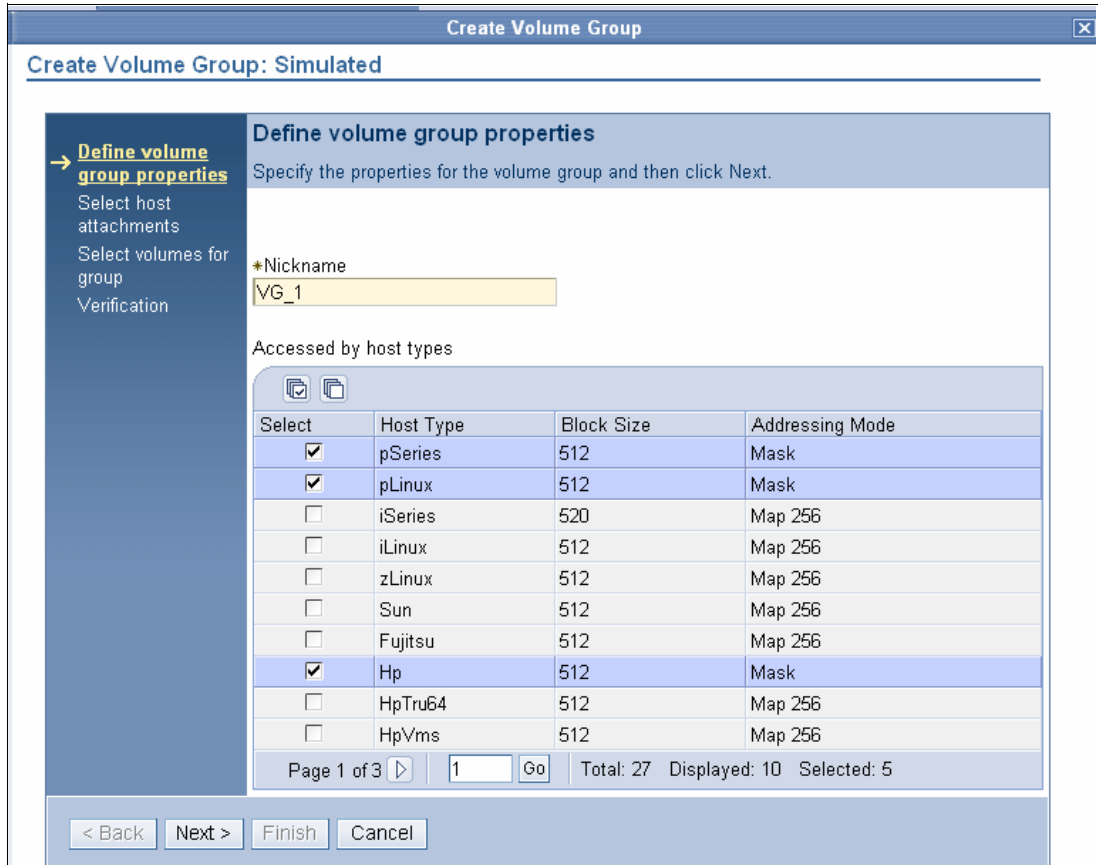


Figure 10-44 The Define volume group properties filled out

4. Select the host attachment you wish to associate the volume group with. See Figure 10-45.

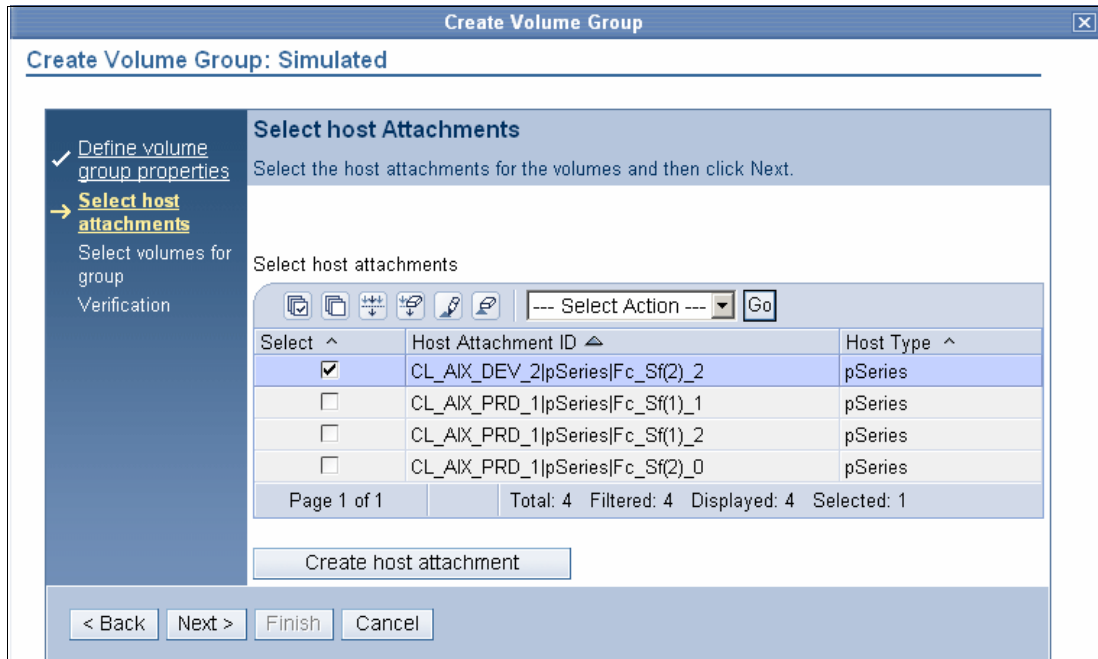


Figure 10-45 The Select host Attachments panel with an attachment selected

5. Select volumes on the Select volumes for group panel shown in Figure 10-46.

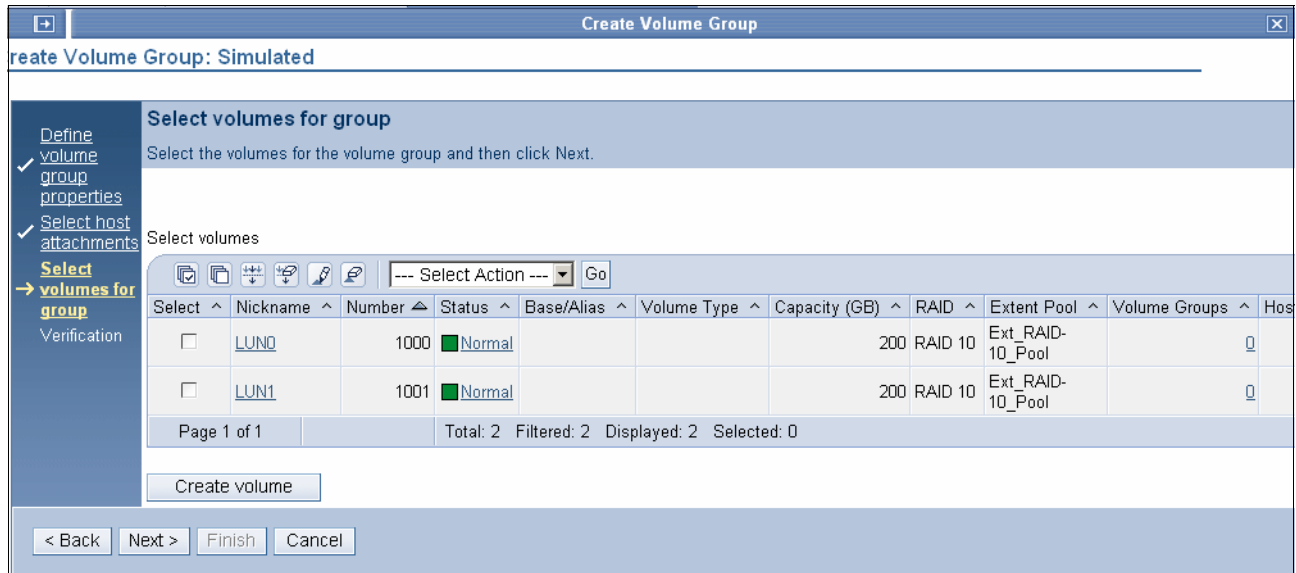


Figure 10-46 The Select volumes for group panel

Click **Next** and **Finish**.

### 10.3.8 Assigning LUNs to the hosts

Under **Simulated Manager**, perform the following steps to configure the volumes:

1. Click **Volumes**.
2. Select the check box next to the volume that you want to assign.
3. Click the Select Action pull-down, and highlight **Add To Volume Group**.
4. Click **Go**.
5. Click the check box next to the desired volume group and click **Apply**.
6. Click **OK**.

You can verify that the volume is now assigned to the desired host volume group by performing the following steps:

1. Click **Host Systems**.
2. Click the check box next to the host nickname.
3. Click the Select Action pull-down, and highlight **Properties**.
4. Click **Go**.

The properties box will be displayed.

### 10.3.9 Deleting LUNs and recovering space in the extent pool

Under **Simulated Manager**, perform the following steps to delete the volumes:

1. Click **Volumes**.
2. Click the check box next to the targeted volume you want to delete.
3. Click on the Select Action pull-down, and highlight **Delete**.



4. Click **Go**.
5. Click **OK**.

### 10.3.10 Creating CKD LCUs

Under **Simulated Manager, zSeries**, perform the following steps:

1. Click **LCUs**.
2. Click the Select Action pull-down and highlight **Create**.
3. Click **Go**.
4. Click the check box next to the LCU ID you wish to create.
5. Click **Next**.
6. In the panel returned, make the following entries:
  - a. Enter the desired **SSID**
  - b. Select the **LCU type**
  - c. Accept the defaults on the other input boxes, unless you are using Copy Services.
7. Click **Next**.
8. Click **Finish**.

### 10.3.11 Creating CKD volumes

Under **Simulated Manager, zSeries**, perform the following steps:

1. Click **Volumes** → **zSeries**.
2. Click the Select Action pull-down, and highlight **Create**.
3. Click **Go**.
4. In the Select Extent pool panel, click the radio button next to the targeted extent pool you want to configure the volume from.
5. Click **Next**.
6. Click the Volume type pull-down and select the Volume type desired.
7. Highlight the LCU number or work with all available LCUs.
8. Click **Next**.
9. In the Define base properties panel do the following:
  - a. Select the radio button next to the addressing policy.
  - b. Enter the quantity of base volumes.
  - c. Enter the base start address.
  - d. Click **Next**.
10. In the next panel returned, do the following:
  - a. Enter the volume Nickname.
  - b. Enter the volume prefix.
  - c. Enter the volume suffix.
11. Click **Next**.

12. Under the Define alias assignments panel, do the following:
  - a. Click the check box next to the LCU number.
  - b. Enter the starting address.
  - c. Specify the order as Ascending or Descending.
  - d. Select the number of aliases per volume, for example, 1 alias to every 4 base volumes, or 2 aliases to every 1 base volume.
  - e. Click **Next**.
13. On the Verification panel, click **Finish**.

### 10.3.12 Displaying the storage unit WWNN

To display the WWNN of the storage unit:

1. Click **Real-time Manager** as shown in Figure 10-47.

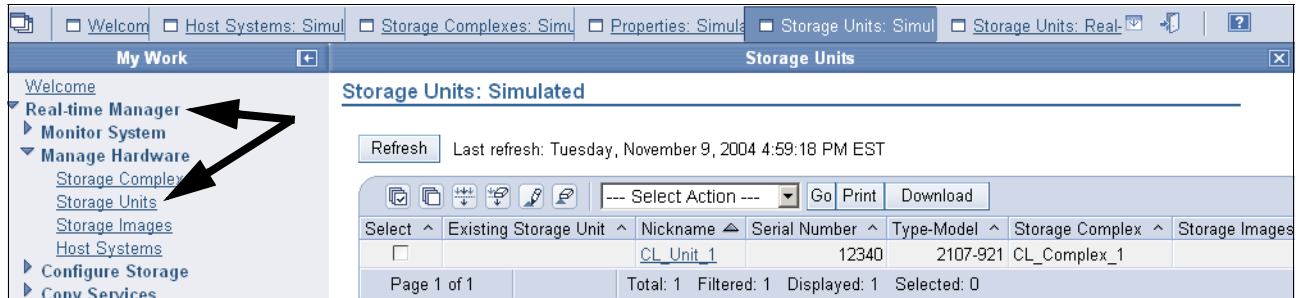


Figure 10-47 The Real-time Manager panel

2. Click **Storage Units**.
3. Select the radio button beside the storage unit name as shown in Figure 10-48.

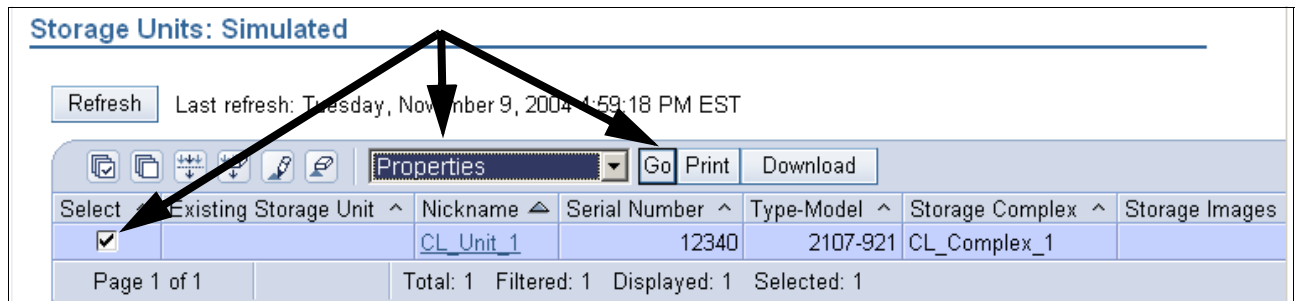


Figure 10-48 View of the storage unit with the Radio button selected and the Properties selected

4. Select **Properties** from the pull-down list.
5. Click **Go**. The General panel shown in Figure 10-49 on page 229 will be returned.

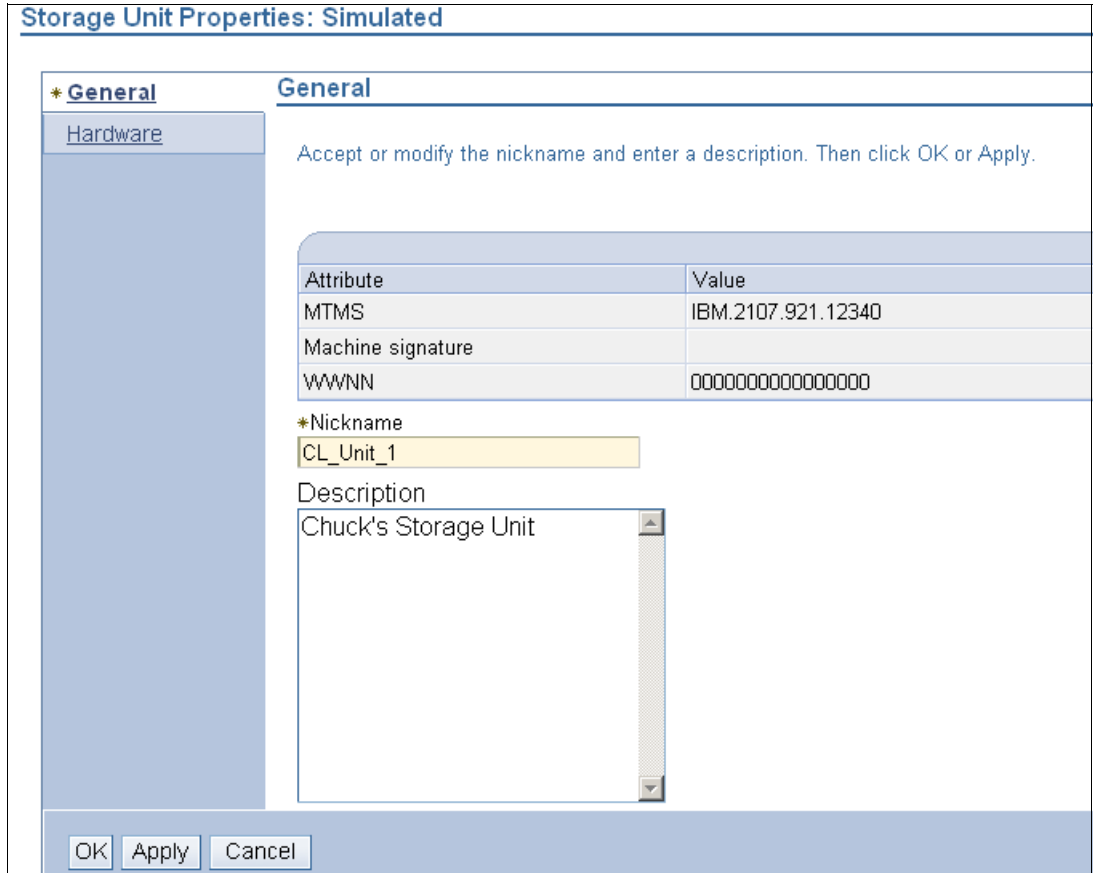


Figure 10-49 View of the WWNN in the General panel

## 10.4 Summary

In this chapter we have discussed the configuration hierarchy, terminology, and concepts. We have recommended an order and methodology for configuring the DS8000 storage server. We have included some logical configuration steps and examples and explained how to navigate the GUI.





## DS CLI

This chapter provides an introduction to the DS Command-Line Interface (DS CLI), which can be used to configure and maintain the DS6000 and DS8000 series. It also describes how the DS CLI can be used to manage Copy Services relationships.

In this chapter we describe:

- ▶ Functionality
- ▶ Supported environments
- ▶ Installation methods
- ▶ Command flow
- ▶ User security
- ▶ Usage concepts
- ▶ Usage examples
- ▶ Mixed device environments and migration
- ▶ DS CLI migration example

## 11.1 Introduction

The IBM TotalStorage DS Command-Line Interface (the DS CLI) is a software package that allows open systems hosts to invoke and manage Copy Services functions as well as to configure and manage all storage units in a storage complex. The DS CLI is a full-function command set. In addition to the DS6000 and DS8000, the DS CLI can also be used to manage Copy Services on the ESS 750s and 800s, provided they are on ESS code versions 2.4.2.x and above. All references in this chapter to the ESS 800 also apply to the ESS 750. Equally, references to the ESS F20 also apply to the ESS E20.

Examples of what you can perform include:

- ▶ Display and change your storage configuration, for example, create and assign volumes.
- ▶ Display existing Copy Services relationships and settings, for example, confirm that remote copy relationships are active and in sync.
- ▶ Create new Copy Services relationships and settings, for example, create a new FlashCopy relationship.

For users of the ESS 800, the DS CLI provides the following *new* capabilities:

- ▶ The ability to create a remote copy relationship between the ESS 800 and the DS8000 or DS6000.
- ▶ The ability to establish dynamic FlashCopy and remote copy relationships on ESS 800 storage servers without using saved tasks.

Prior to the DS CLI, the ESS Copy Services CLI generally did not allow a script to directly invoke a FlashCopy or Remote Mirror and Copy relationship. Instead, a task had to be created and saved first, using the Web Copy Services GUI. A script could then invoke this saved task. Now with the DS CLI, commands can be saved as scripts, which significantly reduces the time to create, edit and verify their content.

The DS CLI uses a syntax that is consistent with other IBM TotalStorage products. All new products will also use this same syntax.

Important reference manuals for users of the DS CLI are the *IBM TotalStorage DS8000 Command-Line Interface User's Guide*, SC26-7625, and *IBM TotalStorage DS6000 Command-Line Interface User's Guide*, SC26-7681. These can be downloaded by going to the relevant section of the following Web site:

<http://www-1.ibm.com/servers/storage/support/disk/index.html>

## 11.2 Functionality

The DS CLI can be used to invoke the following storage configuration tasks:

- ▶ Create userids that can be used with both the DS CLI and the GUI
- ▶ Manage userid passwords
- ▶ Install activation keys for licensed features
- ▶ Manage storage complexes and units
- ▶ Configure and manage storage facility images
- ▶ Create and delete RAID arrays, ranks, and extent pools
- ▶ Create and delete logical volumes

- ▶ Manage host access to volumes
- ▶ Configure host adapter ports

The DS CLI can be used to invoke the following Copy Services functions:

- ▶ FlashCopy - Point-in-time Copy
- ▶ IBM TotalStorage Metro Mirror - Synchronous Peer-to-Peer Remote Copy (PPRC)
- ▶ IBM TotalStorage Global Copy - PPRC-XD
- ▶ IBM TotalStorage Global Mirror - Asynchronous PPRC

**Restriction:** The Copy Services functions in the December 2004 release of the DS CLI will only support the creation of point-in-time copies (FlashCopy). Remote Mirror and Copy functions will be supported in a 2005 release.

## 11.3 Supported environments

The DS CLI will be supported on a very wide variety of open systems operating systems. At present the supported systems are:

- ▶ AIX 5.1, 5.2, 5.3
- ▶ HP-UX 11i v1, v2
- ▶ HP Tru64 version 5.1, 5.1A
- ▶ Linux RedHat 3.0 Advanced Server (AS) and Enterprise Server (ES)
- ▶ SUSE Linux SLES 8, SLES 9
- ▶ Novell Netware 6.5
- ▶ Open VMS 7.3-1, 7.3-2
- ▶ Sun Solaris 7, 8, and 9
- ▶ Windows 2000, Windows Datacenter, and Windows 2003

This list should not be considered final. For the latest list, consult the interoperability Web site located at:

<http://www.ibm.com/servers/storage/disk/ds6000/interop.htm>

or:

<http://www.ibm.com/servers/storage/disk/ds8000/interop.htm>

## 11.4 Installation methods

The DS CLI is supplied and installed via a CD that ships with the machine. The installation does not require a reboot of the open systems host. The DS CLI requires Java™ 1.4.1 or higher. Java 1.4.2 for Windows, AIX, and Linux is supplied on the CD. Many hosts may already have a suitable level of Java installed.

The installation process can be performed via a shell, such as the bash or korn shell, or the Windows command prompt, or via a GUI interface. If performed via a shell, it can be performed silently using a profile file. The installation process also installs software that allows the DS CLI to be completely de-installed should it no longer be required.

The exact install process doesn't really vary by operating system. It consists of:

1. The DS CLI CD is placed in the CD-ROM drive (and mounted if necessary).
2. If using a command line, the user changes to the root directory of the CD. There is a setup command for each supported operating system. The user issues the relevant command and then follows the prompts. If using a GUI, the user navigates to the CD root directory and clicks on the relevant setup executable.
3. The DS CLI is then installed. The default install directory will be:
  - ▶ /opt/ibm/dscli - for all forms of UNIX
  - ▶ C:\Program Files\IBM\dscli - for all forms of Windows
  - ▶ SYS:\dscli - for Novell Netware

## 11.5 Command flow

To understand migration or co-existence considerations, it is important to understand the flow of commands in both the ESS CLI and the DS CLI.

### **ESS Copy Services command flow using ESS Copy Services CLI**

When using the ESS Copy Services CLI with an ESS 800, all commands are issued to a Copy Services Server, present on one cluster of the ESS. Interaction with this server is via either a Web Copy Services GUI interface, or via a CLI interface. To use the CLI, the open systems host needs to have ESS Copy Services CLI software installed. An ESS CLI script will issue commands to the CLI software, which then sends them to the primary Copy Services Server (known as Server A). For backup, a second server can also be defined (known as Server B).

Figure 11-1 on page 235 shows the flow of commands from host to server. When the Copy Services (CS) server receives a command, it determines whether the volumes involved are owned by cluster 1 or cluster 2. This is based on LSS membership (even numbered LSSs belong to cluster 1, odd numbered LSSs belong to cluster 2). The CS server issues the command to the client software on the correct cluster and then reports success or failure back to the CLI software on the open systems host.



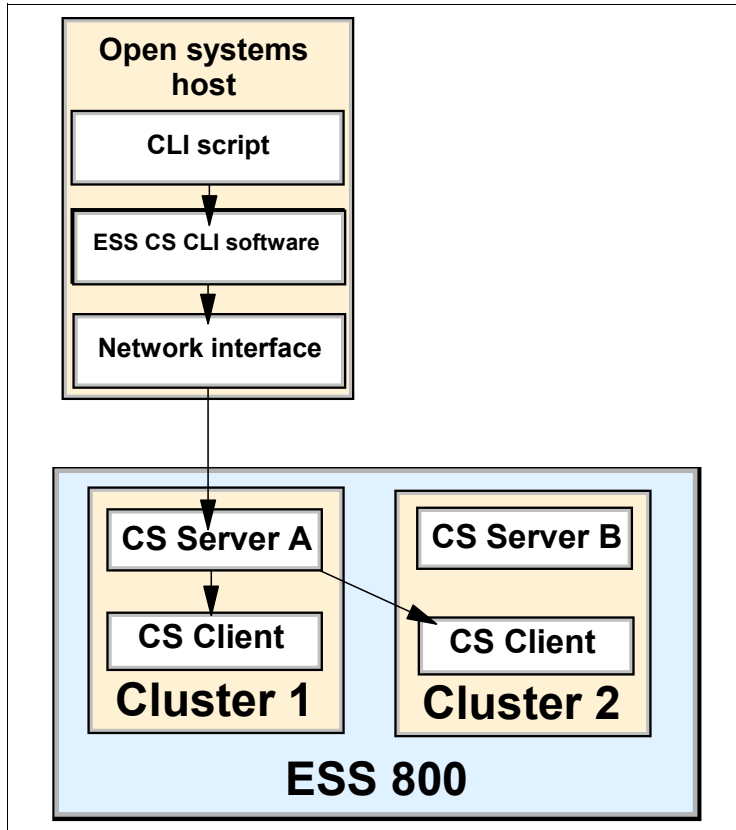


Figure 11-1 Command flow for ESS 800 Copy Services commands

A CS server is now able to manage up to eight F20s and ESS 800s. This means that up to sixteen clusters can be clients of the CS server. All FlashCopy and remote copy commands are sent to the CS server, which then sends them to the relevant client on the relevant ESS.

### DS CLI command flow

Scripts that invoke DS CLI commands issue those commands to the installed DS CLI software on the open systems host. If the command is intended for an ESS 800 volume, then the DS CLI software sends it to the CS server on the ESS 800. If, however, the command is intended for a DS8000, the command is issued to the CLI interpreter of the Storage Hardware Management Console (S-HMC). The S-HMC then interprets the command and issues it to the relevant server in the relevant DS8000 using the redundant internal network that connects the S-HMCs to the DS8000.

### Secure sockets

All DS CLI traffic is encrypted using SSL (secure sockets layer). This means that all traffic between the host server that is running the DS CLI client and the DS CLI server (for example, the S-HMC or the ESS 800 cluster) is secure, including passwords and userids.

### TCP/IP ports

DS CLI servers (such as an S-HMC) use a fixed number of TCP/IP ports to listen on. These ports are listed in Chapter 9, “Configuration planning” on page 157. This is important for planning considerations, where a firewall may exist between the client and the server.

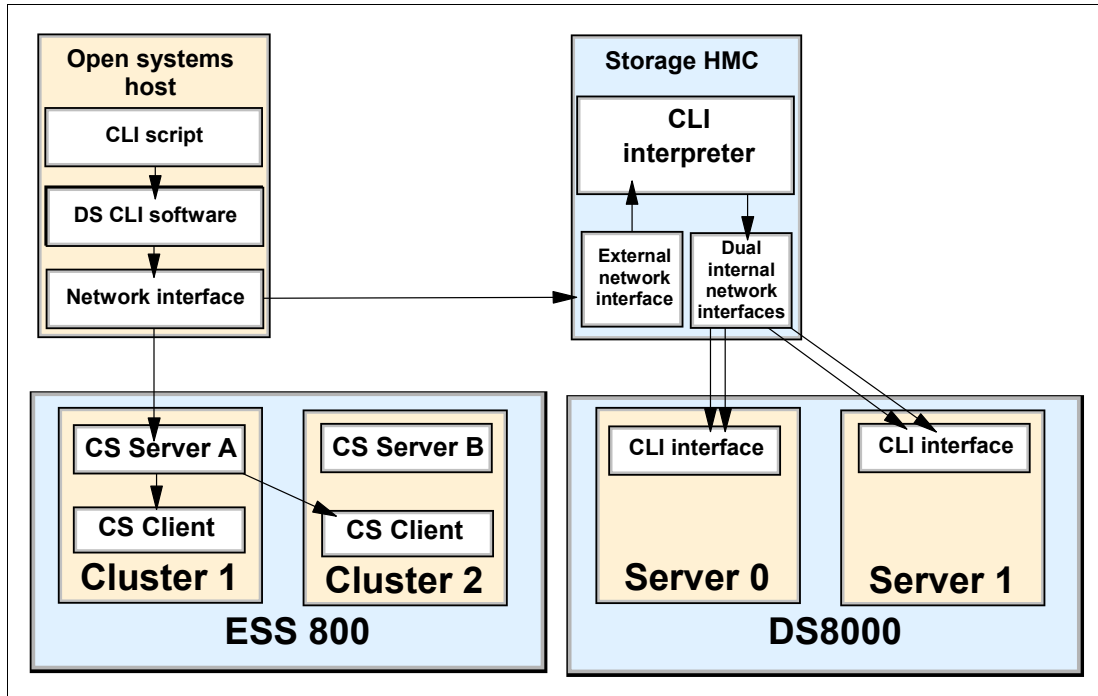


Figure 11-2 DS CLI Copy Services command flow

### DS8000 split network

One thing that you may notice about Figure 11-2 is that the S-HMC has different network interfaces. The external network interface is an Ethernet port that must be accessible from the open systems host network interface. The dual, internal network interfaces use the two internal Ethernet switches within the DS8000 base frame to deliver the commands to the relevant storage server. This means that the DS8000 itself is not on the same network as the open systems host. The S-HMC therefore acts as a bridge between the *external* server network and the *internal* DS8000 network.

Clearly a major benefit of this setup is that the internal network within the DS8000 has no single points of failure. By using a second S-HMC it is possible to create a completely redundant communications network for the DS CLI traffic between a host server and the DS8000 servers.

### DS6000 command flow

If a DS6000 is used, commands are instead issued to the DS Storage Manager PC that has to be supplied and set up when a DS6000 is installed. The DS Storage Manager PC then issues the commands to the relevant DS6000 controller. This command flow is depicted in Figure 11-3 on page 237.

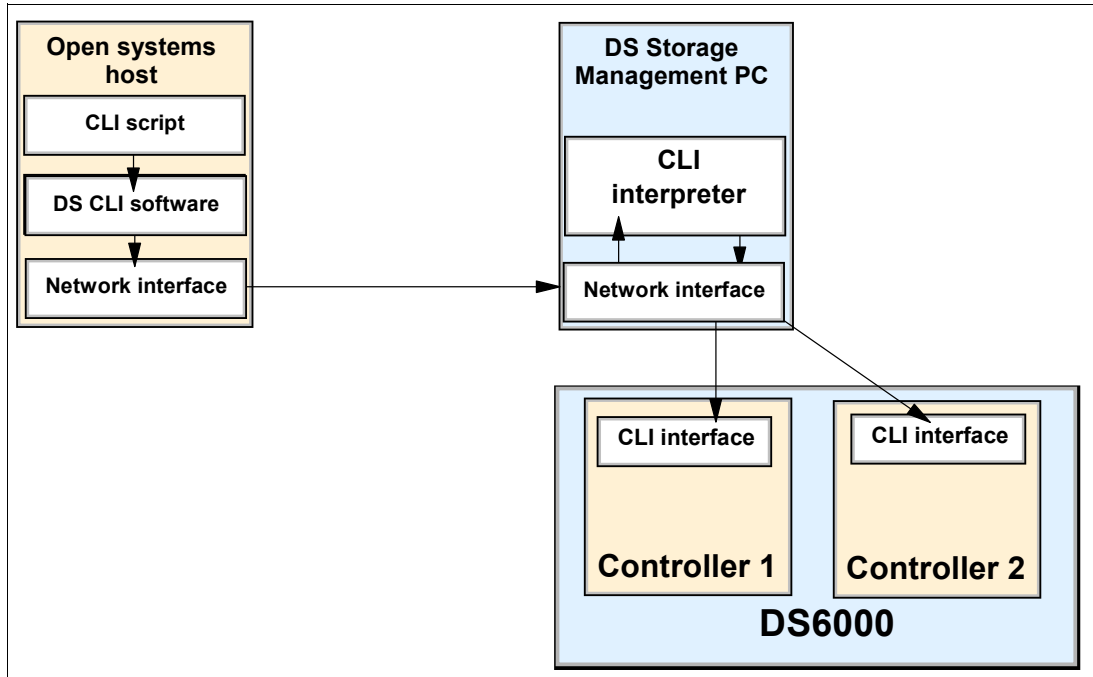


Figure 11-3 Command flow for the DS6000

For the DS6000, it is possible to install a second network interface card within the DS Storage Manager PC. This would allow you to connect it to two separate switches for improved redundancy.

### ESS CLI co-existence

If co-existence with the ESS CS CLI is required, then both the DS CLI and the ESS CLI will have to be installed on the same open systems host, as shown in Figure 11-4 on page 238. Each CLI installs into a separate directory. Depending on how the scripts are written, ESS CLI and DS CLI commands could be issued in the same script.

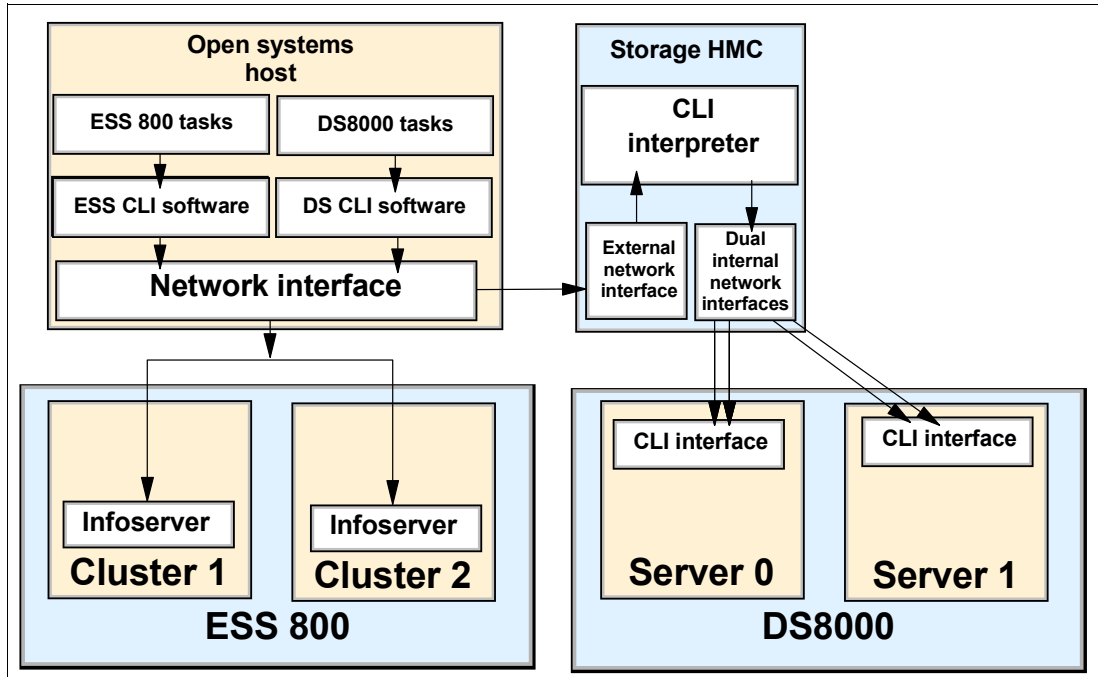


Figure 11-4 CLI co-existence

### Storage management

ESS CLI commands that are used to perform storage management on the ESS 800, are issued to a process known as the *infoserver*. An *infoserver* runs on each cluster, and either *infoserver* can be used to perform ESS 800 storage management. Storage management on the ESS 800 will continue to use ESS CLI commands. Storage management on the DS6000/8000 will use DS CLI commands. This difference in command flow is shown in Figure 11-5.

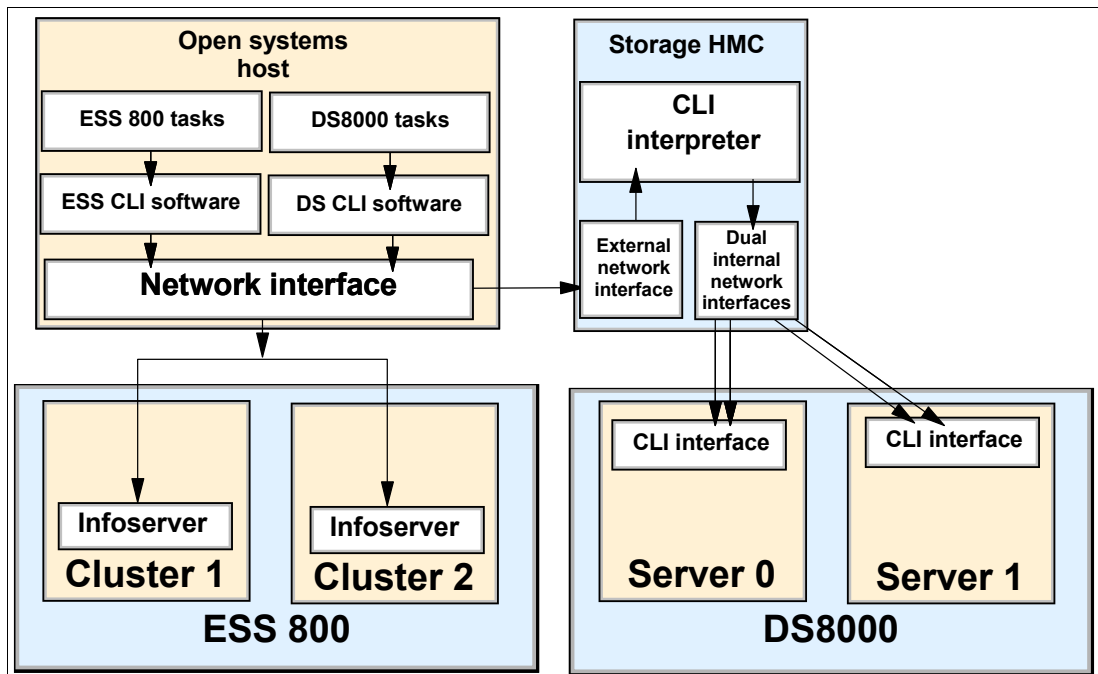


Figure 11-5 Storage management command flow

## 11.6 User security

The DS CLI software must authenticate with the S-HMC or CS Server before commands can be issued. An initial setup task will be to define at least one userid and password whose authentication details are saved in an encrypted file. A profile file can then be used to identify the name of the encrypted password file. Scripts that execute DS CLI commands can use the profile file to get the password needed to authenticate the commands.

User security employs the concept of *groups* to control which functions a particular userid is allowed to perform. A userid can be a member of more than one group. The groups are:

- ▶ `admin` - can perform all tasks - this is the only group that can create and change userids
- ▶ `op_storage` - can perform any configuration task
- ▶ `op_volume` - can configure logical volumes and volume groups
- ▶ `op_copy_services` - can perform Copy Services commands
- ▶ `service` - can perform service commands
- ▶ `monitor` - has read-only access to commands
- ▶ `no_access` - cannot perform any tasks

The functions of these groups are fairly self describing and are fully detailed both in the *IBM TotalStorage DS8000 Command-Line Interface User's Guide*, SC26-7625 and *IBM TotalStorage DS6000 Command-Line Interface User's Guide*, SC26-7681, and the help screens. If a userid is not a member of any group, then it is automatically placed into the `no_access` group to prevent it from performing any functions.

The default userid supplied with an S-HMC or DS Storage Manager is *admin* (whose password is also *admin*). During setup it is advisable that a new userid be created in the admin group. The default userid should then be removed (with the `rmuser` command). Note that userid management can be performed by using either the DS CLI or by using the DS Storage Manager GUI. Userids created by either interface will be usable via either interface.

For an example of how a userid and profile are created, refer to “Procedure to create an encrypted password file” on page 249.

## 11.7 Usage concepts

It is important to understand the various concepts that frame DS CLI usage.

### 11.7.1 Command modes

The DS CLI can be operated in three modes. In the examples that follow, the `lsuser` command is used. The `lsuser` command is used to display which users have been created and to which groups they are a member. For more details on user authentication see “User security” on page 239.

#### **Single command mode**

At a shell prompt, the user specifies a single DS CLI command which is immediately executed, and a return code is presented. To avoid having to enter authentication details, a profile and password file would have to be created first. This is shown in Example 11-1.

*Example 11-1 Using DS CLI via a single command*

---

```
C:\Program Files\IBM\dsccli>dsccli lsuser
Name      Group
=====
admin     admin
```

```
csadmin op_copy_services
exit status of dscli = 0
```

```
C:\Program Files\IBM\dscli>
```

---

It is also possible to include single commands in a script, though this is different from the *script mode* described later. This is because every command that uses the DS CLI would invoke the DS CLI and then exit it. A simple Windows script is shown in Example 11-2.

*Example 11-2 A script to list all users and place their names in a file*

---

```
@ECHO OFF
rem This script is used to list all DS CLI users
rem The lsuser command is executed and the output is sent to file called userlist.txt

dscli lsuser > userlist.txt
echo The user list has been created and placed in userlist.txt
```

---

For those readers familiar with UNIX, a simple example of creating a script is shown in Example 11-3.

*Example 11-3 Creating a DS CLI script*

---

```
/opt/ibm/dscli >echo "dscli lsuser > userlist.txt" > listusers.sh
/opt/ibm/dscli >chmod +x listusers.sh
/opt/ibm/dscli >./listusers.sh
/opt/ibm/dscli >cat userlist.txt
Name      Group
=====
admin     admin

/opt/ibm/dscli>
```

---

**Interactive mode**

In the interactive mode, the user starts the DS CLI program within a shell, and then issues DS CLI commands until the DS CLI is no longer needed. At this point the user exits the DS CLI program. To avoid having to enter authentication details, a profile and password file would have to be created first. The use of the interactive mode is shown in Example 11-4.

*Example 11-4 Using DS CLI in interactive mode*

---

```
C:\Program Files\IBM\dscli>dscli
dscli> lsuser
Name      Group
=====
admin     admin
csadmin   op_copy_services
dscli> exit

exit status of dscli = 0

C:\Program Files\IBM\dscli>
```

---

**Script mode**

The *script mode* allows a user to create a DS CLI script that contains multiple DS CLI commands. These commands are performed one after the other. When the DS CLI executes the last command, it ends and presents a return code. DS CLI scripts in this mode must

contain only DS CLI commands. This is because all commands in the script are executed by a single instance of the DS CLI interpreter. Comments can be placed in the script if they are prefixed by a hash (#). A simple example of a *script mode* script is shown in Example 11-5.

*Example 11-5 DS CLI script mode example*

---

```
# This script issues the 'lsuser' command
lsuser

# end of script
```

---

In this example, the script was placed in a file called listAllUsers.cli, located in the scripts folder within the DS CLI folder. It is then executed by using the **dscli -script** command, as shown in Example 11-6.

*Example 11-6 Executing DS CLI in script mode*

---

```
C:\Program Files\IBM\dscli> dscli -script scripts\listAllUsers.cli
Name      Group
=====
admin     admin
C:\Program Files\IBM\dscli>
```

---

It is possible to create shell or Visual Basic scripts that combine both *script mode* and single commands.

## 11.7.2 Syntax conventions

The DS CLI uses symbols and conventions that are standard in command-line interfaces. These include the ability to input variables from a file and send output to a file. The DS CLI commands are also designed to be case insensitive. This means commands can be entered in either upper, lower, or mixed case, and still work.

## 11.7.3 User assistance

The DS CLI is designed to include several forms of user assistance. The main form of user assistance is via the **help** command. Examples of usage include:

<b>help</b>	Lists all available DS CLI commands
<b>help -s</b>	Lists all available DS CLI commands with brief descriptions of each
<b>help -l</b>	Lists all DS CLI commands with syntax information

If the user is interested in more details about a specific DS CLI command, they can use **-l** (long) or **-s** (short) against a specific command. In Example 11-7, the **-s** parameter is used to get a short description of the **mkflash** command's purpose.

*Example 11-7 Use of the help -s command*

---

```
dscli> help -s mkflash
mkflash      The mkflash command initiates a point-in-time copy from source volumes to
target volumes.
```

---

In Example 11-8, the **-l** parameter is used to get a list of all the parameters that can be used with the **mkflash** command.

*Example 11-8 Use of the help -l command*

```
dscli> help -l mkflash
mkflash [ { -help|-h|-? } ] [-fullid] [-dev storage_image_ID] [-tgtpprc] [-tgtoffline]
[-tgtinhibit] [-freeze] [-record] [-persist] [-nocp] [-wait] [-seqnum Flash_Sequence_Num]
SourceVolumeID:TargetVolumeID
```

## Man pages

A “man page” is available for every DS CLI command. Man pages are most commonly seen in UNIX-based operating systems to give information about command capabilities. This information can be displayed by issuing the relevant command followed by `-h` or `-help` or `-?`, for example:

```
dscli> mkflash -help
```

or

```
dscli> help mkflash
```

## 11.7.4 Return codes

When the DS CLI is exited, an exit status code is provided. This is effectively a return code. If DS CLI commands are issued as separate commands (rather than using script mode), then a return code will be presented for every command. If a DS CLI command fails (for instance, due to a syntax error or the use of an incorrect password), then a failure reason and a return code will be presented. Standard techniques to collect and analyze return codes can be used.

The return codes used by the DS CLI are shown in Table 11-1.

*Table 11-1 DS CLI return codes*

Return code	Category	Description
0	Success	The command was successful.
2	Syntax error	There is a syntax error in the command.
3	Connection error	There was a connection problem to the server.
4	Server error	The DS CLI server had an error.
5	Authentication error	Password or userid details are incorrect.
6	Application error	The DS CLI application had an error.

In Example 11-9 a simple Windows batch file is used to query whether a FlashCopy relationship exists between volumes 1004 and 1005. The batch file then queries the operating system for the return code and provides a verbose response.

*Example 11-9 Sample Windows bat file to test return codes*

```
@ECHO OFF
dscli lsflash -dev IBM.2105-23953 1004:1005
if errorlevel 6 goto level6
if errorlevel 5 goto level5
if errorlevel 4 goto level4
if errorlevel 3 goto level3
if errorlevel 2 goto level2
if errorlevel 0 goto level0

:level6
```



```

echo A DS CLI application error occurred.
goto end
:level5
echo An authentication error occurred. Check the userid and password.
goto end
:level4
echo A DS CLI Server error occurred.
goto end
:level3
echo A connection error occurred. Try pinging 10.0.0.1
echo If this fails call network support on 555-1001
goto end
:level2
echo A syntax error. Check the syntax of the command using online help.
goto end
:level0
echo No errors were encountered.
:end

```

---

Using this sample script, Example 11-10 shows what happens if there is a network problem between the DS CLI client and server (in this example a 2105-800). The DS CLI provides the error code (in this case CMUN00018E) which can be looked up in the DS CLI Users Guide (referred to in “Introduction” on page 232). The DS CLI also provides the exit status (in this example, exit status = 3). Finally, the batch file interprets the return code and provides the user with some additional tips to resolve the problem.

*Example 11-10 Return code examples*

---

```

C:\Program Files\IBM\dsccli> checkflash.bat
CMUN00018E lsflash: Unable to connect to the management console server
exit status of dsccli = 3
A connection error occurred. Try pinging 10.0.0.1
If this fails call network support on 555-1001

C:\Program Files\IBM\dsccli>

```

---

## 11.8 Usage examples

It is not the intent of this section to list every DS CLI command and its syntax. If you need to see a list of all the available commands, or require assistance using DS CLI commands, you are better served by reading the *IBM TotalStorage DS8000 Command-Line Interface User's Guide*, SC26-7625, and *IBM TotalStorage DS6000 Command-Line Interface User's Guide*, SC26-7681. Or you can use the online help. Example 11-11 gives a sample configuration script showing most of the storage management commands that are used on a DS6000 or DS8000.

*Example 11-11 Example of a configuration script*

---

```

# The following command creates a CKD extent pool (CKD extent pool P0 will be created)
mkextpool -dev IBM.2107-9999999 -rankgrp 0 -stgtype ckd ckd_ext_pool0

# The following command creates an array (array A0 will be created)
mkarray -dev IBM.2107-9999999 -raidtype 5 S1

# The following command creates a rank (CKD rank R0 will be created)
mkrank -dev IBM.2107-9999999 -array A0 -stgtype ckd

```

```

# The following command checks the status of the ranks
lsrank -dev IBM.2107-9999999

# The following command assigns rank0 (R0) to extent pool 0 (P0)
chrank -extpool P0 -dev IBM.2107-9999999 R0

# The following command creates an LCU (LCU 02 will be created)
mklcu -dev IBM.2107-9999999 -ss FF02 -qty 1 -id 02
# The following command creates another LCU (LCU 04 will be created)
mklcu -dev IBM.2107-9999999 -ss FF04 -qty 1 -id 04

# The following command creates 32 CKD volumes (0200-021F will be created)
# These ckd volumes are on LCU 02
mkckdvol -dev IBM.2107-9999999 -extpool P0 -cap 3339 -name ckd_vol_#h -qty 32 0200
# The following command creates 32 CKD volumes (0400-041F will be created)
# These ckd volumes are on LCU 04
mkckdvol -dev IBM.2107-9999999 -extpool P0 -cap 3339 -name ckd_vol_#h -qty 32 0400

#The following command lists I/O ports to configure.
lsioport -dev IBM.2107-9999999
# The following commands set two I/O ports to FICON
setioport -topology ficon IBM.2107-9999999/I0010
setioport -topology ficon IBM.2107-9999999/I0011

```

---

## 11.9 Mixed device environments and migration

The Copy Services commands within the DS CLI are designed to interface with both the DS6000 and DS8000 series and the ESS 800. They will not work with the ESS F20. The *storage management commands* within the DS CLI also will not work with ESS 800. This means that customers who currently have a mix of 800s and F20s will have to continue to use the current ESS CLI, but could consider deploying the DS CLI for certain Copy Services functions.

To explain in more detail, for the ESS 800 there are two families of CLI commands. ESS storage management CLI commands are used for storage management and configuration. The ESS Copy Services CLI commands are used to manage and monitor Copy Services relationships (FlashCopy and Remote Mirror and Copy). Currently both kinds of CLI are installed by the same setup file. The DS CLI combines both these functions into one library of commands. Table 11-2 shows which CLI should be used based on which hardware is installed in a particular environment.

Table 11-2 Which CLI to use based on what hardware you have installed

ESS F20	ESS 800	DS6000 and 8000	CLI to use
Installed	Not installed	Not installed	Use ESS CLI only.
Installed	Installed	Not installed	Use ESS CLI for most functions. Consider use of DS CLI for copy functions on the ESS 800.
Not installed	Installed	Installed	Use DS CLI for all Copy Services. Use a combination of ESS CLI and DS CLI for storage management.
Not installed	Not installed	Installed	Use DS CLI only.
Installed	Installed	Installed	Use a combination of ESS CLI and DS CLI.

## Migration considerations

If your environment is currently using the ESS CS CLI to manage Copy Services on your model 800s, you could consider migrating your environment to the DS CLI. Your model 800s will need to be upgraded to a microcode level of 2.4.2 or above.

If your environment is a mix of ESS F20s and ESS 800s, it may be more convenient to keep using only the ESS CLI. This is because the DS CLI cannot manage the ESS F20 at all, and cannot manage storage on an ESS 800. If, however, a DS6000/8000 were to be added to your environment, then you would use the DS CLI to manage remote copy relationships between the DS6000/8000s and the ESS 800s. You could still use the ESS CLI to manage the storage on the F20 and 800, FlashCopy on the F20, and any remote copy relationships between the F20 and the 800.

### 11.9.1 Migration tasks

There are two phases to migrate existing FlashCopy or Remote Mirror and Copy scripts and tasks from the ESS CLI to the DS CLI.

#### **Phase one: Review**

1. Review saved tasks in the ESS 800 Web Copy Services GUI and note the details of every saved task that you wish to migrate. You can also use the ESS CLI to display the contents of each saved task and write them to a file.
2. Review server scripts that perform task set up and execute saved ESS tasks.

#### **Phase two: Perform**

Having performed the review, the scripts need to be changed and created:

1. Translate the contents of the saved ESS 800 tasks into DS CLI commands. A *mini* DS CLI script could be created for every saved task.
2. Translate server scripts that perform task set up and execute saved ESS 800 tasks. This may involve the use of DS CLI commands to perform task setup and the execution of the newly created mini scripts to achieve the same results as the saved tasks.

**Note:** You might consider requesting assistance from IBM in the migration phase. Depending on your geography, IBM can offer CLI migration services to help you ensure the success of your project.

## 11.10 DS CLI migration example

As detailed previously, existing users of the ESS CS CLI with an ESS 800 can consider migrating saved tasks to the DS CLI.

### 11.10.1 Determining the saved tasks to be migrated

Step one is to gather information regarding all saved tasks. This can be done via the GUI or the command line. In this example there are many saved tasks (but only five are shown). Figure 11-6 and Example 11-12 show two ways to get a list of saved tasks on the ESS CS Server. In Figure 11-6 the **Tasks** button has been selected in the ESS 800 Web Copy Services GUI.

<b>Tasks</b> <b>Administration</b> <b>Exit</b>	H_Epath_test16	undefined	No
	H_Epath_test17	undefined	No
	Brocade_pr_1ss10	undefined	No
	Brocade_pr_1ss11	undefined	No
	Flash10041005	Flash copy vol 1004 to 1005	No
<input type="button" value="Run"/> <input type="button" value="Modify"/> <input type="button" value="Group"/> <input type="button" value="Ungroup"/>			

Figure 11-6 A portion of the tasks listed by using the GUI

In Example 11-12, the **list task** command is used. This is an ESS CLI command.

*Example 11-12 Using the list task command to list all saved tasks (only the last five are shown)*

```

arielle@aixserv:/opt/ibm/ESScli > esscli list task -s 10.0.0.1 -u csadmin -p passwd
Wed Nov 24 10:29:31 EST 2004 IBM ESSCLI 2.4.0
Task Name                                Type                                Status
-----
H_Epath_test16                          PPRCEstablishPaths                 NotRunning
H_Epath_test17                          PPRCEstablishPaths                 NotRunning
Brocade_pr_1ss10                         PPRCEstablishPair                  NotRunning
Brocade_pr_1ss11                         PPRCEstablishPair                  NotRunning
Flash10041005                            FCEstablish                         NotRunning

```

### 11.10.2 Collecting the task details

Having collected the names of the saved tasks, the user needs to collect the contents of each task. If viewing tasks via the GUI, you can highlight each task and click the **Information** button to bring up the information panel for each task. An example is shown in Figure 11-7 on page 247.

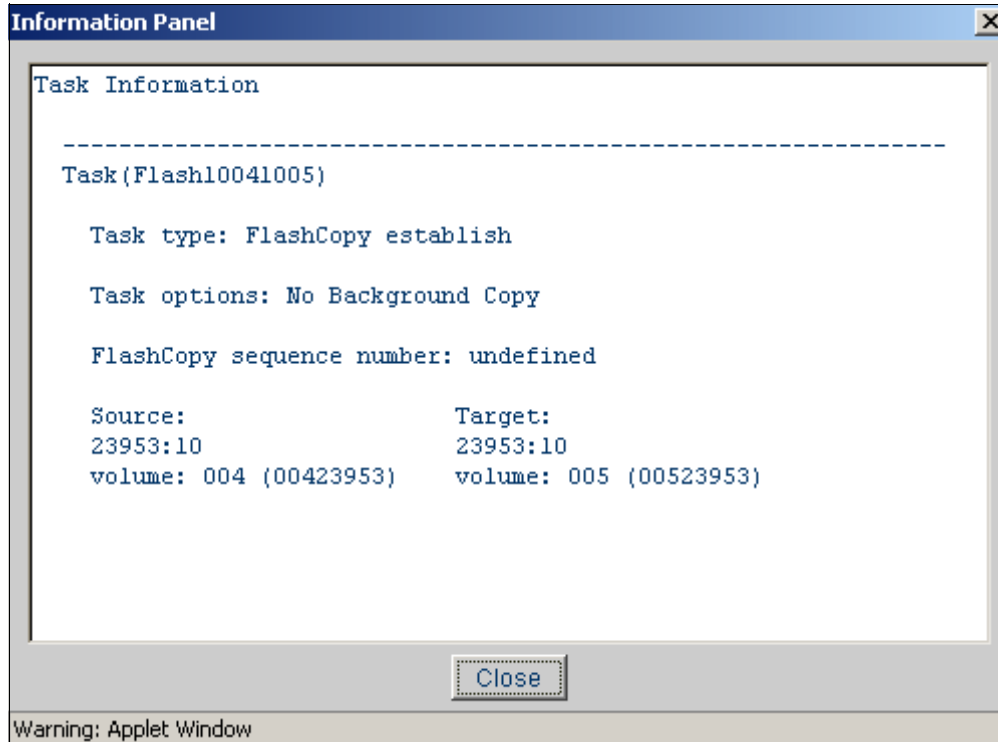


Figure 11-7 Using the GUI to get the contents of a FlashCopy task

It makes more sense, however, to use the ESS CLI **show task** command to list the contents of the tasks, as depicted in Example 11-13.

Example 11-13 Using the command line to get the contents of a FlashCopy task

---

```

mitchell@aixserv:/opt/ibm/ESScli > esscli show task -s 10.0.0.1 -u csadmin -p passw0rd -d
"name=Flash10041005"
Wed Nov 24 10:37:17 EST 2004 IBM ESSCLI 2.4.0
Taskname=Flash10041005
Tasktype=FCEstablish
Options=NoBackgroundCopy
SourceServer=2105.23953
TargetServer=2105.23953
SourceVol          TargetVol
-----
1004                1005

```

---

### 11.10.3 Converting the saved task to a DS CLI command

Having collected the contents of a saved task, it can now be converted into a DS CLI task. Using the data from Example 11-13, each parameter is translated to the correct value for a DS CLI command in Table 11-3.

Table 11-3 Converting a FlashCopy task to DS CLI

ESS CS CLI parameter	Saved task parameter	DS CLI conversion	Explanation
Tasktype	FCEstablish	mkflash	An FCEstablish becomes a mkflash.
Options	NoBackgroundCopy	-nocp	To do a FlashCopy no-copy we use the -nocp parameter.
SourceServer	2105.23953	-dev IBM.2105-23953	The format of the serial number changes. you must use the exact syntax.
TargetServer	2105.23953	N/A	We only need to use the -dev once, so this is redundant.
Source and Target vols	1004 1005	1004:1005	The volume numbers don't change. We simply separate them with a full colon.

So to create the DS CLI command, simply read down the third column:

**mkflash -nocp -dev IBM.2105-23953 1004:1005**

In Example 11-14 the user uses the DS CLI interactive mode. They issue the **mkflash** command and then use **lsflash** to check the success of the command.

Example 11-14 Using interactive dscli mode without profiles

```
sharon@aixsrv:/opt/ibm/dscli > dscli
dscli> mkflash -nocp -dev IBM.2105-23953 1004:1005
CMUC00137I mkflash: FlashCopy pair 1004:1005 successfully created.
dscli> lsflash -dev IBM.2105-23953 1004:1005
ID          SrcLSS SequenceNum Timeout ActiveCopy Recoding Persistent Revertible
=====
1004:1005 10      0          120    Disabled  Disabled Disabled  Disabled
dscli>
```

You can also confirm the status of the FlashCopy by using the Web Copy Services GUI, as shown in Figure 11-8.



Figure 11-8 FlashCopy status via the ESS 800 Web Copy Services GUI

## 11.10.4 Using DS CLI commands via a single command or script

Having translated a saved task into a DS CLI command, you may now want to use a script to execute this task upon request. Since all tasks must be authenticated you will need to create a userid.

### Creating a user ID for use only with ESS 800

When using the DS CLI with an ESS 800, authentication is performed by using a userid and password created with the ESS Specialist. If you have an ESS 800, but not a DS8000 or DS6000 offline configurator, or an S-HMC, then you will not be able to create an encrypted password file. Instead you will need to specify the password and userid in the script file itself. This is no different than with the current ESS CLI, except that userids created using the ESS 800 Web Copy Service Server are not used (the userid used is an ESS Specialist userid).

If you have DS CLI access to a DS offline configuration tool, S-HMC or DS Storage Management console, then you can create an encrypted password file. This will allow you to avoid specifying the password or userid in plain text, in any script or profile.

### Procedure to create an encrypted password file

User management with the DS CLI is via the **mkuser** command to create userids, and the **chuser** command to change passwords. These commands must be issued by a userid in the *admin* group. There is also **rmuser** to remove a userid. An example is:

```
mkuser -group op_copy_services -pw passw0rd -pwfile cs_admin csadmin
```

In this example, a userid called *csadmin* has been created which can be used either for CLI authentication or to log on to the DS Storage Manager GUI. The password is *passw0rd* and the user is a member of the *op\_copy\_services* group. This means the user cannot configure storage or create other users, but can manage Copy Services relationships. A password file called *passwd* was created on the local host (the host on which the **mkuser** command was issued).

Having created the userid called *csadmin*, the password has been saved in an encrypted file called *passwd*. The file is placed in a subfolder of the security folder. If the IP address of the S-HMC is 10.0.0.1 then you will find the password file in `/opt/ibm/dscli/security/10.0.0.1`, or `c:\program files\ibm\dscli\security\10.0.0.1`. If you are moving this file to a different server, you may have to create this folder.

### Setting up a profile

Having created a userid, you will need to edit the profile used by the DS CLI to store the S-HMC IP address (or fully qualified name) and other common parameters. The default profile is located at `C:\Program Files\IBM\dscli\profile\dscli.profile` or `/opt/ibm/dscli/dscli.profile`. An example of a *secure* profile is shown in Example 11-15.

*Example 11-15 Example of a dscli.profile file*

---

```
#DS CLI Profile
#
# Management Console/Node IP Address(es) are specified using the hmc parameter
# hmc1 and hmc2 are equivalent to -hmc1 and -hmc2 command line options.
# hmc1 is cluster 1 of 2105 800 23953
hmc1:10.0.0.1
#hmc2:127.0.0.1

# Password filename is the name of an encrypted password file
# This file is located at /opt/ibm/dscli/security/10.0.0.1
```

```

pwfile: passwd

# Default target Storage Image ID
# If the -dev parameter is needed in a command then it will default to the value here
# "devId" is equivalent to "-dev storage_image_ID"
# the default server that DS CLI commands will be run on is 2105 800 23953

devId: IBM.2105-23953

```

---

If you don't want to create an encrypted password file, or do not have access to a simulator or the real S-HMC, then you can specify the userid and password in plain text. This is done either at the command line or in a script or in the profile. This is not recommended since the password itself is now not as secure. An example of a profile that contains plain text authentication details is shown in Example 11-16.

*Example 11-16 Example of a DS CLI profile that specifies the username and password*

---

```

#DS CLI Profile

# hmc1 is cluster 1 of 2105 800 23953
hmc1:10.0.0.1

#The username to log onto the ESS
username: csadmin

# The password for csadmin:
password: passwd0rd

# Default target Storage Image ID
devId: IBM.2105-23953

```

---

An example of a command where the password is entered in plain text is shown in Example 11-17. In this example the **lsuser** command is issued directly to an S-HMC. Note that the password will still be sent using SSL so a network sniffer would not be able to view it easily. Note also that the syntax between the command and the profile is slightly different.

*Example 11-17 Example of a DS CLI command that specifies the username and password*

---

```

C:\Program Files\IBM\dscli>dscli -hmc1 10.0.0.1 -user admin -passwd passwd0rd lsuser
Name      Group
=====
admin     admin
csadmin   op_copy_services
exit status of dscli = 0

C:\Program Files\IBM\dscli>

```

---

## Issuing a DS CLI command and scripting it

Having created a userid and preferably a password file, and then having edited the default profile, it is now possible to issue DS CLI commands without logging onto the DS CLI interpreter. An example is shown in Example 11-18.

*Example 11-18 Establishing a FlashCopy with a single command*

---

```

anthony@aixsrv:/opt/ibm/dscli > dscli mkflash -nocp 1004:1005
CMUC00137I mkflash: FlashCopy pair 1004:1005 successfully created.
exit status of dscli = 0

```



```
anthony@aixsrv:/opt/ibm/dscli >
```

---

The command can also be placed into a file and that file made executable. An example is shown in Example 11-19.

*Example 11-19 Creating an executable file*

---

```
anthony@aixsrv:/home >echo "/opt/ibm/dscli/dscli mkflash -nocp 1004:1005" > flash1005
anthony@aixsrv:/home >chmod +x flash1005
anthony@aixsrv:/home >./flash1005
anthony@aixsrv:/home >CMUC00137I mkflash: FlashCopy pair 1004:1005 successfully created.
anthony@aixsrv:/home >
```

---

Finally, the command could be issued using script mode. An example of creating and using script mode is shown in Example 11-20.

*Example 11-20 Using script mode*

---

```
arielle@aixsrv:/opt/ibm/dscli >echo "mkflash -nocp 1004:1005" > scripts/flash1005
arielle@aixsrv:/opt/ibm/dscli >dscli -script scripts/flash1005
arielle@aixsrv:/opt/ibm/dscli >CMUC00137I mkflash: FlashCopy pair 1004:1005 successfully
created.
arielle@aixsrv:/opt/ibm/dscli >
```

---

## 11.11 Summary

This chapter has provided some important information about the DS CLI. This new CLI allows considerable flexibility in how DS6000 and DS8000 series storage servers are configured and managed. It also detailed how an existing ESS 800 customer can benefit from the new flexibility provided by the DS CLI.





## Performance considerations

This chapter discusses early performance considerations regarding the DS8000 series.

Disk Magic modelling for DS8000 is available as of December 03, 2004. Contact your IBM sales representative for more information about this tool and the benchmark testing that was done by the Tucson performance measurement lab.

Note that Disk Magic is an IBM internal modelling tool.

We discuss the following topics in this chapter:

- ▶ The challenge with today's disk storage servers
- ▶ How the DS8000 addresses this challenge
- ▶ Specific considerations for open systems and z/OS

## 12.1 What is the challenge?

In recent years we have seen an increasing speed in developing new storage servers which can compete with the speed at which processor development introduces new processors. On the other side, investment protection as a goal to contain Total Cost of Ownership (TCO), dictates inventing smarter architectures that allow for growth at a component level. IBM understood this early on and introduced its Seascape® architecture, and brought the ESS into the marketplace in 1999 based on this architecture.

### 12.1.1 Speed gap between server and disk storage

Disk storage evolved over time from simple structures to a string of disk drives attached to a disk string controller without caching capabilities. The actual disk drive—with its mechanical movement to seek the data, rotational delays and actual transfer rates from the read/write heads to disk buffers—created a speed gap compared to the internal speed of a server with no mechanical speed brakes at all. Development went on to narrow this increasing speed gap between processor memory and disk storage server with more complex structures and data caching capabilities in the disk storage controllers. With cache hits in disk storage controller memory, data could be read and written at channel or interface speeds between processor memory and storage controller memory. These enhanced storage controllers, furthermore, allowed some sharing capabilities between homogenous server platforms like S/390-based servers. Eventually disk storage servers advanced to utilize a fully integrated architecture based on standard building blocks as introduced by IBM with the Seascape architecture. Over time, all components became not only bigger in capacity and faster in speed, but also more sophisticated; for instance, using an improved caching algorithm or enhanced host adapters to handle many processes in parallel.

### 12.1.2 New and enhanced functions

Parallel to this development, new functions were developed and added to the next generation of disk storage subsystems. Some examples of new functions added over time are dual copy, concurrent copy, and eventually various flavors of remote copy and FlashCopy. These functions are all related to managing the data in the disk subsystems, storing the data as quickly as possible, and retrieving the data as fast as possible. Other aspects became increasingly important, like disaster recovery capabilities. Applications demand increasing I/O rates and higher data rates on one hand, but shorter response times on the other hand. These conflicting goals must be resolved, and are the driving force to develop storage servers such as the new DS8000 series.

With the advent of the DS8000 and its server-based structure and virtualization possibilities, another dimension of potential functions within the storage servers is created.

These storage servers grew with respect to functionality, speed, and capacity. Parallel to their increasing capabilities, the complexity grew as well. The art is to create systems which are well balanced from top to bottom, and these storage servers scale very well. Figure 12-1 on page 255 shows an abstract and simplified comparison of the basic components of a host server and a storage server. All components at each level need to be well balanced between each other to provide optimum performance at a minimum cost.

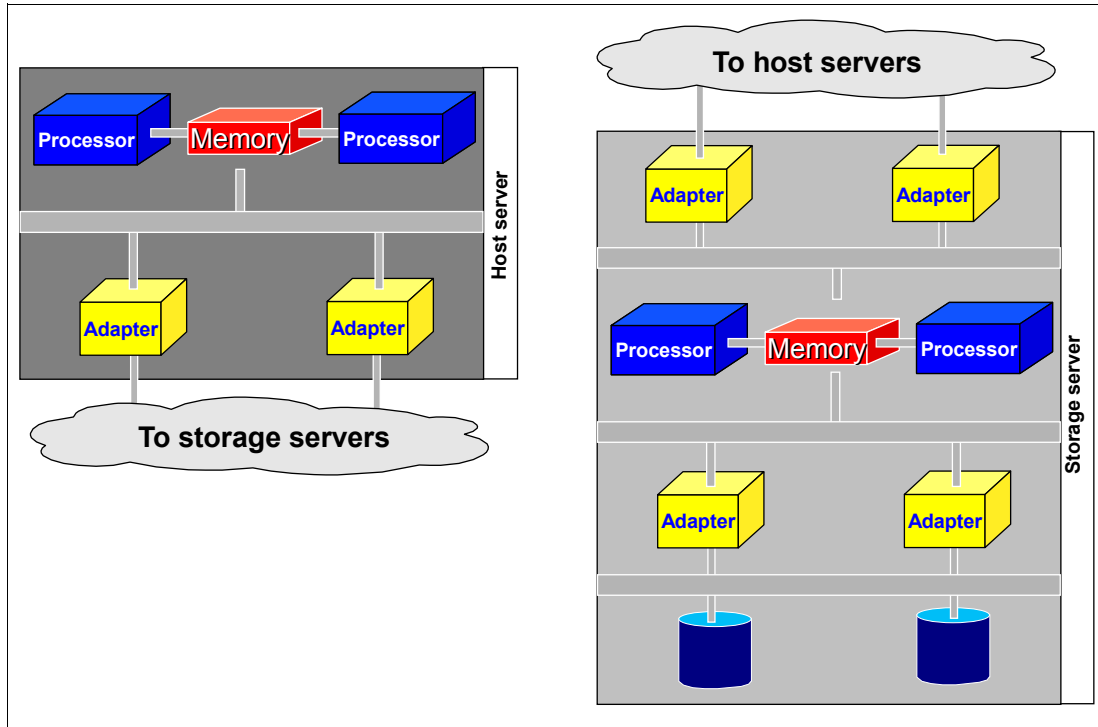


Figure 12-1 Host server and storage server comparison: Balanced throughput challenge

The challenge is obvious: Develop a storage server—from the top with its host adapters down to its disk drives—that creates a balanced system with respect to each component within this storage server and with respect to their interconnectivity with each other. All this must be done while taking into consideration other requirements as well, like investment protection and being competitive not only in performance but also with price and reliability. A further requirement is to keep the life cycle of the product as long as possible due to the substantial development costs for an all new product. Furthermore, provide a storage family approach with a single management interface and the potential to adopt new technology as it develops on a component level. The simple picture suggests that this a more difficult task for a storage server than for a host server. Perhaps it became this way with the evolving complexity of storage servers.

## 12.2 Where do we start?

The IBM Enterprise Storage Server 2105 (ESS) already combined everything mentioned in the previous paragraph when it appeared in the marketplace in 1999. Over time the ESS evolved in many respects to enhance performance and to improve throughput. Despite the powerful design, the technology and implementation used eventually reached its life cycle end.

Looking more closely at the components in the ESS and their enhancements from the E20 to the F20 to the 800 and 800 Turbo II models, some components reached their architectural limits at various levels. This section briefly reviews the most obvious limitations that were encountered over time as the other components were enhanced. It has to be noted that the ESS 800 is still a very competitive disk storage server, which outperforms other storage servers in many respects.

### 12.2.1 SSA backend interconnection

The Storage Serial Architecture (SSA) connectivity with the SSA loops in the lower level of the storage server or backend imposed RAID rank saturation and reached their limit of 40 MB/second for a single stream file I/O. IBM decided not to pursue SSA connectivity, despite its ability to communicate and transfer data within an SSA loop without arbitration.

### 12.2.2 Arrays across loops

Advancing from the F20 to the 800 model, the internal structures and buses increased two fold and so did the back end with striping logical volumes across loops (AAL). This increased the sequential throughput on an SSA loop from 40 MB/sec to 80 MB/sec for a single file operation, distributing the back-end I/O across two loops. Because of the increasing requirements to serve more data and at the same time reduce the application I/O response time, the SSA loop throughput was not sufficient and surfaced more often with RAID rank saturation.

### 12.2.3 Switch from ESCON to FICON ports

The front end got faster when it moved from ESCON at 200 Mbps to FICON at 2 Gbps with an aggregated bandwidth from 32 ESCON ports x 200 Mbps at 6.4 Gbps to 16 FICON ports with 2 Gbps each, yielding 32 Gbps. Note that the pure technology, like 2 Gbps, is not enough to provide good performance. The 2 Gbps FICON implementation in the ESS HA proved to provide industry leading throughput in MB/sec as well as I/Os per second. An ESS FICON port, even today, has the potential to exceed the throughput capabilities of other vendor's FICON ports.

When not properly configured (for example, with four FICON ports in the same HA), these powerful FICON ports have the potential to saturate the respective host bay. Spreading FICON ports evenly across all host bays puts increased pressure on the internals of the ESS below the HA ports.

### 12.2.4 PPRC over Fibre Channel links

With PPRC over ESCON links there was some potential bottleneck when the HA changed from ESCON to FICON. Despite a smart overlap and utilizing multiple PPRC ESCON links for PPRC, the speed difference between FICON/FCP and ESCON channels introduced some imbalance in the ESS. The ESS 800 finally introduced PPRC over FCP links with 2 Gbps. Again this enhancement proves that the move to 2 Gbps technology is only half of the story. With a smart implementation in the FCP port connection for PPRC, the performance of an ESS FCP PPRC port is still not matched today by other implementation efforts.

These performance enhancements into and out of the ESS shifted potential bottlenecks back into the internals of the ESS for very high write I/O rates with 15,000 write I/Os and more per second.

### 12.2.5 Fixed LSS to RAID rank affinity and increasing DDM size

Another growing concern was the fixed affinity of logical subsystems (LSS) to RAID ranks and the respective volume placement. Volumes had to reside within a single SSA loop and even within the same RAID array; later in the A-loop and B-loop, but still bound to a single device adapter (DA) pair.

Even more serious was the addressing issue with the 256 device limit within an LSS and the fixed association to a RAID rank. With the growing disk drive module (DDM) size and

relatively small logical volumes, we ran out of device numbers to address an entire LSS. This happens even earlier when configuring not only real devices (3390B) within an LSS, but also alias devices (3390A) within an LSS in z/OS environments. By the way, an LSS is congruent to an logical control unit (LCU) in this context. An LCU is only relevant in z/OS and the term is not used for open systems operating systems.

## 12.3 How does the DS8000 address the challenge?

The DS8000 overcomes the architectural limits and bottlenecks which developed over time in the ESS due to the increasing number of I/Os and MB/sec.

In this section we go through the different layers and discuss how they have changed to address performance in terms of throughput and I/O rates.

### 12.3.1 Fibre Channel switched disk interconnection at the back end

Because SSA connectivity has not been further enhanced to increase the connectivity speed beyond 40MB/sec, Fibre Channel connected disks were chosen for the DS8000 back end. This technology is commonly used to connect a group of disks in a daisy-chained fashion in a Fibre Channel Arbitrated Loop (FC-AL).

#### FC-AL shortcomings

There are some shortcomings with plain FC-AL. The most obvious ones are:

- ▶ As the term arbitration implies, each individual disk within an FC-AL loop competes with the other disks to get on the loop because the loop supports only one operation at a time.
- ▶ Another challenge which is not adequately solved is the handling of failures within the FC-AL loop, particularly with intermittently failing components on the loops and disks.
- ▶ A third issue with conventional FC-AL is the increasing time it takes to complete a loop operation as the number of devices increases in the loop.

For highly parallel operations, concurrent reads and writes with various transfer sizes, this impacts the total effective bandwidth of an FC-AL structure.

#### How the DS8000 series overcomes FC-AL shortcomings

The DS8000 uses the same Fibre Channel drives as used in conventional FC-AL based storage systems. To overcome the arbitration issue within FC-AL, the architecture is enhanced by adding a switch-based approach and creating FC-AL switched loops, as shown in Figure 12-2 on page 258. Actually it is called a Fibre Channel switched disk subsystem.

These switches use FC-AL protocol and attach FC-AL drives through a point-to-point connection. The arbitration message of a drive is captured in the switch, processed and propagated back to the drive, without routing it through all the other drives in the loop.

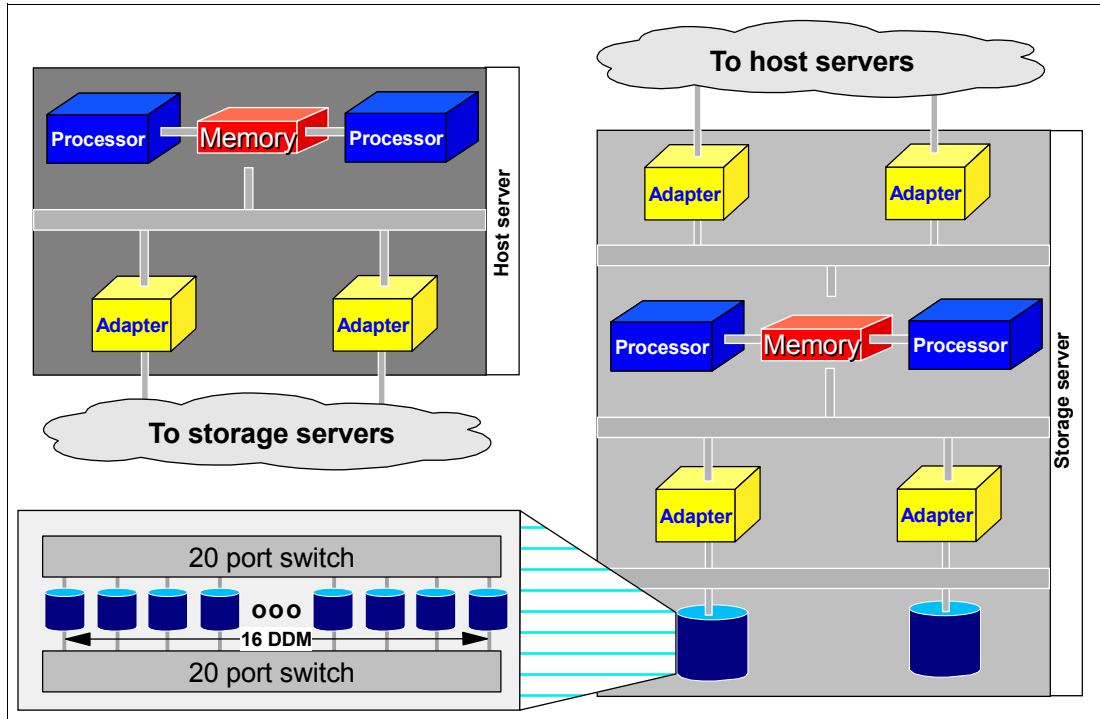


Figure 12-2 Switched FC-AL disk subsystem

Performance is enhanced since both DAs connect to the switched Fibre Channel disk subsystem back end as displayed in Figure 12-3 on page 259. Note that each DA port can concurrently send and receive data.



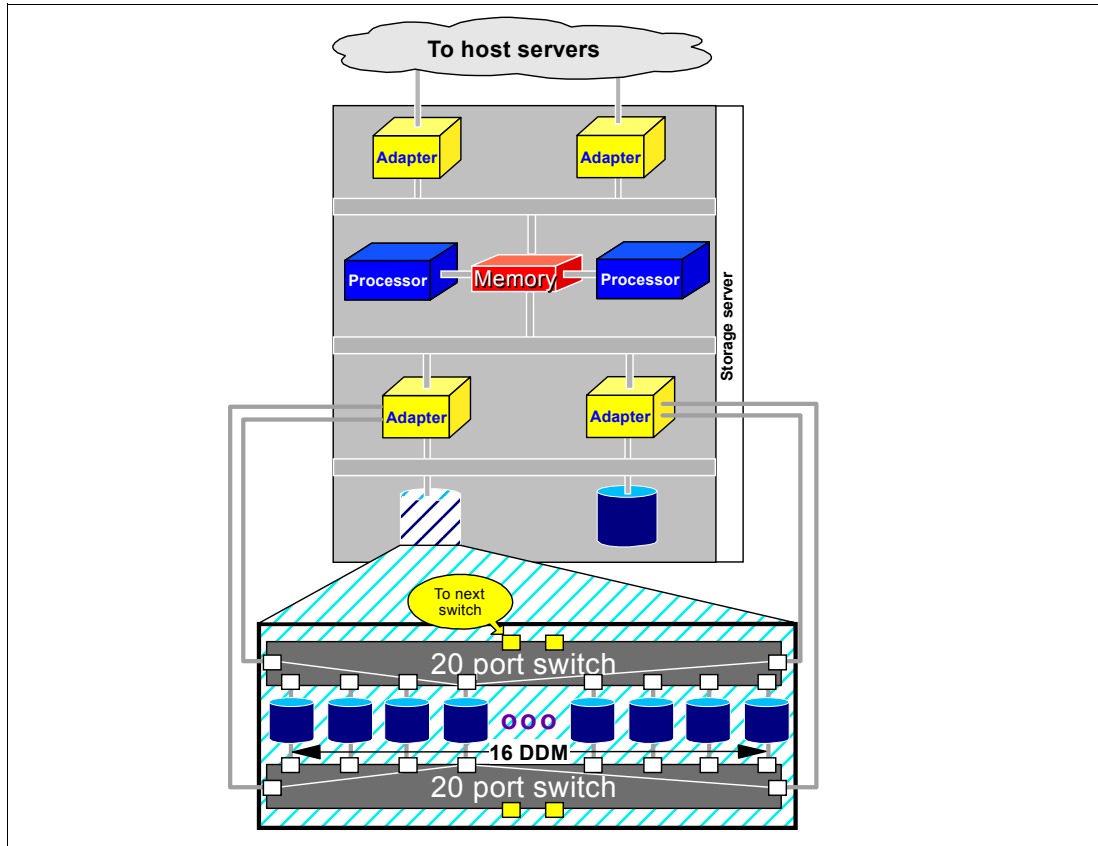


Figure 12-3 High availability and increased bandwidth connecting both DA to two logical loops

These two switched point-to-point loops to each drive, plus connecting both DAs to each switch, accounts for the following:

- ▶ There is no arbitration competition and interference between one drive and all the other drives because there is no hardware in common for all the drives in the FC-AL loop. This leads to an increased bandwidth utilizing the full speed of a Fibre Channel for each individual drive. Note that the external transfer rate of a Fibre Channel DDM is 200 MB/sec.
- ▶ Doubles the bandwidth over conventional FC-AL implementations due to two simultaneous operations from each DA to allow for two concurrent read operations and two concurrent write operations at the same time.
- ▶ Despite the superior performance, don't forget the improved RAS over conventional FC-AL. The failure of a drive is detected and reported by the switch. The switch ports distinguish between intermittent failures and permanent failures. The ports understand intermittent failures which are recoverable and collect data for predictive failure statistics. If one of the switches itself fails, a disk enclosure service processor detects the failing switch and reports the failure using the other loop. All drives can still connect through the remaining switch.

This just outlines the physical structure. A virtualization approach built on top of the high performance architecture contributes even further to enhanced performance. For details see Chapter 5, "Virtualization concepts" on page 83.

### 12.3.2 Fibre Channel device adapter

The DS8000 still relies on eight DDMs to form a RAID-5 or a RAID-10 array. These DDMs are actually spread over two Fibre Channel loops and follow the successful approach to use AAL. With the virtualization approach and the concept of extents, the DAs are mapping the virtualization level over the disk subsystem back end. For more details on the disk subsystem virtualization refer to Chapter 5, “Virtualization concepts” on page 83.

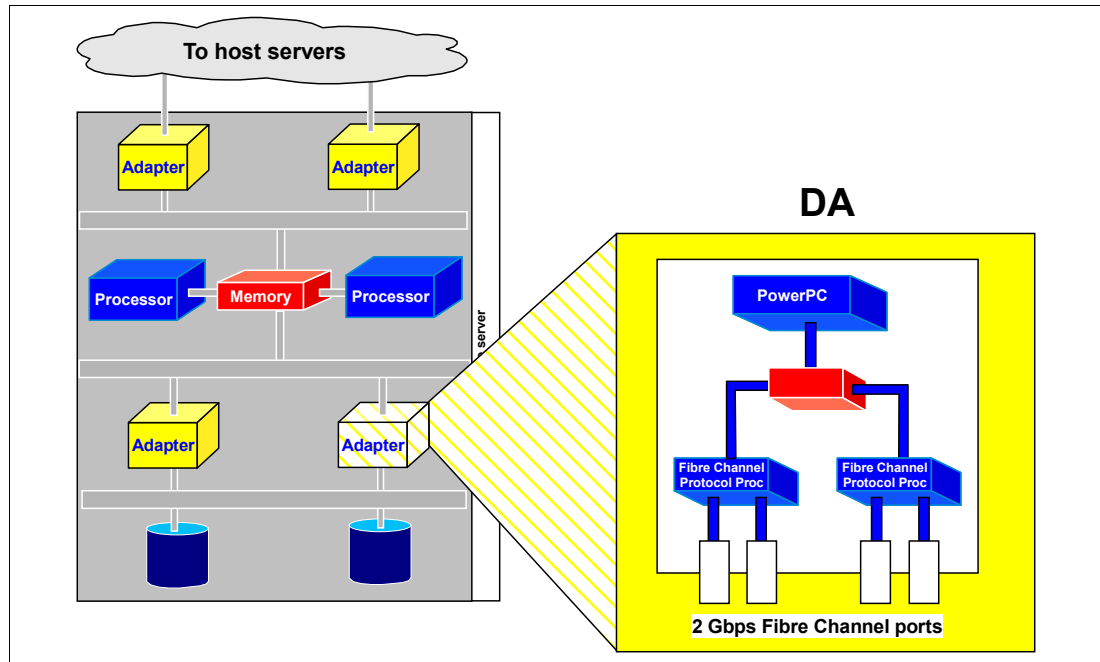


Figure 12-4 Fibre Channel device adapter with 2 Gbps ports

The new RAID device adapter is built on PowerPC technology with four 2 Gbps Fibre Channel ports and high function, high performance ASICs. Each port provides up to five times the throughput of a previous SSA-based DA port. Each single Fibre Channel protocol processor satisfies both Fibre Channel 2 Gbps ports at full speed.

Note that each DA performs the RAID logic and frees up the processors from this task. The actual throughput and performance of a DA is not only determined by the 2 Gbps ports and hardware used, but also by the firmware efficiency.

### 12.3.3 New four-port host adapters

Before looking into the heart of the DS8000 series, we briefly review the new host adapters and their enhancements to address performance. Figure 12-5 on page 261 depicts the new host adapters. These adapters are designed to hold four Fibre Channel ports, which can be configured to support either FCP or FICON. They are also enhanced in their configuration flexibility and provide more logical paths, from 256 with an ESS FICON port to 2,048 per FICON port on the DS8000 series.

Each port continues the tradition of providing industry-leading throughput and I/O rates for FICON and FCP.

Note that a FICON channel can address up to 16,384 devices through a FICON port. The DS8000 series can hold up to 65,280 devices and you need at least four additional FICON channel ports to reach all potential volumes within a DS8000 disk storage server. With four

ports per HA and up to 16 HAs in the smallest family member of the DS8000 series, the DS8100, you can configure up to 64 FICON channel ports. This still provides 16 FICON channel paths to each single device, which is beyond what the zSeries Channel Subsystem provides with its limit of up to eight channel paths per device as a maximum.

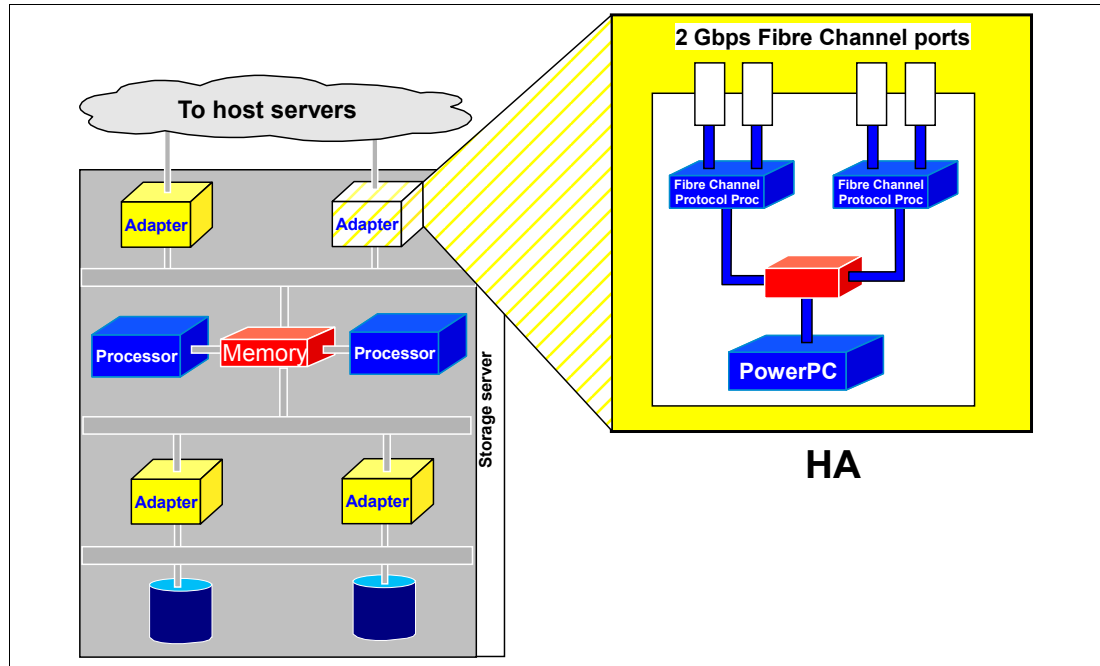


Figure 12-5 Host adapter with 4 Fibre Channel ports

The front end with the 2 Gbps ports scales up to 128 ports for a DS8300. This results in a theoretical aggregated host I/O bandwidth of 128 times 2 Gbps and outperforms an ESS by a factor of eight. The DS8100 still provides four times more bandwidth at the front end than an ESS.

### 12.3.4 POWER5 - Heart of the DS8000 dual cluster design

The DS8000 series incorporates the latest pSeries POWER5 processor technology. The actual processor used is an eServer p5 570 server, which scales from a 1-way to a 16-way SMP using standard 4U building blocks. The first two family members of the DS8000 series utilize two-way and four-way processor complexes. The following sections discuss configuration and performance aspects based on the two-way processor complexes used in the DS8100.

Among the most exciting capabilities the pSeries inherited from zSeries are the dynamic LPAR mode and the micro partitioning capability. This pSeries-based functionality has the potential to be exploited also in future disk storage server enhancements. For details on what the first steps with LPAR technology in the DS8000 look like see Chapter 3, “Storage system LPARs (Logical partitions)” on page 43.

Besides the self-healing features and advanced RAS attributes, the RIO-G structures provide a very high I/O bandwidth interconnect with DAs and HAs to provide system-wide balanced aggregated throughput from top to bottom. A simplified view is in Figure 12-6 on page 262. The smallest processor complex within a DS8100 is the POWER5 p570 two-way SMP processor complex. The dual-processor complex approach allows for concurrent microcode loads, transparent I/O failover and failback support, and redundant, hot-swappable components.

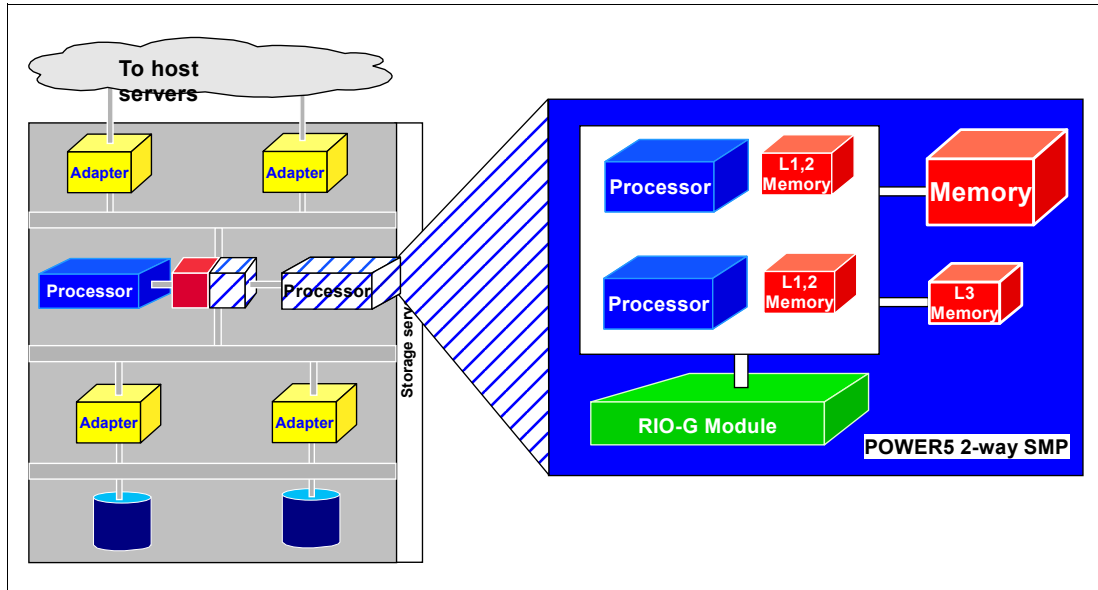


Figure 12-6 Standard pSeries POWER5 p570 2-way SMP processor complexes for DS8100-921

Figure 12-7 provides a less abstract view and outlines some details on the dual 2-way processor complex of a DS8100-921, its gates to host servers through HAs, and its connections to the disk storage back end through the DAs.

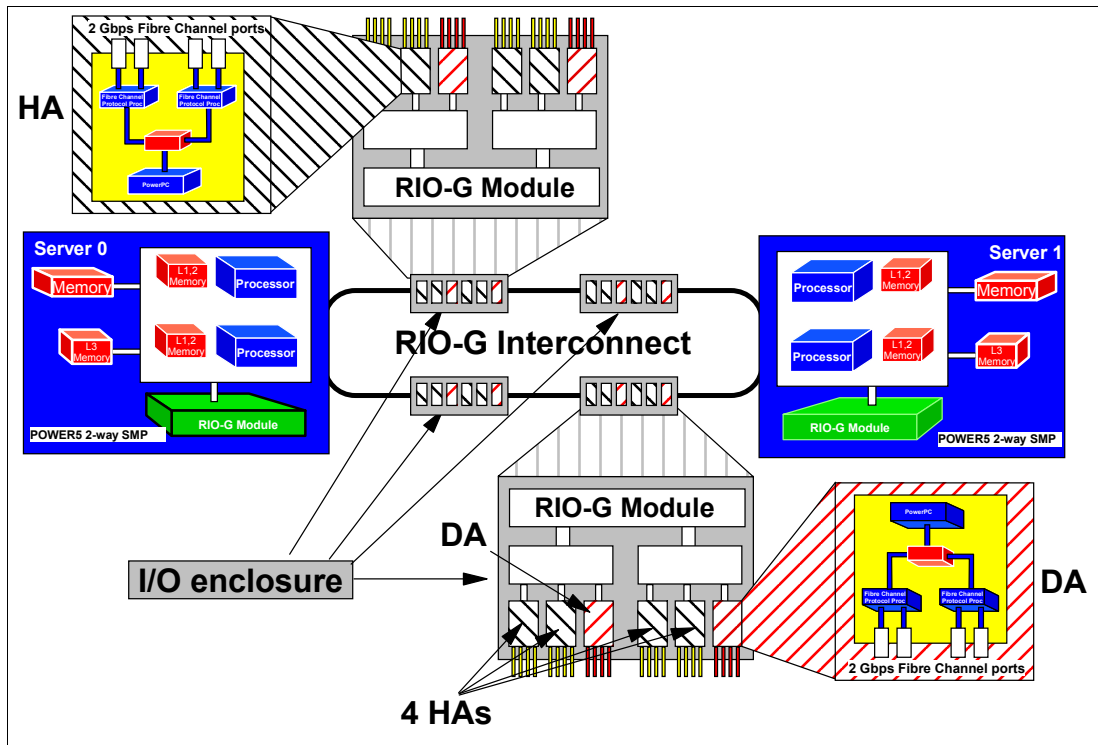


Figure 12-7 DS8100-921 with four I/O enclosures

Each of the two processor complexes is interconnected through the pSeries-based RIO-G interconnect technology and includes up to four I/O enclosures which equally communicate to either processor complex. Note that there is some affinity between the disk subsystem and its individual ranks to either the left processor complex, server 0, or to the right processor

complex, server 1. This affinity is established at creation of an extent pool. For details see Chapter 10, “The DS Storage Manager - logical configuration” on page 189.

Each single I/O enclosure itself contains six Fibre Channel adapters:

- ▶ Two DAs which install in pairs
- ▶ Four HAs which install as required

Each adapter itself contains four Fibre Channel ports.

Although each HA can communicate with each server, there is some potential to optimize traffic on the RIO-G interconnect structure. RIO-G provides a full duplex communication with 1 GB/sec in either direction. There is no such thing as arbitration. Figure 12-7 on page 262 shows that the two left-most I/O enclosures might communicate with server 0, each in full duplex. The two right-most I/O enclosures communicate with server 1, also in full duplex mode. This results in a potential of 8 GB/sec in total just for this single structure. Basically there is no affinity between HA and server. As we see later, the server which owns certain volumes through its DA, communicates with its respective HA when connecting to the host.

### High performance, high availability interconnect to the disk subsystem

Figure 12-8 depicts in some detail how the Fibre Channel switched back-end storage connects to the processor complex.

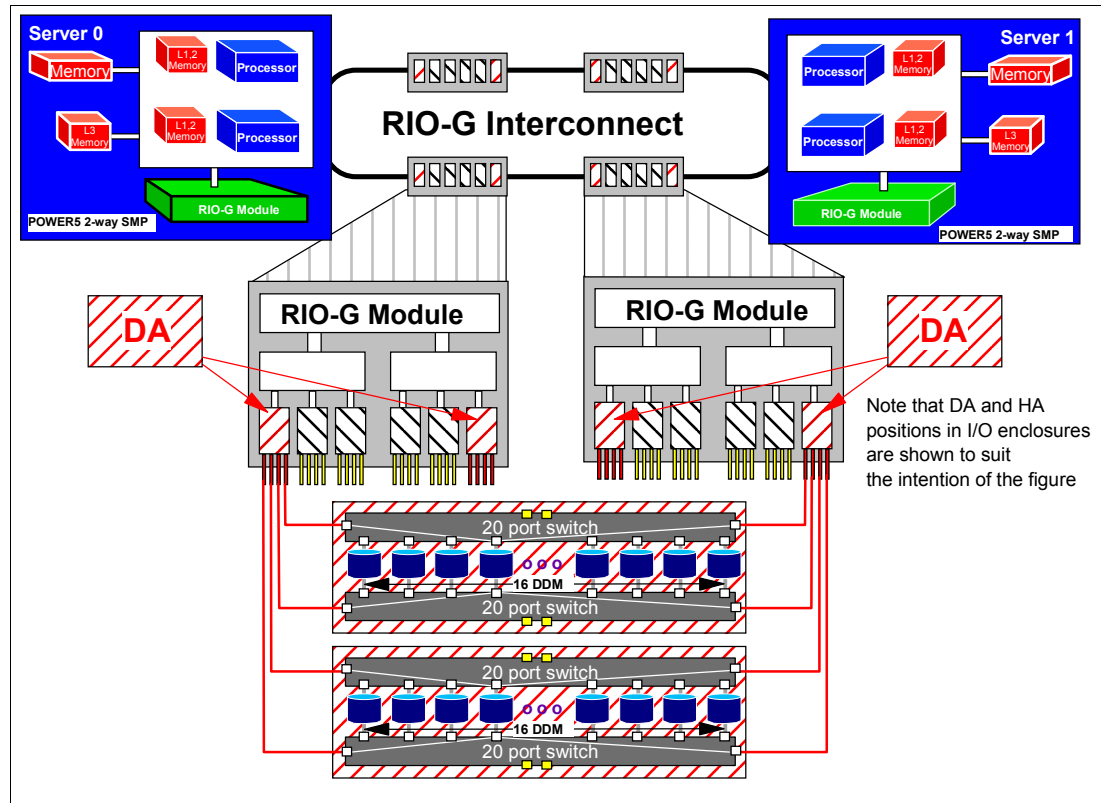


Figure 12-8 Fibre Channel switched backend connect to processor complexes - partial view

All I/O enclosures within the RIO interconnect fabric are equally served from either processor complex.

Each I/O enclosure contains two DAs. Each DA with its four ports connects to four switches to reach out to two sets of 16 drives or disk drive modules (DDMs) each. Note that each 20-port

switch has two ports to connect to the next switch pair with 16 DDMs when vertically growing within a DS8000. As outlined before, this dual two logical loop approach allows for multiple concurrent I/O operations to individual DDMs or sets of DDMs and minimizes arbitration through the DDM/switch port mini loop communication.

### 12.3.5 Vertical growth and scalability

Figure 12-9 shows a simplified view of the basic DS8000 structure and how it accounts for scalability.

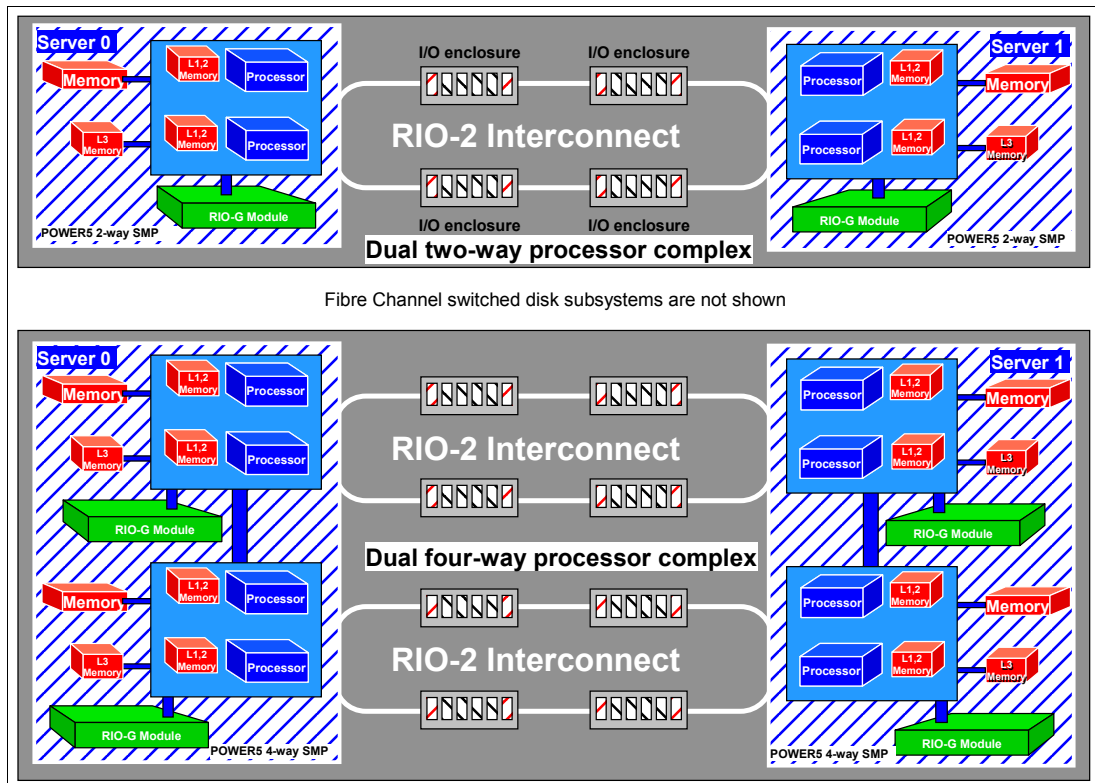


Figure 12-9 DS8100 to DS8300 - scale performance linearly - view without disk subsystems

Although Figure 12-9 does not display the back-end part, it can be derived from the number of I/O enclosures, which suggests that the disk subsystem also doubles, as does everything else, when switching from a DS8100 to an DS8300. Doubling the number of processors and I/O enclosures accounts also for doubling the performance or even more.

Again note here that a virtualization layer on top of this physical layout contributes to additional performance potential.

## 12.4 Performance and sizing considerations for open systems

To determine the most optimal DS8000 layout, the I/O performance requirements of the different servers and applications should be defined up front since they will play a large part in dictating both the physical and logical configuration of the disk subsystem. Prior to designing the disk subsystem, the disk space requirements of the application should be well understood.

### 12.4.1 Workload characteristics

The answers to questions like *how many host connections do I need?*, *how much cache do I need?* and the like always depend on the workload requirements (such as, how many I/Os per second per server, I/Os per second per gigabyte of storage, and so forth).

The information you need, ideally, to conduct detailed modeling includes:

- ▶ Number of I/Os per second
- ▶ I/O density
- ▶ Megabytes per second
- ▶ Relative percentage of reads and writes
- ▶ Random or sequential access characteristics
- ▶ Cache hit ratio

### 12.4.2 Cache size considerations for open systems

Cache sizes in the DS8000 can be either 16, 32, 64, 128, or 256 GB. The 16 GB of system memory is only available on the DS8100 and the 256 GB of system memory is only available on the DS8300.

The factors that have to be considered to determine the proper cache size are:

- ▶ The total amount of disk capacity that the DS8000 will hold
- ▶ The characteristic access density (I/Os per GB) for the stored data
- ▶ The characteristics of the I/O workload (cache friendly, unfriendly, standard; block size; random or sequential; read/write ratio; I/O rate)

If you do not have detailed information regarding the access density and the I/O operations characteristics, but you only know the usable capacity, you can estimate between 2 GB and 4 GB for the size of the cache per 1 TB of storage, as a general rule of thumb.

### 12.4.3 Data placement in the DS8000

Once you have determined the disk subsystem throughput, the disk space and number of disks required by your different hosts and applications, you have to make a decision regarding the data placement.

As is common for data placement and to optimize the DS8000 resources utilization, you should:

- ▶ Equally spread the LUNs across the DS8000 servers.  
Spreading the LUNs equally on rank group 0 and 1 will balance the load across the DS8000 servers.
- ▶ Use as many disks as possible.
- ▶ Distribute across DA pairs and RIO-G loops.
- ▶ Stripe your logical volume across several ranks.
- ▶ Consider placing specific database objects (such as logs) on different ranks.

**Note:** Database logging usually consists of sequences of synchronous sequential writes. Log archiving functions (copying an active log to an archived space) also tend to consist of simple sequential read and write sequences. You should consider isolating log files on separate arrays.

All disks in the storage subsystem should have roughly the equivalent utilization. Any disk that is used more than the other disks will become a bottleneck to performance. A practical method is to make extensive use of volume level striping across disk drives.

#### 12.4.4 LVM striping

Striping is a technique for spreading the data in a logical volume across several disk drives in such a way that the I/O capacity of the disk drives can be used in parallel to access data on the logical volume. The primary objective of striping is very high performance reading and writing of large sequential files, but there are also benefits for random access.

DS8000 logical volumes are composed of extents. An extent pool is a logical construct to manage a set of extents. One or more ranks with the same attributes can be assigned to an extent pool. One rank can be assigned to only one extent pool. To create the logical volume, extents from one extent pool are concatenated. If an extent pool is made up of several ranks, a LUN can potentially have extents on different ranks and so be spread over those ranks.

**Note:** We recommend assigning one rank per extent pool to control the placement of the data. When creating a logical volume in an extent pool made up of several ranks, the extents for this logical volume are taken from the same rank if possible.

However, to be able to create very large logical volumes, you must consider having extent pools that span more than one rank. In this case, you will not control the position of the LUNs and this may lead to an unbalanced implementation, as shown in Figure 12-10 on page 267.

Combining extent pools made up of one rank and then LVM striping over LUNs created on each extent pool, will offer a balanced method to evenly spread data across the DS8000 as shown in Figure 12-10.



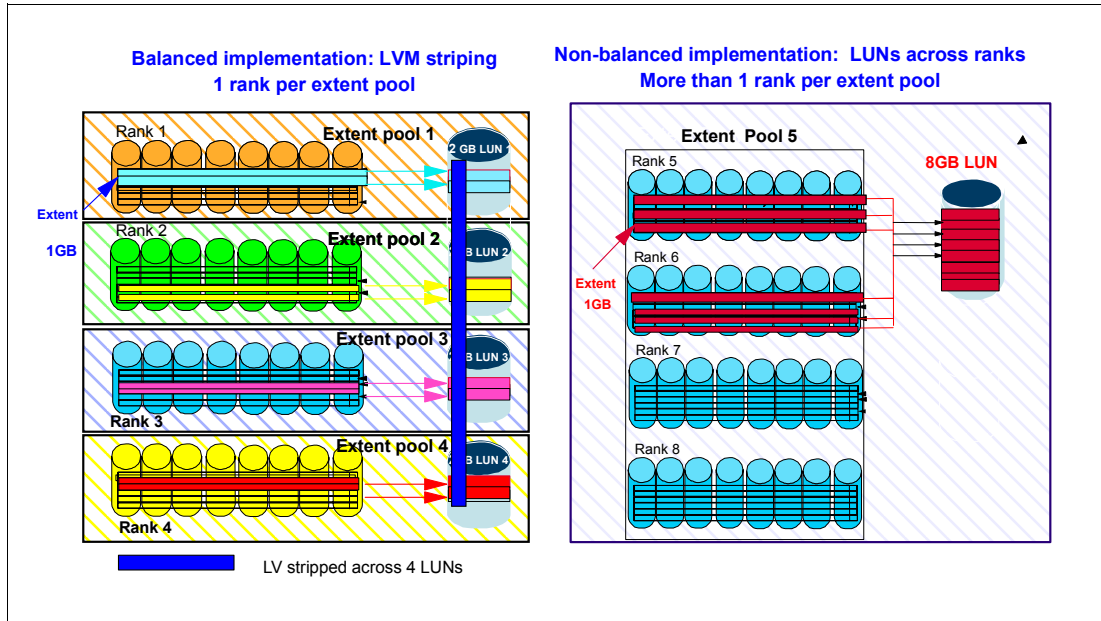


Figure 12-10 Spreading data across ranks

**Note:** The recommendation is to use host striping wherever possible to distribute the access patterns across the physical resources of the DS8000.

### The stripe size

Each striped logical volume that is created by the host's logical volume manager has a stripe size that specifies the fixed amount of data stored on each DS8000 logical volume (LUN) at one time.

**Note:** The stripe size has to be large enough to keep sequential data relatively close together, but not too large so as to keep the data located on a single array.

The recommended stripe sizes that should be defined using your host's logical volume manager are in the range of 4 MB to 64 MB.

You should choose a stripe size close to 4 MB if you have a large number of applications sharing the arrays and a larger size when you have very few servers or applications sharing the arrays.

## 12.4.5 Determining the number of connections between the host and DS8000

When you have determined your workload requirements in terms of throughput, you have to choose the appropriate number of connections to put between your open systems and the DS8000 to sustain this throughput.

A Fibre Channel host port can sustain a maximum of 206 MB/s data transfer. As a general recommendation, you should at least have two FC connections between your hosts and your DS8000.

## 12.4.6 Determining the number of paths to a LUN

When configuring the IBM DS8000 for an open systems host, a decision must be made regarding the number of paths to a particular LUN, because the multipathing software allows (and manages) multiple paths to a LUN. There are two opposing factors to consider when deciding on the number of paths to a LUN:

- ▶ Increasing the number of paths increases availability of the data, protecting against outages.
- ▶ Increasing the number of paths increases the amount of CPU used because the multipathing software must choose among all available paths each time an I/O is issued.

A good compromise is between 2 and 4 paths per LUN.

### ***Subsystem Device Driver (SDD): Dynamic I/O load-balancing***

The Subsystem Device Driver is a pseudo device driver designed to support the multipath configuration environments in the IBM TotalStorage DS8000. It resides in a host system with the native disk device driver as described in Chapter 15, “Open systems support and software” on page 319.

The dynamic I/O load-balancing option (default) of SDD is recommended to ensure better performance because:

- ▶ SDD automatically adjusts data routing for optimum performance. Multipath load balancing of data flow prevents a single path from becoming overloaded, causing input/output congestion that occurs when many I/O operations are directed to common devices along the same input/output path.
- ▶ The path to use for an I/O operation is chosen by estimating the load on each adapter to which each path is attached. The load is a function of the number of I/O operations currently in process. If multiple paths have the same load, a path is chosen at random from those paths.

## 12.4.7 Determining where to attach the host

When determining where to attach multiple paths from a single host system to I/O ports on a host adapter to the storage facility image, the following considerations apply:

- ▶ Choose the attached I/O ports on different host adapters.
- ▶ Spread the attached I/O ports evenly between the four I/O enclosure groups.
- ▶ Spread the I/O ports evenly between the different RIO-G loops.

The DS8000 host adapters have no server affinity, but the device adapters and the rank have server affinity, as illustrated in Figure 12-11 on page 269.

A host is connected through two FC adapters to two DS8000 host adapters located in different I/O enclosures. The host has access to LUN1, which is created in the extent pool 1 controlled by the DS8000 server 0. The host system sends read commands to the storage server. When a read command is executed, one or more logical blocks are transferred from the selected logical drive through a host adapter over an I/O interface to a host.

In the following case the logical device is managed by server 0 and the data is handled by server 0 as shown in Figure 12-11. The read data to be transferred to the host must first be present in server 0's cache. When the data is in the cache it is then transferred through the host adapters to the host.

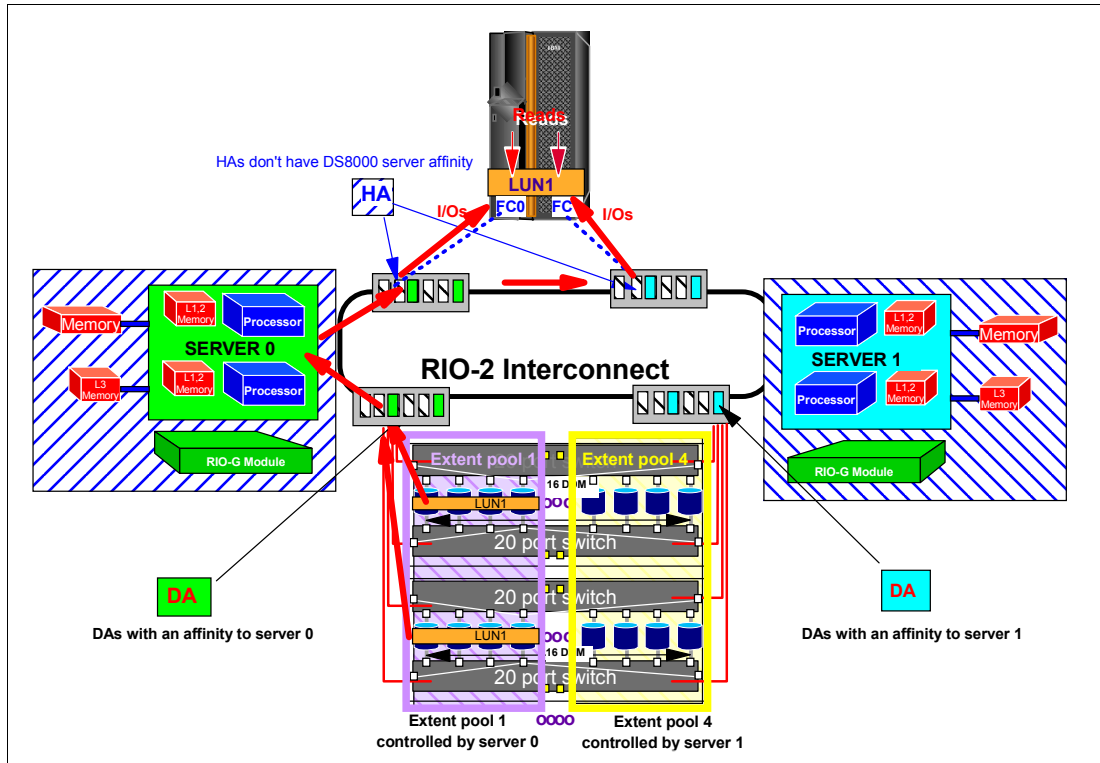


Figure 12-11 Dual port host attachment

## 12.5 Performance and sizing considerations for z/OS

Here we discuss some z/OS-specific topics regarding the performance potential of the DS8000. We also address what to consider when you configure and size a DS8000 to replace older storage hardware in z/OS environments.

### 12.5.1 Connect to zSeries hosts

Figure 12-12 on page 270 displays a configuration fragment on how to connect a DS8000 to FICON hosts.

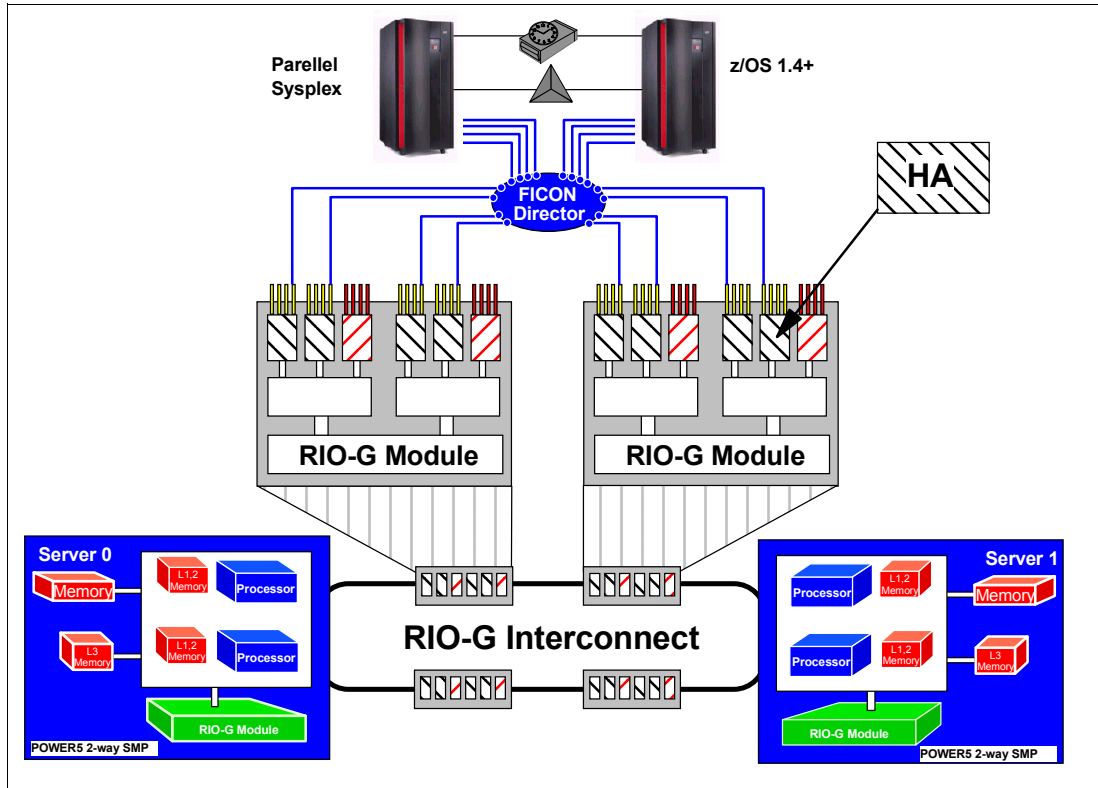


Figure 12-12 DS8100 frontend connectivity example - partial view

Note that this figure only indicates the connectivity to the Fibre Channel switched disk subsystem through its I/O enclosure, symbolized by the rectangles.

Each I/O enclosure can hold up to four HAs. The example in Figure 12-12 shows only eight FICON channels connected to the first two I/O enclosures. Not shown is a second FICON director, which connects in the same fashion to the remaining two I/O enclosures to provide a total of 16 FICON channels in this particular example. The DS8100 disk storage server provides up to 64 FICON channel ports. Again note the very efficient FICON implementation in DS8000 FICON ports.

## 12.5.2 Performance potential in z/OS environments

FICON channels started in the IBM 9672 G5 and G6 servers with 1 Gbps. Eventually these channels were enhanced to FICON Express channels in IBM 2064 and 2066 servers, with double the speed, so they now operate at 2 Gbps.

The example in Figure 12-12 with 16 FICON Express channels has roughly the potential to provide a bandwidth of  $16 \times 175$  MB/sec, equal to about 2.8 GB/sec. This is a very conservative number. Some measurements show up to 206 MB/sec per 2 Gbps FICON Express channel and 3.3 GB/sec aggregated for this particular example with 16 FICON Express channels.

I/O rates with 4 KB blocks are in the range of 6,800 I/Os per second or more per FICON Express channel, again a conservative number. A single FICON Express channel can actually perform up to about 9,000 read hit I/Os per second on the DS8000. This particular example in Figure 12-12, with 16 FICON Express channels, has the potential of over 100,000 I/Os per second. These numbers are rough considerations and rely on 100% read hit rates for

the I/O rate and highly sequential read operations for the MB/sec numbers. They also vary depending on the server type used.

The 2.8 GB/sec sequential read figure—and even significantly more—is achievable with a properly configured DS8300.

A properly configured DS8100 can reach read hit I/O rates with numbers in the 6 digits. The DS8300 does more than double that of the DS8100. These are rough estimates based on the technology and architectural possibilities, and are presented just to help you start imagining what might be possible. More precise figures cannot be stated without access to benchmark results or your own benchmark experience.

What we can expect to see is a significant improvement in response time and throughput with cache-hostile workloads due to the very fast Fibre Channel switched disk subsystems. A new caching algorithm, SARC, is used to guarantee a more efficient cache management than with the old LRU approach. This, in turn, will also positively contribute to better response times for all workloads which show pure locality of reference with the LRU approach.

### 12.5.3 Appropriate DS8000 size in z/OS environments

The potential of the architecture, its implementation and utilized technology allow for some projections at this point (though without having the hard figures at hand). Rules of thumb have the potential to be proven wrong. Therefore, you see here some recommendations on sizing which are rather conservative.

A fully configured ESS 800 Turbo with CKD volumes only and 16 FICON channels is good for the following:

- ▶ Over 30,000 I/Os per second
- ▶ More than 500 MB/sec aggregated sequential read throughput
- ▶ About 350 MB/sec sequential write throughput mirrored in cache

Without discrete DS8000 benchmark figures, a sizing approach to follow could be to propose how many ESS 800s might be consolidated into a DS8000 model. From that you can derive the number of ESS 750s, ESS F20s, and ESS E20s which can collapse into a DS8000. The older ESS models have a known relationship to the ESS 800.

Further considerations are, for example, the connection technology used, like ESCON, FICON, or FICON Express channels, and the number of channels.

Generally speaking, a properly configured DS8100 has the potential to provide the same or better numbers than two ESS 800s. Since the ESS 800 has the performance capabilities of two ESS F20s, a properly configured DS8100 can replace four ESS F20s. As the DS8000 series scales linearly, a well configured DS8300 has the potential to have the same or better numbers concerning I/O rates, sequential bandwidth, and response time than two DS8100 or four ESS 800s. Since the ESS 800 has roughly the performance potential of two ESS F20s, a corresponding number of ESS F20s can be consolidated. This applies also to the ESS 750, which has a similar performance behavior to that of an ESS F20.

Based on customer workload data, an IBM internal modelling tool can project how many ESS 800s to configure and then consolidate the ESS 800s to respective DS8000 models.

#### **Processor memory size considerations for z/OS environments**

Processor memory or cache in the DS8000 contributes to very high I/O rates and helps to minimize I/O response time.

It is not just the pure cache size which accounts for good performance figures. Economical use of cache, like 4 KB cache segments and smart, adaptive caching algorithms, are just as important to guarantee outstanding performance. This is implemented in the DS8000 series.

Processor memory or cache can grow to up to 256 GB in the DS8300 and to 128 GB for the DS8100. This processor memory is subdivided into a data in cache portion, which holds data in volatile memory, and a persistent part of the memory, which functions as NVS to hold DASD fast write (DFW) data until de-staged to disk.

Cache, or processor memory as a DS8000 term, is important to z/OS-based I/Os. Besides the potential for sharing data on a DS8000 processor memory level, it is the main contributor to good I/O performance. When coming from an existing disk storage server environment and you intend to consolidate this environment into DS8000s, follow these recommendations:

- ▶ Choose a cache size for the DS8000 series which has a similar ratio between cache size and disk storage to that of the currently used configuration.
- ▶ When you consolidate multiple disk storage servers, configure the sum of all cache from the source disk storage servers for the target DS8000 processor memory or cache size.

For example, consider replacing four ESS F20s with 3.2 TB and 16 GB cache each with a DS8100. The ratio between cache size and disk storage for the ESS F20 is 0.5% with 16 GB/3.2 TB. The new DS8100 is configured with 18 TB to consolidate 4 x 3.2 TB plus some extra capacity for growth. This would require 90 GB of cache to keep the original cache-to-disk storage ratio. Round up to the next available memory size, which is 128 GB for this DS8100 configuration.

This ratio of 0.5% cache to backstore is considered high performance for z/OS environments. Standard performance suggests a ratio of 0.2% cache to backstore ratio.

### **S/390 or zSeries channel consolidation**

The number of channels plays a role as well when sizing DS8000 configurations and when we know from where we are coming. The total number of channels used where you are coming from has to be considered in the following way:

- ▶ Use the same number of ESCON channels when the DS8000 has to be ESCON-connected as well. Since the maximum number of ESCON channel ports in a DS8100 is 32, and 64 for a DS8300, this also determines the consolidation factor. Note that with ESCON channels only, the potential of the DS8000 series cannot be fully exploited. You might consider taking this opportunity to change from ESCON to FICON to improve performance and throughput by multiple magnitudes compared to an ESCON world.
- ▶ When the connected host uses FICON channels with 1 Gbps technology and it will stay at this speed as determined by the host or switch ports, then keep the same number of FICON ports. So four ESS 800s with eight FICON channels each connected to IBM 9672 G5 or G6 servers might end up in a single DS8300 with 32 FICON channels.
- ▶ When migrating not only to DS8000 models, but also from 1 Gbps FICON to FICON Express channels at 2 Gbps, you can consider consolidating the number of channels to about 2/3 of the original number of channels. Use at least four FICON channels per DS8000. (By the way, when we write about FICON channels, we mean FICON ports in the disk storage servers.)
- ▶ Coming from FICON Express channels, you should keep a minimum of four FICON ports. You might consider using 25% fewer FICON ports in the DS8000 than the aggregated number of FICON 2 Gbps ports from the source environment. For example, when you consolidate two ESS 800s with eight FICON 2 Gbps ports each to a DS8100, plan for a minimum of 12 FICON ports on the DS8100.

Another example, with four ESS F20s each with eight FICON channels, might collapse into about 20 FICON ports when changing to a connectivity speed of 2 Gbps for the target DS8100 or DS8300.

### **Disk array sizing considerations for z/OS environments**

You can determine the number of ranks required not only based on the needed capacity, but also depending on the workload characteristics in terms of access density, read to write ratio, and hit rates.

You can approach this from the disk side and look at some basic disk figures. Fibre Channel disks, for example, at 10k RPM provide an average seek time of approximately 5 ms and an average latency of 3 ms. For transferring only a small block, the transfer time can be neglected. This is an average 8 ms per random disk I/O operation or 125 I/Os per second. A 15k RPM disk provides about 200 random I/Os per second for small block I/Os. A combined number of 8 disks is then good for 1,600 I/Os per second when they spin at 15k per minute. Reduce the number by 12.5% when you assume a spare drive in the 8 pack. Assume further a RAID-5 logic over the 8 packs.

Back at the host side, consider an example with 4,000 I/Os per second, a read to write ratio of 3 to 1, and 50% read cache hits. This leads to the following I/O numbers:

- ▶ 3,000 read I/Os per second.
- ▶ 1,500 read I/Os must read from disk.
- ▶ 1,000 writes with RAID-5 and assuming the worst case results in 4,000 disk I/Os.
- ▶ This totals to 4,500 disk I/Os.

With 15K RPM DDMs you need the equivalent of three 8 packs to satisfy the I/O load from the host for this example. Depending on the required capacity, you then decide the disk capacity, provided each desired disk capacity has 15k RPM. When the access density is less and you need more capacity, follow the example with higher capacity disks, which usually spin at a slower speed like 10k RPM.

In “Fibre Channel device adapter” on page 260 we stated that the disk storage subsystem DA port in a DS8000 has about five times the sequential throughput capability of what an ESS 800 DA port provides. Based on the 2 Gbps Fibre Channel connectivity to a DS8000 disk array this is approximately 200 MB/sec compared to the SSA port of an ESS disk array with 40 MB/sec. A Fibre Channel RAID array provides an external transfer rate of over 200 MB/sec. The sustained transfer rate varies. For a single disk drive, various disk vendors provide on their internet product sites the following numbers:

- ▶ 146 GB DDM with 10K RPM deliver a sustained transfer rate between 38 and 68 MB/sec or 53 MB/sec on average.
- ▶ 73 GB DDM with 15K RPM transfers between 50 and 75 MB/sec or 62.5 MB/sec on average.

The 73 GB DDMs have about 18% more sequential capability than the 146 GB DDM, but 60% more random I/O potential. The I/O characteristic is another aspect to consider when deciding the disk and disk array size. While this discussion is theoretical in approach, it is sufficient to get a first impression.

The IBM internal modelling tool, Disk Magic provides numbers which prove to be more firm. Disk Magic helps to model configurations based on customer workload data. An IBM representative can contact support personnel who will use Disk Magic to configure a DS8000 accordingly.

Use Capacity Magic to find out about usable disk capacity. For information on this internal IBM tool, contact your IBM representative.

## 12.5.4 Configuration recommendations for z/OS

We discuss briefly how to group ranks into extent pools and what the implications are with different grouping approaches. Note the independence of LSSs from ranks. Because an LSS is congruent with a z/OS LCU, we need to understand the implications. It is now possible to have volumes within the very same LCU, which is the very same LSS, but these volumes might reside in different ranks. A horizontal pooling approach assumes that volumes within a logical pool of volumes, like all DB2 volumes, are evenly spread across all ranks. This is independent of how these volumes are represented in LCUs. The following sections assume horizontal volume pooling across ranks, which might be congruent with LCUs when mapping ranks accordingly to LSSs.

### Configure one extent pool for each single rank

Figure 12-12 on page 270 displays some aspects regarding the disk subsystem within a DS8100:

- ▶ Chapter 5, “Virtualization concepts” on page 83 introduced the construct of an extent pool. When defining an extent pool, an affinity is created between this specific extent pool and a server. Due to the virtualization of the Fibre Channel switched disk subsystem you might consider creating as many extent pools as there are RAID ranks in the DS8000. This would then work similar to what is currently in the ESS. With this approach you can control the placement of each single volume and where it ends up in the disk subsystem.
- ▶ In the example in Figure 12-12 on page 270, each rank is in its own extent pool. The evenly numbered extent pools have an affinity to the left server, server 0. The odd number extent pools have an affinity to the right server, server 1. When a rank is subdivided into extents it gets assigned to its own extent pool.
- ▶ Now all volumes which are comprised of extents out of an extent pool have also a respective server affinity when scheduling I/Os to these volumes.
- ▶ This allows you to place certain volumes in specific ranks to avoid potential clustering of many high activity volumes within the same rank. You can create SMS storage groups which are congruent to these extent pools to ease the management effort of such a configuration. But you can still assign multiple storage groups when you are not concerned about the placement of less active volumes.



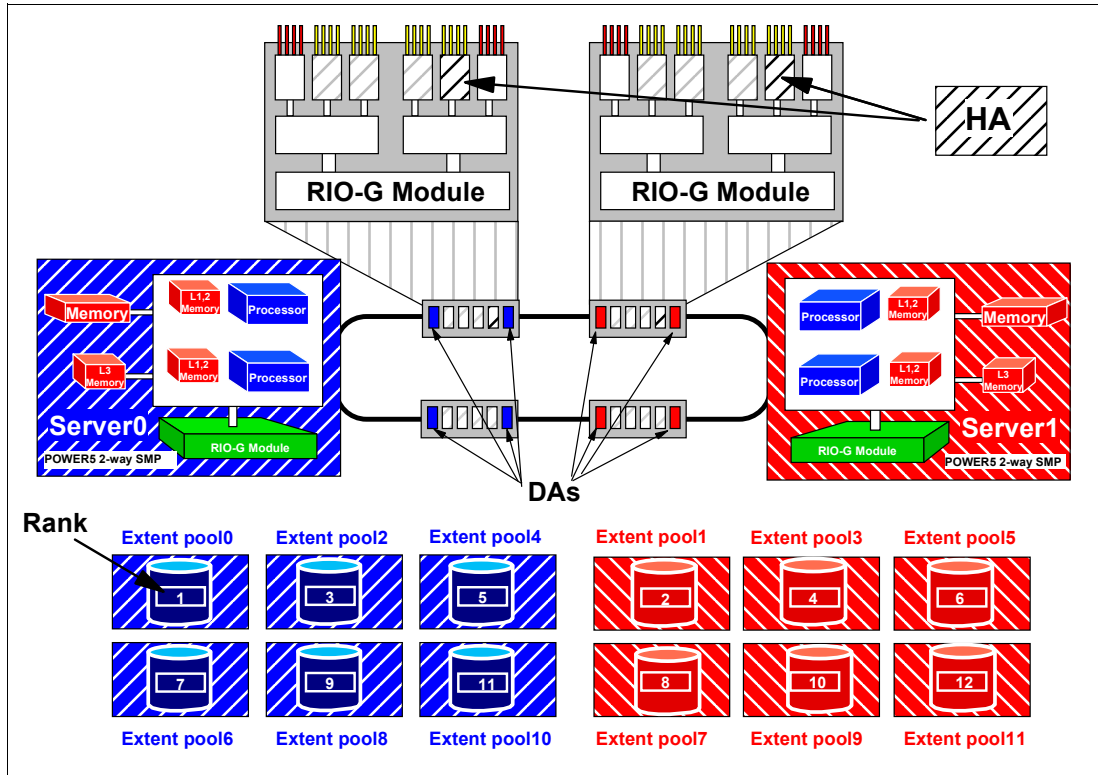


Figure 12-13 Extent pool affinity to processor complex with one extent pool for each rank

Figure 12-12 also indicates that there is no affinity nor a certain preference between HA and processor complexes or servers in the DS8000. In this example either one of the two HAs can address any volume in any of the ranks, which range here from rank number 1 to 12. Note there is an affinity of DAs to the processor complex. As Figure 12-3 on page 259 depicts, a DA pair connects to two switches respectively to two pairs of switches. The first DA of this DA pair connects to the left processor complex or server 0. The second DA of this DA pair connects to the other processor complex or server 1.

### Minimize the number of extent pools

The other extreme is to create just two extent pools when the DS8000 is configured as CKD storage only. You would then subdivide the disk subsystem evenly between both processor complexes or servers as Figure 12-13 shows.

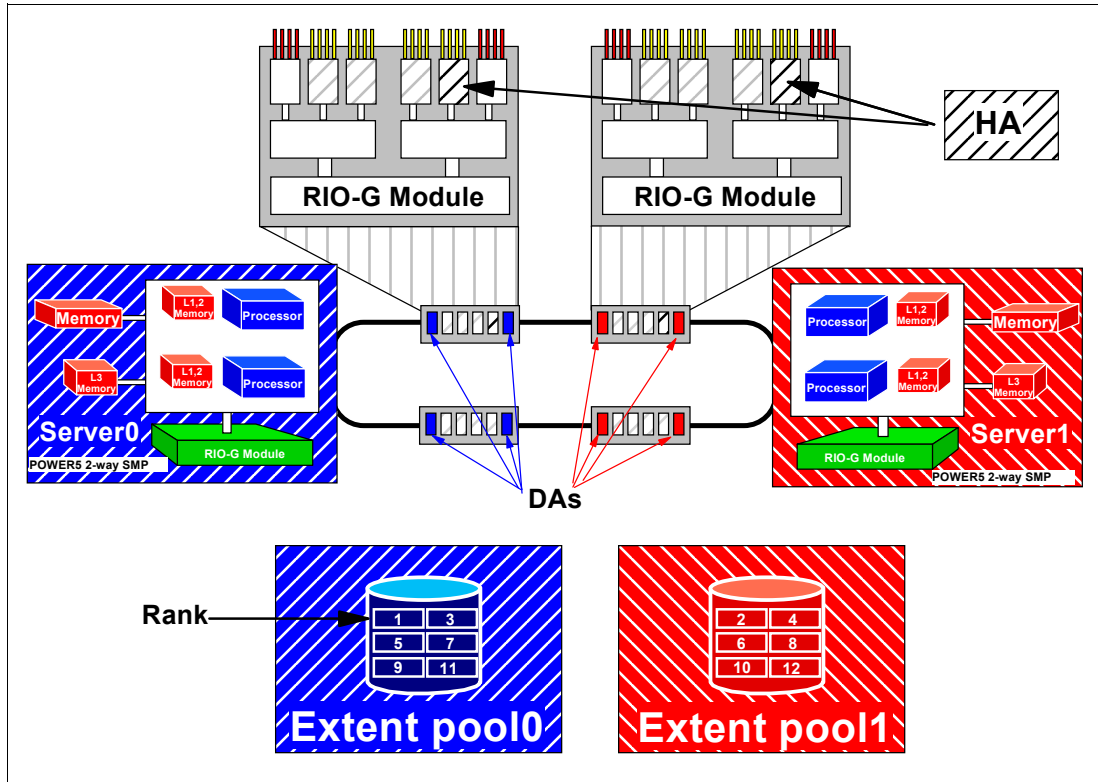


Figure 12-14 Extent pool affinity to processor complex with pooled ranks in two extent pools

Again what is obvious here is the affinity between all volumes residing in extent pool 0 to the left processor complex, server 0, and vice versa for the volumes residing in extent pool 1 and their affinity to the right processor complex or server 1.

When creating volumes there is no straightforward approach to place certain volumes into certain ranks. For example, when you create the first 20 DB2 logging volumes, they would be allocated in a consecutive fashion in the first rank. The concerned RAID site would then host all these 20 logging volumes. Now with AAL and the high performance and large bandwidth capabilities of the Fibre Channel switched disk storage subsystem, this might not be an issue. You may choose to control the placement of your most performance-critical volumes. This might lead to a compromise between both approaches, as Figure 12-14 suggests.

### Plan for a reasonable number of extent pools

Figure 12-15 presents a grouping of ranks into extent pools which follows a similar pattern and discussion as for grouping volumes or volume pools into SMS storage groups.

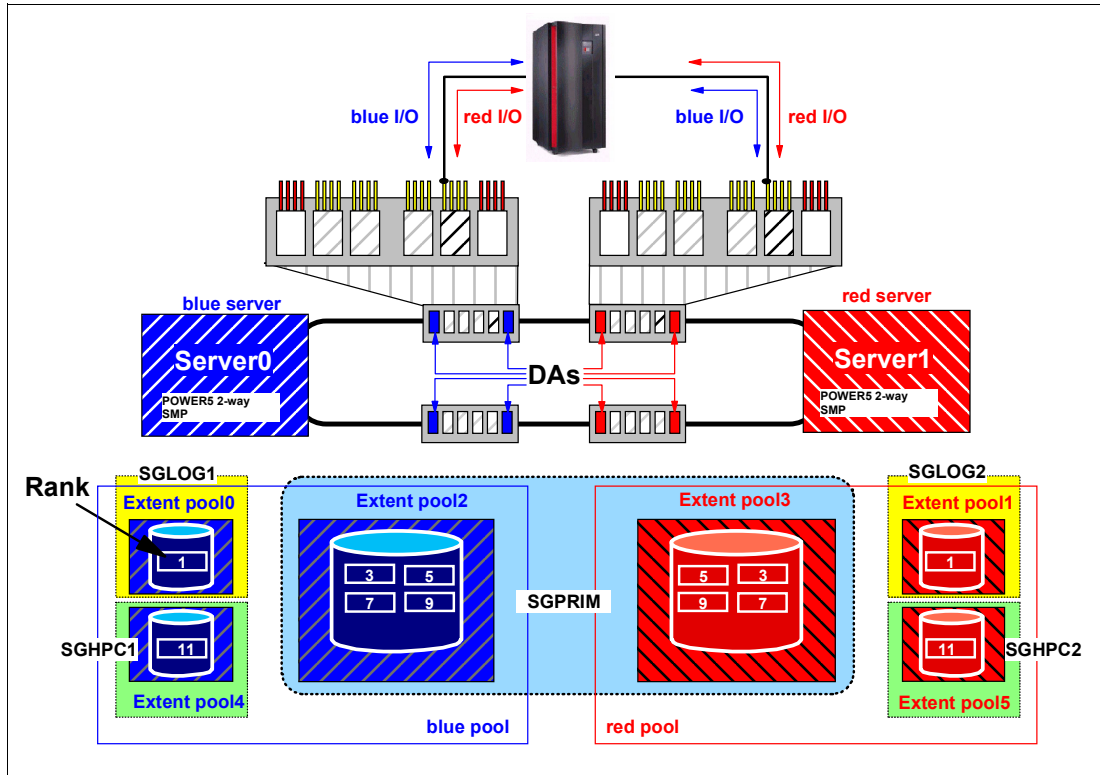


Figure 12-15 Mix of extent pools

Create two general extent pools for all the average workload and the majority of the volumes and subdivide these pools evenly between both processor complexes or servers. These pools contain the majority of the installed ranks in the DS8000. Then you might consider two or four smaller extent pools with dedicated ranks for high performance workloads and their respective volumes. You may consider defining storage groups accordingly which are congruent to the smaller extent pools.

Consider grouping the two larger extent pools into a single SMS storage group. SMS will eventually spread the workload evenly across both extent pools. This allows a system-managed approach to place data sets automatically in the right extent pools. With more than one DS8000 you might consider configuring each DS8000 in a uniform fashion. We recommend grouping all volumes from all the large extent pools into one large SMS storage group, SGPRIM. Cover the smaller, high performance extent pools through discrete SMS storage groups for each DS8000. With two of the configurations displayed in Figure 12-15, this ends up with one storage group, SGPRIM, and six smaller storage groups. SGLOG1 contains Extent pool0 in the first DS8100 and the same extent pool in the second DS8100. Similar considerations are true for SGLOG2. For example, in a dual logging database environment this allows you to assign SGLOG1 to the first logging volume and SGLOG2 for the second logging volume. For very demanding I/O rates and to satisfy a small set of volumes, you might consider keeping Extent pool 4 and Extent pool 5 in both DS8100s separate, through four distinct storage groups, SGHPC1-4.

Figure 12-14 shows, again, that there is no affinity between HA and processor complex or server. Each I/O enclosure connects to either processor complex. But there is an affinity between extent pool and processor complex and, therefore, an affinity between volumes and processor complex. This requires some attention, as outlined previously, when you define your volumes.

## 12.6 Summary

This high performance processor complex configuration is the base for a maximum of host I/O operations per second. The DS8000 Model 921 dual 2-way complex can handle an I/O rate of what about three to four ESS 800s can deliver at maximum speed. This allows you to consolidate three to four ESS 800s into a DS8100. As we saw before, the DS8000 series scales very well and in a linear fashion. Moving on from a 921 dual 2-way processor complex to a 922 dual 4-way processor complex not only doubles the number of POWER5 processors, but also doubles the performance. A properly configured DS8300 allows the same maximum number of host I/O operations per second as what up to six to eight ESS 800s provide.

With the introduction of the smart, switch-based Fibre Channel disk back end to overcome the FC-AL arbitration overhead, the DS8000 provides a comprehensive technology and an unmatched architecture to fulfill very demanding performance requirements, and it provides the highest RAS standards in the industry. This goes along with the 2 Gbps Fibre Channel device adapters (DA). The stage is set for a new record of aggregated bandwidth and I/O rates to the DS8000 high performance disk storage server family. A very promising road map guarantees a bright future for the DS8000 series.

# Implementation and management in the z/OS environment

In this part we discuss considerations for the DS8000 series when used in the z/OS environment. The topics include:

- ▶ z/OS Software
- ▶ Data migration in the z/OS environment





## zSeries software enhancements

This chapter discusses z/OS, z/VM, z/VSE and Transaction Processing Facility (TPF) software enhancements that support the DS8000 series. The enhancements include:

- ▶ Scalability support
- ▶ Large volume support
- ▶ Hardware configuration definition (HCD) to recognize the DS8000 series
- ▶ Performance statistics
- ▶ Resource Management Facility (RMF)

## 13.1 Software enhancements for the DS8000

A number of enhancements have been introduced into the z/OS, z/VM, z/VSE, VSE/ESA and TPF operating systems to support the DS8000. The enhancements are not just to support the DS8000, but also to provide additional benefits that are not specific to the DS8000.

## 13.2 z/OS enhancements

The DS8000 series simplifies system deployment by supporting major server platforms. The DS8000 will be supported on the following releases of the z/OS operating system and functional products:

- ▶ z/OS 1.4 and higher
- ▶ Device support facility (ICKDSF) Release 17
- ▶ Environmental Record Editing and Printing (EREP) 3.5
- ▶ DFSORT™

To exploit the DS8000 in exploitation mode, the Data Facilities Storage Management Subsystem (DFSMS) product of z/Series software is enhanced by way of a Small Programming Enhancement (SPE).

Hosts with operating system software levels prior to z/OS 1.4 are not supported. zLinux will only recognize the DS8000 as a 2105 device.

**Important:** Always review the latest Preventative Service Planning (PSP) 2107DEVICE bucket for software updates.

The PSP information can be found at:

<http://www-1.ibm.com/servers/resourceLink/svc03100.nsf?OpenDatabase>

The software enhancements have been introduced in the following areas:

- ▶ Scalability support
- ▶ Large Volume Support (LVS)
- ▶ Read availability mask support
- ▶ Initial Program Load (IPL) enhancements
- ▶ DS8000 definition to host software
- ▶ Read control unit and device recognition for DS8000
- ▶ New performance statistics
- ▶ Resource Management Facility (RMF)
- ▶ Migration considerations
- ▶ Coexistence considerations

### 13.2.1 Scalability support

The IOS recovery is designed to support a small number of devices per control unit. Today, a unit check is presented on all devices at failover. This does not scale well with a DS8000 that



has the capability to scale up to 63.75K devices. With the current support, we may have CPU or spin lock contention, or exhaust storage below the 16M line at device failover, or both.

Now with z/OS 1.4 and higher with the DS8000 software support, the IOS recovery has been improved by consolidating unit checks at an LSS level instead of each disconnected device. This consolidation will shorten the recovery time as a result of I/O errors. This enhancement is particularly important since the DS8000 has a much higher number of devices compared to the IBM 2105. In the IBM 2105, we have 4K devices and in the DS8000 we have up to 63.75K devices in a storage facility. With the enhanced scalability support, the following is achieved:

- ▶ Common storage (CSA) usage (above and below the 16M line) is reduced.
- ▶ IOS large block pool for error recovery processing and attention, and state change interrupt processing, is located above the 16M line, thus reducing storage demand below the 16M line.
- ▶ Unit control blocks (UCB) are pinned and event notification facility (ENF) signalling during channel path recovery.

### **Benefits of the scalability enhancements**

These scalability enhancements provide additional performance improvements by:

- ▶ Bypassing dynamic pathing validation in channel recovery for reduced recovery I/Os.
- ▶ Reducing elapsed time, by reducing the wait time in channel path recovery.

## **13.2.2 Large Volume Support (LVS)**

As we approach the limit of 64K UCBs, we need to find a way to stay within this limit. Today, with the IBM 2105 volumes, 32,760 cylinders are supported. This gives us the capability to remain within the 64K limit. But as today's storage facilities tend to expand to even larger capacities, we are approaching the 64K limit at a very fast rate. This leaves us no choice but to plan for even larger volumes sizes. Support has been enhanced to expand volumes to 65,520 cylinders, using existing 16 bit cylinder addressing. This is often referred to as 64K cylinder volumes. Components and products such as DADSM/CVAF, DFSMSdss, ICKDSF, and DFSORT, previously shipped with 32,760 cylinders, now also support 65,520 cylinders.

Check point restart processing now supports a checkpoint data set that resides partially or wholly above the 32,760 cylinder boundary.

With the new LVS volumes, the VTOC has the potential to grow very large. Callers such as DFSMSdss will have to read the entire VTOC to find the last allocated DSCB. In cases where the VTOC is very large, performance degradation will be experienced. A new interface is implemented to return the high allocated DSCB on volumes initialized with an INDEX VTOC. DFSMSdss uses this interface to limit VTOC searches and improve performance. The VTOC has to be within the first 64K-1 tracks, while the INDEX can be anywhere on the volume.

**Important:** Global Mirror for z/OS will be limited to 32,760 cylinder volumes only at General Availability.

## **13.2.3 Read availability mask support**

Dynamic CHPID Management (DCM) allows the customer to define a pool of channels that are managed by the system. The channels are added and deleted from control units based on workload importance and availability needs. DCM attempts to avoid single points of failure when adding or deleting a managed channel by not selecting an interface on the control unit on the same I/O card.

Today control unit single point of failure information is specified in a table and must be updated for each new control unit. Instead, we can use the Read Availability Mask (PSF/RSD) command to retrieve the information from the control unit. By doing this, there is no need to maintain a table for this information.

### 13.2.4 Initial Program Load (IPL) enhancements

During the IPL sequence the channel subsystem selects a channel path to read from the SYSRES device. Certain types of I/O errors on a channel path will cause the IPL to fail even though there are alternate channel paths which may work. For example, consider a situation where there is a bad switch link on the first path but good links on the other paths. In this case, you cannot IPL since the same faulty path is always chosen.

The channel subsystem and z/OS is enhanced to retry I/O over an alternate channel path. This will circumvent IPL failures, due to the selection of the same faulty path to read from the SYSRES device.

### 13.2.5 DS8000 definition to host software

The DASD Unit Information Module (UIM) is changed to define the new control unit type of 2107. The attachable device list will include 3380 and 3390 device types that include base and alias Parallel Access Volumes (PAV). HCD users have the option to select 2107 as a control unit type with 3380, 3380A, 3380B, 3390, 3390A, 3390B device types that can be defined to this control unit.

The definition of the 2107 control unit type will allow this storage facility to be uniquely reflected in the Hardware Configuration Manager (HCM). The definition of the 2107 control unit type in the HCD is not required to define an IBM 2107 storage facility to z/Series hosts. Existing IBM 2105 definitions could be used, but the number of LSSs will be limited to the same number as today in the IBM 2105.

The number of LSSs is increased from 16 to 255 per storage facility for the DS8000. The number of devices per LSS is still limited to 256. The number of CKD logical volumes is increased from 4K to 64K devices per storage facility.

### 13.2.6 Read control unit and device recognition for DS8000

The host system will inform the attached DS8000 of its capabilities, such that it supports native DS8000 control unit and devices. The DS8000 will then only return information that is supported by the attached host system using the self-description data, such as read data characteristics (RDC), sense ID, and read configuration data (RCD).

The following commands now display device type 2107 in their output:

- ▶ DEVSERV QDASD and PATHS command responses
- ▶ IDCAMS LISTDATA COUNTS, DSTATUS, STATUS and IDCAMS SETCACHE

Figure 13-1 on page 285, displays output from the DEVSERV QDASD command, which shows the device type as a 2107.

```

DS QD,9882,RDC,DCE
IEE459I 11.36.47 DEVSERV QDASD 053
UNIT VOLSER SCUTYPE DEVTYPE   CYL  SSID SCU-SERIAL DEV-SERIAL EF-CHK
9882 IN9882 2107921 2107000 65520 1011 0113-00511 0113-00511 **OK**
  READ DEVICE CHARACTERISTIC
2107E833900A5E80 FFF72024FFF0000F E000E5A205940222 1309067400000000
0000000000000000 24241F02DFEE0001 0677080F007F4A00 003C000000000000
  DCE AT V020D8A30
3878807100C457C0 00000000024B60D0 D800FFF0FFEF2424 1FF70800004A0000
00FC24DC9400F0FE 021F3C1E00038000 0000000000000000
***          1 DEVICE(S) MET THE SELECTION CRITERIA
***          0 DEVICE(S) FAILED EXTENDED FUNCTION CHECKING

```

Figure 13-1 Display output from DEVSERV QDASD command

Figure 13-2 displays output from the DEVSERV PATHS command, which shows the device type as a 2107.

```

DS P,9882
IEE459I 11.38.36 DEVSERV PATHS 055
UNIT DTYPE  M CNT VOLSER  CHPID=PATH STATUS
      RTYPE  SSID CFW TC   DFW  PIN  DC-STATE CCA  DDC  ALT  CU-TYPE
9882,33903 ,0,000,IN9882,29=+ 2B=+
      2107  1011 Y  YS.  YY.   N   SIMPLEX  02  02          2107
***** SYMBOL DEFINITIONS *****
O = ONLINE          + = PATH AVAILABLE

```

Figure 13-2 Display output from DEVSERV PATHS command

The following DFSMS components and products are updated to recognize real control unit and real device identifiers:

- ▶ Device support - system initialization
- ▶ DFSMSdss
- ▶ System Data Mover (SDM)
- ▶ Interactive Storage Management Facility (ISMF)
- ▶ Device Support Facility (ICKDSF)
- ▶ DFSORT
- ▶ EREP

## 13.2.7 New performance statistics

There are two new sets of performance statistics that will be reported by the DS8000. Since a logical volume is no longer allocated on a single RAID rank with a single RAID type or single device adapter pair, the performance data will be provided with a new set of rank performance statistics and extent pool statistics. The RAID RANK reports will no longer be reported by RMF and IDCAMS LISTDATA batch reports. RMF and IDCAMS LISTDATA is enhanced to report the new logical volume statistics that will be provided on the DS8000. These reports will consist of back-end counters that capture the activity between the cache and the ranks in the DS8000 for each individual logical volume. The new rank and extent pool statistics will be disk-system-wide instead of volume-wide only.

## LISTDATA COUNT

The RAID RANK counters report will not be available on the DS8000. These counters are being replaced with the new RANK and Extent Pool statistics that will be in the RMF reports. Figure 13-3 shows the RAID RANK Counters output that is available on the ESS Model 800.

```
2105 STORAGE CONTROL
                                SUBSYSTEM COUNTERS REPORT
                                RAID RANK COUNTERS
                                SUBSYSTEM ID X'3210'

RAID RANK ID X'0F00'
DEVICE ADAPTER ID X'14'
NUMBER OF HDDS IN RAID RANK      7
HDD SECTOR SIZE                   524

                                RAID RANK OPERATIONS
REQUESTS          I/O REQUESTS  RESP/TIME  SECTOR REQUESTS
  READ              4008062      0.015      422528304
  WRITE             1742359      0.019      209103590
IDCAMS SYSTEM SERVICES                                TIME: 11:36:18
IDC0001I FUNCTION COMPLETED, HIGHEST CONDITION CODE WAS 0

                                LEGEND
                                RAID RANK COUNTERS LEGEND
SUBSYSTEM ID      - SUBSYSTEM TO WHICH THE RAID RANK IS ATTACHED
NUMBER OF HDDS    - NUMBER OF HARD DISK DRIVES IN THE RAID RANK
HDD SECTOR SIZE   - SIZE IN BYTES OF A PHYSICAL HDD IO REQUEST
RANK OPERATIONS  - OPERATIONS ASSOCIATED WITH A RAID RANK
  I/O REQUESTS    - NUMBER OF I/O REQUESTS
  RESP/TIME       - AVERAGE RESPONSE IN (MS)
  SECTOR REQUESTS - NUMBER OF PHYSICAL HDD IO REQUESTS
  READ           - OPERATIONS CONTAINING AT LEAST ONE SEARCH OR READ COMMAND B
  WRITE          - OPERATIONS CONTAINING AT LEAST ONE WRITE COMMAND
IDCAMS SYSTEM SERVICES                                TIME: 11:40:04
IDC0001I FUNCTION COMPLETED, HIGHEST CONDITION CODE WAS 0
```

Figure 13-3 RAID RANK counters report that will not be available on DS8000

Figure 13-4 shows the LISTDATA COUNTS report output for the DS8000. This report shows the Segment Pool number and the back-end information.

```

LISTDATA COUNTS VOLUME(IN9882) UNIT(3390) DEVICE
IDCAMS SYSTEM SERVICES                                TIME: 11:18:19
                2107 STORAGE CONTROL
                SUBSYSTEM COUNTERS REPORT
                VOLUME IN9882  DEVICE ID X'02'
                SUBSYSTEM ID X'1011'
                CHANNEL OPERATIONS
                .....SEARCH/READ..... WRITE.....
                TOTAL  CACHE READ          TOTAL  DASDFW CACHE WRITE
REQUESTS
  NORMAL                2374          2368          146          141          141
  SEQUENTIAL             38            38            525          513          513
  CACHE FAST WRITE       0              0              0            N/A           0
TOTALS                  2412          2406          671          654          654
REQUESTS                CHANNEL OPERATIONS
  INHIBIT CACHE LOADING                0
  BYPASS CACHE                          0
TRANSFER OPERATIONS          DASD/CACHE  CACHE/DASD
  NORMAL                       17          2650
  SEQUENTIAL                    35            N/A
DASD FAST WRITE RETRIES                0
DEVICE STATUS      CACHING:          ACTIVE
                  DASD FAST WRITE: ACTIVE
                  DUPLEX PAIR:      NOT ESTABLISHED
DATA TRANSFERS     BYTES    RESP/TIME
  UNITS             128KB    16MS
  READ              405      330
  WRITE             144      223
SEGMENT POOL NUMBER X'0001'
BACK END DATA TRANSFERS  BYTES    RESP/TIME    REQUESTS
  UNITS                   128KB    16MS
  READ                    4493824  365          188
  WRITE                   45194476  38384        2778
IDCAMS SYSTEM SERVICES                                TIME: 11:18:19
IDC0001I FUNCTION COMPLETED, HIGHEST CONDITION CODE WAS 0

```

Figure 13-4 LISTDATA COUNTS report of DS8000

## LISTDATA STATUS

Figure 13-5 displays the output from the LISTDATA STATUS report. The output is the same, except that 2107 is now displayed for the storage control.

```

LISTDATA STATUS VOLUME(SHE200) UNIT(3390)
IDCAMS SYSTEM SERVICES                                TIME: 12:57:00

                2107 STORAGE CONTROL
                SUBSYSTEM STATUS REPORT
                VOLUME SHE200  DEVICE ID X'00'
                SUBSYSTEM ID X'3205'
                .....CAPACITY IN BYTES.....
                SUBSYSTEM STORAGE          NONVOLATILE STORAGE
CONFIGURED                8388608K                196608K
AVAILABLE                 7541792K                N/A
PINNED                    0                        0
OFFLINE                   8192K                    N/A
RETURNED STATUS:  0-19 01000B01 000000C0 00000080 00000073 14200000
                  20-39 00000000 20000003 00030000 00000000 00003205
SUBSYSTEM CACHING STATUS:  ACTIVE
SD CACHING CONDITIONS:   CACHE FAST WRITE ACTIVE
NVS STATUS:              ACTIVE
DEVICES WITH STATISTICS  12
STATISTIC SETS/DEVICE    1
DEVICE STATUS
  FOR DEVICE ID X'00'    CACHING:          ACTIVE
                        DASD FAST WRITE:  ACTIVE
                        DUPLEX PAIR:     NOT ESTABLISHED

IDCAMS SYSTEM SERVICES                                TIME: 12:57:00
IDC0001I FUNCTION COMPLETED, HIGHEST CONDITION CODE WAS 0

```

Figure 13-5 LISTDATA STATUS report

## LISTDATA PINNED

Figure 13-6 shows the output from a LISTDATA PINNED report. The output is the same, except that 2107 is now displayed for the storage control.

```

LISTDATA PINNED VOLUME(SHE200) UNIT(3390) DEVICE
IDCAMS SYSTEM SERVICES

                2107 STORAGE CONTROL
                PINNED TRACK REPORT
                VOLUME SHE200  DEVICE ID X'00'
                SUBSYSTEM ID X'3205'
CCHH      TYPE                DATA SET NAME
00090009  RETRIABLE           IBMUSER.DATASET.NUMBER1
0011000A  NON-RETRIABLE       IBMUSER.DATASET.NUMBER2
0013000B  NVS, CACHE AVAILABLE  IBMUSER.DATASET.NUMBER3
0014000C  CACHE COPY DEFECTIVE      IBMUSER.DATASET.NUMBER4

LISTDATA PINNED VOLUME(EV9LIA) UNIT(3390) ALL
IDC11560I NO PINNED TRACKS EXIST FOR VOLUME EV9LIA

```

Figure 13-6 LISTDATA PINNED report

**Note:** Rank and extent pool statistics will not be reported by IDCAMS LISTDATA batch reports.

## SETCACHE

The DASD fast write attributes cannot be changed to OFF status on the DS8000. Figure 13-7 on page 289, displays the messages you will receive when the IDCAMS SETCACHE parameters of DEVICE, DFW, SUBSYSTEM, or NVS with OFF are specified.

```
SETCACHE DEVICE OFF FILE(FILEX)
IDC31562I THE DEVICE PARAMETER IS NOT AVAILABLE FOR THE SPECIFIED
IDC31562I SUBSYSTEM OR DEVICE
IDC3003I FUNCTION TERMINATED. CONDITION CODE IS 12

    SETCACHE DFW OFF FILE(FILEX)
IDC31562I THE DASDFASTWRITE PARAMETER IS NOT AVAILABLE FOR THE
IDC31562I SPECIFIED SUBSYSTEM OR DEVICE
IDC3003I FUNCTION TERMINATED. CONDITION CODE IS 12

    SETCACHE SUBSYSTEM OFF FILE(FILEX)
IDC31562I THE SUBSYSTEM PARAMETER IS NOT AVAILABLE FOR THE SPECIFIED
IDC31562I SUBSYSTEM OR DEVICE
IDC3003I FUNCTION TERMINATED. CONDITION CODE IS 12

    SETCACHE NVS OFF FILE(FILEX)
IDC31562I THE NVS PARAMETER IS NOT AVAILABLE FOR THE SPECIFIED
IDC31562I SUBSYSTEM OR DEVICE
IDC3003I FUNCTION TERMINATED. CONDITION CODE IS 12
```

Figure 13-7 SETCACHE options

All other parameters should be accepted as they are today on the IBM 2105. For example, setting device caching ON is accepted, but has no affect on the subsystem.

### 13.2.8 Resource Management Facility (RMF)

RMF support for the DS8000 is added via an SPE (APAR number OA06476, PTFs UA90079 and UA90080). RMF is enhanced to provide Monitor I and III support for the IBM TotalStorage DS family. The ESS Disk Systems Postprocessor report now contains two new sections: Extent Pool Statistics and Rank Statistics. These statistics are generated from SMF record 74 subtype 8:

- ▶ The ESS Extent Pool Statistics section provides capacity and performance information about allocated disk space. For each extent pool, it shows the real capacity and the number of real extents.
- ▶ The ESS Rank Statistics section provides measurements about read and write operations in each rank of an extent pool. It also shows the number of arrays and the array width of all ranks. These values show the current configuration. The wider the rank, the more performance capability it has. By changing these values in your configuration, you can influence the throughput of your work.

Also, new response and transfer statistics are available with the Postprocessor Cache Activity report generated from SMF record 74 subtype 5. These statistics are provided at the subsystem level in the Cache Subsystem Activity report and at the volume level in the Cache Device Activity report. In detail, RMF provides the average response time and byte transfer rate per read and write requests. These statistics are shown for the I/O activity (called host adapter activity) and transfer activity from hard disk to cache and vice-versa (called disk activity).

## 13.2.9 Migration considerations

A DS8000 will be supported as an IBM 2105 for z/OS systems without the DFSMS and z/OS SPE installed. This will allow customers to *roll* the SPE to each system in a sysplex without having to take a sysplex-wide outage. An IPL will have to be taken to activate the DFSMS and z/OS portions of this support.

## 13.2.10 Coexistence considerations

Support for the DS8000 running in 2105 mode on systems without this SPE installed will be provided. It will consist of the recognition of the DS8000 real control unit type and device codes when it runs in 2105 emulation on these down-level systems.

Input/Output definition files (IODF) created by HCD may be shared on systems that do not have this SPE installed. Additionally, existing IODF files that define IBM 2105 control unit records for a 2107 subsystem should be able to be used as long as 16 or fewer logical subsystems are configured in the DS8000.

## 13.3 z/VM enhancements

z/VM is an IBM operating system that supplies a virtual machine to each logged-on user. The DS8000 will be supported on z/VM 4.4 and higher.

**Important:** Always review the latest Preventative Service Planning (PSP) 2107DEVICE bucket for software updates.

The PSP information can be found at:

<http://www-1.ibm.com/servers/resourceLink/svc03100.nsf?OpenDatabase>

## 13.4 z/VSE enhancements

z/VSE is a system that consists of a basic operating system (VSE/Advanced Functions) and any IBM-supplied and user-written programs required to meet the data processing needs of a user. VSE and the hardware that it controls form a complete computing system. The DS8000 will be supported on:

- ▶ z/VSE 3.1 and higher
- ▶ VSE/ESA 2.7 and higher

**Important:** Always review the latest Preventative Service Planning (PSP) 2107DEVICE bucket for software updates.

The PSP information can be found at:

<http://www-1.ibm.com/servers/resourceLink/svc03100.nsf?OpenDatabase>

VSE/ESA does not support 64K LVS for the DS8000.



## 13.5 TPF enhancements

TPF is an IBM platform for high volume, online transaction processing. It is used by industries demanding large transaction volumes, such as airlines and banks. The DS8000 will be supported on TPF 4.1 and higher.

**Important:** Always review the latest Preventative Service Planning (PSP) 2107DEVICE bucket for software updates.

The PSP information can be found at:

<http://www-1.ibm.com/servers/resourceLink/svc03100.nsf?OpenDatabase>





## Data migration in zSeries environments

This chapter describes several methods for migrating data from existing disk storage servers onto the DS8000 disk storage server images. This includes migrating data from the ESS 2105 as well as from other disk storage servers to the new DS8000 disk storage server images. The focus is on z/OS environments. The following topics are covered from a planning standpoint:

- ▶ Data migration objectives in z/OS environments
- ▶ Data migration based on physical migration
- ▶ Data migration based on logical migration
- ▶ Combination of physical and logical data migration
- ▶ z/VM and VSE/ESA data migration

This chapter does not provide a detailed step-by-step migration process description, which would fill another book. There is a more detailed outline for a system-managed storage environment.

## 14.1 Define migration objectives in z/OS environments

Data migration is an important activity that needs to be planned well to ensure the success of the DS8000 implementation. Because today's business environment does not allow you to interrupt data processing services, it is crucial to make the data migration onto the new storage servers as smooth as possible. The configuration changes and the actual data migration ought to be transparent to the users and applications, with no or only minimal impact on data availability. This requires you to plan for non-disruptive migration methods and also to guarantee data integrity at any time during the migration process.

### 14.1.1 Consolidate storage subsystems

In the course of a data migration you might consider consolidating volumes and reducing the number of storage servers.

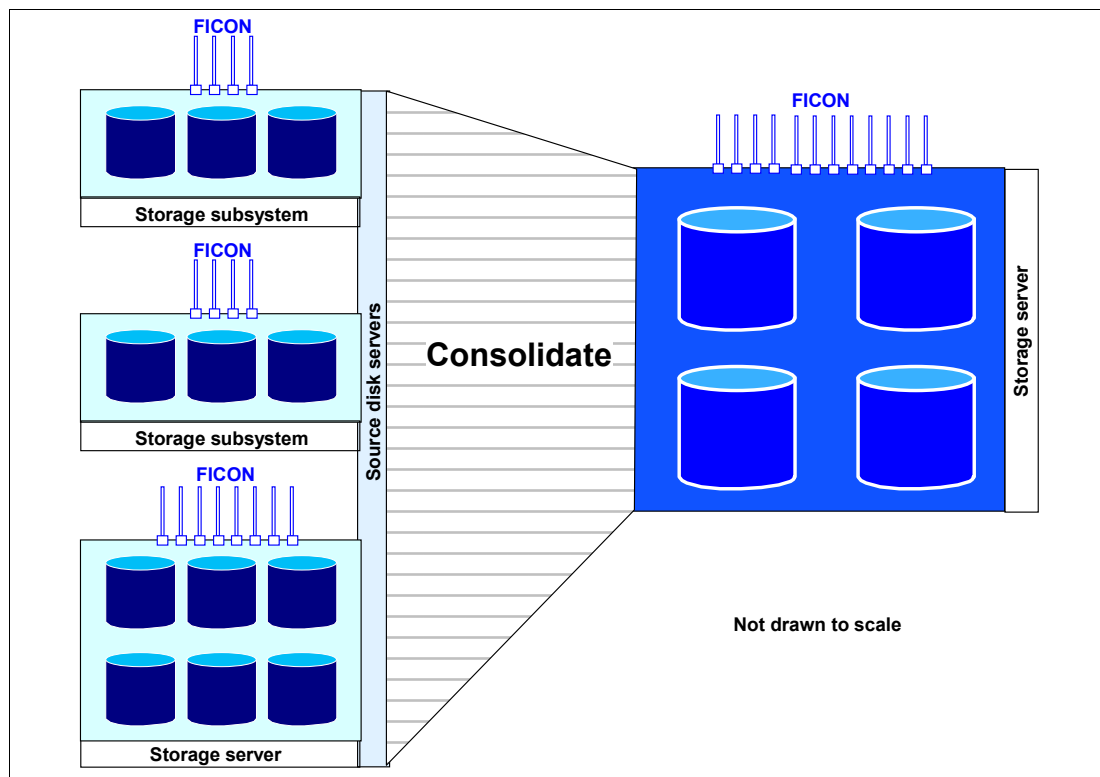


Figure 14-1 Consolidation opportunities when migrating to DS8000

A DS8100 or DS8300 can initially contain up to 64 logical control units (LCUs), and this will later increase to 255 LCUs within a DS8000 storage server image. This allows the DS8100 or DS8300 to simulate up to 255 times an IBM 3390-6 control unit image with 256 devices within each single control unit image. 255 times 256 equals the maximum number of devices a storage server image can manage, which is 65,280 volumes. This suggests consolidation of multiple control units into a single DS8100 or DS8300 disk storage server.

On the other hand, you might consider having multiple physical footprints to disperse volumes for availability. Possible target configurations may conflict with recommendations from software platforms, where physical placement of specific data sets into separate physical hardware is recommended for availability. For example, the manual *z/OS V1 R6.0 MVS Setting up a Sysplex*, SA22-7625-09, states:

“To avoid a single point of failure in a sysplex, IBM recommends that, for all couple data sets, you create an alternate couple datasets on a different device, control unit, and channel from the primary.”

You may interpret this to mean not just having separate LCUs, but rather separate physical control units, if possible.

Besides careful planning and depending on the configuration complexity, it might take weeks or months to complete the planning and to perform the actual migration.

## 14.1.2 Consolidate logical volumes

Another aspect of consolidation within a data migration effort might be to plan for larger volume sizes and consolidate or fold more than one source volume to a new and bigger target volume. Up until now, most customers relied on a standard size volume, which is a 3390-3 with 3,339 cylinders per 3390-3. With the affinity of volumes to RAID arrays and the underlying 8-packs, the increasing DDM size in the ESS 2105 made it impossible to utilize the full RAID array capacity of a RAID rank with 146 GB and even larger DDMs, when choosing 3390-3 volumes. The full RAID array capacity could only be utilized when changing to bigger volume sizes to reduce the number of volumes per RAID array or LSS.

This is not necessary any more with the DS8000 models due to their flexibility to configure back-end storage independently from the DDM sizes. There is no affinity any longer between an LSS and physical disk sets. The allocation of a volume happens in extents or increments of the size of a 3390-1 volume or 1113 cylinders, and a 3390-3 consists of exactly three extents from an extent pool. A 3390-9 with 10017 cylinders is comprised of exactly 9 extents out of an extent pool. There is no affinity any longer to a disk set for a logical volume like a 3390-3 or any other sized 3390 volume. Despite the fact that it is not required any longer to move from a 3390-3 volume type to some bigger volume type with the DS8000, you might still want to plan in the course of a data migration to reduce the number of volumes.

### Considerations for new logical volume size

Although the newly announced DS8000 storage servers now support logical 3390 volumes of up to 65520 cylinders or 982800 tracks, you might still want to plan for a standard sized logical volume which is smaller than 65520 cylinders. Consider that full volume operations will take longer to copy or dump when the volume size is increased. This also applies to the first full initial volume replication for Metro Mirror, when creating the pairs, as well as for XRC. A compromise has to be planned for, which might be different from configuration to configuration.

In a pure system-managed storage environment with no full volume operations any more, except for migration perhaps, a volume size of 30051 cylinders might be fine for most of the data. This is the space of nine 3390-3 volumes or exactly 27 extents out of an extent pool. This would guarantee that all space is fully utilized when staying with increments of the standard extent size, which is 1113 cylinders or the equivalent of a 3390-1 model. When you plan for a larger volume size consider a multiple of a 3390-3 model, if possible.

With installations still performing full volume operations to a significant extent, you might want to plan for smaller volumes. For example, to full dump nine volumes of 3390-3 in parallel will most likely have a shorter elapsed time than dumping a single volume with the capacity of nine 3390-3 volumes. In such an environment a standard 3390-9 volume might still be appropriate. Although a 3390-9 volume has three times the capacity of a 3390-3, when changing from ESCON to FICON the throughput increases roughly by a factor of 10 and shortens the elapsed time, especially for highly sequential I/O.

Note that some data set types cannot grow beyond 64K tracks. When coming from 3390-3 and staying with 50,085 tracks of a model 3 this limit of 64K tracks extent allocation is not an issue. When you already work with larger volumes you are familiar with these considerations. But it may be a surprise to you without this experience.

Data set types that can exploit more than 64K tracks are:

- ▶ Physical sequential extended format, DSORG=PSE
- ▶ VSAM
- ▶ PDSE
- ▶ HFS
- ▶ Page data sets

Note that extended format data sets are required to reside on system-managed volumes. The extended format attribute is usually assigned through a data class construct. Example 14-1 displays a list of data classes which have the extended attribute.

*Example 14-1 Data Classes with EF attribute*

```

                                DATA CLASS LIST
Command ===>

CDS Name : SYS1.DFSMS.SCDS

Enter Line Operators below:

      LINE      DATACLAS LAST TIME      EXTENDED
      OPERATOR  NAME      MODIFIED  DATA SET NAME TYPE ADDRESSABILITY  COMPACTION
      ---(1)---- --(2)--- --(25)--- -----(26)----- -----(27)----- ---(28)---
              EXT      10:45      EXTENDED REQUIRED YES              NO
              EXTCOMP  07:30      EXTENDED REQUIRED YES              YES
              STRIPE   06:59      EXTENDED REQUIRED NO               ----
              XRCJ     07:09      EXTENDED REQUIRED NO               ----
      -----      -----      -----      -----      -----      BOTTOM OF DATA -----

```

Data set types that cannot grow beyond 64K tracks are:

- ▶ Physical sequential, DSORG=PS
- ▶ Partitioned, DSORG=PO
- ▶ Direct, DSORG=DA
- ▶ JES spool

Data set types that must be allocated in the first 64K tracks on the volume are:

- ▶ VTOC
- ▶ VTOC Index
- ▶ JES spool without APAR OW49317

Furthermore, consider adjusting the VTOC size when moving to large volumes. It all depends on the number of data sets allocated on the volume. You might consider defining the same size for VTOC, VTOC Index, and VVDS for all volumes with the same capacity, although the sizing requirements for database volumes are usually different than for TSO volumes. There are some further recommendations in *Device Support Facilities User's Guide and Reference, Release 17, GC35-0033-27, Appendix C. VTOC Index*, and in the following sections.

### Dynamic Parallel Access Volumes required for large volumes

Utilizing big volumes will require you to use dynamic Parallel Access Volumes (PAV) to allow many concurrent accesses to the very same volume concurrently. On a big volume, with 9 times the capacity of a 3390-3, we see about 9 times as many concurrent I/Os as what we

see on a single 3390-3. Or, to put it differently, we may see on a single volume as many concurrent I/Os as we see on nine 3390-3 volumes. Despite the PAV support, it still might be necessary to balance disk storage activities across disk storage server images.

With the DS8000 you will have the flexibility to define LSSs of exactly the size you desire rather than being constrained by RAID rank topology. This means that you define the number of PAVs or alias devices that you need, rather than a number dictated by the RAID rank size. Assuming you decide to create an LSS with 256 devices each, then the volume size you decide upon determines how many alias devices to configure for WLM management. Table 14-1 provides a proposal for a configuration with batch and transaction workload using about 50 percent of the total disk space. It further assumes that all volumes are evenly and horizontally spread across all LSSs and that all these volumes are system-managed.

*Table 14-1 Suggested numbers of base and alias devices in an LSS with 256 devices*

Volume size in cyl	Number base dev	Number alias dev	Capacity/LSS
3,339 (3390-3)	170 - 192	86 - 64	550 - 480 GB
10,017 (3390-9)	128 - 170	128 - 86	1 - 1.5 TB
30,051 (3390-9+)	86 - 128	170 - 128	2.3 - 3.4 TB
30.051 (3390-9+)	86 - 128	170 - 128	2.3 - 3.4 TB

In a FICON environment this is supposed to be a conservative ratio between base and alias volumes, which has the potential to provide a perfect compromise between utilized device numbers and minimizing IOS queueing time. IOS queueing time is usually an indication of volume contention due to more than one I/O request at a time to the very same volume.

Note that the configured LSS capacity is determined by the number of base devices chosen and has no link any more to the actual rank size. The numbers in Table 14-1 are rounded. Note also that the number of devices per LSS is still limited to a maximum of 256.

### 14.1.3 Keep source and target volume size at the current size

When the number of volumes does not reach the current zSeries limit of 64K volumes or is significantly below this limit, you might stay with 3390-3 as a standard and avoid the additional migration effort at this time. Introducing solutions based on Remote Mirror and Copy, the number of devices to plan for can quickly reach the limits of number of devices supported in current zSeries servers.

Volume consolidation is still a bit painful because it requires logical data set movement to properly maintain catalog entries for most of the data. Only the first volume can be copied through a full volume operation to a larger target volume. After that full copy operation, the VTOC on the target volume needs to be adjusted to hold many more entries than the first source volume. Another consideration for the first full volume copy operation is that the volume names must be maintained on the new volume because full physical volume operations do not maintain catalog entries. Otherwise you would not be able to locate the data sets any more in a system-managed environment, which always goes through the catalog to locate data sets and orients data set location solely on volume serial numbers.

Referring to certain volumes or volume lists in JCL may need changes to the JCL to modify or remove these volumes or lists of volumes. Independent of whether these volumes are system-managed with the guaranteed space attribute or non-managed volumes, JCL most likely needs to be adjusted to reflect the new volume names.

### 14.1.4 Summary of data migration objectives

To summarize the objective of data migration, it might be feasible to not just migrate the data from existing storage subsystems to the new storage server images, but also to consolidate source storage subsystems to one or fewer target storage servers. A second migration layer might be to consolidate multiple source volumes to larger target volumes, which is also called volume folding. The latter is in general more difficult to do and requires data migration on a data set level. Some down time is needed to move the remaining data sets, which are usually open and active 24 hours every day. If you, for example, consolidate a storage group with 1,000 volumes to 200 volumes, the service interruptions could be few and long or frequent and brief.

## 14.2 Data migration based on physical migration

Physical migration here refers to physical full volume operations, which in turn require the same device geometry on the source and target volume. The target volume capacity is equal to the source volume capacity or larger. The device geometry is defined by the track capacity and the number of tracks per cylinder. The same device geometry means that source and target devices have the same track capacity and the same number of tracks per cylinder. Usually this is not an issue because over time the device geometry of the IBM 3390 volume has become a quasi standard and most installations have used this standard. For organizations still using other device geometry (for example, 3380), it might be worthwhile to consider a device geometry conversion, if possible. This requires moving the data on a logical level, which is on a data set level, and allows for reblocking during the migration from 3380 to 3390.

Utilizing physical full volume operations is possible through the following software-, microcode-, and hardware-based functions:

- ▶ Software-based
  - DFSMSdss
  - TDMF
  - FDRPAS
- ▶ Software- and hardware-based:
  - zSeries Piper - uses currently a zSeries Multiprise® server with ESCON attachment only
  - z/OS Global Mirror (XRC)
- ▶ Hardware- and microcode-based:
  - Global Mirror
  - Global Copy
  - FlashCopy in combination with either Global Mirror or Global Copy, or both
  - Metro/Global Copy

The following section discusses DFSMSdss and the Remote Copy-based approaches in more detail.

### 14.2.1 Physical migration with DFSMSdss and other storage software

Full volume copy through the COPY command copies all data between like devices from a source volume to a target volume. The target volume might be bigger than the source but cannot be smaller than the source volume. You have to keep the same volume name and the same volume serial number (VOLSER) on the target volume; otherwise, the data set cannot be located any more via catalog locates. This is achieved through the COPYVOLID



parameter. When the target volume is larger than the source volume it is usually necessary to adjust the VTOC size on the target volume with the ICKDSF REFORMAT REFVTOC command to make the entire volume size accessible to the system.

DFSMSdss also provides full DUMP and full RESTORE commands. With the DUMP command an entire volume is copied to tape cartridges and can then be restored from tape via the RESTORE command to the new source volume. During that time all data sets on that volume are not available to the application in order to keep data consistency between when the DUMP is run and when the RESTORE command completes. The advantage of this method is that it creates a copy which offers fail-back capabilities. When source and target disk servers are not available at the same time for migration, this might be a feasible approach to migrate the data over to the new hardware.

DFSMSdss is optimized to read and write data sequentially as fast as possible. Besides optimized channel programs, which always use the latest enhancements the hardware and microcode provides, DFSMSdss also allows a highly parallel I/O pattern, which is achieved either through the PARALLEL keyword within a single job/step or you can submit more than one DFSMSdss job and run several DFSMSdss jobs in parallel.

TDMF and FDRPAS provide concurrent full volume migration capabilities, which are best described as remote copy functions for migration based on software that allows a controlled switch-over to the new target volume. As a general rule, these might be considered when the number of volumes to be migrated is in the hundreds rather than in the range of thousands of volumes to be migrated. With large migration tasks, the number of volumes has to be broken down to smaller volume sets so that the migration can happen in a controlled fashion. This lengthens the migration period, so if possible, other approaches might be considered.

Both software products are usually associated with fees or service-based fees. When the number of volumes is in the range of up to a few hundred, then standard software like DFSMSdss is an option. But DFSMSdss-based migration does not automatically switch over to the target volumes and usually requires some weekend efforts to complete. DFSMSdss is standard software and part of z/OS, so there are no extra costs for software.

The choice of which software approach to take depends on the business requirements and service levels which the data center has to follow. The least disruptive approach is to provide software packages that switch in a controlled and transparent manner over to the target device, like TDMF and FDRPAS do. When brief service interruptions can be tolerated, then the standard software is still a popular solution.

## 14.2.2 Software- and hardware-based data migration

Piper z/OS (an IBM IGS service) and z/OS Global Mirror are tools for data migration that are based on software, which in turn relies on specific hardware or microcode support. This section outlines these two popular approaches to migrate data.

### Data migration with Piper for z/OS

IBM offers a migration service using the Piper tool, which is a combination of FDRPAS as the software used in a migration server (which is part of the service offering) that is connected to the customer configuration during the migration.

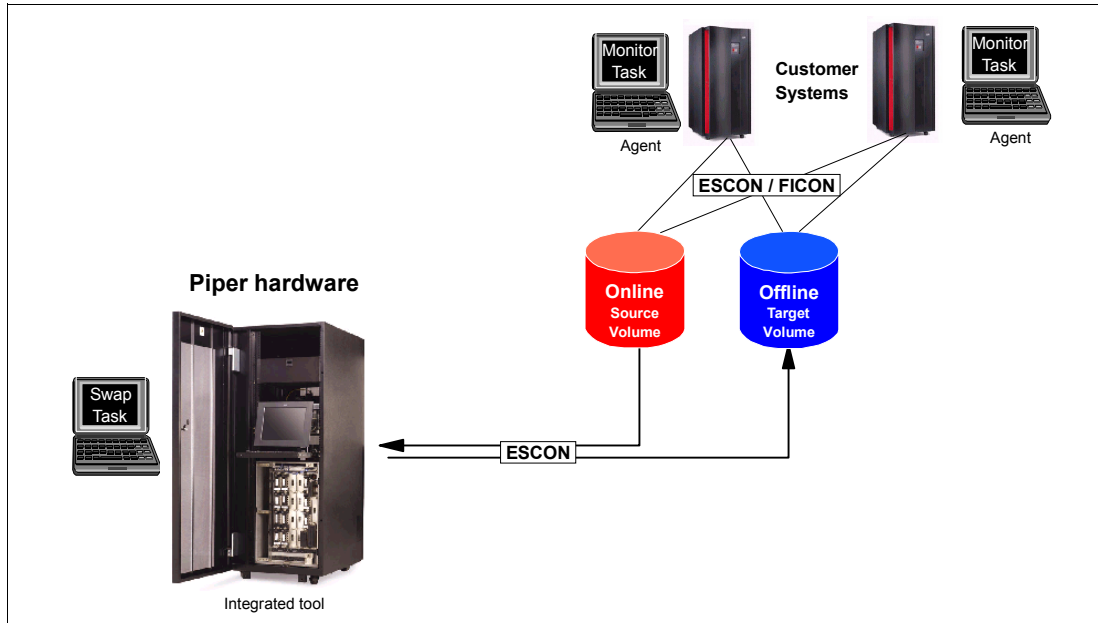


Figure 14-2 Piper for z/OS environment configuration

Currently this server is a Multiprise 3000, which can connect through ESCON channels only. This will exclude this approach to migrate data to the DS6000, which only provides a pure FICON or Fibre Channel connectivity. It can be used, though, for the DS8000, which still allows you to connect to the ESCON infrastructure through supported ESCON host adapters. IBM plans to enhance the Piper server with FICON channel capable hardware to allow migration to FICON-only environments. Currently the IBM Piper migration service offering includes the following, which IBM provides:

- ▶ S/390 Multiprise 3000
- ▶ An ESCON director with 16 ports to connect to the customer z/Series-based fabric
- ▶ Preloaded FDRPAS migration software
- ▶ 19 inch rack enclosure
- ▶ Agent tasks which need to be installed in the customer systems

An FDRPAS master task runs on the Piper CPU, which coordinates all activities. Monitor tasks are required on the customer's systems to monitor and coordinate with the swap task in the Piper CPU.

The advantages of this Piper-based migration offering are:

- ▶ Simple installation without the need for IPLs on the customer side.
- ▶ Transparent data migration without interruption to connected application hosts and no application down time.
- ▶ Parameter-controlled activity which can be dynamically modified at any time to pace the migration.
- ▶ Suspend/resume of migration at any time without exposing data integrity.
- ▶ Migration during usual business hours for convenient management of the migration process.
- ▶ Independent of hardware vendor and suited for all S/390 or zSeries attached disk storage.
- ▶ Supports Parallel Access Volume handling.

Most of these benefits also apply to migration efforts controlled by the customer when utilizing TDMF or FDRPAS in customer-managed systems.

Piper for z/OS is an IGS service offering which relieves the customer of the actual migration process and requires customer involvement only in the planning and preparation phase. The actual migration is transparent to the customer's application hosts and Piper even manages a concurrent switch-over to the target volumes. Piper is neutral to the disk storage vendors and works for all devices supported under z/OS or OS/390.

### Data migration with z/OS Global Mirror

Another alternative is z/OS Global Mirror (XRC). XRC is an asynchronous solution that has a mode for disaster recovery (DR) solutions as well as a particular migration mode of SESSIONTYPE(MIGRATE). This mode does not require the customer to plan for a JOURNALs configuration at the secondary site, which is mandatory for DR solutions based on XRC.

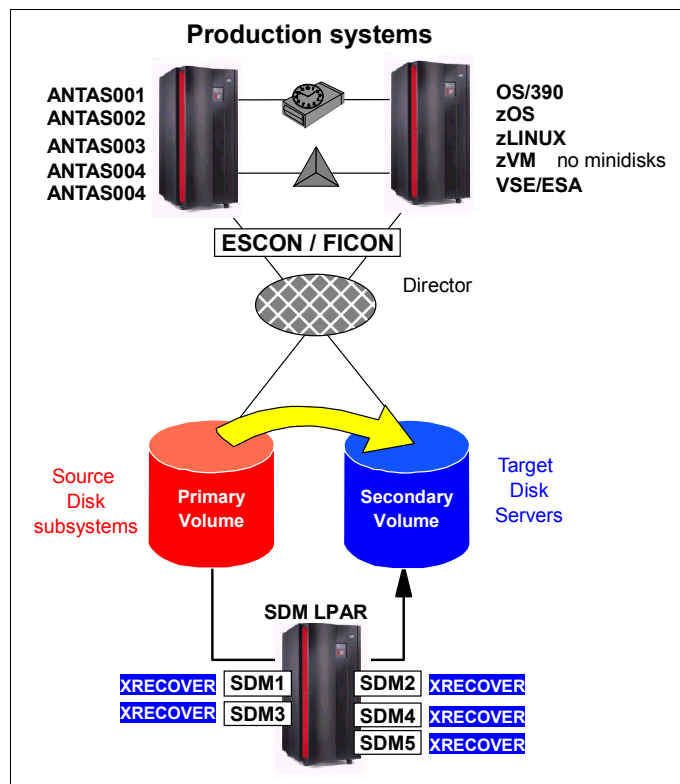


Figure 14-3 Data Migration with z/OS Global Mirror (XRC)

XRC allows for an online data migration approach, which is almost non-disruptive. The application only needs to shut down for the actual switch to the target volumes in the new disk storage server. It can restart immediately to connect to the new disk storage server after the XRC secondary volumes have been relabeled by a single XRC command per XRC session, XRECOVER.

In addition to transparent data replication, the advantage of XRC is extreme scalability. Figure 14-3 shows that XRC can run either in existing system images or in a dedicated LPAR. Each image can host up to five System Data Movers (SDM). An SDM is an address space (ANTAS00x) that is started by a respective XSTART command. A reasonable number of XRC volume pairs that a single SDM can manage is in the range of 1,500 to 2,000 volume pairs. With up to five SDMs within a system image, this totals approximately 10,000 volume pairs.

This requires an adequate bandwidth for the connectivity between the disk storage servers to the system image which hosts the SDMs. Because XRC in migration mode stores the data through, it mainly requires channel bandwidth and SDM tends to monopolize its channels. Therefore, the approach with dedicated channel resources is an advantage over a shared channel configuration and would almost not impact the application I/Os.

XRC requires disk storage subsystems which support XRC primary volumes through the microcode. Currently only IBM- or HDS-based controllers support XRC as a primary or source disk subsystem. As an exception, this does not apply to the IBM RVA storage controller, which does not support XRC as a primary XRC device. Also, EMC does not provide XRC support at the XRC primary site.

### 14.2.3 Hardware- or microcode-based migration

Hardware- and microcode-based migration through remote copy is usually only possible between like hardware, so using remote copy through microcode is not possible with different disks from vendor A at the source site and disks from vendor B at the target site. Therefore, we discuss only what is possible for IBM disk storage servers using remote copy or Peer-to-Peer Remote Copy (PPRC) and its variations.

Remote copy approaches with Global Mirror, Metro Mirror, Metro/Global Copy, and Global Copy allow the primary and secondary site to be any combination of ESS 750s, ESS 800s and DS6000s or DS8000s.

Although IBM did not announce support of Metro/Global Copy for the DS8000, it will work for certain migration scenarios, as outlined in the next section, when the DS8000 disk server receives the data as the Global Copy secondary disk server. In this case the DS8000 holds the Global Copy secondary volumes within a plain Global Copy relationship with the respective ESS as the Global Copy primary disk server.

#### Bridge from ESCON to FICON with Metro/Global Copy

The ESS Model E20 and Model F20 do not support PPRC over Fibre Channel links, but only PPRC based on PPRC ESCON links. In contrast, the newly announced disk storage servers support only PPRC over Fibre Channel links and do not support PPRC ESCON links. The ESS Model 800 supports both PPRC link technologies.

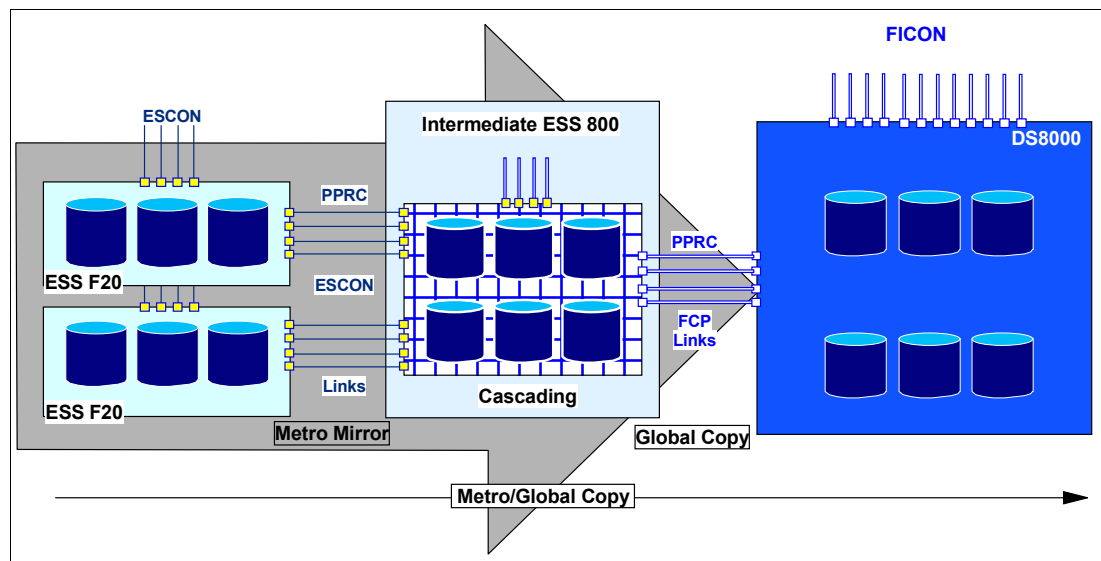


Figure 14-4 Intermediate ESS 800 used to migrate data with PPRC over ESCON from older models

To utilize the advantage of PPRC with concurrent data migration on a physical volume level from older ESS models like the ESS F20, an ESS 800 (or in less active configurations an ESS 750) might be used during the migration period to bridge from a PPRC ESCON link storage server to the new disk storage server which supports only PPRC over FCP links. The approach is Metro/Global Copy with the ESS 800 in between hosting the cascaded volumes, which are PPRC secondary volumes for Metro Mirror and at the same time also PPRC primary volumes for the Global Copy configuration. It is recommended that you connect the intermediate ESS to a host whether it is over ESCON channels or FICON channels. This allows the ESS 800 to off-load messages to the host as well as to manage the PPRC volumes within the intermediate ESS during the migration period.

The actual setup and management might be performed through the ESS GUI with its Copy Services application. Another possibility is to manage such a cascaded configuration with host-based software like ICKDSF or TSO commands, when the TSO command support for cascaded volumes is available. Otherwise use a combination of TSO commands and ICKDSF for just defining the cascaded bit when setting up Global Copy between the intermediate ESS 800 and the target DS8000 disk server.

Dynamic Address Switching (P/DAS) for a non-disruptive application I/O switch from the old hardware to the volumes in the new hardware is not possible in this configuration because P/DAS requires the PPRC secondary volumes to be in a DUPLEX state. PPRC-XD stays, by definition, always in PENDING state. Theoretically it is possible to switch from PPRC-XD to Synchronous PPRC and replicate the data twice over Synchronous PPRC, which may impose significant impact to write I/O to the old hardware. It is usually quicker and less difficult, when a brief application down time is acceptable, to switch from the old hardware to the volumes in the new hardware.

Again this approach is only possible from IBM ESS to IBM DS6000 or IBM DS8000 disk storage servers and it requires the same size or larger PPRC secondary volumes with the same device geometry.

### Data migration with Metro Mirror or Global Copy

A variation to the approach discussed above is to use straightforward Metro Mirror or Global Copy from the ESS 750 or ESS 800 to the DS8000.

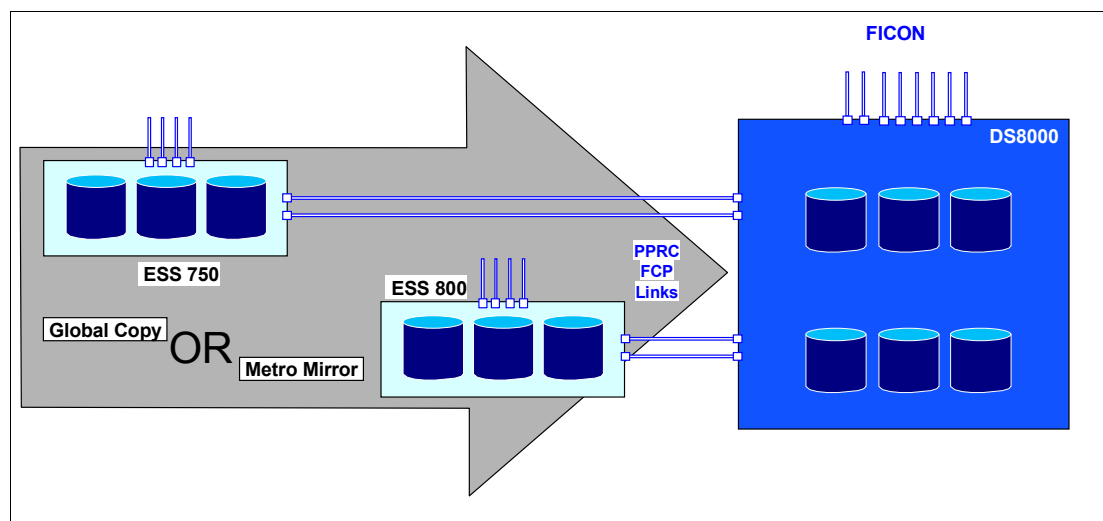


Figure 14-5 Metro Mirror or Global Copy from ESS 750 or ESS 800 to DS8000

Metro Mirror (Synchronous PPRC) provides data consistency at any time once the volumes are in full DUPLEX state, although through its synchronous approach it imposes a slight

impact to the application write I/Os at the source storage subsystems from where we migrate the data. This assumes a local data migration and that the distance is within the supported Metro Mirror distance for PPRC over FCP links. You can switch the application I/Os any time from the old to the new disk configuration. This requires you to quiesce or shut down the application server and restart the application servers after terminating the PPRC configuration. The restart uses a modified I/O definition file but the volume serial number will stay the same and all data will be located correctly through catalog locate processing.

Global Copy (PPRC-XD) on the other side does not impact the application write I/Os due to its asynchronous data replication, but the drawback here is that it does not guarantee data consistency at the receiving site. Before switching from the source equipment to the new target equipment, make sure all data is replicated to the receiving site. This can be forced by dynamically switching from Global Copy to Metro Mirror. When all primary volumes are in full DUPLEX state the source and target disk servers contain the same data at any time. At this point prepare and execute the switchover to the new disk storage server.

Instead of switching from Global Copy to Metro Mirror, you might stop the applications and shut down the application servers. Then check that all data is replicated to the target disk server. This might be a bit labor-intensive in a large environment without the help of automation scripts. Basically you would check each individual primary volume (= source volume) that all data is copied over. Through the current Copy Services application on an ESS 800 this would look like Figure 14-6.

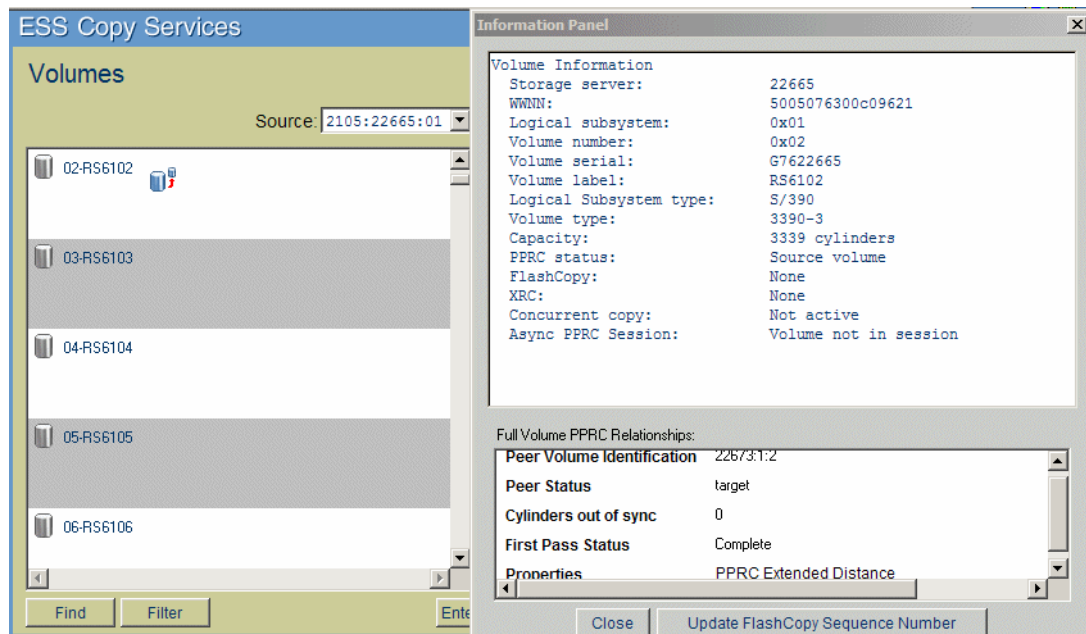


Figure 14-6 Check with Global Copy whether all data replicated to the new volume

This approach is not really practical. ICKDSF allows you to query the status of a Global Copy primary volume and displays the amount of data which is not yet replicated, as shown in Example 14-2.

*Example 14-2 Check through ICKDSF query commands that all data is replicated*

```
PPRCOPY DDNAME(DD02) QUERY
ICK00700I DEVICE INFORMATION FOR 6102 IS CURRENTLY AS FOLLOWS:
      PHYSICAL DEVICE = 3390
      STORAGE CONTROLLER = 2105
      STORAGE CONTROL DESCRIPTOR = E8
```

```

        DEVICE DESCRIPTOR = 0A
        ADDITIONAL DEVICE INFORMATION = 4A000035
ICK04030I DEVICE IS A PEER TO PEER REMOTE COPY VOLUME

```

QUERY REMOTE COPY - VOLUME

DEVICE	LEVEL	STATE	PATH	STATUS	(PRIMARY)		(SECONDARY)	
					SSID SER #	CCA LSS	SSID SER #	CCA LSS
6102	PRIMARY	PENDING	XD	ACTIVE	6100 22665	02 01	8100 22673	02 01

PATHS	SAID/DEST	STATUS	DESCRIPTION
1	00A4 0020	13	ESTABLISHED Fibre Channel PATH

IF PENDING/SUSPEND: **COUNT OF TRACKS REMAINING TO BE COPIED = 2147**

```

ICK02206I PPRCOPY QUERY FUNCTION COMPLETED SUCCESSFULLY
ICK00001I FUNCTION COMPLETED, HIGHEST CONDITION CODE WAS 0
        00:33:54    11/19/04

```

After all applications are stopped or shut down, Global Copy eventually replicates all data across and the respective count is zero, as shown in Example 14-3.

*Example 14-3 All data is replicated*

**PPRCOPY DDNAME(DD02) QUERY**

```

ICK00700I DEVICE INFORMATION FOR 6102 IS CURRENTLY AS FOLLOWS:
        PHYSICAL DEVICE = 3390
        STORAGE CONTROLLER = 2105
        STORAGE CONTROL DESCRIPTOR = E8
        DEVICE DESCRIPTOR = 0A
        ADDITIONAL DEVICE INFORMATION = 4A000035
ICK04030I DEVICE IS A PEER TO PEER REMOTE COPY VOLUME

```

QUERY REMOTE COPY - VOLUME

DEVICE	LEVEL	STATE	PATH	STATUS	(PRIMARY)		(SECONDARY)	
					SSID SER #	CCA LSS	SSID SER #	CCA LSS
6102	PRIMARY	PENDING	XD	ACTIVE	6100 22665	02 01	8100 22673	02 01

PATHS	SAID/DEST	STATUS	DESCRIPTION
1	00A4 0020	13	ESTABLISHED Fibre Channel PATH

IF PENDING/SUSPEND: **COUNT OF TRACKS REMAINING TO BE COPIED = 0**

```

ICK02206I PPRCOPY QUERY FUNCTION COMPLETED SUCCESSFULLY
ICK00001I FUNCTION COMPLETED, HIGHEST CONDITION CODE WAS 0
        00:34:10    11/19/04

```

As these examples demonstrate, it is a bit hard to read and to find in the ICKDSF line output. Probably some REXX-based procedure might have to scan through the ICKDSF SYSPRINT output, which is directed to a data set. ICKDSF asks for a JCL DD statement for each single volume to query when the volume is ONLINE to the system. This is not very handy. So, we

are back to the TSO CQUERY command which does not need any additional JCL statements and does not care whether the volume is ONLINE or OFFLINE to the system. TSO provides a nicely formatted output, as the following examples display, which still might be directed to an output data set, so some REXX procedure could find the specific numbers.

*Example 14-4 TSO CQUERY to identify data replication status on primary volume*

```
***** PPRC REMOTE COPY CQUERY - VOLUME *****
*
*                               (PRIMARY) (SECONDARY) *
*                               SSID CCA LSS SSID CCA LSS*
*DEVICE  LEVEL      STATE      PATH STATUS  SERIAL#    SERIAL#    *
*-----  -
* 6102  PRIMARY..  PENDING.XD  ACTIVE..  6100 02 01  8100 02 01 *
*          CRIT(NO).....          CGRPLB(NO). 000000022665 000000022673*
* PATHS  PFCA SFCA STATUS: DESCRIPTION
* -----
*   1   00A4 0020   13   PATH ESTABLISHED...
*          ---- ----   00   NO PATH.....
*          ---- ----   00   NO PATH.....
*          ---- ----   00   NO PATH.....
* IF STATE = PENDING/SUSPEND:   TRACKS OUT OF SYNC =   491
*                               TRACKS ON VOLUME   =  50085
*                               PERCENT OF COPY COMPLETE =  99%
* SUBSYSTEM          WNNN          LIC LEVEL
* -----
* PRIMARY....  5005076300C09621          2.4.01.0062
* SECONDARY.1  5005076300C09629
*****
ANTP0001I CQUERY COMMAND COMPLETED FOR DEVICE 6102. COMPLETION CODE: 00
```

A quick way is to open the data set which received the SYSTSPRT output from TSO in batch and exclude all data. Then an F COMPLETE ALL would only display a single line per volume and you could quickly spot when a volume is not 100% complete. This manual approach might be acceptable for a one-time effort when migrating data in small or medium sized configurations.

*Example 14-5 All data is replicated*

```
***** PPRC REMOTE COPY CQUERY - VOLUME *****
*
*                               (PRIMARY) (SECONDARY) *
*                               SSID CCA LSS SSID CCA LSS*
*DEVICE  LEVEL      STATE      PATH STATUS  SERIAL#    SERIAL#    *
*-----  -
* 6102  PRIMARY..  PENDING.XD  ACTIVE..  6100 02 01  8100 02 01 *
*          CRIT(NO).....          CGRPLB(NO). 000000022665 000000022673*
* PATHS  PFCA SFCA STATUS: DESCRIPTION
* -----
*   1   00A4 0020   13   PATH ESTABLISHED...
*          ---- ----   00   NO PATH.....
*          ---- ----   00   NO PATH.....
*          ---- ----   00   NO PATH.....
*                               PERCENT OF COPY COMPLETE = 100%
* SUBSYSTEM          WNNN          LIC LEVEL
* -----
* PRIMARY....  5005076300C09621          2.4.01.0062
* SECONDARY.1  5005076300C09629
*****
ANTP0001I CQUERY COMMAND COMPLETED FOR DEVICE 6102. COMPLETION CODE: 00
```



Again all these approaches to utilize microcode-based mirroring capabilities require the right hardware as source and target disk servers.

For completeness it is pointed out that Global Mirror is also an option to migrate data from an ESS 750 or ESS 800 to a DS8000. This might apply to certain cases at the receiving site which require consistent data at any time, although Global Copy is used for the actual data movement. Please note that the consistent copy in the new disk server is not concurrent with the primary copy except if the application is stopped and all data is replicated.

Another approach to migrate data beyond the scope of volumes is to use software which migrates data on a logical or data set level and locates the data sets through their respective catalog entries. This approach is outlined in the following section.

## 14.3 Data migration based on logical migration

Data migration based on logical migration is a data set by data set migration which maintains catalog entries according to the data movement between volumes and, therefore, is not a volume-based migration. This is the cleanest way to migrate data and also allows device conversion from, for example, 3380 to 3390. It also supports transparently multivolume data sets. Logical data migration is a software-only approach and does not rely on certain volume characteristics nor on device geometries.

The following software products and components support logical data migration:

- ▶ DFSMS allocation management
- ▶ Allocation management by CA-ALLOC
- ▶ DFSMSdss
- ▶ DFSMSShsm™
- ▶ FDR
- ▶ System utilities like:
  - IDCAMS with REPRO, EXPORT/IMPORT commands
  - IEBCOPY to migrate Partitioned Data Sets (PDS) or Partitioned Data Sets Extended (PDSE)
  - ICEGENER as part of DFSORT which can handle sequential data but not VSAM data sets, which also applies to IEBGENER
- ▶ CA-Favor
- ▶ CA-DISK or ASM2
- ▶ Database utilities for data which is managed by certain database managers like DB2 or IMS™. CICS® as a transaction manager usually uses VSAM data sets.

### 14.3.1 Data Set Services Utility

DFSMSdss is a common utility which can be used, at this time, not for physical full volume operations, but for data set level operations. Pointing to certain input volumes, DFSMSdss can also move data sets in a logical fashion off of certain source volumes.

*Example 14-6 DFSMSdss for logical data migration from certain input volumes*

---

```
//MIGRATE EXEC PGM=ADRSSU
//SYSPRINT DD SYSOUT=*
//* ----- FROM VOLUMES ----- ***
//IN001 DD UNIT=3390,VOL=SER=AAAAAA,DISP=SHR
//IN002 DD UNIT=3390,VOL=SER=BBBBBB,DISP=SHR
//IN003 DD UNIT=3390,VOL=SER=CCCCCC,DISP=SHR
//SYSIN DD *
COPY DS(INC(**) EXCLUDE(SYS1.VTOCIX.*,SYS1.VVDS.*)) -
```

```
LIDD(IN001,IN002,IN003) -  
DELETE PURGE CATALOG SELECTMULTI(ANY) SPHERE-  
WAIT(0,0) ADMIN OPT(3) CANCELERROR  
/*  
//
```

---

Example 14-6 depicts how to migrate all data sets from certain volumes. The keyword is LOGINDDDNAME, or LIDD, which identifies the volumes from where the data is to be picked. There is no output volume specified, although it is also possible to distribute all data sets from the input or source volumes to a larger output volume or to more than one output volume. Example 14-6 assumes a system-managed environment. Here the system would automatically place the output data sets according to what the Automatic Class Selection Routine (ACS) assigns. The next section is more specific on how this approach works.

### 14.3.2 Hierarchical Storage Manager, DFSMSHsm

If the number of volumes is rather small and the source volume's data sets are all managed by DFSMS/MVS, which implies that the data set is managed through Management Classes, you might consider utilizing DFSMSHsm to migrate all data sets off the source volumes. During recall the data set would then be allocated onto the new volumes, provided that the source volumes are disabled for new allocations in a system-managed environment. This approach is not practical for large data migration scenarios.

Another variation of DFSMSHsm to migrate data on a logical data set level is to utilize Aggregate Backup and Recovery (ABARS). ABARS allows you, through powerful data selection filters, to copy all data sets onto cartridges (ABACKUP). Subsequent ARECOVER then restores the aggregate. This is nothing but the group of data sets which have been selected before through filtering, put back onto the new DS8000 storage servers. For more information, see the redbook *DFSMSHsm ABARS and Mainstar Solutions*, SG24-5089.

### 14.3.3 System utilities

System utilities don't play an import role any longer in large scale migrations since DFSMSdss incorporated the capability to manage all data set types for copy and move operations in a very efficient way. In some instances DFSMSdss still calls some system utility functions, though. IDCAMS REPRO and EXPORT/IMPORT still have some popularity. During both operations, REPRO or EXPORT/IMPORT which includes a REPRO step, perform data set reorganization. System utilities are an integral part of z/OS and free of charge.

In a VSE/ESA environment, VSE/VSAM functions are still used, such as REPRO or EXPORT/IMPORT, to copy and move data sets. There are also other software vendor utilities to manage data migration. An elegant migration approach that can be used when disks are system-managed is outlined in the following section. This approach assumes that the old and new disk servers can be configured concurrently to the concerned host servers.

### 14.3.4 Data migration within the System-managed storage environment

System-managed storage (SMS) manages volumes and storage groups (SG) and allocates data sets based on a policy or rules for certain storage groups (SG). Certain volumes within an SG or within more than one SG are eligible for new data set allocations. These SMS volumes and SMS SGs also maintain a status that decides whether this volume or that SG is eligible to receive new data set allocations.

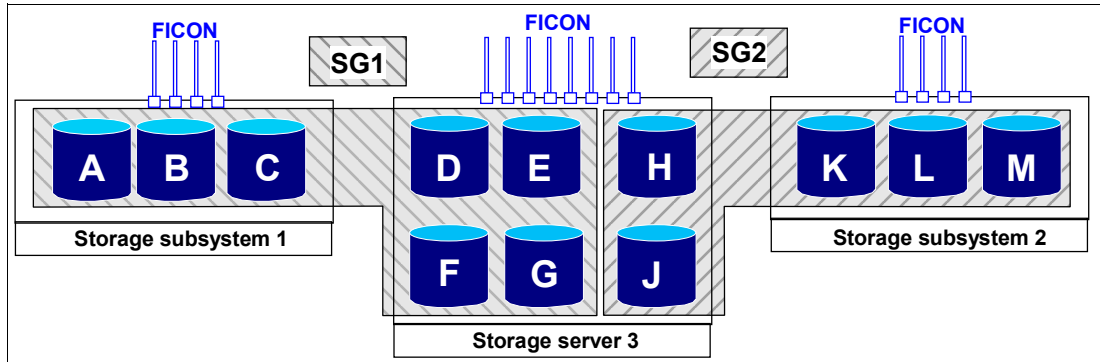


Figure 14-7 SMS Storage Groups - migration source environment

To explain this approach, Figure 14-7 contains two SGs, SG1 and SG2, which are distributed over three storage controllers. Assume these three storage controllers are going to be consolidated into a new DS8000 storage server, and that the number of volumes will be consolidated from 12 down to four, with their respective capacity as displayed in Figure 14-7.

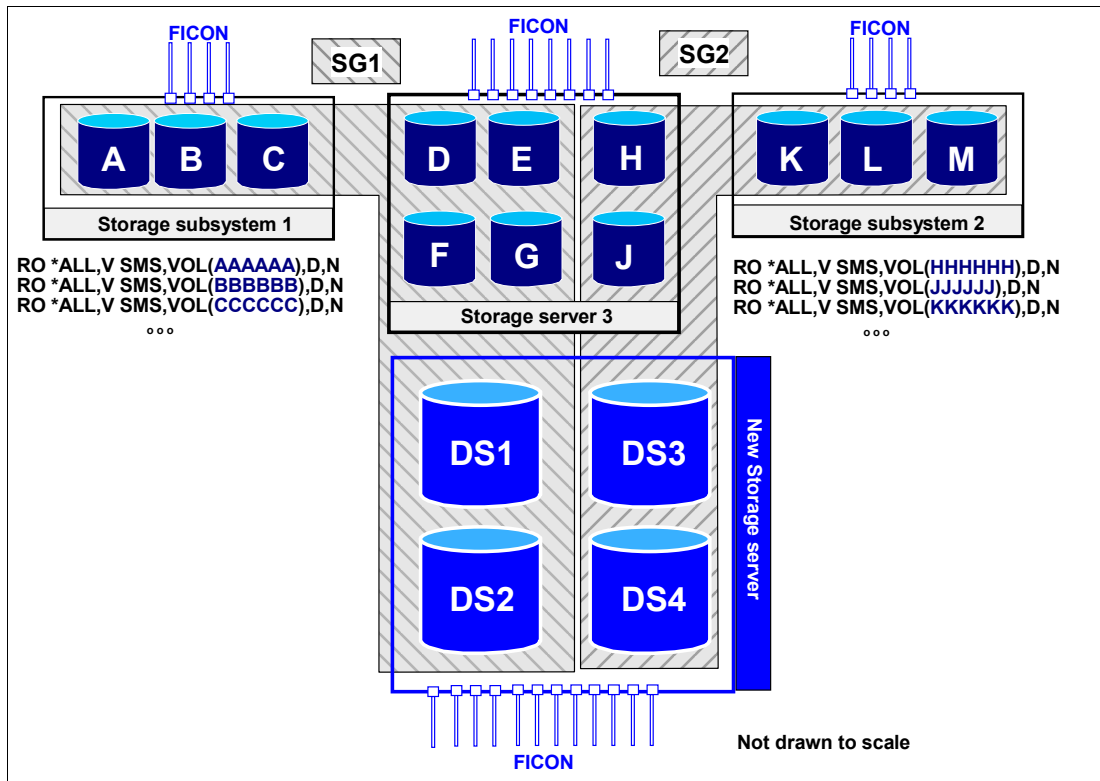


Figure 14-8 Utilize SMS Storage Group and Volume status to direct all new allocation to new volumes

When both the old hardware and the new hardware can be installed and connected to the host servers, the new volumes are integrated into the existing SGs, SG1 and SG2. Then change the SMS volume's status to disable volumes A to G for new allocations. This is possible through an SMS system command and is propagated to all systems within this SMS-plex. DISABLE, NEW or D,N implies that all new allocations which are directed to SG1 can no longer happen on volumes A to G but must go to volumes DS1 or DS2. This allows you to gradually migrate the data from the old devices A through G onto the new and larger devices DS1 and DS2. Example 14-7 shows a pre-defined batch job to modify the volume status of all of the old volumes in SG2 which are displayed in Figure 14-8.

*Example 14-7 Execute SMS modify commands through pre-defined job*

---

```
//DISNEW JOB , ' DISABLE NEW ',MSGCLASS=T,CLASS=B,
//          MSGLEVEL=(1,1),REGION=OM,USER=&SYSUID
/*JOBPARM S=VSL1
// COMMAND 'V SMS,VOL(HHHHHH,*),D,N'
// COMMAND 'V SMS,VOL(JJJJJJ,*),D,N'
// COMMAND 'V SMS,VOL(KKKKKK,*),D,N'
// COMMAND 'V SMS,VOL(LLLLLL,*),D,N'
// COMMAND 'V SMS,VOL(MMMMMM,*),D,N'
// COMMAND 'D SMS,SG(SG2),LISTVOL'
//* ----- ***
//STEP1 EXEC PGM=IEFBR14
//
```

---

This SMS modify command approach has a drawback: It holds only until the next IPL. Then SMS will read again what the volume and SG status is from the respective Source Control Data Set (SCDS) and will reload the Active Control Data Set (ACDS) accordingly. To make this change permanent you have to change the SCDS through ISMF panels or execute jobs through NaviQuest to change the SCDS.

We go briefly through the ISMF panel sequence to make the SMS volume status complete and permanent.

Example 14-8 is the first ISMF Storage Group Application panel where you choose the actual SCDS, the storage group, here XC, and select option 4 to alter the SMS volume status.

*Example 14-8 Select SMS storage group in SCDS*

---

```
STORAGE GROUP APPLICATION SELECTION
Command ==>

To perform Storage Group Operations, Specify:
CDS Name . . . . . 'SYS1.DFSMS.SCDS'
                                     (1 to 44 character data set name or 'Active' )
Storage Group Name   XC              (For Storage Group List, fully or
                                     partially specified or * for all)
Storage Group Type   (VIO, POOL, DUMMY, COPY POOL BACKUP,
                     OBJECT, OBJECT BACKUP, or TAPE)

Select one of the following options :
  4 1. List   - Generate a list of Storage Groups
    2. Define - Define a Storage Group
    3. Alter  - Alter a Storage Group
    4. Volume - Display, Define, Alter or Delete Volume Information

If List Option is chosen,
  Enter "/" to select option      Respecify View Criteria
                                  Respecify Sort Criteria

Use ENTER to Perform Selection;
Use HELP Command for Help; Use END Command to Exit.
```

---

The next panel which appears is in Example 14-9. Option 3 allows you to alter the SMS volume status. Here you select the volumes which ought to receive the change.

*Example 14-9 Select respective volume range(s) from Storage Group*

---

```
STORAGE GROUP VOLUME SELECTION
Command ==>
```

```
CDS Name . . . . . : SYS1.DFSMS.SCDS
Storage Group Name : XC
Storage Group Type : POOL
```

Select One of the following Options:

- 3 1. Display - Display SMS Volume Statuses (Pool & Copy Pool Backup only)
2. Define - Add Volumes to Volume Serial Number List
3. Alter - Alter Volume Statuses (Pool & Copy Pool Backup only)
4. Delete - Delete Volumes from Volume Serial Number List

Specify a Single Volume (in Prefix), or Range of Volumes:

```
Prefix From To Suffix Type
===> XC 6510 6514 X
===>
===>
===>
```

Use ENTER to Perform Selection;

Use HELP Command for Help; Use END Command to Exit.

The panel in Example 14-9 on page 310 provides a list in the lower third of the screen where you can specify ranges of VOLSERS, which will then be changed all at once. This is, therefore, more powerful than the SMS system commands, which are on a single volume level only.

The following panel in this sequence, shown in Example 14-10, is a confirmation panel that displays the current SMS volume status which we plan to alter. It also shows the system names which belong to the SMS-plex. A SMS-plex is usually congruent with the underlying Parallel Sysplex®.

*Example 14-10 Display current SMS volume status before altering*

SMS VOLUME STATUS ALTER

Page 1 of 2

Command ===>

```
SCDS Name . . . . . : SYS1.DFSMS.SCDS
Storage Group Name . : XC
Volume Serial Numbers : XC6510 - XC6514
```

To ALTER SMS Volume Status, Specify:

System/Sys Group Name	SMS Vol Status	System/Sys Group Name	SMS Vol Status	( Possible SMS Vol Status for each: NOTCON, ENABLE, DISALL, DISNEW, QUIALL, QUINEW )
MCECEBC	===> ENABLE	MZBCVS2	===> ENABLE	
	===>		===>	
	===>		===>	
	===>		===>	* SYS GROUP = sysplex
	===>		===>	minus systems in the
	===>		===>	sysplex explicitly
	===>		===>	defined in the SCDS
	===>		===>	

In this panel we overwrite the SMS volume status with the desired status change. This shows in the following panel, shown in Example 14-11 on page 312.

*Example 14-11 Indicate SMS volume status change for all connected system images*

---

```
SMS VOLUME STATUS ALTER                Page 1 of 2
Command ==>>>

SCDS Name . . . . . : SYS1.DFSMS.SCDS
Storage Group Name . : XC
Volume Serial Numbers : XC6510 - XC6514

To ALTER SMS Volume Status, Specify:

System/Sys      SMS Vol      System/Sys      SMS Vol      ( Possible SMS Vol
Group Name      Status      Group Name      Status      Status for each:
-----
MCECEBC  ==>> disnew      MZBCVS2  ==>> disnew      NOTCON, ENABLE,
           ==>>                                DISALL, DISNEW,
           ==>>                                QUIALL, QUINEW )
           ==>>
           ==>>                                * SYS GROUP = sysplex
           ==>>                                minus systems in the
           ==>>                                sysplex explicitly
           ==>>                                defined in the SCDS
           ==>>
```

---

After pressing Enter and PF3 to validate and perform the SMS volume status change, a validation panel confirms that the requested change did happen. This is shown in Example 14-12.

*Example 14-12 Confirmation about SMS volume status change*

---

```
STORAGE GROUP VOLUME SELECTION        ALL VOLUMES ALTERED
Command ==>>>

CDS Name . . . . . : SYS1.DFSMS.SCDS
Storage Group Name : XC
Storage Group Type : POOL

Select One of the following Options:
  3  1. Display - Display SMS Volume Statuses (Pool & Copy Pool Backup only)
     2. Define  - Add Volumes to Volume Serial Number List
     3. Alter   - Alter Volume Statuses (Pool & Copy Pool Backup only)
     4. Delete  - Delete Volumes from Volume Serial Number List

Specify a Single Volume (in Prefix), or Range of Volumes:
      Prefix  From    To    Suffix  Type
==>> XC      6510   6514   _____ X
==>>
==>>
==>>

Use ENTER to Perform Selection;
Use HELP Command for Help; Use END Command to Exit.
```

---

In this example all volumes that were selected through the filtering in the previous panel no longer allow any new allocation on these volumes. But this happens only after the updated SCDS is activated and copied into the Active Control Data Set (ACDS). A way to activate a new SMS configuration is through the ISMF Primary Menu panel with option 8. Under the CDS APPLICATION SELECTION, choose option 4 and then 5 to finally perform the change.

To show how powerful, meaningful naming conventions for VOLSERS might be combined with the selection capabilities in ISMF, Example 14-13 shows an example of how to change the status of 920 volumes at once. This example assumes a contiguous number range for the volume serial numbers.

*Example 14-13 Indicate SMS volume status change for 920 volumes*

---

```

STORAGE GROUP VOLUME SELECTION
Command ==>

CDS Name . . . . . : SYS1.DFSMS.SCDS
Storage Group Name : XC
Storage Group Type : POOL

Select One of the following Options:
 3 1. Display - Display SMS Volume Statuses (Pool & Copy Pool Backup only)
    2. Define - Add Volumes to Volume Serial Number List
    3. Alter - Alter Volume Statuses (Pool & Copy Pool Backup only)
    4. Delete - Delete Volumes from Volume Serial Number List

Specify a Single Volume (in Prefix), or Range of Volumes:
  Prefix From To Suffix Type
====> DB2 001 300 X
====> IMS 001 350 x
====> LRG 001 150 X
====> TMP 001 120 x

Use ENTER to Perform Selection;
Use HELP Command for Help; Use END Command to Exit.
```

---

Through this approach in utilizing the SMS volume status, the data gradually moves to the new environment. This is not the fastest approach, but it is the approach with the least effort required. To speed up the process you run DFSMSdss full volume logical copy operations on all source volumes as suggested by the storage administrator, but do not specify target volumes. An example is given in Example 14-6 on page 307. Then allocation runs through the standard SMS allocation and automatically selects the correct target volume, and moves the data sets, which are not allocated at the time the job executes. Another option is to run DFSMSdss migration not on a volume level, which allows more control, but plan for DFSMSdss migration jobs on the entire SG level, perhaps over the weekend. Figure 14-14 illustrates this. You might specify for STORGRP a list of storage group names to address multiple storage groups at the same time.

*Example 14-14 DFSMSdss moves all data sets out of a storage group*

---

```

/* COMMAND 'V SMS,VOL(AAAAAA,*),D,N'
/* COMMAND 'D SMS,SG(SG1),LISTVOL'
/* ----- ***
//MOVEDATA EXEC PGM=ADRDSU
//SYSPRINT DD SYSOUT=*
//SYSIN DD *
COPY STORGRP(SG1) -
DS(INC(**)) -
EXCLUDE(SYS1.VTOCIX.*,SYS1.VVDS.*)) -
DELETE PURGE SELECTMULTI(ANY) SPHERE -
WAIT(00,00) ADMIN OPT(3) CANCELERROR
/*
/* ----- ***
//AGAIN EXEC PGM=IEBGENER
```

```

//SYSPRINT DD DUMMY
//SYSUT1 DD DSN=WB.MIGRATE.CNTL(DSS#SG1),DISP=SHR
//SYSUT2 DD SYSOUT=(A,INTRDR)
//SYSIN DD DUMMY
//* ----- JOB END ----- ***
//

```

---

You might keep the job repeatedly executing through the second step AGAIN, where the same job is read into the system again through the internal MVS reader.

Eventually there remain a few data sets on the source volumes which are always open. These data sets require you to stop the concerned application, close and unallocate these data sets, and then run the job in Figure 14-14 once more.

Verify at the end of this logical data set migration that all data has been removed from the source disk server with the IEHLIST utility's LISTVTOC command.

Again this approach requires you to have the old and new equipment connected at the same time and most likely over an extended period, except if you push the migration jobs through like in Example 14-14 on page 313, in which you can run more than one instance concurrently.

### 14.3.5 Summary of logical data migration based on software utilities

Problems encountered when not using an allocation manager like system-managed storage are less flexibility when using esoteric unit names, or complex and time-consuming tasks in maintaining hard-coded JCL volume names, which need to be changed when creating new volumes on new disk storage servers. It is recommended that you use system-managed volumes to overcome the limitations with esoteric unit names and hard-coded volume names in JCL.

Logical data migration is difficult and can be time-consuming, and it usually requires system down time. System-managed storage allows for a less difficult data migration, when it is on a logical level, in order to consolidate not just disk storage servers, but also volumes moving to larger target volumes.

## 14.4 Combine physical and logical data migration

The following approach combines physical and logical data migration:

- ▶ Physical full volume copy to larger capacity volume when both volumes have the same device geometry (same track size and same number of tracks per cylinder).
- ▶ Use COPYVOLID to keep the original volume label and to not confuse catalog management. You can still locate the data on the target volume through standard catalog search.
- ▶ Adjust the VTOC of the target volume to make the larger volume size visible to the system with the ICKDSF REFORMAT command, to refresh REFVTOC, or expand the VTOC EXT VTOC, which requires you to delete and rebuild the VTOC index using EXTINDEX in the REFORMAT command.
- ▶ Then perform the logical data set copy operation to the larger volumes. This allows you to use either DFSMSdss logical copy operations, as outlined before, or the system-managed data approach.



When a level is reached where no data moves any more because the remaining data sets are in use all the time, some down time has to be scheduled to perform the movement of the remaining data. This might require you to run DFSMSDss jobs from a system which has no active allocations on the volumes which need to be emptied.

## 14.5 z/VM and VSE/ESA data migration

DFSMS/VM® provides a set of software utility and command functions which are suitable for data migration.

- ▶ DASD Dump Restore (DDR) provides a physical copy and is suited to copy data between devices with the same device geometry. DDR cannot be utilized for a device migration, like from 3380 to 3390.
- ▶ DIRMAINT CMDISK command moves data between minidisks in VM from any device type to any device type which is supported by VM.
- ▶ CMS COPYFILE command offers a logical copy on a file level from any minidisk device to any minidisk device which VM supports and is suited to migrate to a different device type.
- ▶ Another logical migration approach is possible through the CP PTAPE command. PTAPE dumps spool files to tape and then re-loads these files back onto the new disk storage.

Last, but not least, PPRC might be considered when moving the data from any ESS model to the new storage server DS8000. Because PPRC is a host server independent approach, similar considerations apply as outlined under “Hardware- or microcode-based migration” on page 302.

z/OS Global Mirror under a z/OS image also allows you to move z/VM full mini disks between different storage servers and would allow you to connect to the source disk server through ESCON and to the target disk storage server with FICON.

In VSE/ESA data migration you might consider the following approaches:

- ▶ Physical volume copy from all ESS models to the new DS8000 disk storage server through PPRC as outlined under “Hardware- or microcode-based migration” on page 302.
- ▶ Logical copy operations under VSE/ESA which allow data movement from source storage servers to new DS8000 disk storage servers are the following:
  - VSE FASTCOPY to move data between volumes with the same device geometry.
  - VSE DITTO to copy individual files which allow a device migration.
  - VSE IDCAMS commands REPRO or EXPORT/IMPORT to move VSAM files between any device types.

There are other software vendor products as well which provide the capability to migrate data onto new storage servers, for example, CA Favor.

## 14.6 Summary of data migration

The route which an installation takes to migrate data to one or more DS8000 series storage servers depends on requirements for service levels, whether application down time is acceptable, available tools and cost estimates with certain budget limits.

When coming from an ESS 750 or ESS 800, Metro Mirror seems to be the natural choice, and allows for concurrent data migration with almost no application impact. There would be no

impact when the actual switch uses P/DAS, although it is quicker and easier to allow for a brief service interruption and quickly switch to the new disk storage server. Because Metro Mirror provides data consistency at any time, the switch-over to the new disk server is simple and does not require further efforts to ensure data consistency at the receiving site. It is feasible to use the GUI-based approach because migration is usually a one time effort. Command-line interfaces such as TSO commands are an alternative and can be automated to some extent with REXX procedures.

When the source disk servers are not compatible with the DS8000 and the migration is based on a full volume, physical level, then there are options which depend on various circumstances which are different for each installation. When TDMF or FDRPAS is available and the customer is used to managing volume migration using these software tools, then it is a likely approach to use these tools for a larger scale migration as well. In other circumstances it might be feasible to include the migration as part of a total package when installing and implementing DS8000 disk storage servers. This is possible through IGS services which rely on FDRPAS, TDMF or Piper for z/OS. There is always DFSMSdss, which is still popular, but this approach usually requires an outage of the specific application systems.

Finally, the migration might be used as an opportunity to consolidate volumes at the same time. After a decision has been made about the new volume size, which is preferably a multiple of a 3390-1 or 1113 cylinders, it is required to logically move the data sets. One option is to rely on SMS-based allocation in a system-managed environment, combined with DFSMSdss and logical data set copies targeted from individual source volume sets or entire SMS storage groups. A variation here might be a combination of full volume physical copy for the very first volume copied to a larger target volume, followed by further logical-based copy operations with DFSMSdss after adjusting the VTOC and VTOC index information.

After the migration and storage consolidation you will be using a disk storage server technology that will serve you with promising performance and excellent scalability combined with rich functionality and high availability.

# Implementation and management in the open systems environment

In this part we discuss considerations for the DS8000 series when used in an open systems environment. The topics include:

- ▶ Open systems support and software
- ▶ Data migration in open systems





## Open systems support and software

In this chapter we describe how the DS8000 fits into your open systems environment. In particular, we discuss:

- ▶ The extent of the open systems support
- ▶ Where to find detailed and accurate information
- ▶ Major changes from the ESS 2105
- ▶ Software provided by IBM with the DS8000
- ▶ IBM solutions and services that integrate DS8000 functionality

## 15.1 Open systems support

The scope of open systems support of the new DS8000 model is based on that of the ESS 2105, with some exceptions:

- ▶ No parallel SCSI attachment support
- ▶ Some new operating systems were added
- ▶ Some legacy operating systems and the corresponding servers were removed
- ▶ Some legacy HBAs were removed

New versions of operating systems, servers, file systems, host bus adapters, clustering products, SAN components, and application software are constantly announced in the market. Every modification to any of these components that can affect the interoperability with the storage system must be tested. The integrity of the customer data always has the highest priority. A new version or product can be added to the list of supported environments only after it is proven that all components work with each other flawlessly.

Information about the supported environments changes frequently. Therefore, you are strongly advised always to refer to the online resources listed in 15.1.2, “Where to look for updated and detailed information” on page 320.

### 15.1.1 Supported operating systems and servers

Table 15-1 provides an overview of the open system platforms, operating systems, and high availability clustering applications that are generally supported for attachment to the DS8000. However, support is given for specific models and operating system versions only. Details about the allowed combinations can be found in the resources listed in 15.1.2, “Where to look for updated and detailed information” on page 320.

*Table 15-1 Platforms, operating systems and applications supported with DS8000*

Server platforms	Operating systems	Clustering applications
IBM pSeries, RS/6000®, IBM BladeCenter™ JS20	AIX, Linux	IBM HACMP™ (AIX only)
IBM iSeries	OS/400, i5/OS™, Linux, AIX	IBM HACMP (AIX only)
HP PARisc, Itanium II	HP UX	HP MC/Serviceguard
HP Alpha	OpenVMS, Tru64 UNIX	HP TruCluster
Intel IA-32, IA-64, IBM BladeCenter HS20 and HS40	Microsoft Windows, VMWare, Novell Netware, Linux	Microsoft Cluster Service, Novell Netware Cluster Services
SUN	Solaris	Sun Cluster
Apple Macintosh	OS X	
Fujitsu PrimePower	Solaris	
SGI	IRIX	

### 15.1.2 Where to look for updated and detailed information

This section provides a list of online resources where detailed and up-to-date information about supported configurations, recommended settings, device driver versions, and so on, can be found. Due to the high innovation rate in the IT industry, the support information is

updated frequently. Therefore it is advisable to visit these resources regularly and check for updates.

### **The DS8000 Interoperability Matrix**

The *DS8000 Interoperability Matrix* always provides the latest information about supported platforms, operating systems, HBAs and SAN infrastructure solutions. It contains detailed specifications about models and versions. It also lists special support items, such as boot support, and exceptions. It can be found at:

<http://www.ibm.com/servers/storage/disk/ds8000/pdf/ds8000-matrix.pdf>

### **The IBM HBA Search Tool**

For information about supported Fibre Channel HBAs and the recommended or required firmware and device driver levels for all IBM storage systems, you can visit the *IBM HBA Search Tool* site, sometimes also referred to as the *Fibre Channel host bus adapter firmware and driver level matrix*:

<http://knowledge.storage.ibm.com/HBA/HBASearchTool>

For each query, select one storage system and one operating system only, otherwise the output of the tool will be ambiguous. You will be shown a list of all supported HBAs together with the required firmware and device driver levels for your combination. Furthermore, you can select a detailed view for each combination with more information, quick links to the HBA vendors' Web pages and their IBM supported drivers, and a guide to the recommended HBA settings.

### **The DS8000 Host Systems Attachment Guide**

The *DS8000 Host Systems Attachment Guide*, SC26-7628, guides you in detail through all the steps that are required to attach an open system host to your DS8000 storage system. It is available at:

<http://www.ibm.com/servers/storage/disk/ds8000>

### **The TotalStorage Proven™ program**

IBM has introduced the *TotalStorage Proven* program to help clients identify storage solutions and configurations that have been pre-tested for interoperability. It builds on IBM's already extensive interoperability efforts to develop and deliver products and solutions that work together with third party products.

The TotalStorage Proven Web site provides more detail on the program, as well as the list of pre-tested configurations:

<http://www.ibm.com/servers/storage/proven/index.html>

### **HBA vendor resources**

All of the Fibre Channel HBA vendors have Web sites that provide information about their products, facts and features, as well as support information. These sites will be useful when the IBM resources are not sufficient, for example, when troubleshooting an HBA driver. Please be aware that IBM cannot be held responsible for the content of these sites.

#### ***QLogic Corporation***

The Qlogic Web site can be found at:

<http://www.qlogic.com>

QLogic maintains a page that lists all the HBAs, drivers, and firmware versions that are supported for attachment to IBM storage systems:

[http://www.qlogic.com/support/oem\\_detail\\_all.asp?oemid=22](http://www.qlogic.com/support/oem_detail_all.asp?oemid=22)

### **Emulex Corporation**

The Emulex home page is:

<http://www.emulex.com>

They also have a page with content specific to IBM storage systems:

<http://www.emulex.com/ts/docoem/framibm.htm>

### **JNI / AMCC**

AMCC took over the former JNI, but still markets FC HBAs under the JNI brand name. JNI HBAs are supported for DS8000 attachment to Sun systems. The home page is:

<http://www.amcc.com>

Their IBM storage specific support page is:

<http://www.jni.com/OEM/oem.cfm?ID=4>

### **Atto**

Atto supplies HBAs which IBM supports for Apple Macintosh attachment to the DS8000. Their home page is:

<http://www.attotech.com>

They have no IBM storage specific page. Their support page is:

<http://www.attotech.com/support.html>

Downloading drivers and utilities for their HBAs requires registration.

## **Platform and operating system vendors' pages**

The platform and operating system vendors also provide lots of support information to their customers. Go there for general guidance about connecting their systems to SAN-attached storage. However, be aware that in some cases you will not find information that is intended to help you with third party (from their point of view) storage systems, especially when they also have storage systems in their product portfolio. You may even get misleading information about interoperability and support from IBM. It is beyond the scope of this book to list all the vendors' Web sites.

### **15.1.3 Differences to the ESS 2105**

For the DS8000 the support matrix went through a cleanup process. The changes are described in this section. For details see the resources listed in the previous section.

- ▶ There are no parallel SCSI adapters for the DS8000. Therefore all parallel SCSI support had to be dropped. No host system can be connected to the DS8000 via parallel SCSI.
- ▶ Older HBA models, especially all 1 Gbit/s models were also removed from the support matrix. Some new models were added.
- ▶ Legacy SAN infrastructure solutions, like hubs and gateways, are not supported.
- ▶ There is no support for IBM TotalStorage SAN Volume Controller storage software for Cisco MDS 9000 (SVC4MDS) at initial GA.



- ▶ Some legacy operating systems and operating system versions were dropped from the support matrix. These are either versions which were withdrawn from marketing or support that are not marketed or supported by their vendors or are not seen as significant enough anymore to justify the testing effort necessary to support them. Major examples include:
  - IBM AIX 4.x, OS/400 V5R1, Dynix/ptx
  - Microsoft Windows NT®
  - Sun Solaris 2.6, 7
  - HP UX 10, 11, Tru64 4.x, OpenVMS 5.x
  - Novell Netware 4.x
  - All professional Linux distributions, SUSE SLES7
- ▶ Some new operating systems and versions were added, including:
  - Apple Macintosh OS X
  - IBM AIX 5.3
  - IBM i5 OS V5R3
  - VMWare 2.5.0 (first quarter of 2005)
  - Redhat Enterprise Linux 3 IA-64
  - SUSE Linux Enterprise Server 9 (first quarter of 2005)

### 15.1.4 Boot support

For most of the supported platforms and operating systems you can use the DS8000 as a boot device. The *DS8000 Interoperability Matrix* provides detailed information about boot support. Refer to “The DS8000 Interoperability Matrix” on page 321.

The *DS8000 Host Systems Attachment Guide*, SC26-7628, helps you with the procedures necessary to set up your host in order to boot from the DS8000. See “The DS8000 Host Systems Attachment Guide” on page 321.

The *SDD User's Guide*, SC26-7637, also helps with identifying the optimal configuration and lists the steps required to boot from multipathing devices. For more information refer to 15.2, “Subsystem Device Driver” on page 324.

### 15.1.5 Additional supported configurations (RPQ)

There is a process for cases where a desired configuration is not represented in the support matrix. This process is called *Request for Price Quotation* (RPQ). Clients should contact their IBM storage sales specialist or IBM Business Partner for submission of an RPQ. Initiating the process does not guarantee the desired configuration will be supported. This depends on the technical feasibility and the required test effort. A configuration that equals or is similar to one of the already approved ones is more likely to get approved than a completely different one.

### 15.1.6 Differences in interoperability between the DS8000 and DS6000

The DS8000 and DS6000 have the same open systems support matrix. There are only a few exceptions, with respect to the timing. DS6000 will support some operating systems not at initial GA, but in the first quarter of 2005:

- ▶ OpenVMS
- ▶ Tru64

- ▶ Novell Netware
- ▶ VMware 2.5.0
- ▶ SuSE SLES9

## 15.2 Subsystem Device Driver

To ensure maximum availability most customers choose to connect their open systems hosts through more than one Fibre Channel path to their storage systems. With an intelligent SAN layout this protects you from failures of FC HBAs, SAN components, and host ports in the storage subsystem.

Some operating systems, however, can't deal natively with multiple paths to a single disk; they see the same disk multiple times. This puts the data integrity at risk, because multiple write requests can be issued to the same data and nothing takes care of the correct order of writes.

To utilize the redundancy and increased I/O bandwidth you get with multiple paths, you need an additional layer in the operating system's disk subsystem to recombine the multiple disks seen by the HBAs into one logical disk. This layer manages path failover, should a path become unusable, and balancing of I/O requests across the available paths.

For most operating systems that are supported for DS8000 attachment, IBM makes the IBM Subsystem Device Driver (SDD) available at no additional charge, to provide the following functionality:

- ▶ Enhanced data availability through automatic path failover and fallback
- ▶ Increased performance through dynamic I/O load-balancing across multiple paths
- ▶ Ability for concurrent download of licensed internal code
- ▶ User configurable path-selection policies for the host system

SDD can be downloaded from:

<http://www.ibm.com/servers/storage/support/software/sdd/downloading.html>

When you click the **Subsystem Device Driver downloads** link, you will be presented a list of all operating systems for which SDD is available. Selecting one leads you to the download packages, the *SDD User's Guide*, SC30-4096, and additional support information. The user's guide contains all the information that is needed to install, configure, and use SDD for all supported operating systems.

For some operating systems, additional information about SDD can be found in Appendix A, "Open systems operating systems specifics" on page 343.

**Note:** SDD and RDAC, the multipathing solution for the IBM TotalStorage DS4000 series, can coexist on most operating systems, as long as they manage separate HBA pairs. Refer to the documentation of your DS4000 series storage system for detailed information.

IBM AIX alternatively offers MPIO, a native multipathing solution. It allows the use of *Path Control Modules* (PCMs) for optimal storage system integration. IBM provides SDDPCM, a PCM with the SDD full functionality. See "IBM AIX" on page 347, for more detail.

IBM OS/400 V5R3 doesn't use SDD. It provides native multipath support since V5R3. For details refer to Appendix B, "Using DS8000 with iSeries" on page 373.

## 15.3 Other multipathing solutions

Some operating systems come with native multipathing software, for example:

- ▶ SUN StorEdge Traffic Manager for Sun Solaris
- ▶ HP PVLinks for HP UX
- ▶ IBM AIX native multipathing (MPIO) (see “IBM AIX” on page 347)
- ▶ IBM OS/400 V5R3 multipath support (see Appendix B, “Using DS8000 with iSeries” on page 373)

In addition there are third party multipathing solutions, such as Veritas DMP, which is part of Veritas Volume Manager.

Most of these solutions are also supported for DS8000 attachment, although the scope may vary. There may be limitations for certain host bus adapters or operating system versions. Always consult the DS8000 Interoperability Matrix for the latest information.

## 15.4 DS CLI

The DS Command-Line Interface (DS CLI) provides a command set to perform almost all monitoring and configuration tasks on the DS8000. This includes the ability to:

- ▶ Verify and change the storage unit configuration, with some limitations
- ▶ Check the current logical storage and Copy Services configuration
- ▶ Create new logical storage and Copy Services configuration settings
- ▶ Modify or delete logical storage and Copy Services configuration settings

**Limitation:** Some basic configuration tasks cannot be performed using the DS CLI, such as creating a storage complex or adding a storage unit to a storage complex.

The DS CLI allows you to invoke and manage logical storage configuration tasks and Copy Services functions from an open systems host through batch processes and scripts.

It is part of the DS8000 Licensed Internal Code and is delivered with the Customer Software Packet. It is closely tied to the installed version of code and therefore not available for download. The CLI code must be obtained from the same software bundle as the current microcode update. When DS8000 code updates occur, the DS CLI must also be updated.

The DS CLI is available for most of the supported operating systems. The DS8000 Interoperability Matrix contains a complete list.

There are some pre-requisites for the host system that the DS CLI is running on:

- ▶ Java 1.4.1 or later must be installed.
- ▶ ksh (Korn shell) or bash (Bourne again shell) must be available. Install shield does not support the sh shell.

The *DS8000 Command-Line Interface User's Guide*, SC26-7625 contains detailed installation instructions for all supported host operating systems.

The DS CLI can be used in two modes:

- ▶ Single command mode: You invoke the DS CLI program in an operating system shell or command prompt and pass the command it is to execute directly as a parameter. The

command will be passed directly to the S-HMC for immediate execution. The return code of the DS CLI program corresponds to the return code of the command it executed. This mode can be used for scripting.

- ▶ Interactive mode: You start the DS CLI program on your host. It provides you with a shell environment that allows you to enter commands and send them to the S-HMC for immediate execution.

There also is a section in this book describing the usage of the DS CLI, including some examples (see Chapter 11, “DS CLI” on page 231).

The *DS8000 Command-Line Interface User's Guide*, SC26-7625, contains a complete command reference.

## 15.5 IBM TotalStorage Productivity Center

The IBM TotalStorage Productivity Center (TPC) is an open storage management solution that helps to reduce the effort of managing complex storage infrastructures, to increase storage capacity utilization, and to improve administrative efficiency. It is designed to enable an agile storage infrastructure that can respond to on-demand storage needs.



Figure 15-1 IBM TotalStorage Productivity Center

TPC is the integration point for storage and fabric management, and replication, as depicted in Figure 15-1. It provides a *launchpad* for the following IBM TotalStorage Open Software Family and Tivoli products:

- ▶ IBM Tivoli Storage Resource Manager
- ▶ IBM Tivoli SAN Manager
- ▶ IBM Tivoli Storage Manager
- ▶ IBM TotalStorage Multiple Device Manager

These products allow for the management of data through its life cycle, device configuration, performance, replication, storage network fabric, data backup, data availability, and data recovery, as well as enterprise policies for managing host, application, database, and file system data.

As a component of the IBM TotalStorage Productivity Center, Multiple Device Manager is designed to reduce the complexity of managing SAN storage devices by allowing administrators to configure, manage, and monitor storage from a single console.

The devices managed are not restricted to IBM brand products. In fact, any device compliant with the *Storage Network Industry Association (SNIA) Storage Management Initiative Specification (SMI-S)* can be managed with the IBM TotalStorage Multiple Device Manager. The protocol utilized to enable the central management is called the *Common Information Model (CIM)*. This open standards based protocol uses XML to define CIM objects and to manage storage devices over HTTP transactions. Figure 15-2 shows the MDM main panel.

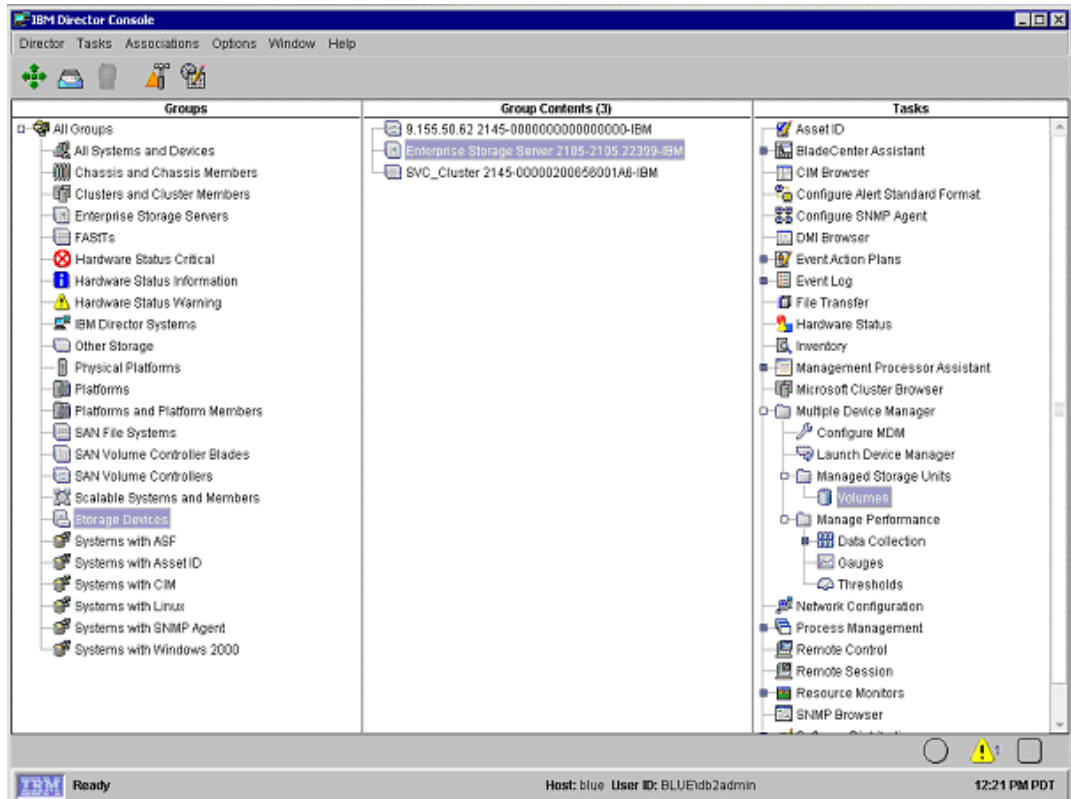


Figure 15-2 MDM main panel

For more information about the IBM TotalStorage Multiple Device Manager refer to the redbook *IBM TotalStorage Multiple Device Manager Usage Guide*, SG24-7097.

Updated support summaries, including specific software, hardware, and firmware levels supported, are maintained at:

<http://www.ibm.com/storage/support/mdm>

The IBM TotalStorage Multiple Device Manager is composed of three subcomponents:

- ▶ Device Manager
- ▶ TPC for Disk
- ▶ TPC for Replication

They are described in the following sections.

## 15.5.1 Device Manager

The Device Manager (DM) builds on the *IBM Director* technology. It uses the *Service Level Protocol* (SLP) to discover supported storage systems on the SAN. The SLP enables the discovery and selection of generic services accessible through an IP network. The DM then uses managed objects to manage these devices.

DM also provides a subset of configuration functions for the managed devices, primarily LUN allocation and assignment. Its functionality includes aggregation and grouping of devices and provides policy based actions across multiple storage devices. These services communicate with the CIM Agents that are associated with the particular devices to perform the required configuration. Devices that are not SMI-S compliant are not supported. The DM also interacts and provides SAN management functionality when the IBM Tivoli SAN Manager is installed.

The DM health monitoring keeps you aware of hardware status changes in the discovered storage devices. You can drill down to the status of the hardware device, if applicable. This enables you to understand which components of a device are malfunctioning and causing an error status for the device. Figure 15-3 shows an example of a DM view.

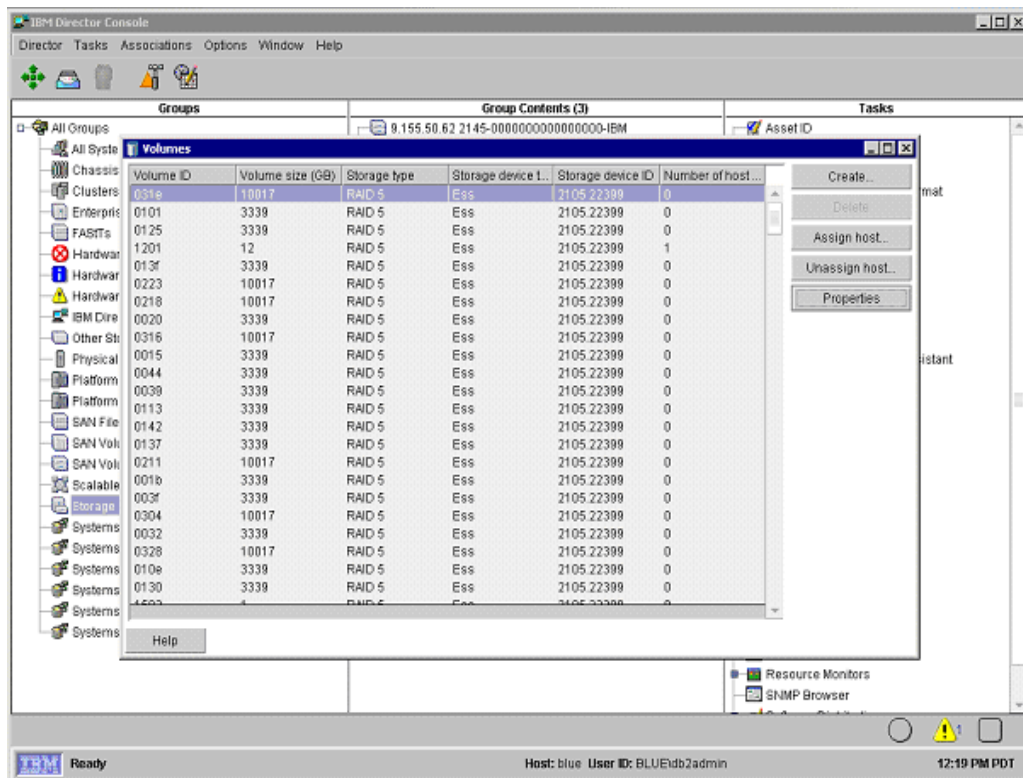


Figure 15-3 Sample Device Manager view

In summary, the Device Manager is responsible for:

- ▶ Discovery of supported devices (SMI-S compliant)
- ▶ Data collection (asset, availability, configuration)
- ▶ Providing a topographical view of storage

## 15.5.2 TPC for Disk

*TPC for Disk*, formerly known as *MDM Performance Manager*, provides the following functions:

- ▶ Collect performance data from devices.
- ▶ Configure performance thresholds.
- ▶ Monitor performance metrics across storage subsystems from a single console.
- ▶ Receive timely alerts to enable event action based on customer policies.
- ▶ View performance data from the performance manager database.
- ▶ Enable storage optimization through identification of the best performing volumes.

TPC for Disk collects data from IBM or non-IBM networked storage devices that implement SMI-S. A performance collection task collects performance data from one or more storage groups of one device type. It has individual start and stop times, and a sampling frequency. The sampled data is stored in DB2 database tables.

You can use TPC for Disk to set performance thresholds for each device type. Setting thresholds for certain criteria enables TPC for Disk to notify you when a certain threshold has been exceeded, so that you can take action before a critical event occurs. You can also specify actions to be taken automatically. These may be just to log the occurrence or to trigger an event. The settings can vary by individual device.

You can view performance data from the performance manager database in both graphical and tabular forms. This is shown in Figure 15-4.

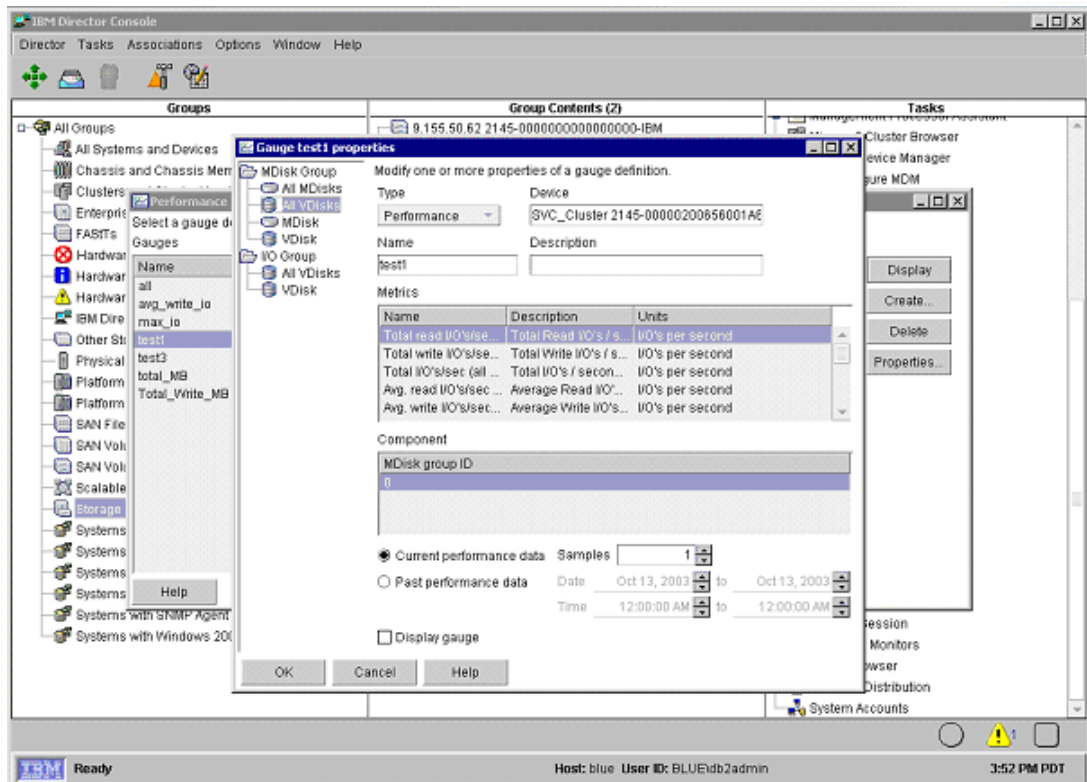


Figure 15-4 Sample screenshot of TPC for Disk



The *Volume Performance Advisor* is an automated tool to help the storage administrator pick the best possible placement of a new LUN to be allocated from a performance perspective. It uses the historical performance statistics collected from the managed devices and locates unused storage capacity on the SAN that exhibits the best estimated performance characteristics. Allocation optimization involves several variables that are user controlled, such as the required performance level and the time of day/week/month of prevalent access. This function is fully integrated with the DM function.

For detailed information about the installation and the use of TPC for Disk, refer to the redbook *IBM TotalStorage Multiple Device Manager Usage Guide*, SG24-7097.

### 15.5.3 TPC for Replication

*TPC for Replication*, formerly known as *MDM Replication Manager*, provides a single point of control for all replication activities. Given a set of source volumes to be replicated, it will find the appropriate targets, perform all the configuration actions required, and ensure the source and target volumes relationships are set up.

TPC for Replication administers and configures the copy services functions of the managed storage systems and monitors the replication actions. It can manage two types of copy services: the Continuous Copy, also known as Peer-to-Peer Remote Copy (PPRC) and the Point-in-time Copy, also known as FlashCopy.

It supports replication sessions, which ensure that data on multiple related heterogeneous volumes is kept consistent, provided that the underlying hardware supports the necessary primitive operations. Multiple pairs are handled as a consistent unit, *Freeze-and-Go* functions can be performed when mirroring errors occur. It is designed to control and monitor the data replication operations in large-scale customer environments.

TPC for Replication provides a user interface for creating, maintaining, and using volume groups and for scheduling copy tasks. It populates the lists of volumes using the DM interface. An administrator can also perform all tasks with the TPC for Replication command-line interface.

## 15.6 Global Mirror Utility

The *DS8000 Global Mirror Utility* (GMU) is a standalone tool to provide a management layer for IBM TotalStorage Global Mirror failover and failback (FO/FB). It provides clients with a set of twelve basic commands that utilize the DS Open-API to accomplish either a planned or unplanned FO/FB sequence.

The purpose of the GMU is to automate the complex sequence of steps necessary to set up and manage Global Mirror relationships for a large number of LUNs. There is no limitation to the number of volumes managed by GMU.

In the event of a planned or unplanned FO/FB, the system administrator issues a few GMU commands to recover consistent data on the remote site before he starts host I/O activities. The GMU commands can be incorporated in user scripts or utilities to further automate datacenter operations. However, IBM does not support such scripts or utilities without a specific service arrangement.

The GMU is distributed on a separate CD with the DS8000 Licensed Internal Code (LIC). It includes installation instructions and a user guide. It consists of two components, a server and a client. The server receives requests from the client and communicates over TCP/IP to



the DS8000 Storage Hardware Management Console for the execution of the commands. The commands allow you to:

- ▶ Create, modify, start, stop, and resume a Global Mirror session
- ▶ Manage failover and failback operations including managing consistency
- ▶ Perform planned outages

To monitor the DS8000 volume status and the Global Mirror session status, you can either use the DS Storage Manager or DS CLI.

## 15.7 Enterprise Remote Copy Management Facility (eRCMF)

eRCMF is a multi-site disaster recovery solution, managing IBM Total Storage Remote Mirror and Copy as well as FlashCopy functions, while ensuring data consistency across multiple machines. It is a scalable, flexible solution for the DS8000 with the following functions:

- ▶ When a site failure occurs or may be occurring, eRCMF splits the two sites in a manner that allows the backup site to be used to restart the applications. This needs to be fast enough that when the split occurs, operations on the production site are not impacted.
- ▶ It manages the states of the Metro Mirror and FlashCopy relationships, so that the customer knows when the data is consistent and can control on which site the applications run.
- ▶ It offers easy commands to place the data in the state it needs to be in. For example, if the data is out of sync, the command **resync** will cause eRCMF to scan the specified volumes and issue commands to bring the volumes back into sync.
- ▶ It offers a tool to execute eRCMF configuration checks. This check is intended to verify that the eRCMF configuration matches the physical DS8000 setup in the customer environment. This is required to discover configuration changes that affect the eRCMF configuration as well. Regular checks support customers in keeping the eRCMF configuration up-to-date with their actual environment; otherwise, full eRCMF management functionality is not given.

eRCMF is a IBM Global Services offering. More information about eRCMF can be found at:

<http://www-1.ibm.com/services/us/index.wss/so/its/a1000110>

## 15.8 Summary

The new DS8000 enterprise disk subsystem offers broad support and superior functionality for all major open system host platforms. In addition, IBM provides software packages for many platforms that are needed to exploit all of the functionality. IBM also integrated DS8000 functionality into its family of storage management software products that help to reduce the effort of managing complex storage infrastructures, to increase storage capacity utilization and to improve administrative efficiency.





## Data migration in the open systems environment

In this chapter we discuss important concepts for the migration of data to the new DS8000:

- ▶ Data migration considerations
- ▶ Data migration and consolidation
- ▶ Comparison of the different methods

## 16.1 Introduction

The term data migration has a very diverse scope. We use it here solely to describe the process of moving data from one type of storage to another, or to be exact, from one type of storage to a DS8000. In many cases, this process is not only comprised of the mere copying of the data, but also includes some kind of consolidation.

With our focus on storage, we distinguish three kinds of consolidation, also illustrated in Figure 16-1:

- ▶ The consolidation of distributed, direct-attached storage to shared, SAN-attached disk storage
- ▶ The consolidation of many small volumes into a few larger ones
- ▶ The consolidation of several small storage systems into a few larger ones

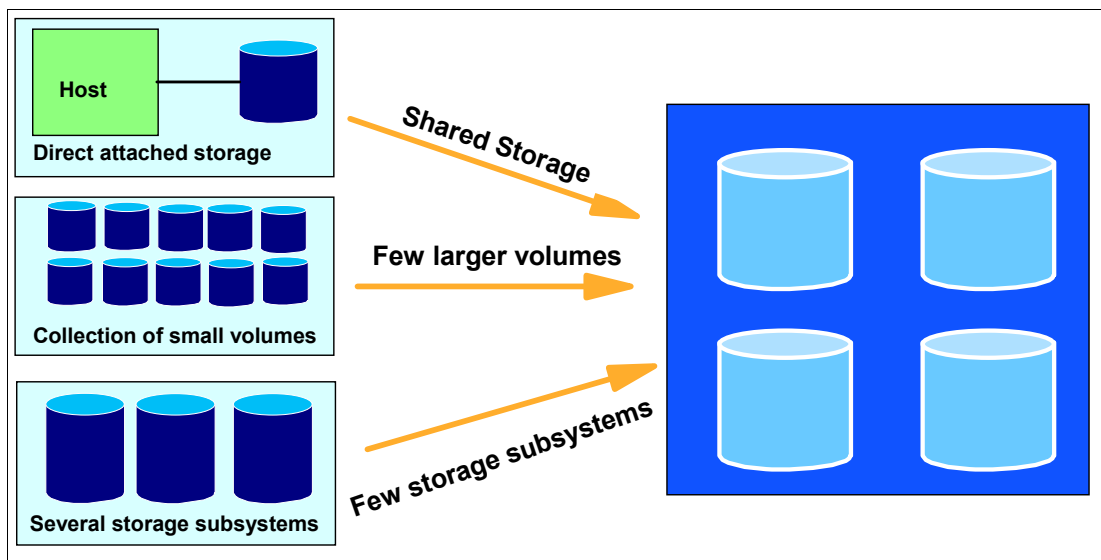


Figure 16-1 Different ways of consolidation

Very often, the set goal of a consolidation effort is a combination of more than one of these types. The DS8000, with its exceptional scalability and performance, makes an ideal target system for storage consolidation.

There are many different methods for data migration. To decide what is best in your case, gather information about the following items:

- ▶ The source and target storage make and type
- ▶ The amount of data to be migrated
- ▶ The amount of time available for the migration
- ▶ The ability to connect both source and target storage at the same time
- ▶ The availability of spare disk or tape capacity for temporary storage
- ▶ The format of the data itself
- ▶ The consolidation goals
- ▶ Can the migration be disruptive, and for how long
- ▶ The distance between source and target

We describe the most common methods in the next section. Be aware that, in a heterogeneous IT environment, you will most likely have to choose more than one method.

**Note:** When discussing disruptiveness, we don't consider any interruptions that may be caused by adding the new DS8000 LUNs to the host and later by removing the old storage. They vary too much from operating system to operating system, even from version to version. However, they have to be taken into account, too.

Once it is decided which method (or methods) to use, the migration process starts with a very careful planning phase. Items to be taken into account during migration planning include:

- ▶ Availability of all the required hardware and software
- ▶ Installation of the new and removal of the old storage
- ▶ Installation of drivers required for the new storage
- ▶ A test of the new environment
- ▶ The storage configuration before, during, and after the migration
- ▶ The time schedule of the whole process, including the scheduling of outages
- ▶ Does everyone involved have the necessary skills to perform their tasks?

Additional information about data migration to the DS8000 can be found in the *DS8000 Introduction and Planning Guide*, GC35-0495, available for download at:

[http://www-1.ibm.com/servers/storage/disk/ds8000/pdf/DS8000\\_planning\\_guide.pdf](http://www-1.ibm.com/servers/storage/disk/ds8000/pdf/DS8000_planning_guide.pdf)

Exceptional care must be exercised in data sharing (clustered) environments. If data is shared between more than one host, all of them have to be made aware of the changes in the configuration, even if the migration is only performed by one of them. Refer to the documentation of your clustering solution for ways to propagate configuration changes throughout the cluster.

**Note:** IBM Global Services can assist you in all phases of the migration process with professional skill and methods.

## 16.2 Comparison of migration methods

There are numerous methods that can be used to migrate data from one storage system to another. We briefly describe the most common ones and list their advantages and disadvantages in the following sections.

**Note:** The IBM iSeries platform with the OS/400 and i5/OS operating systems has a different approach to data management than the other open systems. Therefore different strategies for data migration apply. Refer to Appendix B, "Using DS8000 with iSeries" on page 373, for more information.

### 16.2.1 Host operating system-based migration

Data migration using tools that come with the host operating system has these advantages:

- ▶ No additional cost for software or hardware
- ▶ System administrators are used to using the tools

Reasons against using these methods could include:

- ▶ Different methods are necessary for different data types and operating systems
- ▶ Strong involvement of the system administrator is necessary

Today the majority of data migration tasks are performed with one of the methods discussed in the following sections.

### Basic copy commands

Using copy commands is the simplest way to move data from one storage system to another, for example:

- ▶ **copy**, **xcopy**, drag and drop for Windows
- ▶ **cp**, **cpio** for UNIX

These commands are available on every system supported for DS8000 attachment, but work only with data organized in file systems. Data can be copied between file systems of different sizes. Therefore this method can be used for the consolidation of small volumes into larger ones. Figure 16-2 outlines the process.

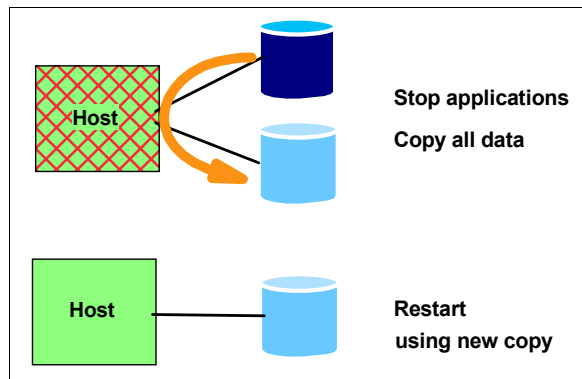


Figure 16-2 Migration with copy commands

The most significant disadvantage of this method is the disruptiveness. To preserve data consistency, the applications writing to the data which is migrated have to be interrupted for the duration of the copy process. Furthermore, some copy commands cannot preserve advanced metadata, such as access control lists or permissions.

**Attention:** If your storage systems are attached through multiple paths, make sure that the multipath drivers for the old and the new storage system can coexist on one host. If not, you have to revert the host to a single path configuration before you attach the new storage system. You can change back to a multipath configuration after the migration is complete.

This is valid for all migration methods where source and target are attached to the host at the same time.

### Copy raw devices

For raw data there are tools that allow you to read and write disk devices directly, such as **dd** for UNIX. They copy the data and its organizational structure (metadata) without having any intelligence about it. Therefore they cannot be used for consolidation of small volumes into larger ones. Special care has to be taken when data and its metadata are kept in separate places. They both have to be copied and realigned on the target system. By themselves, they are useless.

This method also requires the disruption of applications writing to the data for the complete process.

### Online copy and synchronization with rsync

**rsync** is an open source tool that is available for all major open system platforms, including Windows and Novell Netware.

Its original purpose is the remote mirroring of file systems with as few network requirements as possible. Once the initial copy is done, it keeps the mirror up to date by only copying changes. Additionally, the incremental copies are compressed.

**rsync** can also be used for local mirroring and therefore for data migration. It allows you to copy the data while applications are writing to it and thus minimizes the disruption. As for the normal copy commands, **rsync** works on file systems only. It also allows consolidation. Figure 16-3 shows the steps required.

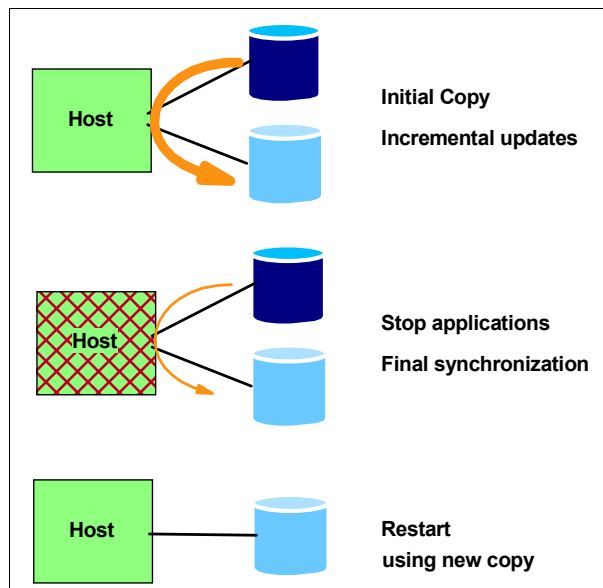


Figure 16-3 Data migration with rsync

You set up the initial copy during normal operation, allowing enough time for the copy to complete before the planned switch to the new storage (cut over time). Once the initial copy is complete, you keep it up to date with incremental **rsync** runs, for instance once a day, during off-peak hours. At the cut over time, you stop the applications using the data, let **rsync** do a final synchronization, and restart the applications using the migrated data. The duration of the disruption depends only on the amount of data that has changed since the last re-synchronization.

More information about **rsync** can be found on the **rsync** project home page:

<http://samba.org/rsync/>

### Migration using volume management software

Logical Volume Managers (LVMs) are available for all open systems (for Windows it is called Disk Manager). The LVM creates a layer of storage virtualization within the operating system. The most basic functionality every LVM provides is to:

- ▶ Extend logical volumes across several physical disks
- ▶ Stripe data across several physical disks to enhance performance

- ▶ Mirror data for higher availability and migration

The LUNs provided by the DS8000 appear to the LVM as physical SCSI disks.

Usually the process is to set up a mirror of the data on the old disks to the new LUNs, wait until it is synchronized and split it at the cut over time. Some LVMs provide commands that automate this process.

The biggest advantage of using the LVM for data migration is that the process can be totally non-disruptive, as long as the operating system allows you to add and remove LUNs dynamically. Due to the virtualization nature of LVM, it also allows for all kinds of consolidation. Figure 16-4 shows the process.

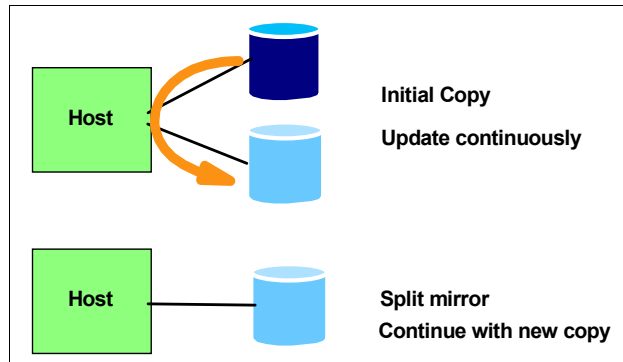


Figure 16-4 Migration using LVM mirroring

The major disadvantage is that the LVM mirroring method requires a lot of system administrator intervention and attention. Production data is manipulated while production is running. Carelessness can lead to the outage that one wanted to avoid by selecting this method.

## Backup and restore

Every serious IT operation will have ways to back up and restore data. They can be used for data migration. We list this method here because it shares the common advantages and disadvantages with the methods discussed previously, although the tools will not always be provided natively by the operating system.

All open system platforms and many applications provide native backup and restore capabilities. They may not be very sophisticated sometimes, but they are often suitable in smaller environments. In larger data centers it is customary to have a common backup solution across all systems. Either can be used for data migration.

The backup and restore option allows for consolidation because the tools are usually aware of the data structures they handle.

One significant difference to most of the other methods discussed here, is that it does not require the source and target storage systems to be connected to the hosts at the same time. Figure 16-5 on page 339 illustrates this method.



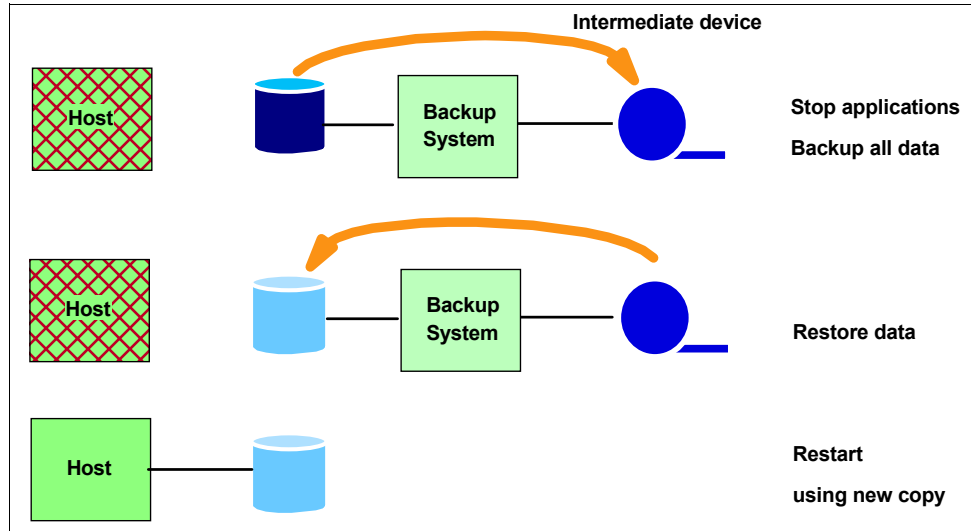


Figure 16-5 Migration using backup and restore

The major disadvantage is again the disruptiveness. The applications that write to the data to be migrated must be stopped for the whole migration process. Backup and restore to and from tape usually takes longer than direct copy from disk to disk. The duration of the disruption can be reduced somewhat by using incremental backups.

## 16.2.2 Subsystem-based data migration

The DS8000 provides remote copy functionality, which also can be used to migrate data:

- ▶ IBM TotalStorage Metro Mirror, formerly known as PPRC, for distances up to 300km
- ▶ IBM TotalStorage Global Copy, formerly known as PPRC Extended Distance, for longer distances
- ▶ A combination of Metro Mirror and Global Copy with an intermediate device in certain cases

**Restriction:** The combination of Metro Mirror and Global Copy is not available at GA.

These methods are host system agnostic and can therefore be used with only minimum system administrator attention. They also do not add any additional CPU load to the host systems, and they don't require the host system to be connected to both storage systems at the same time.

The necessary disruption is minimal. The initial copy is started during normal operation. Once it is complete, the target is kept up-to-date by only copying changes made to the source. At the cut over time, the applications are stopped and the mirror is allowed to reach synchronization. Then the target system is connected to the host instead of the source system and the applications can be restarted with the new copy.

**Important:** The source storage system must be removed from the host completely, not only physically, but also logically, including all configuration data.

However, the copy functions do not allow for the consolidation of smaller volumes into larger ones, since they are not aware of the structure of the data.

## Metro Mirror and Global Copy

From a local data migration point of view both methods are on par with each other, with Global Copy having a smaller impact on the subsystem performance and Metro Mirror requiring almost no time for the final synchronization phase. It is advisable to use Global Copy instead of Metro Mirror, if the source system is already at its performance limit even without remote mirroring. Figure 16-6 outlines the migration steps.

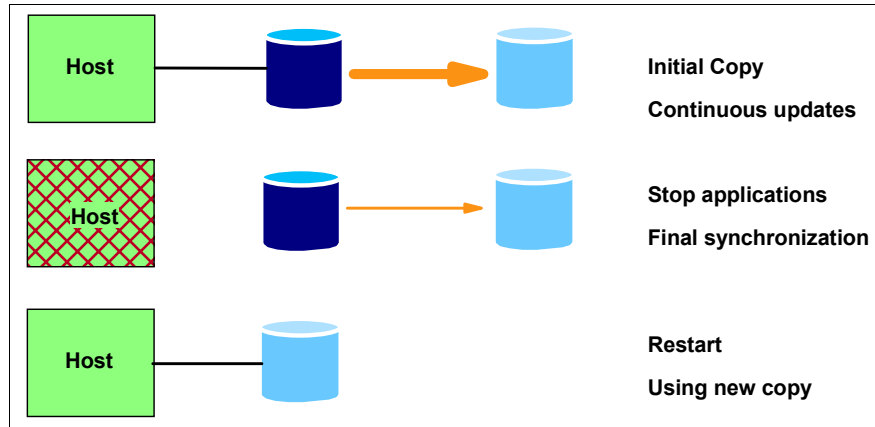


Figure 16-6 Migration with Metro Mirror or Global Copy

The remote copy functionality can be used to migrate data in either direction between the Enterprise Storage Server (ESS) 750 or 800 and the new DS8000 and DS6000 storage systems. The ESS E20 and F20 lack the support for remote copy over Fibre Channel and can therefore not be mirrored directly to a DS8000.

## Combination of Metro Mirror and Global Copy

A cascading Metro Mirror and Global Copy solution is useful in two cases:

- ▶ The source system is already mirrored for disaster tolerance and mirroring is mandatory for production. Then a Global Copy relationship can be used to migrate the data from the secondary volumes of the Metro Mirror pair to the new machine.
- ▶ Data must be migrated from an older ESS E20 or F20. Here a Metro Mirror using ESCON connectivity is used to mirror the data to an intermediate ESS 800, which in turn will copy the data to the DS8000 with Global Copy.

Figure 16-7 shows the setup and steps to take for this method.

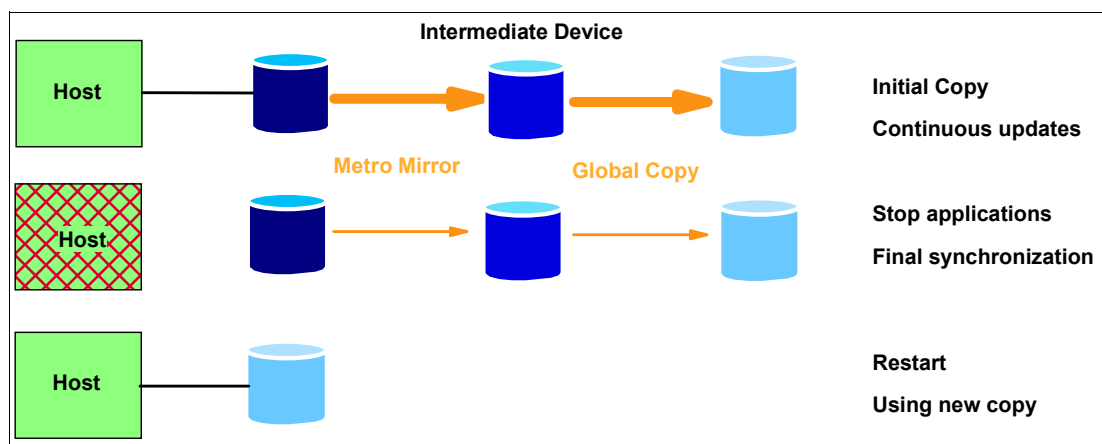


Figure 16-7 Migration with Metro Mirror, Global Copy and an intermediate device

### 16.2.3 IBM Piper migration

Piper is a hardware and software solution to move data between disk systems while production is ongoing. It is used in conjunction with IBM migration services. Piper is available for mainframe and open systems environments. Here we discuss the open systems version only. For mainframe environments see Chapter 14, “Data migration in zSeries environments” on page 293.

The Piper hardware consists of a portable rack enclosure containing Fibre Channel routers, a SAN Switch, and non-disruptive power supplies.

Piper migration can be performed independently of the host operating system and the source disk system. It doesn't generate extra host workload. However, it does not support the consolidation of small volumes into larger ones, because it is not aware of the data structure.

Since Piper has to be in the data path, between the host and the source storage system, it can only be used to migrate Fibre Channel attached storage. It also requires two short interruptions of data access, one to bring it into the data path, another one to take it out, after the data has been moved. Figure 16-8 illustrates how Piper works.

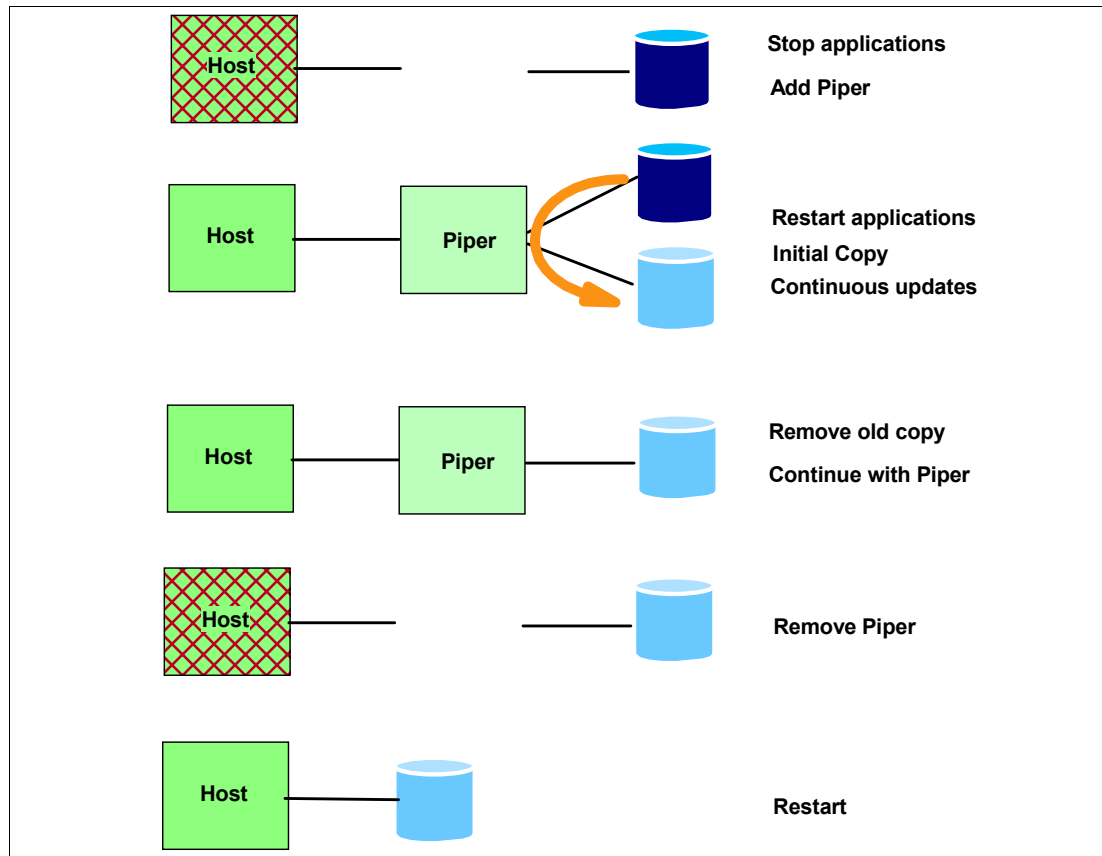


Figure 16-8 Piper migration

For open systems migration, no additional software has to be installed on the host. IBM migration services use some scripts to determine the exact storage configuration that has to be duplicated on the target system.

For more information about the Piper offerings refer to:

[http://www.ibm.com/servers/storage/services/featured/hardware\\_assist.html](http://www.ibm.com/servers/storage/services/featured/hardware_assist.html)

## 16.2.4 Other migration applications

There are a number of applications available from other vendors that can assist in data migration. We don't discuss them here in detail. Some examples include:

- ▶ Softek Data Replicator for Open
- ▶ NSI Double-Take
- ▶ XoSoft WANSync

There also are storage virtualization products which can be used for data migration in a similar manner to the Piper tool. They are installed on a server which forms a virtualization engine that resides in the data path. Examples are:

- ▶ IBM SAN Volume Controller
- ▶ Falconstore IPStore
- ▶ Datacore SANSymphony

An important question to ask, before deciding among these methods, is whether the virtualization engine can be removed from the data path after the migration has completed.

## 16.3 IBM migration services

This is the easiest way to migrate data, because IBM will assist you throughout the complete migration process. In several countries IBM offers a migration service. Check with your IBM sales representative about migration services for your specific environment and needs.

Businesses today require efficient and accurate data migration. IBM provides technical specialists at your location to plan and migrate your data to your DS8000 disk system.

This migration is accomplished using either native operating system mirroring, remote mirroring, or the Piper migration tool to replicate your data to the DS8000 disk system with minimum interruption to service. In addition, IBM will provide a Migration Control Book which specifies the activities performed during these services.

The benefits of IBM migration services include:

- ▶ Minimized downtime and no data loss.
- ▶ Superior data protection that preserves data updates throughout the migration, allowing the process to be interrupted if needed.
- ▶ A *Migration Control Book* that details the work performed, so your staff can use it for post migration management.

## 16.4 Summary

This chapter shows that there are many ways to accomplish data migration. Thorough analysis of the current environment, evaluation of the requirements and planning are necessary. Once you decide on one or more migration methods, refer to the documentation of the tools you want to use to define the exact sequence of steps to take.

Special care must be exercised when data is shared between more than one host.

IBM Global Services can assist you in all stages to ensure a successful and smooth migration.



# A

## Open systems operating systems specifics

In this appendix, we describe the particular issues of some operating systems with respect to the attachment to a DS8000. The following subjects are covered:

- ▶ Planning considerations
- ▶ Common tools
- ▶ IBM AIX
- ▶ Linux on various platforms
- ▶ Microsoft Windows
- ▶ HP OpenVMS

## General considerations

In this section we cover some topics that are not specific to a single operating system. This includes available documentation, some planning considerations, and common tools.

### The DS8000 Host Systems Attachment Guide

The *DS8000 Host Systems Attachment Guide*, SC26-7628 provides instructions to prepare a host system for DS8000 attachment. For all supported platforms and operating systems it covers:

- ▶ Installation and configuration of the FC HBA
- ▶ Peculiarities of the operating system with regard to storage attachment
- ▶ How to prepare a system that boots from the DS8000 (when supported)

It varies in detail for the different platforms.

Many more publications are available from IBM and other vendors. Refer to Chapter 15, “Open systems support and software” on page 319 and to the operating system-specific sections in this chapter.

## Planning

Thorough planning is necessary to ensure that your new DS8000 will perform efficiently in your data center. In this section we raise the questions you will have to ask and the things you have to consider before you start. We don't cover all the items in detail because this is not an implementation book.

For a more detailed discussion of these considerations related to performance, refer to Chapter 12, “Performance considerations” on page 253.

### Capacity planning considerations

For proper sizing of your DS8000 storage subsystem more than the total required capacity has to be known. Consider the following questions:

- ▶ What is the capacity for each host system or even each application? What are this system's performance requirements (I/Os, throughput)?
- ▶ How much capacity is needed for fixed block (open systems) and how much for CKD (mainframe) data?
- ▶ Do you need advanced copy functions?
- ▶ Which DS8000 model is adequate to achieve capacity and performance requirements?
- ▶ What is the number of disk drives needed, their size and speed?
- ▶ With the usable capacity known, what is the raw capacity to order?
- ▶ What is the number of Fibre Channel attachment needed?
- ▶ Do you have to plan for future expansion? Is Capacity on Demand an option?

### Data placement considerations

A DS8000 logical volume is composed of extents. These extents are striped across all disks in an array (or rank, which is equivalent). To create the logical volume, extents from one extent pool are concatenated. Within a given extent pool there is no control over the placement of

the data, even if this pool spans several ranks. If possible, the extents for one logical volume are taken from the same rank.

To get higher throughput values than a single array can deliver, it is necessary to stripe the data across several arrays. This can only be achieved through striping on the host level.

To achieve maximum granularity and control for data placement, you will have to create an extent pool for every single rank.

However, some operating systems support only a limited number of attached disks, or make it difficult for the administrator to combine several physical disks into one big volume. In the DS8000 logical volumes cannot span several extent pools. To be able to create very large logical volumes, you must consider having extent pools that include more than one rank.

## UNIX performance monitoring tools

Some tools are worth discussing because they are available for almost all UNIX variants and system administrators are accustomed to using them. You may have to administer a server, and these are the only tools you have available to use. These tools offer a quick way to tell whether a system is I/O-bound:

- ▶ **iostat**
- ▶ **sar** (System Activity Report)
- ▶ **vmstat** (Virtual Memory Statistics)

## IOSTAT

The base tool for evaluating I/O performance of disk devices for UNIX operating systems is **iostat**. Although available on most UNIX platforms, **iostat** varies in its implementation from system to system.

The **iostat** command is useful to determine whether a system's I/O load is balanced or whether a single volume is becoming a performance bottleneck. The tool reports I/O statistics for TTY devices, disks, and CD-ROMs. It monitors I/O device throughput and utilization by observing the time the disks are active in relation to their average transfer rates.

**Tip:** I/O activity monitors, such as **iostat**, have no way of knowing whether the disk they are seeing is a single physical disk or a logical disk striped upon multiple physical disks in a RAID array. Therefore, some performance figures reported for a device, for example, %busy, could appear high.

Example A-1 shows a sample **iostat** output, taken on an AIX host. It shows disk device statistics since the last reboot.

*Example: A-1 AIX iostat output.*

---

```
#iostat
Disks:      % tm_act   Kbps    tps    Kb_read  Kb_wrtn
hdisk0      0.0         0.3     0.0    29753    48076
hdisk1      0.1         0.1     0.0    11971    26460
hdisk2      0.2         0.8     0.1    91200   108355
cd0         0.0         0.0     0.0     0        0
```

---

The output reports the following:

- ▶ The %tm\_act column indicates the percentage of the measured interval time that the device was busy.
- ▶ The Kbps column shows the average data rate, read and write data combined, of this device.
- ▶ The tps column shows the transactions per second. Note that an I/O transaction can have a variable transfer size. This field may also appear higher than would normally be expected for a single physical disk device.
- ▶ The Kb\_read and Kb\_wrtn columns show the total amount of data read and written.

**iostat** can also be issued for continuous monitoring with a given number of iterations and a monitoring period. It will then print a report like that in Example A-1 on page 345 for every period, with the values calculated for exactly this period. In most cases this mode is more useful, because bottlenecks mostly appear only during peak times and are not reflected in an overall average. Be aware that the first in the series of reports represents the average since boot and should be discarded.

Example A-2 shows an **iostat** report from SUN Solaris. You see an example of a device that appears to be very busy (sd1). The r/s column shows 124.3 reads per second; the %b column shows 90 percent busy. The svc\_t column, however, shows a service time of 15.7 ms, still quite reasonable for 124 I/Os per second. Depending on the application layout, this report could lead to the conclusion that the I/O load of this system is unbalanced. Some disks get a lot more I/O requests than others. A consequence of this could be to move certain parts of a database from the busiest disks to less used ones.

*Example: A-2 SUN Solaris iostat output*

---

```
#iostat -x
extended disk statistics
disk  r/s  w/s  Kr/s  Kw/s  wait  actv  svc_t  %w  %b
fd0   0.0  0.0   0.0   0.0  0.0  0.0   0.0  0  0
sd1  124.3 14.5 3390.9 399.7 0.0  2.0  15.7  0 90
sd2   0.7  0.4  13.9   4.0  0.0  0.0   7.8  0  1
sd3   0.4  0.5   2.5   3.8  0.0  0.1   8.1  0  1
sd6   0.0  0.0   0.0   0.0  0.0  0.0   5.8  0  0
sd8   0.3  0.2   9.4   9.6  0.0  0.0   8.6  0  1
sd9   0.7  1.3  12.4  21.3  0.0  0.0   5.2  0  3
```

---

The implementation of the **iostat** command is different for every UNIX variant. It also offers many different options and parameters. Refer to your system documentation and the **iostat** manpage for more information.

## System Activity Report (SAR)

The *System Activity Report (SAR)* provides a quick way to tell if a system is *I/O bound*. SAR has numerous options, providing paging, TTY, CPU busy, and many other statistics.

One way you can run **sar** is by specifying a sampling interval and the number of times you want it to run.

This is shown in Example A-3 on page 347. It displays CPU usage information, sampled five times with a two second interval. To check whether a system is I/O bound, the important column to look at is %wio. The %wio indicates the time spent waiting on I/O from all disks, both internal and external. Here, too, the first line represents the average since boot time and should be discarded.



### Example: A-3 SAR Sample Output

---

```
# sar -u 2 5
AIX aixtest 3 4 001750154C00 2/5/03
17:58:15  %usr  %sys  %wio %idle
17:58:17   43    9    1   46
17:58:19   35   17    3   45
17:58:21   36   22   20   23
17:58:23   21   17    0   63
17:58:25   85   12    3    0
Average    44   15    5   35
```

---

As a general rule of thumb, a server with over 40 percent waiting on I/O is spending too much time waiting for I/O. However, you also have to take the type of workload into account. If you are running a video file server, serving I/O will be the primary activity of the machine and you will expect high %wio values.

A system with very busy CPUs can mask I/O wait. The definition of %wio is: Idle with some processes waiting for I/O (only block I/O, raw I/O, or VM pageins/swapins indicated). If the system is CPU busy and also is waiting for I/O, the accounting will increment the CPU busy values, but not the %wio column.

The other column headings in the example indicate:

- ▶ %usr: Time system spent executing application code
- ▶ %sys: Time system spent executing operating system calls
- ▶ %idle: Time the system was idle with no outstanding I/O requests

The implementation of the **sar** command is different for the various UNIX variants. However, the output of **sar -u** is the same for all.

There are other modes to use **sar**, which we will not discuss further:

- ▶ Ongoing system activity accounting via cron
- ▶ Display previously captured data

**sar** offers many different options and parameters. Refer to your system documentation and the **sar** manpage for more information.

## VMSTAT

The **vmstat** utility is a useful tool for taking a quick snapshot or overview of the system performance. It is easy to see what is happening with regard to the CPUs, paging, swapping, interrupts, I/O wait, and much more. There are several reports that **vmstat** can provide. They vary slightly between the different versions of UNIX. Refer to your system documentation and the **vmstat** manpage for more information.

## IBM AIX

This section covers items specific to the IBM AIX operating system. It is not intended to repeat the information that is contained in other publications. We focus on topics that are not covered in the well known literature or are important enough to be repeated here.

## Other publications

Apart from the *DS8000 Host Systems Attachment Guide*, SC26-7628, there are two redbooks that cover pSeries storage attachment:

- ▶ *Practical Guide for SAN with pSeries*, SG24-6050, covers all aspects of connecting an AIX host to SAN-attached storage. However, it is not quite up-to-date; the last release was in 2002.
- ▶ *Fault Tolerant Storage - Multipathing and Clustering Solutions for Open Systems for the IBM ESS*, SG24-6295, focuses mainly on high availability and covers SDD and HACMP topics. It also is from 2002.

Much of the technical information for pSeries or AIX also covers external storage, since SAN attachment became standard procedure in almost all data centers of size with a claim to availability.

## The AIX host attachment scripts

AIX needs some file sets installed to support DS8000 disks. They prepare the *Object Data Manager* (ODM) with information about the subsystem. That way the DS8000 volumes are identified properly and all performance parameters optimized.

For the ESS under AIX there was a set of Host Attachment scripts which could be downloaded from the Web. For the DS8000 using AIX SDD 1.6.0.0, these scripts have been consolidated into one called FCP Host Attachment Script. The most up-to-date version for the DS8000 can be downloaded from:

<http://www-1.ibm.com/support/dlsearch.wss?rs=540&q=host+scripts&tc=ST52G7&dc=D417>

## Finding the World Wide Port Names

In order to allocate DS8000 disks to a pSeries server, the World Wide Port Name (WWPN) of each of the pSeries Fibre Channel adapters have to be registered in the DS8000. You can use the **lscfg** command to find out these names, as shown in Example A-4.

*Example: A-4 Finding Fibre Channel adapter WWN*

---

```
lscfg -vl fcs0
fcs0          U1.13-P1-I1/Q1  FC Adapter

Part Number.....00P4494
EC Level.....A
Serial Number.....1A31005059
Manufacturer.....001A
Feature Code/Marketing ID...2765
FRU Number.....      00P4495
Network Address.....10000000C93318D6
ROS Level and ID.....02C03951
Device Specific.(Z0).....2002606D
Device Specific.(Z1).....00000000
Device Specific.(Z2).....00000000
Device Specific.(Z3).....03000909
Device Specific.(Z4).....FF401210
Device Specific.(Z5).....02C03951
Device Specific.(Z6).....06433951
Device Specific.(Z7).....07433951
Device Specific.(Z8).....20000000C93318D6
Device Specific.(Z9).....CS3.91A1
Device Specific.(ZA).....C1D3.91A1
```

```
Device Specific.(ZB).....C2D3.91A1
Device Specific.(YL).....U1.13-P1-I1/Q1
```

---

You can also print the WWPN of an HBA directly by running

```
lscfg -v1 <fcs#> | grep Network
```

The # stands for the instance of each FC HBA you want to query.

## Managing multiple paths

It is a common and recommended practice to assign a DS8000 volume to the host system through more than one path, to ensure availability in case of a SAN component failure and to achieve higher I/O bandwidth. AIX will discover a separate hdisk for each path to a DS8000 logical volume.

To utilize the path redundancy and increased I/O bandwidth, you need an additional software layer in the AIX disk subsystem to recombine the multiple hdisks into one device.

### Subsystem device driver (SDD)

The IBM Subsystem Device Driver (SDD) software is a host-resident pseudo device driver designed to support the multipath configuration environments in IBM products. SDD resides in the host system with the native disk device driver and manages redundant connections between the host server and the DS8000. SDD is available for AIX 5.1, 5.2, and 5.3.

Refer to 15.2, “Subsystem Device Driver” on page 324 for download information and installation and usage documentation.

#### **Determine the installed SDD level**

To determine whether SDD is installed, and at which level, you can use the `lslpp -l` command, as shown in Example A-5.

*Example: A-5 lslpp -l “\*sdd\*”*

---

Fileset	Level	State	Description
-----			
Path: /usr/lib/objrepos devices.sdd.52.rte	1.5.1.2	COMMITTED	IBM Subsystem Device Driver for AIX V52
Path: /etc/objrepos devices.sdd.52.rte	1.5.1.2	COMMITTED	IBM Subsystem Device Driver for AIX V52

---

#### **Useful SDD commands**

The `datapath query device` command displays information about all vpath devices. It is useful to determine the number of paths to each SDD vpath device and their status. See Example A-6.

*Example: A-6 datapath query device command*

---

```
{yli4642:root}/home/redbook -> datapath query device
```

```
Total Devices : 2
```

```
DEV#: 0 DEVICE NAME: vpath3 TYPE: 2107 POLICY: Optimized
SERIAL: 10522873
```

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi2/hdisk17	OPEN	NORMAL	0	0
1	fscsi2/hdisk19	OPEN	NORMAL	27134	0
2	fscsi3/hdisk21	OPEN	NORMAL	0	0
3	fscsi3/hdisk23	OPEN	NORMAL	27352	0

DEV#: 1 DEVICE NAME: vpath4 TYPE: 2107 POLICY: Optimized  
SERIAL: 20522873

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi2/hdisk18	CLOSE	NORMAL	25734	0
1	fscsi2/hdisk20	CLOSE	NORMAL	0	0
2	fscsi3/hdisk22	CLOSE	NORMAL	25500	0
3	fscsi3/hdisk24	CLOSE	NORMAL	0	0

This **lsvpcfg** command helps verify the vpath configuration state. See Example A-7.

*Example: A-7 lsvpcfg command*

```
vpath0 (Available pv) 018FA067 = hdisk1 (Available) hdisk3 (Available) hdisk5 (Available)
vpath1 (Available ) 019FA067 = hdisk2 (Available) hdisk4 (Available) hdisk6 (Available)
```

The **hd2vp** command converts a volume group made of hdisk devices to an SDD vpath device volume group. Run **hd2vp <volume group name>** for each volume group to convert.

## Multipath I/O (MPIO)

AIX MPIO is an enhancement to the base OS environment that provides native support for multi-path Fibre Channel storage attachment. MPIO automatically discovers, configures, and makes available every storage device path. The storage device paths are managed to provide high availability and load balancing of storage I/O. MPIO is part of the base kernel and is available for AIX 5.2 and AIX 5.3.

The base functionality of MPIO is limited. It provides an interface for vendor-specific *Path Control Modules* (PCMs) which allow for implementation of advanced algorithms.

IBM provides a PCM for DS8000 that enhances MPIO with all the features of the original SDD. It is called SDDPCM and is available from the SDD download site (refer to 15.2, “Subsystem Device Driver” on page 324)

There are some reasons to prefer MPIO (with SDDPCM) to traditional SDD:

- ▶ Performance improvements due to direct integration with AIX
- ▶ Better integration if different storage systems are attached
- ▶ Easier administration through native AIX commands

**Important:** If you choose to use MPIO with SDDPCM instead of SDD, you have to remove the regular DS8000 Host Attachment Script and install the MPIO version of it. This script identifies the DS8000 volumes to the operating system as MPIO manageable. Of course, you can't have SDD and MPIO with SDDPCM on a given server at the same time.

**Note:** At initial GA, there will be no AIX MPIO support for the DS8000.

For basic information about MPIO see the “Multiple Path I/O” section in the *AIX 5L System Management Concepts: Operating System and Devices* guide:

[http://publib16.boulder.ibm.com/pseries/en\\_US/aixbman/admnconc/hotplug\\_mgmt.htm#mpioconcepts](http://publib16.boulder.ibm.com/pseries/en_US/aixbman/admnconc/hotplug_mgmt.htm#mpioconcepts)

The management of MPIO devices is described in the “Managing MPIO-Capable Devices” section of *System Management Guide: Operating System and Devices for AIX 5L*:

[http://publib16.boulder.ibm.com/pseries/en\\_US/aixbman/baseadm/manage\\_mpio.htm](http://publib16.boulder.ibm.com/pseries/en_US/aixbman/baseadm/manage_mpio.htm)

**Restriction:** A point worth considering when deciding between SDD and MPIO is, that the IBM TotalStorage SAN Volume Controller does not support MPIO at this time. For updated information refer to:

<http://www.ibm.com/servers/storage/software/virtualization/svc/index.html>

### **Determine the installed SDDPCM level**

You use the same command as for SDD, `ls1pp -l "*sdd*"`, to determine the installed level of SDDPCM. It will also tell you whether you have SDD or SDDPCM installed.

SDDPCM software provides useful commands such as:

- ▶ `pcmpath query device` to check the configuration status of the devices
- ▶ `pcmpath query adapter` to display information about adapters
- ▶ `pcmpath query essmap` to display each device, path, location, and attributes

### **Useful MPIO commands**

The `lspath` command displays the operational status for the paths to the devices, as shown in Example A-8. It can also be used to read the attributes of a given path to an MPIO-capable device.

*Example: A-8 lspath command result*

---

```
{part1:root}/ -> lspath |pg
Enabled hdisk0   scsi0
Enabled hdisk1   scsi0
Enabled hdisk2   scsi0
Enabled hdisk3   scsi7
Enabled hdisk4   scsi7
...
Missing hdisk9   fscsi0
Missing hdisk10  fscsi0
Missing hdisk11  fscsi0
Missing hdisk12  fscsi0
Missing hdisk13  fscsi0
...
Enabled hdisk96  fscsi2
Enabled hdisk97  fscsi6
Enabled hdisk98  fscsi6
Enabled hdisk99  fscsi6
Enabled hdisk100 fscsi6
```

---

The `chpath` command is used to perform change operations on a specific path. It can either change the operational status or tunable attributes associated with a path. It cannot perform both types of operations in a single invocation.

The `rmpath` command unconfigures or undefines, or both, one or more paths to a target device. It is not possible to unconfigure (undefine) the last path to a target device using the `rmpath` command. The only way to do this is to unconfigure the device itself (for example, use the `rmdev` command).

Refer to the manpages of the MPIO commands for more information.

## LVM configuration

In AIX, all storage is managed by the *AIX Logical Volume Manager (LVM)*. It virtualizes physical disks to be able to dynamically create, delete, resize, and move logical volumes for application use. To AIX our DS8000 logical volumes appear as physical SCSI disks. There are some considerations to take into account when configuring LVM.

### LVM striping

Striping is a technique for spreading the data in a logical volume across several physical disks in such a way that all disks are used in parallel to access data on one logical volume. The primary objective of striping is to increase the performance of a logical volume beyond that of a single physical disk.

In the case of a DS8000, LVM striping can be used to distribute data across more than one array (rank).

Refer to Chapter 12, “Performance considerations” on page 253 for a more detailed discussion of methods to optimize performance.

### LVM mirroring

LVM has the capability to mirror logical volumes across several physical disks. This improves availability, because in case a disk fails, there will be another disk with the same data. When creating mirrored copies of logical volumes, make sure that the copies are indeed distributed across separate disks.

With the introduction of SAN technology, LVM mirroring can even provide protection against a site failure. Using long wave Fibre Channel connections, a mirror can be stretched up to a 10 km distance.

Another application for LVM mirroring is online (non-disruptive) data migration. See Chapter 16, “Data migration in the open systems environment” on page 333.

## AIX access methods for I/O

AIX provides several modes to access data in a file system. It may be important for performance to choose the right access method.

### ***Synchronous I/O***

Synchronous I/O occurs while you wait. An application’s processing cannot continue until the I/O operation is complete. This is a very secure and traditional way to handle data. It ensures consistency at all times, but can be a major performance inhibitor. It also doesn’t allow the operating system to take full advantage of functions of modern storage devices, such as queueing, command reordering, and so on.

### ***Asynchronous I/O***

Asynchronous I/O operations run in the background and do not block user applications. This improves performance, because I/O and application processing run simultaneously. Many applications, such as databases and file servers, take advantage of the ability to overlap processing and I/O. They have to take measures to ensure data consistency, though. You can configure, remove, and change asynchronous I/O for each device using the **chdev** command or SMIT.

**Tip:** If the number of async I/O (AIO) requests is high, then the recommendation is to increase *maxservers* to approximately the number of simultaneous I/Os there might be. In most cases, it is better to leave the *minservers* parameter to the default value since the AIO kernel extension will generate additional servers if needed. By looking at the CPU utilization of the AIO servers, if the utilization is even across all of them, that means that they're all being used; you may want to try increasing their number in this case. Running **psstat -a** will allow you to see the AIO servers by name, and running **ps -k** will show them to you as the name *kproc*.

### **Direct I/O**

An alternative I/O technique called Direct I/O bypasses the Virtual Memory Manager (VMM) altogether and transfers data directly from the user's buffer to the disk and vice versa. The concept behind this is similar to raw I/O in the sense that they both bypass caching at the file system level. This reduces CPU overhead and makes more memory available to the database instance, which can make more efficient use of it for its own purposes.

Direct I/O is provided as a file system option in JFS2. It can be used either by mounting the corresponding file system with the **mount -o dio** option, or by opening a file with the `O_DIRECT` flag specified in the `open()` system call. When a file system is mounted with the **-o dio** option, all files in the file system use Direct I/O by default.

Direct I/O benefits applications that have their own caching algorithms by eliminating the overhead of copying data twice, first between the disk and the OS buffer cache, and then from the buffer cache to the application's memory.

For applications that benefit from the operating system cache, Direct I/O should not be used, because all I/O operations would be synchronous. Direct I/O also bypasses the JFS2 read-ahead. Read-ahead can provide a significant performance boost for sequentially accessed files.

### **Concurrent I/O**

In 2003, IBM introduced a new file system feature called *Concurrent I/O* (CIO) for JFS2. It includes all the advantages of Direct I/O and also relieves the serialization of write accesses. It improves performance for many environments, particularly commercial relational databases. In many cases, the database performance achieved using Concurrent I/O with JFS2 is comparable to that obtained by using raw logical volumes.

A method for enabling the concurrent I/O mode is to use the **mount -o cio** option when mounting a file system.

## **Boot device support**

The DS8100 and DS8300 are supported as boot devices on RS/6000 and pSeries that support Fibre Channel boot capability. This support is not currently available for the IBM eServer BladeCenter. Refer to *DS8000 Host Systems Attachment Guide*, SC26-7628, for additional information.

## **AIX on IBM iSeries**

With the announcement of the IBM iSeries i5, it is now possible to run AIX in a partition on the i5. This can be either AIX 5L V5.2 or V5.3. All supported functions of these operating system levels are supported on i5, including HACMP for high availability and external boot from Fibre Channel devices.

The DS8000 requires the following i5 I/O adapters to attach directly to an i5 AIX partition:

- ▶ 0611 Direct Attach 2 Gigabit Fibre Channel PCI
- ▶ 0625 Direct Attach 2 Gigabit Fibre Channel PCI-X

It is also possible for the AIX partition to have its storage *virtualized*, whereby a partition running OS/400 hosts the AIX partition's storage requirements. In this case, if using DS8000, it would be attached to the OS/400 partition using either of the following I/O adapters:

- ▶ 2766 2 Gigabit Fibre Channel Disk Controller PCI
- ▶ 2787 2 Gigabit Fibre Channel Disk Controller PCI-X

For more information on OS/400 support for DS8000, please see Appendix B, “Using DS8000 with iSeries” on page 373.

For more information on running AIX in an i5 partition, please refer to the i5 Information Center at:

[http://publib.boulder.ibm.com/infocenter/series/v1r2s/en\\_US/index.htm?info/iphathat/iphathat1par/kickoff.htm](http://publib.boulder.ibm.com/infocenter/series/v1r2s/en_US/index.htm?info/iphathat/iphathat1par/kickoff.htm)

**Note:** AIX will not run in a partition on earlier 8xx and prior iSeries systems.

## Monitoring I/O performance

### **iostat**

The **iostat** command is used to monitor system input/output device loading by observing the time the physical disks are active in relation to their average transfer rates. It also reports on CPU use. It provides data on the activity of physical volumes, not for file systems or logical volumes. Refer to “UNIX performance monitoring tools” on page 345 for more information.

### **filemon**

The **filemon** command monitors the performance of the file system, and reports the I/O activity with regard to files, virtual memory segments, logical volumes, and physical volumes.

Normally, **filemon** runs in the background while one or more applications are being executed and monitored. It automatically starts and monitors a trace of the program's file system and I/O events in real time. By default, the trace is started immediately, but it can be deferred until the user issues a **trcon** command. Tracing can be turned on and off with **tron** and **troff** as desired, while **filemon** is running. After stopping with **trcstop**, **filemon** generates an I/O activity report and exits. It writes its report to standard output or to a specified file. The report begins with a summary of the I/O activity for each of the levels being monitored and ends with detailed I/O activity statistics for each of the levels being monitored.

Example A-9 shows the output file of the following command:

```
filemon -v -o fmon.out -0 a1; sleep 30; trcstop
```

This monitors the activity at all file system levels for 30 seconds and writes a verbose report to the file **fmon.out**.

*Example: A-9 Filemon output file*

---

```
Wed Nov 17 16:59:43 2004
System: AIX part1 Node: 5 Machine: 00CFC02D4C00
Cpu utilization: 50.5%
```

```
Most Active Files
```



```

-----
#MBs #opns #rds #wrs file volume:inode
-----
0.3 1 70 0 unix <major=0,minor=5>:34096
0.0 1 2 0 ksh.cat <major=0,minor=5>:46237
0.0 1 2 0 cmdtrace.cat <major=0,minor=5>:45847
0.0 1 2 0 hosts <major=0,minor=4>:516
0.0 7 2 0 SWservAt <major=0,minor=4>:594
0.0 7 2 0 SWservAt.vc <major=0,minor=4>:595

```

Most Active Segments

```

-----
#MBs #rpgs #wpgs segid segtype volume:inode
-----
0.0 1 0 26fecd ???
0.0 1 0 23fec7 ???

```

Most Active Logical Volumes

```

-----
util #rblk #wblk KB/s volume description
-----
0.39 7776 11808 1164.0 /dev/u021v /u02
0.01 0 16896 1004.3 /dev/u041v /u04
0.00 0 3968 235.9 /dev/u031v /u03
0.00 16 0 1.0 /dev/hd2 /usr

```

Most Active Physical Volumes

```

-----
util #rblk #wblk KB/s volume description
-----
0.87 568 808 81.8 /dev/hdisk65 IBM MPIO FC 2107
0.87 496 800 77.0 /dev/hdisk79 IBM MPIO FC 2107
0.87 672 776 86.1 /dev/hdisk81 IBM MPIO FC 2107
0.86 392 960 80.4 /dev/hdisk63 IBM MPIO FC 2107
0.86 328 776 65.6 /dev/hdisk83 IBM MPIO FC 2107
0.86 528 624 68.5 /dev/hdisk69 IBM MPIO FC 2107
0.86 480 656 67.5 /dev/hdisk55 IBM MPIO FC 2107
0.86 408 536 56.1 /dev/hdisk73 IBM MPIO FC 2107
0.86 456 720 69.9 /dev/hdisk77 IBM MPIO FC 2107
0.86 440 720 68.9 /dev/hdisk59 IBM MPIO FC 2107
...

```

Detailed Physical Volume Stats (512 byte blocks)

```

-----
VOLUME: /dev/hdisk65 description: IBM MPIO FC 2107
reads: 37 (0 errs)
  read sizes (blks): avg 15.4 min 8 max 16 sdev 2.2
  read times (msec): avg 6.440 min 0.342 max 10.826 sdev 3.301
  read sequences: 37
  read seq. lengths: avg 15.4 min 8 max 16 sdev 2.2
writes: 52 (0 errs)
  write sizes (blks): avg 15.5 min 8 max 16 sdev 1.9
  write times (msec): avg 0.809 min 0.004 max 2.963 sdev 0.906
  write sequences: 52
  write seq. lengths: avg 15.5 min 8 max 16 sdev 1.9
seeks: 89 (100.0%)
  seek dist (blks): init 10875128,
                   avg 5457737.4 min 16 max 22478224 sdev 4601825.0
  seek dist (%tot blks):init 27.84031,

```

```
                                avg 13.97180 min 0.00004 max 57.54421 sdev 11.78066
time to next req(msec): avg 89.470 min 0.003 max 949.025 sdev 174.947
throughput:                    81.8 KB/sec
utilization:                   0.87
```

...

---

## Linux

Linux is an open source UNIX-like kernel, originally created by Linus Torvalds. The term “Linux” is often used to mean the whole operating system, GNU/Linux. The Linux kernel, along with the tools and software needed to run an operating system, are maintained by a loosely organized community of thousands of, mostly, volunteer programmers.

There are several organizations (distributors) that bundle the Linux kernel, tools, and applications to form a “distribution,” a package that can be downloaded or purchased and installed on a computer. Some of these distributions are commercial, others are not.

### Support issues that distinguish Linux from other operating systems

Linux is different from the other, proprietary, operating systems in many ways:

- ▶ There is no one person or organization that can be held responsible or called for support.
- ▶ Depending on the target group, the distributions differ largely in the kind of support that is available.
- ▶ Linux is available for almost all computer architectures.
- ▶ Linux is rapidly changing.

All these factors make it difficult to promise and provide generic support for Linux. As a consequence, IBM has decided on a support strategy that limits the uncertainty and the amount of testing.

IBM only supports the major Linux distributions that are targeted at enterprise customers:

- ▶ RedHat Enterprise Linux
- ▶ SUSE Linux Enterprise Server
- ▶ RedFlag Linux

These distributions have release cycles of about one year, are maintained for five years and require the user to sign a support contract with the distributor. They also have a schedule for regular updates. These factors mitigate the issues listed previously. The limited number of supported distributions also allows IBM to work closely with the vendors to ensure interoperability and support. Details about the supported Linux distributions can be found in the DS8000 Interoperability Matrix:

<http://www.ibm.com/servers/storage/disk/ds8000/pdf/ds8000-matrix.pdf>

See also “The DS8000 Interoperability Matrix” on page 321.

There are exceptions to this strategy when the market demand justifies the test and support effort.

## Existing reference material

There is a lot of information available that helps you set up your Linux server to attach it to a DS8000 storage subsystem.

### **The DS8000 Host Systems Attachment Guide**

The *DS8000 Host Systems Attachment Guide*, SC26-7628, provides instructions to prepare an Intel IA-32 based machine for DS8000 attachment, including:

- ▶ How to install and configure the FC HBA
- ▶ Peculiarities of the Linux SCSI subsystem
- ▶ How to prepare a system that boots from the DS8000

It is not very detailed with respect to the configuration and installation of the FC HBA drivers.

### **Implementing Linux with IBM Disk Storage**

The redbook *Implementing Linux with IBM Disk Storage*, SG24-6261, covers several hardware platforms and storage systems. It is not yet updated with information about the DS8000. The details provided for the attachment to the IBM Enterprise Storage Server (ESS 2105) are mostly valid for DS8000, too. Read it for information regarding storage attachment:

- ▶ Via FCP to an IBM eServer zSeries running Linux
- ▶ To an IBM eServer pSeries running Linux
- ▶ To an IBM eServer BladeCenter running Linux

It can be downloaded from:

<http://publib-b.boulder.ibm.com/abstracts/sg246261.html>

### **Linux with zSeries and ESS: Essentials**

The redbook, *Linux with zSeries and ESS: Essentials*, SG24-7025, provides a lot of information about Linux on IBM eServer zSeries and the ESS. It also describes in detail how the Fibre Channel (FCP) attachment of a storage system to zLinux works. It does not, however, describe the actual implementation. This information is at:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247025.pdf>

### **Getting Started with zSeries Fibre Channel Protocol**

The redpaper *Getting Started with zSeries Fibre Channel Protocol* is an older publication (last updated in 2003) which provides an overview of Fibre Channel (FC) topologies and terminology, and instructions to attach open systems (fixed block) storage devices via FCP to an IBM eServer zSeries running Linux. It can be found at:

<http://www.redbooks.ibm.com/redpapers/pdfs/redp0205.pdf>

### **Other sources of information**

Numerous hints and tips, especially for Linux on zSeries, are available on the IBM Redbooks technotes page:

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

IBM eServer zSeries dedicates its own Web page to storage attachment via FCP:

[http://www.ibm.com/servers/eserver/zseries/connectivity/ficon\\_resources.html](http://www.ibm.com/servers/eserver/zseries/connectivity/ficon_resources.html)

The zSeries connectivity support page lists all supported storage devices and SAN components that can be attached to a zSeries server. There is an extra section for FCP attachment:

<http://www.ibm.com/servers/eserver/zseries/connectivity/#fcp>

The whitepaper *ESS Attachment to United Linux 1 (IA-32)* is available at:

<http://www.ibm.com/support/docview.wss?uid=tss1td101235>

It is intended to help users to attach a server running an enterprise-level Linux distribution based on United Linux 1 (IA-32) to the IBM 2105 Enterprise Storage Server. It provides very detailed step by step instructions and a lot of background information about Linux and SAN storage attachment.

Another whitepaper, *Linux on IBM eServer pSeries SAN - Overview for Customers* describes in detail how to attach SAN storage (ESS 2105 and FASTT) to a pSeries server running Linux:

[http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux\\_san.pdf](http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_san.pdf)

Most of the information provided in these publications is valid for DS8000 attachment, although much of it was originally written for the ESS 2105.

## Important Linux issues

Linux treats SAN-attached storage devices like conventional SCSI disks. The Linux SCSI I/O subsystem has some peculiarities that are important enough to be described here, even if they show up in some of the publications listed in the previous section.

### Some Linux SCSI basics

Within the Linux kernel, device types are defined by *major numbers*. The instances of a given device type are distinguished by their *minor number*. They are accessed through special device files. For SCSI disks, the device files `/dev/sdx` are used, with *x* being a letter from a through z for the first 26 SCSI disks discovered by the system and continuing with aa, ab, ac, and so on, for subsequent disks. Due to the mapping scheme of SCSI disks and their partitions to major and minor numbers, each major number allows for only 16 SCSI disk devices. Therefore we need more than one major number for the SCSI disk device type. Table A-1 shows the assignment of special device files to major numbers.

Table A-1 Major numbers and special device files

Major number	First special device file	Last special device file
8	<code>/dev/sda</code>	<code>/dev/sdp</code>
65	<code>/dev/sdq</code>	<code>/dev/sdaf</code>
66	<code>/dev/sdag</code>	<code>/dev/sdav</code>
...		
71	<code>/dev/sddi</code>	<code>/dev/sddx</code>
128	<code>/dev/sddy</code>	<code>/dev/sden</code>
129	<code>/dev/sdeo</code>	<code>/dev/sdfd</code>
...		
135	<code>/dev/sdig</code>	<code>/dev/sdiv</code>

Each SCSI device can have up to 15 partitions, which are represented by the special device files /dev/sda1, /dev/sda2, and so on. The mapping of partitions to special device files and major and minor numbers is shown in Table A-2.

Table A-2 Minor numbers, partitions and special device files

Major number	Minor number	Special device file	Partition
8	0	/dev/sda	all of 1st disk
8	1	/dev/sda1	1st partition of 1st disk
	...		
8	15	/dev/sda15	15th partition of 1st disk
8	16	/dev/sdb	all of 2nd disk
8	17	/dev/sdb1	1st partition of 2nd disk
	...		
8	31	/dev/sdb15	15th partition of 2nd disk
8	32	/dev/sdc	all of 3rd disk
	...		
8	255	/dev/sdp15	15th partition of 16th disk
65	0	/dev/sdq	all of 16th disk
65	1	/dev/sdq1	1st partition on 16th disk
...	...		

### Missing device files

The Linux distributors do not always create all the possible special device files for SCSI disks. If you attach more disks than there are special device files available, Linux will not be able to address them. You can create missing device files with the **mknod** command. The **mknod** command requires four parameters in a fixed order:

- ▶ The name of the special device file to create
- ▶ The type of the device, b stands for a block device, c for a character device
- ▶ The major number of the device
- ▶ The minor number of the device

Refer to the manpage of the **mknod** command for more details. Example A-10 shows the creation of special device files for the 17th SCSI disks and its first three partitions.

*Example: A-10 Create new special device files for SCSI disks*

---

```

mknod /dev/sdq b 65 0
mknod /dev/sdq1 b 65 1
mknod /dev/sdq2 b 65 2
mknod /dev/sdq3 b 65 3

```

---

After creating the device files you may have to change their owner, group, and file permission settings to be able to use them. Often, the easiest way to do this is by duplicating the settings of existing device files, as shown in Example A-11. Be aware that after this sequence of commands, all special device files for SCSI disks have the same permissions. If an application requires different settings for certain disks, you have to correct them afterwards.

*Example: A-11 Duplicating the permissions of special device files*

---

```
knox:~ # ls -l /dev/sda /dev/sda1
rw-rw---- 1 root disk 8, 0 2003-03-14 14:07 /dev/sda
rw-rw---- 1 root disk 8, 1 2003-03-14 14:07 /dev/sda1
knox:~ # chmod 660 /dev/sd*
knox:~ # chown root:disk /dev/sda*
```

---

## Managing multiple paths

If you assign a DS8000 volume to a Linux system through more than one path, it will see the same volume more than once. It will also assign more than one special device file to it. To utilize the path redundancy and increased I/O bandwidth, you need an additional layer in the Linux disk subsystem to recombine the multiple disks seen by the system into one, to manage the paths and to balance the load across them.

The IBM multipathing solution for DS8000 attachment to Linux on Intel IA-32 and IA-64 architectures, IBM pSeries and iSeries is the *IBM Subsystem Device Driver* (SDD) (see 15.2, “Subsystem Device Driver” on page 324). SDD for Linux is available in the Linux RPM package format for all supported distributions from the SDD download site. It is proprietary and binary only. It only works with certain kernel versions with which it was tested. The README file on the SDD for Linux download page contains a list of the supported kernels.

The version of the *Linux Logical Volume Manager* that comes with all current Linux distributions does not support its physical volumes being placed on SDD vpath devices.

SDD is not available for Linux on zSeries. SUSE Linux Enterprise Server 8 for zSeries comes with built-in multipathing provided by a patched Logical Volume Manager. Today there is no multipathing support for Redhat Enterprise Linux for zSeries.

## Limited number of SCSI devices

Due to the design of the Linux SCSI I/O subsystem in the Linux Kernel version 2.4, the number of SCSI disk devices is limited to 256. Attaching devices through more than one path reduces this number. If, for example, all disks were attached through 4 paths, only up to 64 disks could be used.

**Important:** The latest update to the SUSE Linux Enterprise Server 8, Service Pack 3, uses a more dynamic method of assigning major numbers and allows the attachment of up to 2304 SCSI devices.

## SCSI device assignment changes

Linux assigns special device files to SCSI disks in the order they are discovered by the system. Adding or removing disks can change this assignment. This can cause serious problems if the system configuration is based on special device names (for example, a file system that is mounted using the /dev/sda1 device name). You can avoid some of them by using:

- ▶ Disk Labels instead of device names in /etc/fstab
- ▶ LVM Logical Volumes instead of /dev/sd.. devices for file systems

- ▶ SDD, which creates a persistent relationship between a DS8000 volume and a vpath device regardless of the /dev/sd.. devices

## RedHat Enterprise Linux (RH-EL) multiple LUN support

RH-EL by default is not configured for multiple LUN support. It will only discover SCSI disks addressed as LUN 0. The DS8000 provides the volumes to the host with a fixed Fibre Channel address and varying LUN. Therefore RH-EL 3 will see only one DS8000 volume (LUN 0), even if more are assigned to it.

Multiple LUN support can be added with an option to the SCSI midlayer Kernel module `scsi_mod`. To have multiple LUN support added permanently at boot time of the system, add the following line to the file `/etc/modules.conf`:

```
options scsi_mod max_scsi_luns=128
```

After saving the file, rebuild the module dependencies by running:

```
depmod -a
```

Now you have to rebuild the InitialRAMDisk, using the command:

```
mkinitrd <initrd-image> <kernel-version>
```

Issue `mkinitrd -h` for more help information. A reboot is required to make the changes effective.

## Fibre Channel disks discovered before internal SCSI disks

In some cases, when the Fibre Channel HBAs are added to a RedHat Enterprise Linux system, they will be automatically configured in a way that they are activated at boot time, before the built-in parallel SCSI controller that drives the system disks. This will lead to shifted special device file names of the system disk and can result in the system being unable to boot properly.

To prevent the FC HBA driver from being loaded before the driver for the internal SCSI HBA you have to change the `/etc/modules.conf` file:

- ▶ Locate the lines containing `scsi_hostadapterx` entries where `x` is a number.
- ▶ Reorder these lines: first come the lines containing the name of the internal HBA driver module, then the ones with the FC HBA module entry.
- ▶ Renumber the lines: no number for the first entry, 1 for the second, 2 for the 3rd, and so on.

After saving the file, rebuild the module dependencies by running:

```
depmod -a
```

Now you have to rebuild the InitialRAMDisk, using the command:

```
mkinitrd <initrd-image> <kernel-version>
```

Issue `mkinitrd -h` for more help information. If you reboot now, the SCSI and FC HBA drivers will be loaded in the correct order.

Example A-12 on page 362 shows how the `/etc/modules.conf` file should look like with two Adaptec SCSI controllers and two QLogic 2340 FC HBAs installed. It also contains the line that enables multiple LUN support. Note that the module names will be different with different SCSI and Fibre Channel adapters.

*Example: A-12 Sample /etc/modules.conf*

---

```
scsi_hostadapter aic7xxx
scsi_hostadapter1 aic7xxx
scsi_hostadapter2 qla2300
scsi_hostadapter3 qla2300
options scsi_mod max_scsi_luns=128
```

---

## Adding FC disks dynamically

The commonly used way to discover newly attached DS8000 volumes is to unload and reload the Fibre Channel HBA driver. However, this action is disruptive to all applications that use Fibre Channel attached disks on this particular host.

A Linux system can recognize newly attached LUNs without unloading the FC HBA driver. The procedure slightly differs depending on the installed FC HBAs.

In case of QLogic HBAs issue the command:

```
echo "scsi-qlascan" > /proc/scsi/qla2300/<adapter-instance>
```

With Emulex HBAs, issue the command:

```
sh force_lpfsc_scan.sh "lpfc<adapter-instance>"
```

This script is not part of the regular device driver package and must be downloaded separately:

[http://www.emulex.com/ts/downloads/linuxfc/re1/201g/force\\_lpfsc\\_scan.sh](http://www.emulex.com/ts/downloads/linuxfc/re1/201g/force_lpfsc_scan.sh)

It requires the tool **dfc** to be installed under `/usr/sbin/lpfc`.

In both cases the command must be issued for each installed HBA, with the `<adapter-instance>` being the SCSI instance number of the HBA.

After the FC HBAs re-scan the fabric, you can make the new devices available to the system with the command:

```
echo "scsi add-single-device s c t l" > /proc/scsi/scsi
```

The quadruple `s c t l` is the physical address of the device:

- ▶ `s` is the SCSI instance of the FC HBA
- ▶ `c` is the channel (in our case always 0)
- ▶ `t` is the target address (usually 0, except if a volume is seen by a HBA more than once)
- ▶ `l` is the LUN

The new volumes are added after the already existing ones. The following examples illustrate this. Example A-13 shows the original disk assignment as it existed since the last system start.

*Example: A-13 SCSI disks attached at system start time*

---

```
/dev/sda - internal SCSI disk
/dev/sdb - 1st DS8000 volume, seen by HBA 0
/dev/sdc - 2nd DS8000 volume, seen by HBA 0
/dev/sdd - 1st DS8000 volume, seen by HBA 1
/dev/sde - 2nd DS8000 volume, seen by HBA 1
```

---



Example A-14 shows the SCSI disk assignment after one more DS8000 volume is added.

*Example: A-14 SCSI disks after dynamic addition of another DS8000 volume*

---

```
/dev/sda - internal SCSI disk
/dev/sdb - 1st DS8000 volume, seen by HBA 0
/dev/sdc - 2nd DS8000 volume, seen by HBA 0
/dev/sdd - 1st DS8000 volume, seen by HBA 1
/dev/sde - 2nd DS8000 volume, seen by HBA 1
/dev/sdf - new DS8000 volume, seen by HBA 0
/dev/sdg - new DS8000 volume, seen by HBA 1
```

---

The mapping of special device files is now different than it would have been if all three DS8000 volumes had been already present when the HBA driver was loaded. In other words: if the system is now restarted, the device ordering will change to what is shown in Example A-15. See also “SCSI device assignment changes” on page 360.

*Example: A-15 SCSI disks after dynamic addition of another DS8000 volume and reboot*

---

```
/dev/sda - internal SCSI disk
/dev/sdb - 1st DS8000 volume, seen by HBA 0
/dev/sdc - 2nd DS8000 volume, seen by HBA 0
/dev/sdd - new DS8000 volume, seen by HBA 0
/dev/sde - 1st DS8000 volume, seen by HBA 1
/dev/sdf - 2nd DS8000 volume, seen by HBA 1
/dev/sdg - new DS8000 volume, seen by HBA 1
```

---

## Gaps in the LUN sequence

The QLogic HBA driver cannot deal with gaps in the LUN sequence. When it tries to discover the attached volumes, it probes for the different LUNs, starting at LUN 0 and continuing until it reaches the first LUN without a device behind it.

When assigning volumes to a Linux host with QLogic FC HBAs, make sure LUNs start at 0 and are in consecutive order. Otherwise the LUNs after a gap will not be discovered by the host. Gaps in the sequence can occur when you assign volumes to a Linux host that are already assigned to another server.

The Emulex HBA driver behaves differently: it always scans all LUNs up to 127.

## Linux on IBM iSeries

Since OS/400 V5R1, it has been possible to run Linux in an iSeries partition. On iSeries models 270 and 8xx, the primary partition must run OS/400 V5R1 or higher and Linux is run in a secondary partition. For later i5 systems (models i520, i550, i570 and i595), Linux can run in any partition.

The DS8000 requires the following iSeries I/O adapters to attach directly to an iSeries or i5 Linux partition:

- ▶ 0612 Linux Direct Attach PCI
- ▶ 0626 Linux Direct Attach PCI-X

It is also possible for the Linux partition to have its storage *virtualized*, whereby a partition running OS/400 hosts the Linux partition's storage requirements. In this case, if using the DS8000, they would be attached to the OS/400 partition using either of the following I/O adapters:

- ▶ 2766 2 Gigabit Fibre Channel Disk Controller PCI

► 2787 2 Gigabit Fibre Channel Disk Controller PCI-X

For more information on OS/400 support for DS8000, please see Appendix B, “Using DS8000 with iSeries” on page 373.

More information on running Linux in an iSeries partition can be found in the iSeries Information Center at:

<http://publib.boulder.ibm.com/series/v5r2/ic2924/index.htm>

For running Linux in an i5 partition check, the i5 Information Center at:

[http://publib.boulder.ibm.com/infocenter/series/v1r2s/en\\_US/info/iphbi/iphbi.pdf](http://publib.boulder.ibm.com/infocenter/series/v1r2s/en_US/info/iphbi/iphbi.pdf)

## Troubleshooting and monitoring

### The /proc pseudo file system

The /proc pseudo file system is maintained by the Linux kernel and provides dynamic information about the system. The directory /proc/scsi contains information about the installed and attached SCSI devices.

The file /proc/scsi/scsi contains a list of all attached SCSI devices, including disk, tapes, processors, and so on. Example A-16 shows a sample /proc/scsi/scsi file.

*Example: A-16 Sample /proc/scsi/scsi file*

---

```
knox:~ # cat /proc/scsi/scsi
Attached devices:
Host: scsi0 Channel: 00 Id: 00 Lun: 00
  Vendor: IBM-ESXS Model: DTN036C1UCDY10F Rev: S25J
  Type: Direct-Access ANSI SCSI revision: 03
Host: scsi0 Channel: 00 Id: 08 Lun: 00
  Vendor: IBM Model: 32P0032a S320 1 Rev: 1
  Type: Processor ANSI SCSI revision: 02
Host: scsi2 Channel: 00 Id: 00 Lun: 00
  Vendor: IBM Model: 2107921 Rev: .545
  Type: Direct-Access ANSI SCSI revision: 03
Host: scsi2 Channel: 00 Id: 00 Lun: 01
  Vendor: IBM Model: 2107921 Rev: .545
  Type: Direct-Access ANSI SCSI revision: 03
Host: scsi2 Channel: 00 Id: 00 Lun: 02
  Vendor: IBM Model: 2107921 Rev: .545
  Type: Direct-Access ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 00
  Vendor: IBM Model: 2107921 Rev: .545
  Type: Direct-Access ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 01
  Vendor: IBM Model: 2107921 Rev: .545
  Type: Direct-Access ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 02
  Vendor: IBM Model: 2107921 Rev: .545
  Type: Direct-Access ANSI SCSI revision: 03
```

---

There also is an entry in /proc for each HBA, with driver and firmware levels, error counters, and information about the attached devices. Example A-17 on page 365 shows the condensed content of the entry for a QLogic Fibre Channel HBA.

*Example: A-17 Sample /proc/scsi/qla2300/x*

---

```
knox:~ # cat /proc/scsi/qla2300/2
QLogic PCI to Fibre Channel Host Adapter for ISP23xx:
    Firmware version: 3.01.18, Driver version 6.05.00b9
Entry address = c1e00060
HBA: QLA2312 , Serial# H28468
Request Queue = 0x21f8000, Response Queue = 0x21e0000
Request Queue count= 128, Response Queue count= 512
.
.
Login retry count = 012
Commands retried with dropped frame(s) = 0

SCSI Device Information:
scsi-qla0-adapter-node=200000e08b0b941d;
scsi-qla0-adapter-port=210000e08b0b941d;
scsi-qla0-target-0=5005076300c39103;

SCSI LUN Information:
(Id:Lun)
( 0: 0): Total reqs 99545, Pending reqs 0, flags 0x0, 0:0:81,
( 0: 1): Total reqs 9673, Pending reqs 0, flags 0x0, 0:0:81,
( 0: 2): Total reqs 100914, Pending reqs 0, flags 0x0, 0:0:81,
```

---

## Performance monitoring with iostat

The **iostat** command can be used to monitor the performance of all attached disks. It is shipped with every major Linux distribution, but not necessarily installed by default. It reads data provided by the kernel in `/proc/stats` and prints it in human readable format. See the manpage of **iostat** for more details.

## The generic SCSI tools

The SUSE Linux Enterprise Server comes with a set of tools that allow low-level access to SCSI devices. They are called the *sg tools*. They talk to the SCSI devices through the generic SCSI layer, which is represented by special device files `/dev/sg0`, `/dev/sg0`, and so on.

By default SLES 8 provides sg device files for up to 16 SCSI devices (`/dev/sg0` through `/dev/sg15`). Additional sg device files can be created using the command **mknod**. After creating new sg devices you should change their group setting from `root` to `disk`. Example A-18 shows the creation of `/dev/sg16`, which would be the first one to create.

*Example: A-18 Creation of new device files for generic SCSI devices*

---

```
mknod /dev/sg16 c 21 16
chgrp disk /dev/sg16
```

---

Useful sg tools are:

- ▶ **sg\_inq /dev/sgx** prints SCSI Inquiry data, such as the volume serial number.
- ▶ **sg\_scan** prints the `/dev/sg` → `scsihost`, channel, target, LUN mapping.
- ▶ **sg\_map** prints the `/dev/sd` → `/dev/sg` mapping.
- ▶ **sg\_readcap** prints the block size and capacity (in blocks) of the device.
- ▶ **sginfo** prints SCSI inquiry and mode page data; it also allows manipulating the mode pages.

## Microsoft Windows 2000/2003

**Note:** Because Windows NT is no longer supported by Microsoft (and DS8000 support is provided on RPQ only), we do not discuss Windows NT here.

DS8000 supports FC attachment to Microsoft Windows 2000/2003 servers. For details regarding operating system versions and HBA types see the *DS8000 Interoperability Matrix*, available at:

<http://www.ibm.com/servers/storage/disk/ds8000/interop.html>

The support includes cluster service and acting as a boot device. Booting is supported currently with host adapters QLA23xx (32 bit or 64 bit) and LP9xxx (32 bit only). For a detailed discussion about SAN booting (advantages, disadvantages, potential difficulties, and troubleshooting) we highly recommend the Microsoft document *Boot from SAN in Windows Server 2003 and Windows 2000 Server*, available at:

<http://www.microsoft.com/windowsserversystem/storage/technologies/bootfromsan/bootfromsaninwindows.mspx>

### HBA and operating system settings

Depending on the host bus adapter type, several HBA and driver settings may be required. Refer to the *DS8000 Host Systems Attachment Guide*, SC26-7628, for the complete description of these settings. Although the volumes can be accessed with other settings too, the values recommended there have been tested for robustness.

To ensure optimum availability and recoverability when you attach a storage unit to a Windows 2000/2003 host system, we recommend setting the TimeOutValue value associated with the host adapters to 60 seconds. The operating system uses the TimeOutValue parameter to bind its recovery actions and responses to the disk subsystem. The value is stored in the Windows registry at:

```
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Disk\TimeOutValue
```

The value has the data type REG-DWORD and should be set to 0x0000003c hexadecimal (60 decimal).

### SDD for Windows

An important task with a Windows host is the installation of the SDD multipath driver. Ensure that SDD is installed before adding additional paths to a device. Otherwise, the operating system could lose the ability to access existing data on that device. For details, refer to the *IBM TotalStorage Multipath Subsystem Device Driver User's Guide*, SC30-4096. Here we highlight only some important items:

- ▶ SDD does not support I/O load balancing with Windows 2000 server clustering (MSCS). For Windows 2003, SDD 1.6.0.0 (or later) is required for load balancing with MSCS.
- ▶ When booting from the FC storage systems, special restrictions apply:
  - With Windows 2000, you should not use the same HBA as both the FC boot device and the clustering adapter. The reason for this is the usage of SCSI bus resets by MSCS to break up disk reservations during quorum arbitration. Because a bus reset cancels all pending I/O operations to all FC disks visible to the host via that port, an MSCS-initiated bus reset may cause operations on the C:\ drive to fail.

- With Windows 2003, MSCS uses target resets. See the Microsoft technical article *Microsoft Windows Clustering: Storage Area Networks at:*

<http://www.microsoft.com/windowsserver2003/techinfo/overview/san.msp>

Windows Server 2003 will allow for boot disk and the cluster server disks hosted on the same bus. However, you would need to use Storport miniport HBA drivers for this functionality to work. This is *not* a supported configuration in combination with drivers of other types (for example, SCSI port miniport or Full port drivers).

- If you reboot a system with adapters while the primary path is in a failed state, you must manually disable the BIOS on the first adapter and manually enable the BIOS on the second adapter. You cannot enable the BIOS for both adapters at the same time. If the BIOS for both adapters is enabled at the same time and there is a path failure on the primary adapter, the system will stop with an INACCESSIBLE\_BOOT\_DEVICE error upon reboot.

## Windows Server 2003 VDS support

With Windows Server 2003 Microsoft introduced the *Virtual Disk Service (VDS)*. It unifies storage management and provides a single interface for managing block storage virtualization. This interface is vendor and technology neutral, and is independent of the layer where virtualization is done, operating system software, RAID storage hardware, or other storage virtualization engines.

VDS is a set of APIs which uses two sets of providers to manage storage devices. The built-in *VDS software providers* enable you to manage disks and volumes at the operating system level. *VDS hardware providers* supplied by the hardware vendor enable you to manage hardware RAID arrays. Windows Server 2003 components that work with VDS include the Disk Management MMC snap-in, the **DiskPart** command-line tool, and the **DiskRAID** command-line tool, which is available in the Windows Server 2003 Deployment Kit. Figure A-1 shows the VDS architecture.

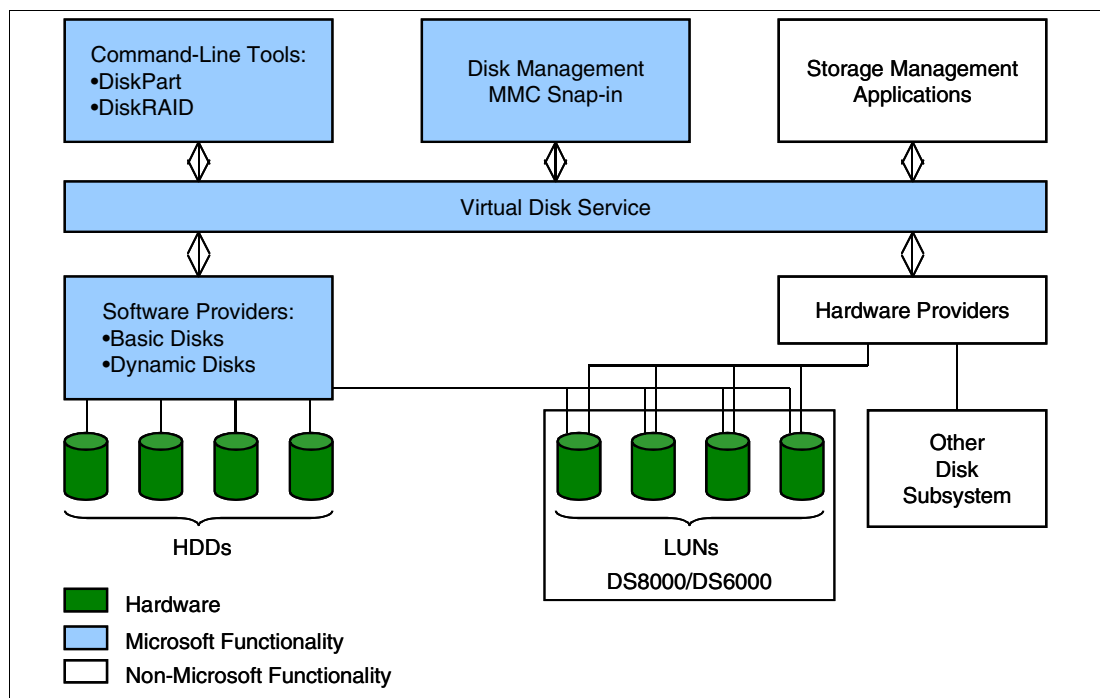


Figure A-1 Microsoft VDS Architecture

For a detailed description of VDS, refer to the *Microsoft Windows Server 2003 Virtual Disk Service Technical Reference* at:

[http://www.microsoft.com/Resources/Documentation/windowsserv/2003/all/techref/en-us/W2K3TR\\_vds\\_intro.asp](http://www.microsoft.com/Resources/Documentation/windowsserv/2003/all/techref/en-us/W2K3TR_vds_intro.asp)

The DS8000 can act as a VDS hardware provider. The implementation is based on the DS Common Information Model (CIM) agent, a middleware application that provides a CIM-compliant interface. The Microsoft Virtual Disk Service uses the CIM technology to list information and manage LUNs. See the *IBM TotalStorage DS Open Application Programming Interface Reference*, GC35-0493, for information on how to install and configure VDS support.

The following sections present examples of VDS integration with advanced functions of the DS8000 storage systems that became possible with the implementation of the DS CIM agent.

### ***Geographically Dispersed Sites***

Geographically Dispersed Sites (GDS) for MSCS is designed to provide high availability and a disaster recovery solution for clustered Microsoft Server environments. It integrates Microsoft Cluster Service (MSCS) and the Metro Mirror (PPRC) feature of the DS8000. It is designed to allow Microsoft Cluster installations to span geographically dispersed sites and help protect clients from site disasters or storage system failures. This solution is offered through IBM storage services.

For more details about GDS refer to:

[http://www.ibm.com/servers/storage/solutions/business\\_continuity/pdf/IBM\\_TotalStorage\\_GDS\\_Whitepaper.pdf](http://www.ibm.com/servers/storage/solutions/business_continuity/pdf/IBM_TotalStorage_GDS_Whitepaper.pdf)

### ***Volume Shadow Copy Service***

The Volume Shadow Copy Service provides a mechanism for creating consistent point-in-time copies of data, known as shadow copies. It integrates IBM TotalStorage FlashCopy to produce consistent shadow copies, while also coordinating with business applications, file-system services, backup applications and fast-recovery solutions.

For more information refer to:

[http://www.microsoft.com/resources/documentation/WindowsServ/2003/all/techref/en-us/w2k3tr\\_vss\\_how.asp](http://www.microsoft.com/resources/documentation/WindowsServ/2003/all/techref/en-us/w2k3tr_vss_how.asp)

## **HP OpenVMS**

DS8000 supports FC attachment of OpenVMS Alpha systems with operating system Version 7.3 or newer. For details regarding operating system versions and HBA types, see the *DS8000 Interoperability Matrix*, available at:

<http://www.ibm.com/servers/storage/disk/ds8000/interop.html>

The support includes clustering and multiple paths (exploiting the OpenVMS built-in multipathing). Boot support is available via Request for Price Quotations (RPQ). The DS API and the DS CLI are currently not available for OpenVMS.

## **FC port configuration**

The OpenVMS FC driver has some limitations in handling FC error recovery. The operating system may react to some situations with MountVerify conditions which are not recoverable. Affected processes may hang and eventually stop.

Instead of writing a special OpenVMS driver, it has been decided to handle this in the DS8000 host adapter microcode. As a result, DS8000 FC ports cannot be shared between OpenVMS and non-OpenVMS hosts.

**Important:** The DS8000 FC ports used by OpenVMS hosts must not be accessed by any other operating system, not even accidentally. The OpenVMS hosts have to be defined for access to these ports only, and it must be ensured that no foreign HBA (without definition as an OpenVMS host) is seen by these ports. Conversely, an OpenVMS host must have access only to DS8000 ports configured for OpenVMS compatibility.

You must dedicate storage ports for only the OpenVMS host type. Multiple OpenVMS systems can access the same port. Appropriate zoning must be enforced from the beginning. Wrong access to storage ports used by OpenVMS hosts may clear the OpenVMS-specific settings for these ports. This might remain undetected for a long time—until some failure happens, and by then I/Os might be lost. It is worth mentioning that OpenVMS is the only platform with such a restriction (usually, different open systems platforms can share the same DS8000 FC adapters).

## Volume configuration

OpenVMS Fibre Channel devices have device names according to the schema:

`$1$DGA<n>`

with the following elements:

- ▶ The first portion `$1$` of the device name is the allocation class (a decimal number in the range 1–255). FC devices always have the allocation class 1.
- ▶ The following two letters encode the drivers where the first letter denotes the device class (D = disks, M = magnetic tapes) and the second letter the device type (K = SCSI, G = Fibre Channel). So all Fibre Channel disk names contain the code DG.
- ▶ The third letter denotes the adapter channel (from range A to Z). Fibre Channel devices always have the channel identifier A.
- ▶ The number `<n>` is the User-Defined ID (UDID), a number from the range 0–32767 which is provided by the storage system in response to an OpenVMS-special SCSI inquiry command (from the range of command codes reserved by the SCSI standard for vendor's private use).

OpenVMS does not identify a Fibre Channel disk by its path or SCSI target/LUN like other operating systems. It relies on the UDID. Although OpenVMS uses the WWID to control all FC paths to a disk, a Fibre Channel disk which does not provide this additional UDID cannot be recognized by the operating system.

In the DS8000, the volume name acts as the UDID for OpenVMS hosts. If the character string of the volume name evaluates to an integer in the range 0–32767, then this integer is replied as the answer when an OpenVMS host asks for the UDID.

The DS management utilities do not enforce UDID rules. They accept incorrect values that are not valid for OpenVMS. It is possible to assign the same UDID value to multiple DS8000 volumes. However, because the UDID is in fact the device ID seen by the operating system, several consistency rules have to be fulfilled. These rules are described in detail in the OpenVMS operating system documentation (see *HP Guidelines for OpenVMS Cluster Configurations*):

- ▶ Every FC volume must have a UDID that is unique throughout the OpenVMS cluster that accesses the volume. The same UDID may be used in a different cluster or for different stand-alone host.
- ▶ If the volume is planned for MSCP serving, then the UDID range is limited to 0–9999 (by operating system restrictions in the MSCP code).

OpenVMS system administrators tend to use elaborate schemes for assigning UDIDs, coding several hints about physical configuration into this logical ID, for instance odd/even values or reserved ranges to distinguish between multiple data centers, storage systems, or disk groups. Thus they must be able to provide these numbers without additional restrictions imposed by the storage system. In the DS8000 UDID is implemented with full flexibility, which leaves the responsibility about restrictions to the customer.

## Command Console LUN

HP StorageWorks FC controllers use LUN 0 as *Command Console LUN (CCL)* for exchanging commands and information with in-band management tools. This concept is similar to the Access LUN of IBM TotalStorage DS4000 (FAStT) controllers.

Because the OpenVMS FC driver has been written with StorageWorks controllers in mind, OpenVMS always considers LUN 0 as CCL, never presenting this LUN as disk device. On HP StorageWorks HSG and HSV controllers, you cannot assign LUN 0 to a volume.

The DS8000 assigns LUN numbers per host using the lowest available number. The first volume that is assigned to a host becomes this host's LUN 0, the next volume is LUN 1, and so on.

Because OpenVMS considers LUN 0 as CCL, the first DS8000 volume assigned to the host cannot be used even when a correct UDID has been defined. So we recommend creating the first OpenVMS volume with a minimum size as a *dummy volume* for usage as the CCL. Multiple OpenVMS hosts, even in different clusters, that access the same storage system, can share the same volume as LUN 0, because there will be no other activity to this volume. In large configurations with more than 256 volumes per OpenVMS host or cluster, it might be necessary to introduce another *dummy volume* (when LUN numbering starts again with 0).

Defining a UDID for the CCL is not required by the OpenVMS operating system. OpenVMS documentation suggests that you always define a unique UDID since this identifier causes the creation of a CCL device visible for the OpenVMS command **show device** or other tools. Although an OpenVMS host cannot use the LUN for any other purpose, you can display the multiple paths to the storage device, and diagnose failed paths. Fibre Channel CCL devices have the OpenVMS device type GG.

## OpenVMS volume shadowing

OpenVMS disks can be combined in host-based mirror sets, called OpenVMS *shadow sets*. This functionality is often used to build disaster-tolerant OpenVMS clusters.

The OpenVMS shadow driver has been designed for disks according to DEC's *Digital Storage Architecture (DSA)*. This architecture, forward-looking in the 1980s, includes some requirements which are handled by today's SCSI/FC devices with other approaches. Two such things are the forced error indicator and the atomic revector operation for bad-block replacement.

When a DSA controller detects an unrecoverable media error, a spare block is revector to this logical block number, and the contents of the block are marked with a forced error. This



causes subsequent read operations to fail, which is the signal to the shadow driver to execute a repair operation using data from another copy.

However, there is no forced error indicator in the SCSI architecture, and the revector operation is nonatomic. As a substitute, the OpenVMS shadow driver exploits the SCSI commands READ LONG (READL) and WRITE LONG (WRITE L), optionally supported by some SCSI devices. These I/O functions allow data blocks to be read and written together with their disk device error correction code (ECC). If the SCSI device supports READL/WRITE L, OpenVMS shadowing emulates the DSA forced error with an intentionally incorrect ECC. For details see Scott H. Davis, Design of VMS Volume Shadowing Phase II — Host-based Shadowing, Digital Technical Journal Vol. 3 No. 3, Summer 1991, archived at:

<http://research.compaq.com/wr1/DECarchives/DTJ/DTJ301/DTJ301SC.TXT>

The DS8000 provides volumes as SCSI-3 devices and thus does not implement a forced error indicator. It also does not support the READL and WRITE L command set for data integrity reasons.

Usually the OpenVMS SCSI Port Driver recognizes if a device supports READL/WRITE L, and the driver sets the NOFE (no forced error) bit in the Unit Control Block. You can verify this setting with the SDA utility: After starting the utility with the **analyze/system** command, enter the **show device** command at the SDA prompt. Then the NOFE flag should be shown in the device's characteristics.

The OpenVMS command for mounting shadow sets provides a qualifier **/override=no\_forced\_error** to support non-DSA devices. To avoid possible problems (performance loss, unexpected error counts, or even removal of members from the shadow set), we recommend you apply this qualifier.





# B

## Using DS8000 with iSeries

In this appendix, the following topics are discussed:

- ▶ Supported environment
- ▶ Logical volume sizes
- ▶ Protected versus unprotected volumes
- ▶ Multipath
- ▶ Adding units to OS/400 configuration
- ▶ Sizing guidelines
- ▶ Migration
- ▶ Linux and AIX support

## Supported environment

Not all hardware and software combinations for OS/400 support the DS8000. This section describes the hardware and software pre-requisites for attaching the DS8000.

### Hardware

The DS8000 is supported on all iSeries models which support Fibre Channel attachment for external storage. Fibre Channel was supported on all model 8xx onwards. AS/400 models 7xx and prior only supported SCSI attachment for external storage, so they cannot support the DS8000.

There are two Fibre Channel adapters for iSeries. Both support the DS8000:

- ▶ 2766 2 Gigabit Fibre Channel Disk Controller PCI
- ▶ 2787 2 Gigabit Fibre Channel Disk Controller PCI-X

Each adapter requires its own dedicated I/O processor.

The iSeries Storage Web page provides information about current hardware requirements, including support for switches. This can be found at:

[http://www-1.ibm.com/servers/eserver/iseries/storage/storage\\_hw.html](http://www-1.ibm.com/servers/eserver/iseries/storage/storage_hw.html)

### Software

The iSeries must be running V5R2 or V5R3 (i5/OS) of OS/400. In addition, at the time of writing, the following PTFs are required:

- ▶ V5R2
  - MF33327, MF33301, MF33469, MF33302, SI14711 and SI14754
- ▶ V5R3
  - MF33328, MF33845, MF33437, MF33303, SI14690, SI14755 and SI14550

Prior to attaching the DS8000 to iSeries, you should check for the latest PTFs, which may have superseded those shown here.

## Logical volume sizes

OS/400 is supported on DS8000 as Fixed Block storage. Unlike other Open Systems using FB architecture, OS/400 only supports specific volume sizes and these may not be an exact number of extents. In general, these relate to the volume sizes available with internal devices, although some larger sizes are now supported for external storage only. OS/400 volumes are defined in decimal Gigabytes ( $10^9$  bytes).

Table B-1 gives the number of extents required for different iSeries volume sizes.

Table B-1 OS/400 logical volume sizes

Model type		OS/400 Device size (GB)	Number of LBAs	Extents	Unusable space (GiB)	Usable space%
Unprotected	Protected					
2107-A01	2107-A81	8.5	16,777,216	8	0.00	100.00
2107-A02	2107-A82	17.5	34,275,328	17	0.66	96.14

Model type		OS/400 Device size (GB)	Number of LBAs	Extents	Unusable space (GiB)	Usable space%
Unprotected	Protected					
2107-A05	2107-A85	35.1	68,681,728	33	0.25	99.24
2107-A04	2107-A84	70.5	137,822,208	66	0.28	99.57
2107-A06	2107-A86	141.1	275,644,416	132	0.56	99.57
2107-A07	2107-A87	282.2	551,288,832	263	0.13	99.95

**Note:** In Table B-1, GiB represents “Binary Gigabytes” ( $2^{30}$  bytes) and GB represents “Decimal Gigabytes” ( $10^9$  bytes)

When creating the logical volumes for use with OS/400, you will see that in almost every case, the OS/400 device size doesn’t match a whole number of extents, and so some space will be wasted. You should use the figures in Table B-1 on page 374 in conjunction with Figure 9-8 on page 176 to see how much space will be wasted for your specific configuration. You should also note that the #2766 and #2787 Fibre Channel Disk Adapters used by iSeries can only address 32 LUNs, so creating more, smaller LUNs will require more IOAs and their associated IOPs. For more sizing guidelines for OS/400, refer to “Sizing guidelines” on page 396.

## Protected versus unprotected volumes

When defining OS/400 logical volumes, you must decide whether these should be *protected* or *unprotected*. This is simply a notification to OS/400 – it does not mean that the volume is protected or unprotected. In reality, all DS8000 LUNs are protected, either RAID-5 or RAID-10. Defining a volume as unprotected means that it is available for OS/400 to mirror that volume to another of equal capacity — either internal or external. If you do not intend to use OS/400 (host based) mirroring, you should define your logical volumes as protected.

Under some circumstances, you may wish to mirror the OS/400 Load Source Unit (LSU) to a LUN in the DS8000. In this case, only one LUN should be defined as unprotected; otherwise, when mirroring is started to mirror the LSU to the DS8000 LUN, OS/400 will attempt to mirror all unprotected volumes.

## Changing LUN protection

It is not possible to simply change a volume from protected to unprotected, or vice versa. If you wish to do so, you must delete the logical volume. This will return the extents used for that volume to the extent pool. You will then be able to create a new logical volume with the correct protection. This is unlike ESS E20, F20, and 800, where the entire array containing the logical volume had to be reformatted.

However, before deleting the logical volume on the DS8000, you must first remove it from the OS/400 configuration (assuming it was still configured). This is an OS/400 task which is disruptive if the disk is in the System ASP or User ASPs 2-32 because it requires an IPL of OS/400 to completely remove the volume from the OS/400 configuration. This is no different from removing an internal disk from an OS/400 configuration. Indeed, deleting a logical volume on the DS8000 is similar to physically removing a disk drive from an iSeries. Disks can be removed from an Independent ASP with the IASP varied off without IPLing the system.

## Adding volumes to iSeries configuration

Once the logical volumes have been created and assigned to the host, they will appear as *non-configured units* to OS/400. This may be some time after being created on the DS8000. At this stage, they are used in exactly the same way as non-configured internal units. There is nothing particular to external logical volumes as far as OS/400 is concerned. You should use the same functions for adding the logical units to an Auxiliary Storage Pool (ASP) as you would for internal disks.

### Using 5250 interface

Adding disk units to the configuration can be done either using the green screen interface with Dedicated Service Tools (DST) or System Service Tools (SST), or with the iSeries Navigator GUI. The following example shows how to add a logical volume in the DS8000 to the System ASP, using green screen SST.

1. Start System Service Tools STRSST and sign on.
2. Select option **3**, Work with disk units as shown in Figure B-1.

```

                                     System Service Tools (SST)

Select one of the following:

    1. Start a service tool
    2. Work with active service tools
    3. Work with disk units
    4. Work with diskette data recovery
    5. Work with system partitions
    6. Work with system capacity
    7. Work with system security
    8. Work with service tools user IDs

Selection
    3

F3=Exit      F10=Command entry      F12=Cancel
```

Figure B-1 System Service Tools menu

3. Select Option **2**, Work with disk configuration as shown in Figure B-2 on page 376.

```

                                     Work with Disk Units

Select one of the following:

    1. Display disk configuration
    2. Work with disk configuration
    3. Work with disk unit recovery

Selection
    2

F3=Exit      F12=Cancel
```

Figure B-2 Work with Disk Units menu

- When adding disk units to a configuration, you can add them as empty units by selecting Option 2 or you can choose to allow OS/400 to balance the data across all the disk units. Normally, we recommend balancing the data. Select Option 8, Add units to ASPs and balance data as shown in Figure B-3.

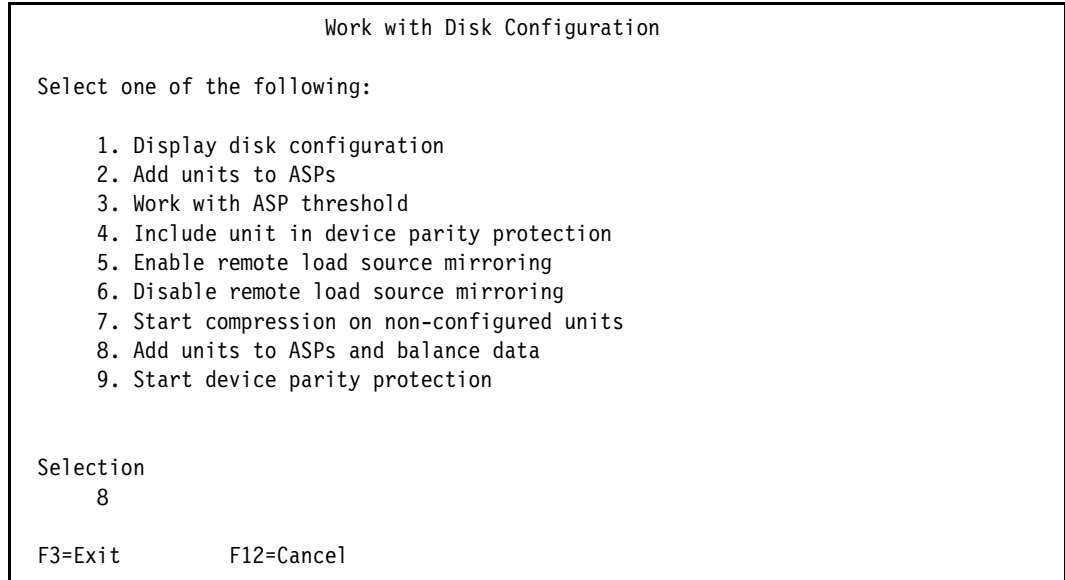


Figure B-3 Work with Disk Configuration menu

- Figure B-4 on page 377 shows the Specify ASPs to Add Units to panel. Specify the ASP number next to the desired units. Here we have specified ASP1, the System ASP. Press Enter.

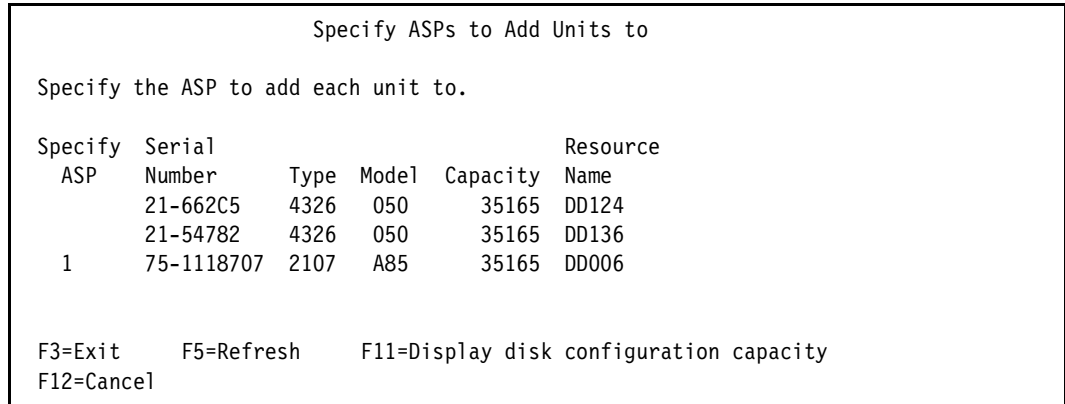


Figure B-4 Specify ASPs to Add Units to

- The Confirm Add Units panel will appear for review as shown in Figure B-5. If everything is correct, press **Enter** to continue.

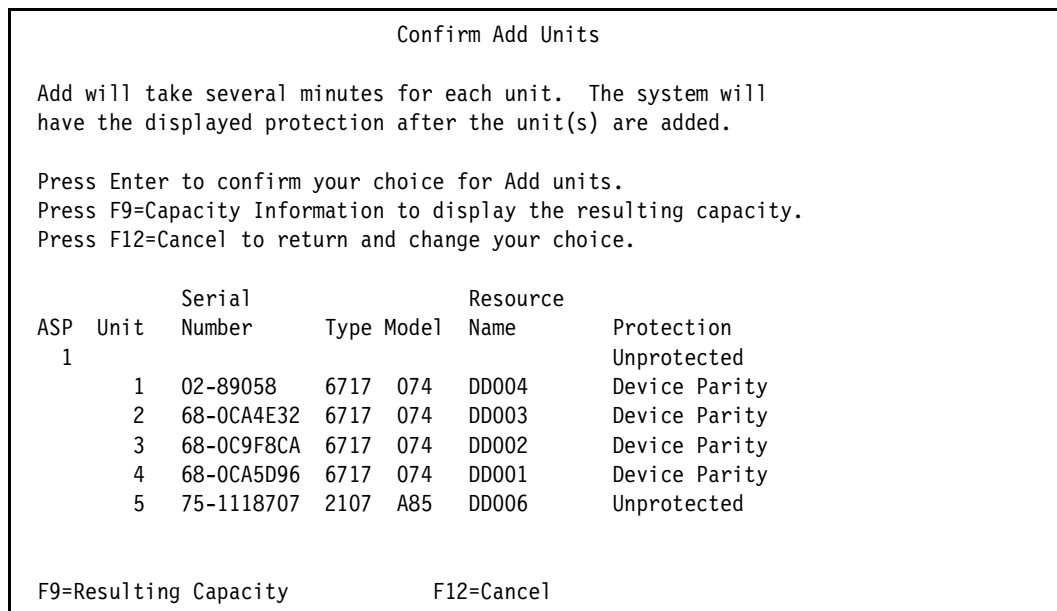


Figure B-5 Confirm Add Units

- Depending on the number of units you are adding, this step could take some time. When it completes, display your disk configuration to verify the capacity and data protection.

## Adding volumes to an Independent Auxiliary Storage Pool

Independent Auxiliary Storage Pools (IASPs) can be switchable or private. Disks are added to an IASP using the iSeries navigator GUI. In this example, we are adding a logical volume to a private (non-switchable) IASP.

- Start iSeries Navigator. Figure B-6 on page 378 shows the initial panel.

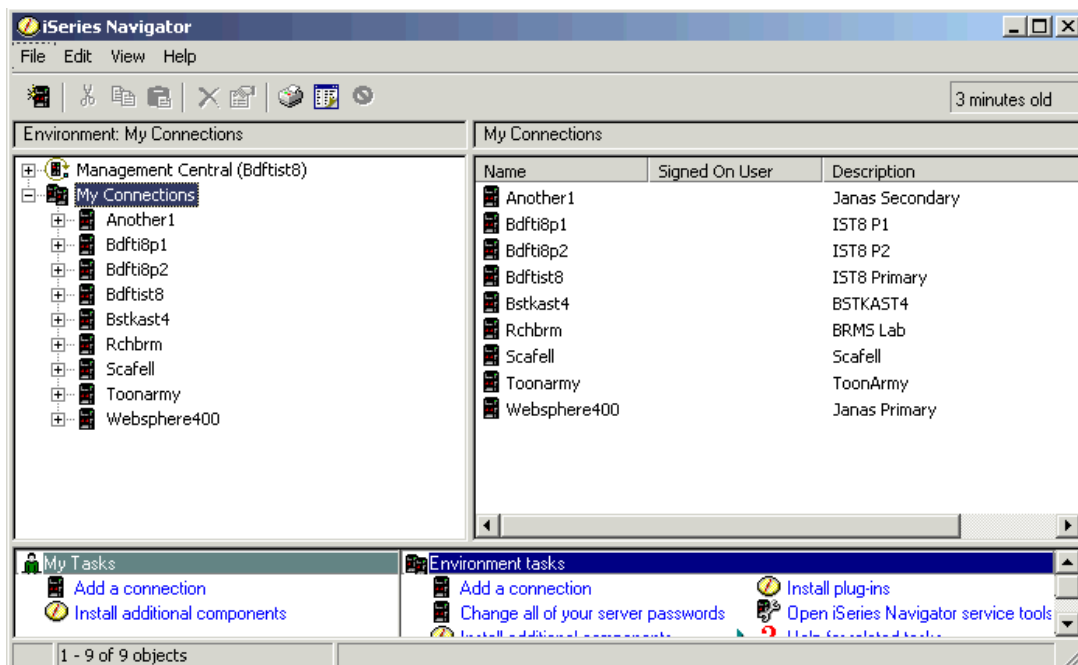


Figure B-6 iSeries Navigator initial panel



- Expand the iSeries to which you wish to add the logical volume and sign on to that server as shown in Figure B-7.

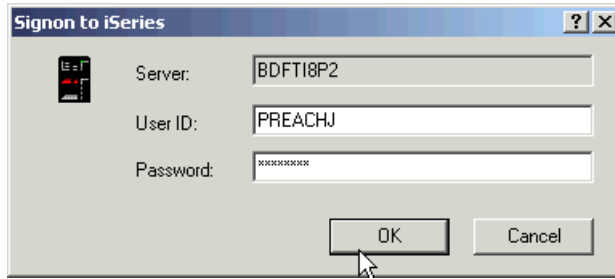


Figure B-7 iSeries Navigator Signon to iSeries window

- Expand **Configuration and Service, Hardware, and Disk Units** as shown in Figure B-8 on page 379.

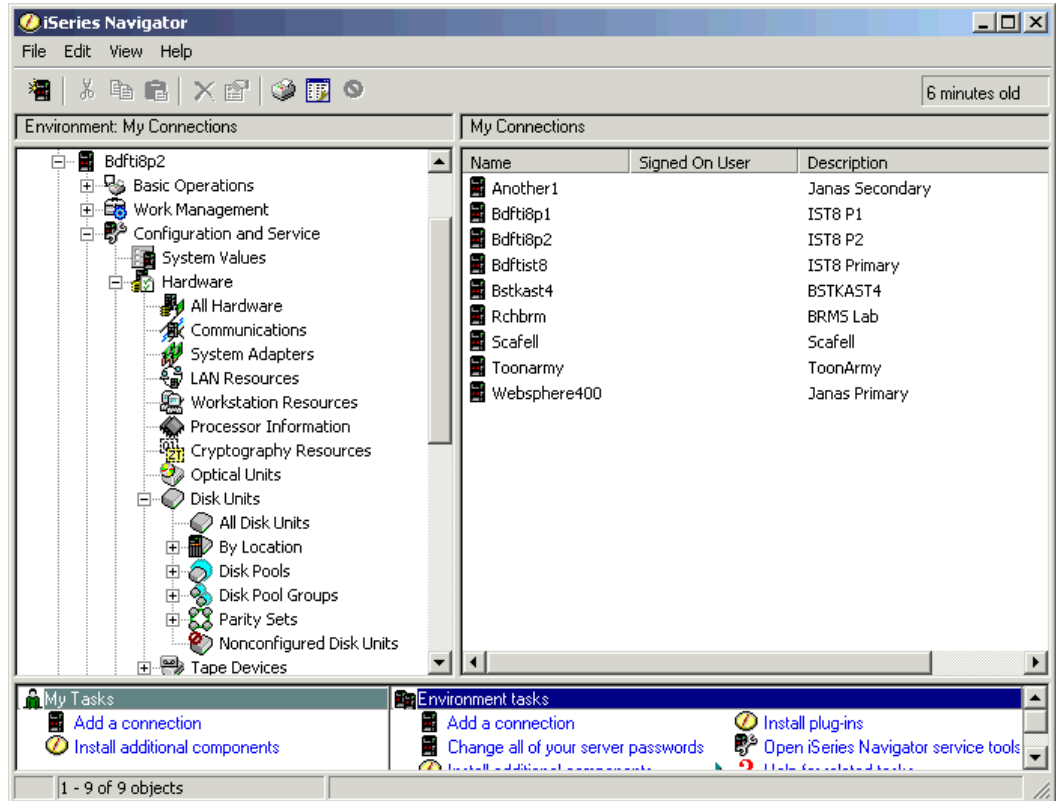


Figure B-8 iSeries Navigator Disk Units

- You will be asked to sign on to SST as shown in Figure B-9. Enter your Service tools ID and password and press **OK**.

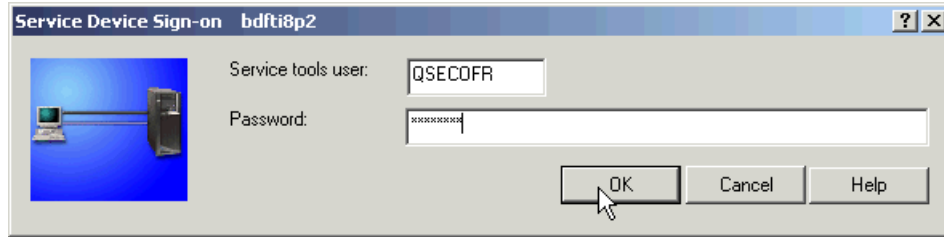


Figure B-9 SST Signon

5. Right-click **Disk Pools** and select **New Disk Pool** as shown in Figure B-10 on page 380.

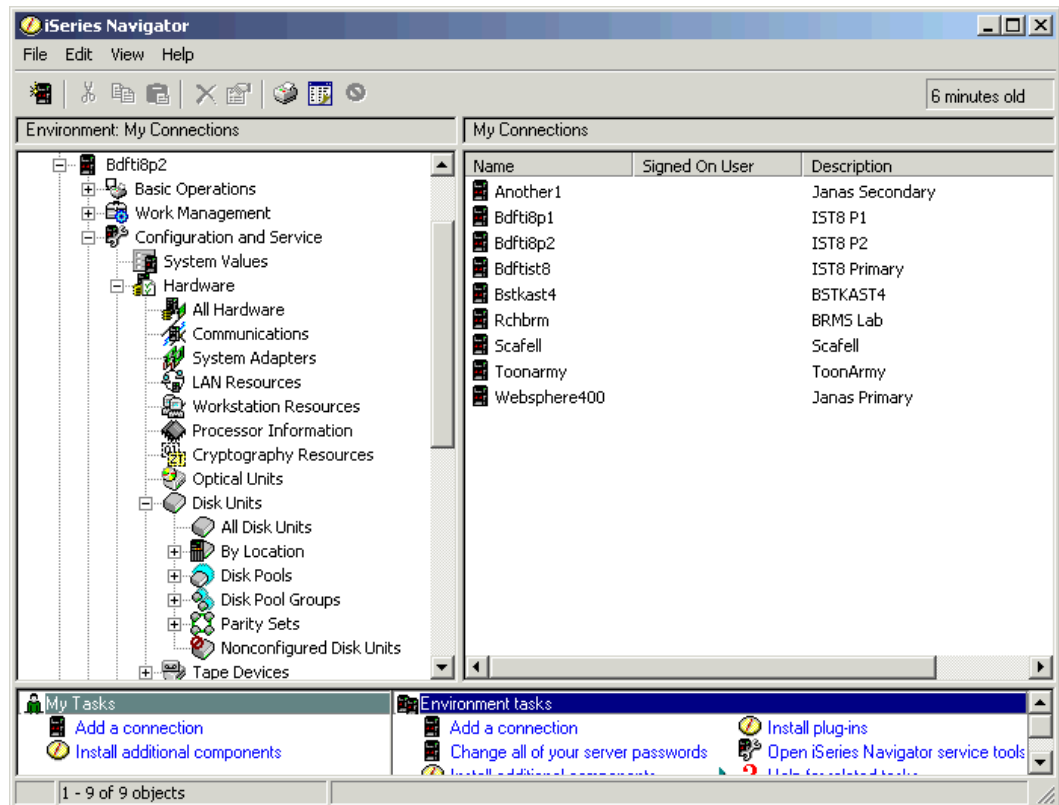


Figure B-10 Create a new disk pool

6. The New Disk Pool wizard appears as shown in Figure B-11. Click **Next**.

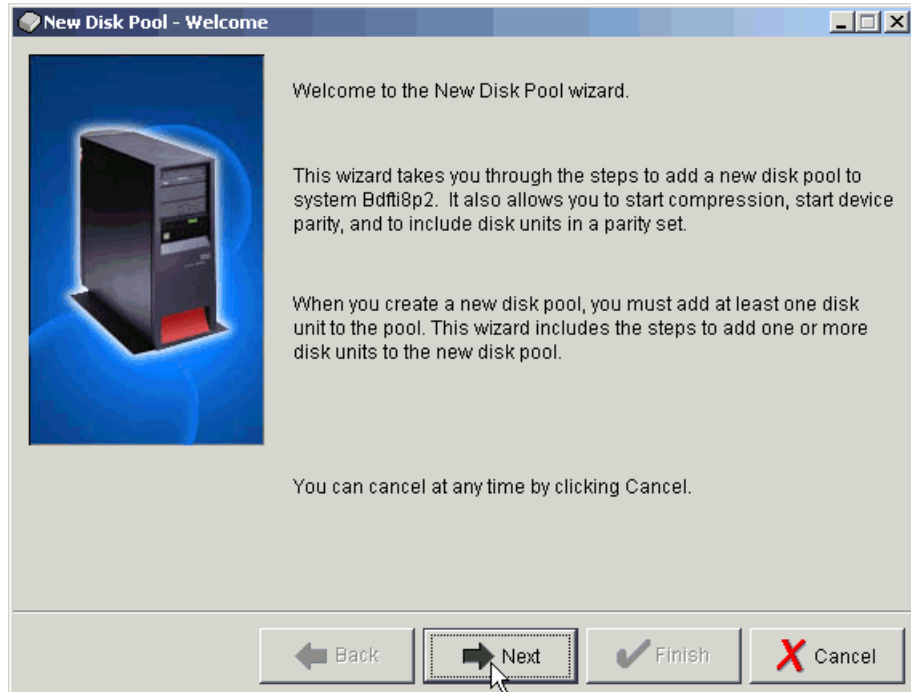


Figure B-11 New disk pool - welcome

7. On the New Disk Pool dialog shown in Figure B-12, select **Primary** from the pull-down for the Type of disk pool, give the new disk pool a name and leave Database to default to **Generated by the system**. Ensure the disk protection method matches the type of logical volume you are adding. If you leave it unchecked, you will see all available disks. Select **OK** to continue.

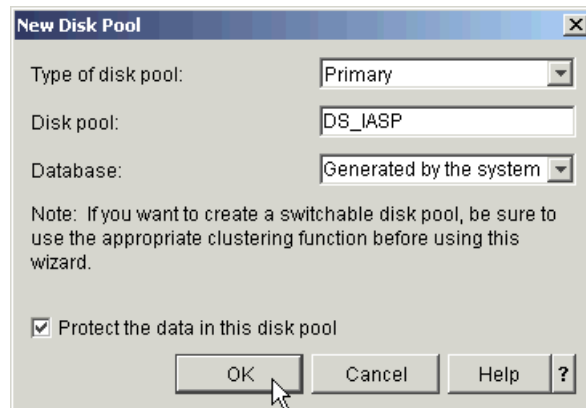


Figure B-12 Defining a new disk pool

8. A confirmation panel like that shown in Figure B-13 will appear to summarize the disk pool configuration. Select **Next** to continue.

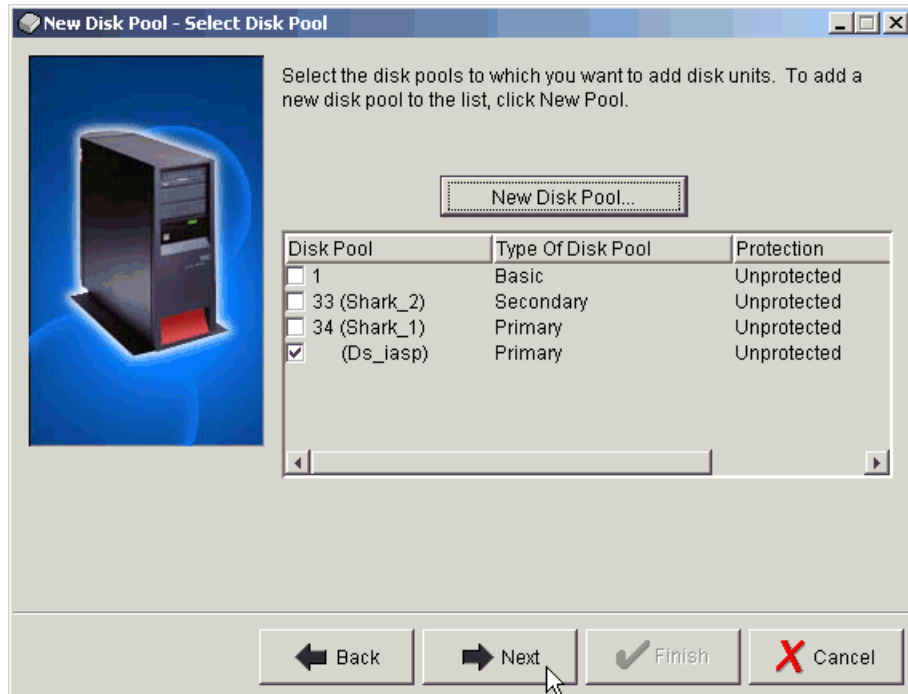


Figure B-13 Confirm disk pool configuration

- Now you need to add disks to the new disk pool. On the Add to disk pool screen, click the **Add disks** button as shown in Figure B-14 on page 382.

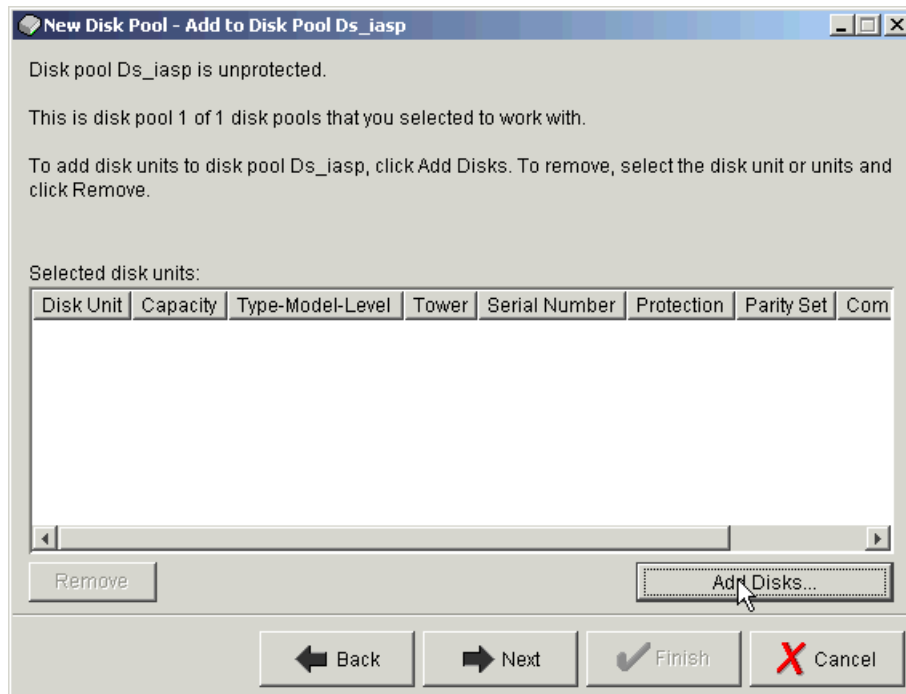


Figure B-14 Add disks to Disk Pool

- A list of non-configured units similar to that shown in Figure B-15 will appear. Highlight the disks you want to add to the disk pool and click **Add**.

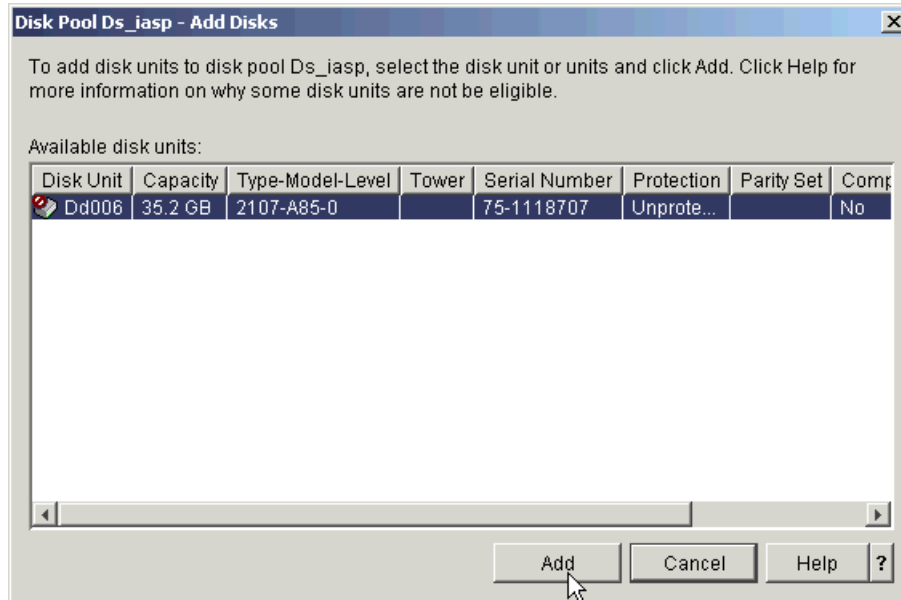


Figure B-15 Choose the disks to add to the Disk Pool

11. A confirmation screen appears as shown in Figure B-16 on page 383. Click **Next** to continue.

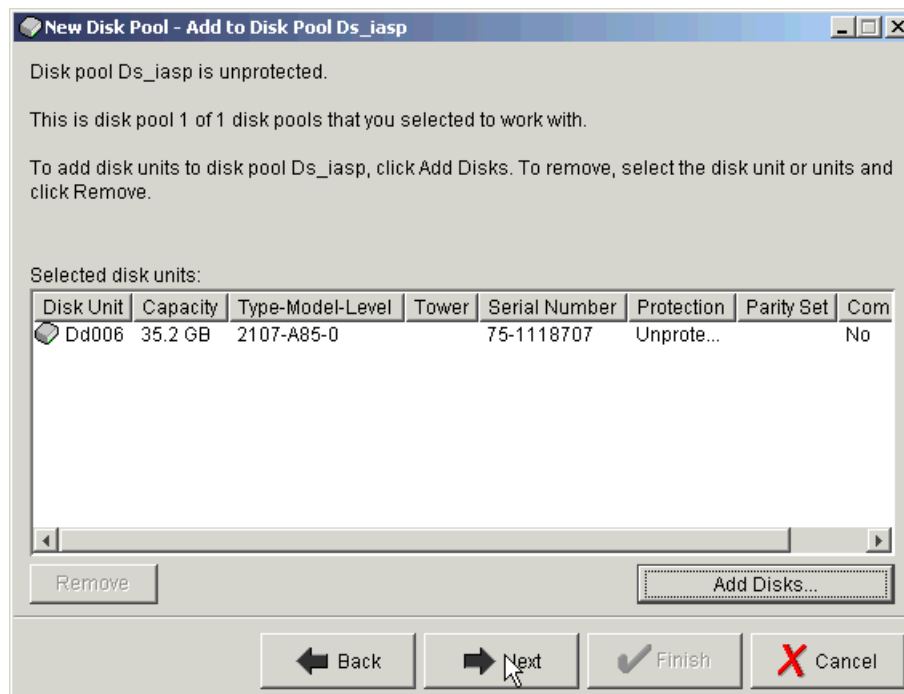


Figure B-16 Confirm disks to be added to Disk Pool

12. A summary of the Disk Pool configuration similar to Figure B-17 appears. Click **Finish** to add the disks to the Disk Pool.

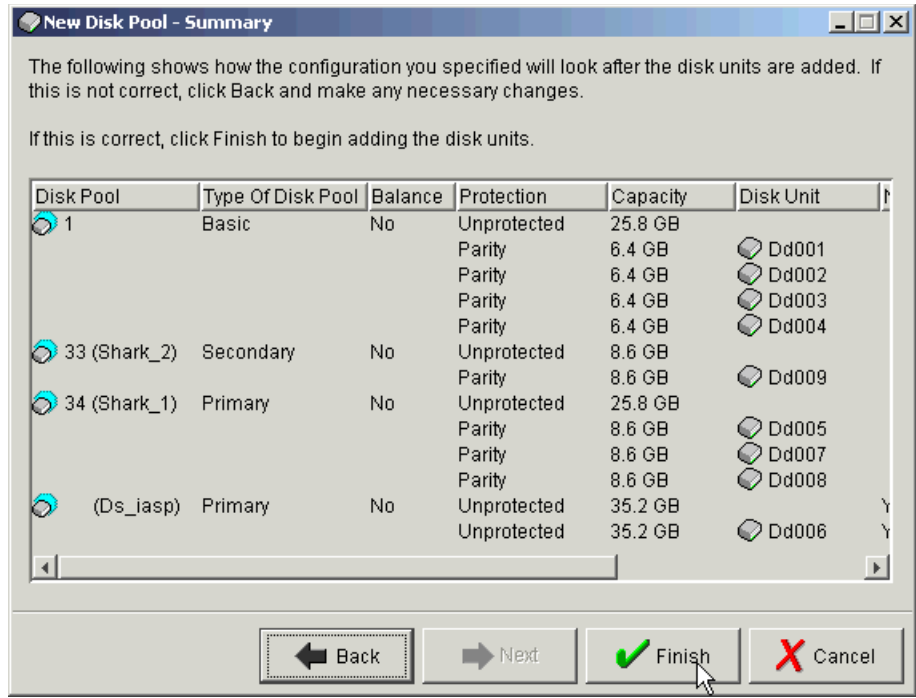


Figure B-17 New Disk Pool Summary

13. Take note of and respond to any message dialogs which appear. After taking action on any messages, the New Disk Pool Status panel shown in Figure B-18 on page 384 will appear showing progress. This step may take some time, depending on the number and size of the logical units being added.

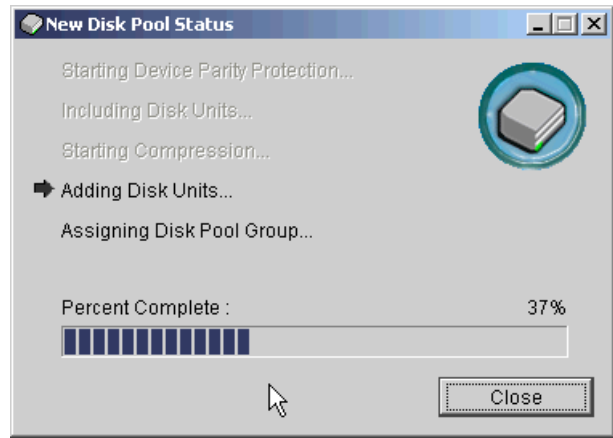


Figure B-18 New Disk Pool Status

14. When complete, click **OK** on the information panel shown in Figure B-19.

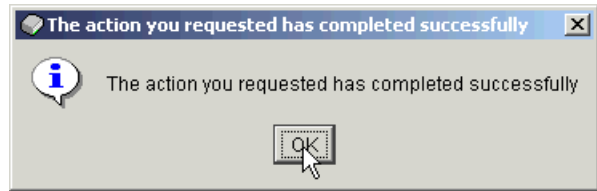


Figure B-19 Disks added successfully to Disk Pool

15. The new Disk Pool can be seen on iSeries Navigator **Disk Pools** in Figure B-20.

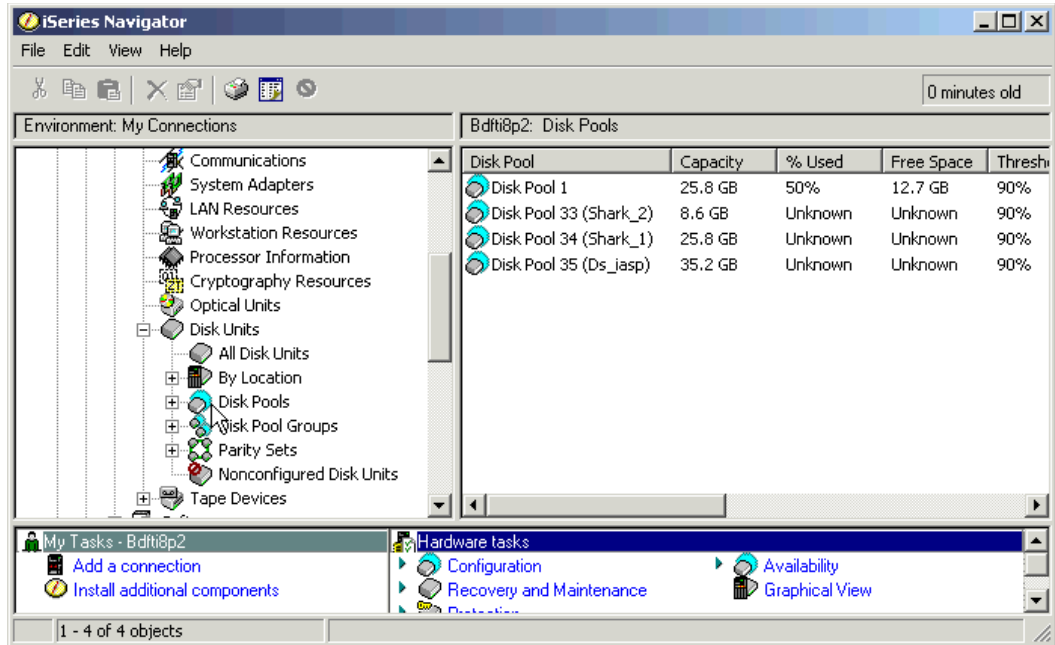


Figure B-20 New Disk Pool shown on iSeries Navigator

16. To see the logical volume, as shown in Figure B-21, expand **Configuration and Service, Hardware, Disk Pools** and click the disk pool you just created.

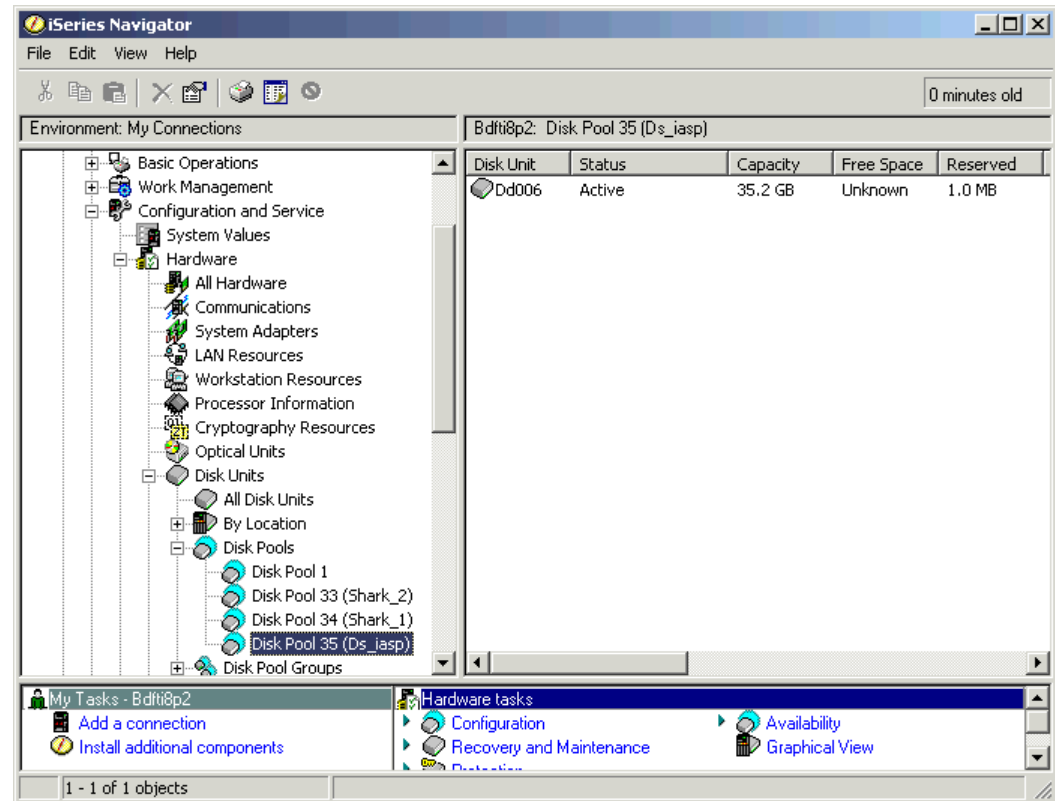


Figure B-21 New logical volume shown on iSeries Navigator

# Multipath

Multipath support was added for external disks in V5R3 of i5/OS (also known as OS/400 V5R3). Unlike other platforms which have a specific software component, such as *Subsystem Device Driver* (SDD), multipath is part of the base operating system. At V5R3, up to eight connections can be defined from multiple I/O adapters on an iSeries server to a single logical volume in the DS8000. Each connection for a multipath disk unit functions independently. Several connections provide availability by allowing disk storage to be utilized even if a single path fails.

Multipath is important for iSeries because it provides greater resilience to SAN failures, which can be critical to OS/400 due to the single level storage architecture. Multipath is not available for iSeries internal disk units but the likelihood of path failure is much less with internal drives. This is because there are fewer interference points where problems can occur, such as long fiber cables and SAN switches, as well as the increased possibility of human error when configuring switches and external storage, and the concurrent maintenance on the DS8000 which may make some paths temporarily unavailable.

Many iSeries customers still have their entire environment in the System ASP and loss of access to any disk will cause the system to fail. Even with User ASPs, loss of a UASP disk will eventually cause the system to stop. Independent ASPs provide isolation such that loss of disks in the IASP will only affect users accessing that IASP while the rest of the system is unaffected. However, with multipath, even loss of a path to disk in an IASP will not cause an outage.

Prior to multipath being available, some customers used OS/400 mirroring to two sets of disks, either in the same or different external disk subsystems. This provided implicit dual-path as long as the mirrored copy was connected to a different IOP/IOA, BUS, or I/O tower. However, this also required two copies of data. Since disk level protection is already provided by RAID-5 or RAID-10 in the external disk subsystem, this was sometimes seen as unnecessary.

With the combination of multipath and RAID-5 or RAID-10 protection in the DS8000, we can provide full protection of the data paths and the data itself without the requirement for additional disks.

## Avoiding single points of failure

In Figure B-22, there are fifteen single points of failure, excluding the iSeries itself and the DS8000 storage facility. Failure points 9-12 will not be present if you do not use an *Inter Switch Link* (ISL) to extend your SAN. An outage to any one of these components (either planned or unplanned) would cause the system to fail if IASPs are not used (or the applications within an IASP if they are).



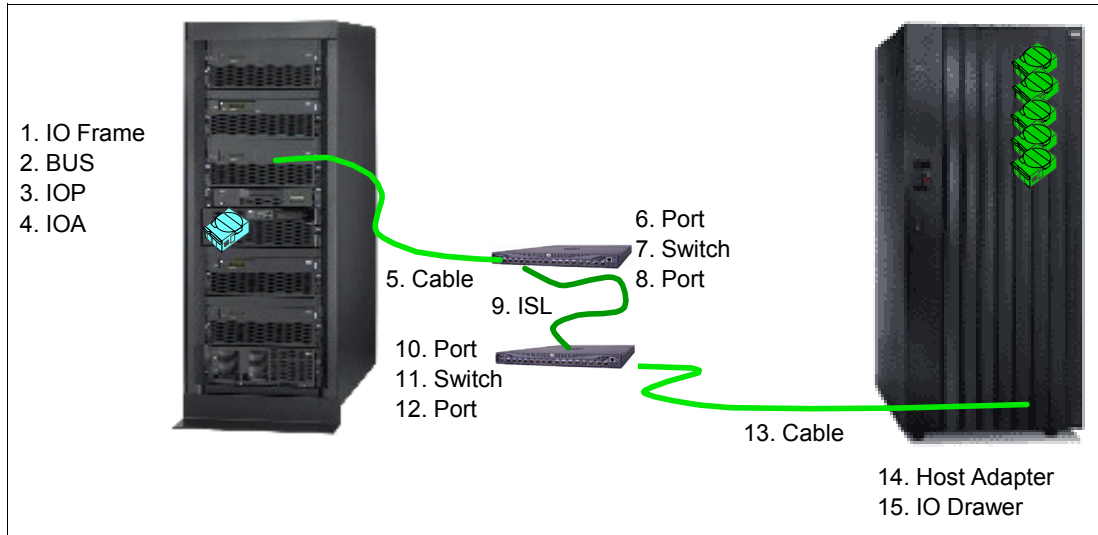


Figure B-22 Single points of failure

When implementing multipath, you should provide as much redundancy as possible. As a minimum, multipath requires two IOAs connecting the same logical volumes. Ideally, these should be on different buses and in different I/O racks in the iSeries. If a SAN is included, separate switches should also be used for each path. You should also use Host Adapters in different I/O drawer pairs in the DS8000. Figure B-23 on page 387 shows this.

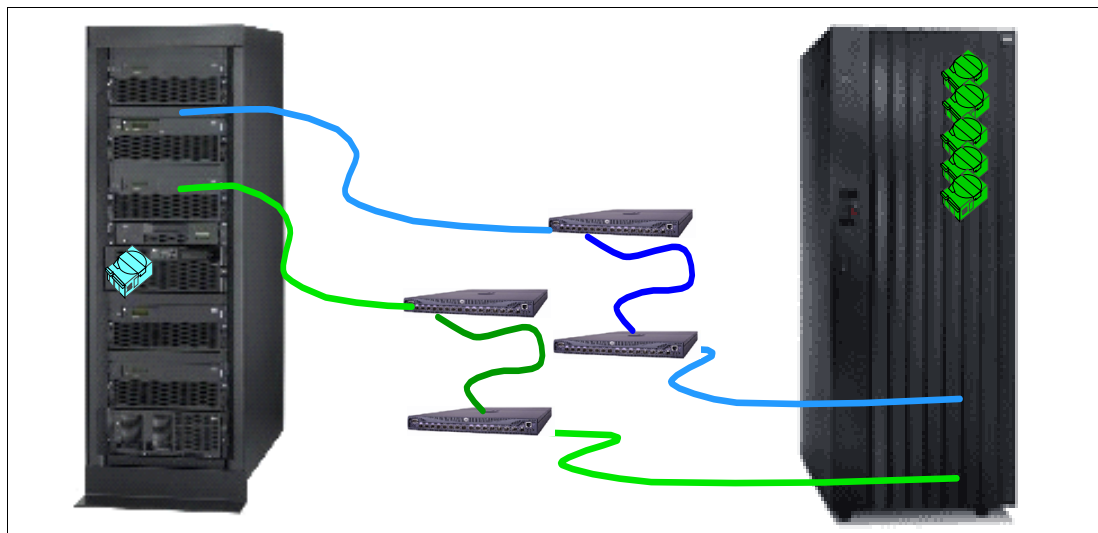


Figure B-23 Multipath removes single points of failure

Unlike other systems, which may only support two paths (dual-path), OS/400 V5R3 supports up to eight paths to the same logical volumes. As a minimum, you should use two, although some small performance benefits may be experienced with more. However, since OS/400 multipath spreads I/O across all available paths in a *round-robin* manner, there is no *load balancing*, only *load sharing*.

## Configuring multipath

iSeries has two I/O adapters that support DS8000:

- ▶ 2766 2 Gigabit Fibre Channel Disk Controller PCI

► 2787 2 Gigabit Fibre Channel Disk Controller PCI-X

Both can be used for multipath and there is no requirement for all paths to use the same type of adapter. Both adapters can address up to 32 logical volumes. This does not change with multipath support. When deciding how many I/O adapters to use, your first priority should be to consider performance throughput of the IOA since this limit may be reached before the maximum number of logical units. See “Sizing guidelines” on page 396 for more information on sizing and performance guidelines.

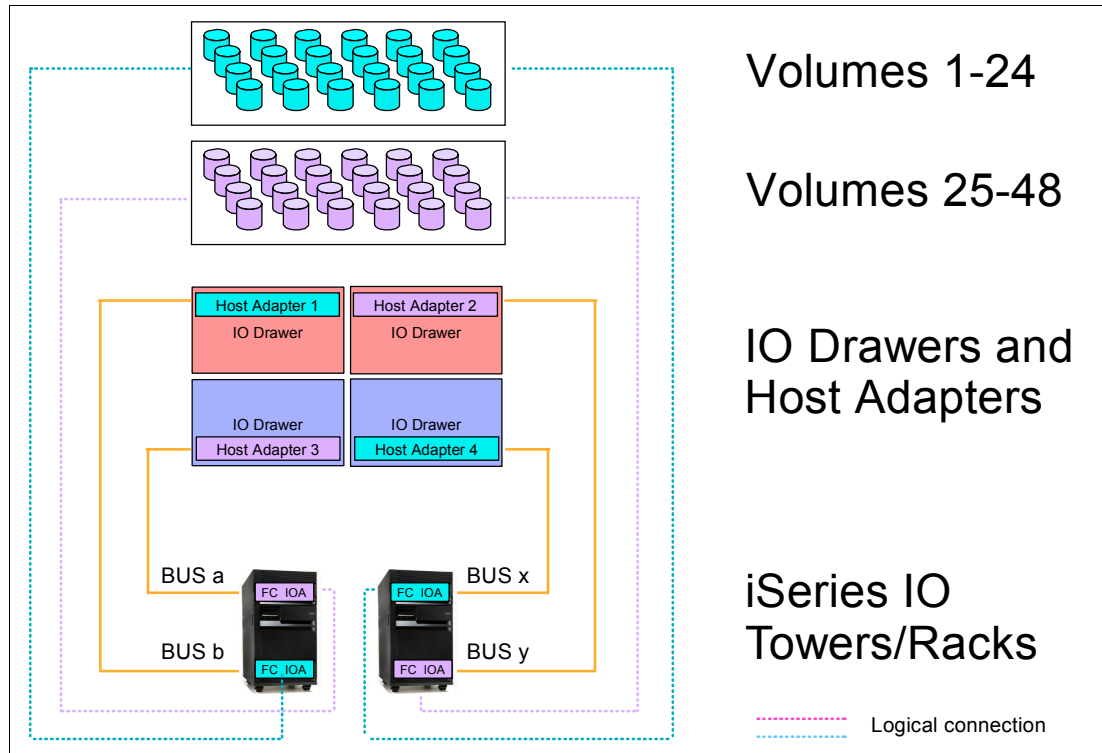


Figure B-24 Example of multipath with iSeries

Figure B-24 shows an example where 48 logical volumes are configured in the DS8000. The first 24 of these are assigned via a host adapter in the top left I/O drawer in the DS8000 to a Fibre Channel I/O adapter in the first iSeries I/O tower or rack. The next 24 logical volumes are assigned via a host adapter in the lower left I/O drawer in the DS8000 to a Fibre Channel I/O adapter on a different BUS in the first iSeries I/O tower or rack. This would be a valid single path configuration.

To implement multipath, the first group of 24 logical volumes is also assigned to a Fibre Channel I/O adapter in the second iSeries I/O tower or rack via a host adapter in the lower right I/O drawer in the DS8000. The second group of 24 logical volumes is also assigned to a Fibre Channel I/O adapter on a different BUS in the second iSeries I/O tower or rack via a host adapter in the upper right I/O drawer.

## Adding multipath volumes to iSeries using 5250 interface

If using the green screen 5250 interface, sign on to SST and perform the following steps as described in “Using 5250 interface” on page 376.

1. Option 3, Work with disk units.
2. Option 2, Work with disk configuration.

3. Option 8, Add units to ASPs and balance data.

You will then be presented with a panel similar to Figure B-25 on page 389. The values in the Resource Name column show DDxxx for single path volumes and DMPxxx for those which have more than one path. In this example, the 2107-A85 logical volume with serial number 75-1118707 is available through more than one path and reports in as DMP135.

4. Specify the ASP to which you wish to add the multipath volumes.

Specify ASPs to Add Units to						
Specify the ASP to add each unit to.						
Specify ASP	Serial Number	Type	Model	Capacity	Resource Name	
	21-662C5	4326	050	35165	DD124	
	21-54782	4326	050	35165	DD136	
1	75-1118707	2107	A85	35165	DMP135	
F3=Exit      F5=Refresh      F11=Display disk configuration capacity						
F12=Cancel						

Figure B-25 Adding multipath volumes to an ASP

**Note:** For multipath volumes, only one path is shown. In order to see the additional paths, see “Managing multipath volumes using iSeries Navigator” on page 392.

5. You will then be presented with a confirmation screen as shown in Figure B-26. Check the configuration details and if correct, press **Enter** to accept.

Confirm Add Units						
Add will take several minutes for each unit. The system will have the displayed protection after the unit(s) are added.						
Press Enter to confirm your choice for Add units.						
Press F9=Capacity Information to display the resulting capacity.						
Press F12=Cancel to return and change your choice.						
ASP	Unit	Serial Number	Type	Model	Resource Name	Protection
1						Unprotected
	1	02-89058	6717	074	DD004	Device Parity
	2	68-OCA4E32	6717	074	DD003	Device Parity
	3	68-OC9F8CA	6717	074	DD002	Device Parity
	4	68-OCA5D96	6717	074	DD001	Device Parity
	5	75-1118707	2107	A85	DMP135	Unprotected
F9=Resulting Capacity      F12=Cancel						

Figure B-26 Confirm Add Units

## Adding volumes to iSeries using iSeries Navigator

The iSeries Navigator GUI can be used to add volumes to the System, User or Independent ASPs. In this example, we are adding a multipath logical volume to a private (non-switchable) IASP. The same principles apply when adding multipath volumes to the System or User ASPs.

Follow the steps outlined in “Adding volumes to an Independent Auxiliary Storage Pool” on page 378.

When you get to the point where you will select the volumes to be added, you will see a panel similar to that shown in Figure B-27. Multipath volumes appear as DMPxxx. Highlight the disks you want to add to the disk pool and click **Add**.

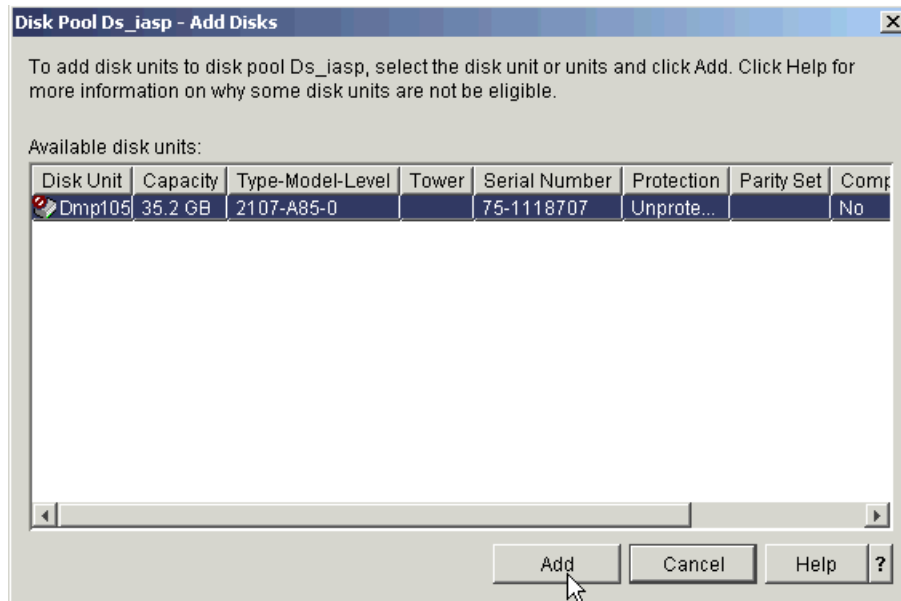


Figure B-27 Adding a multipath volume

**Note:** For multipath volumes, only one path is shown. In order to see the additional paths, see “Managing multipath volumes using iSeries Navigator” on page 392.

The remaining steps are identical to those in “Adding volumes to an Independent Auxiliary Storage Pool” on page 378.

When you have completed these steps, the new Disk Pool can be seen on iSeries Navigator **Disk Pools** in Figure B-28 on page 391.

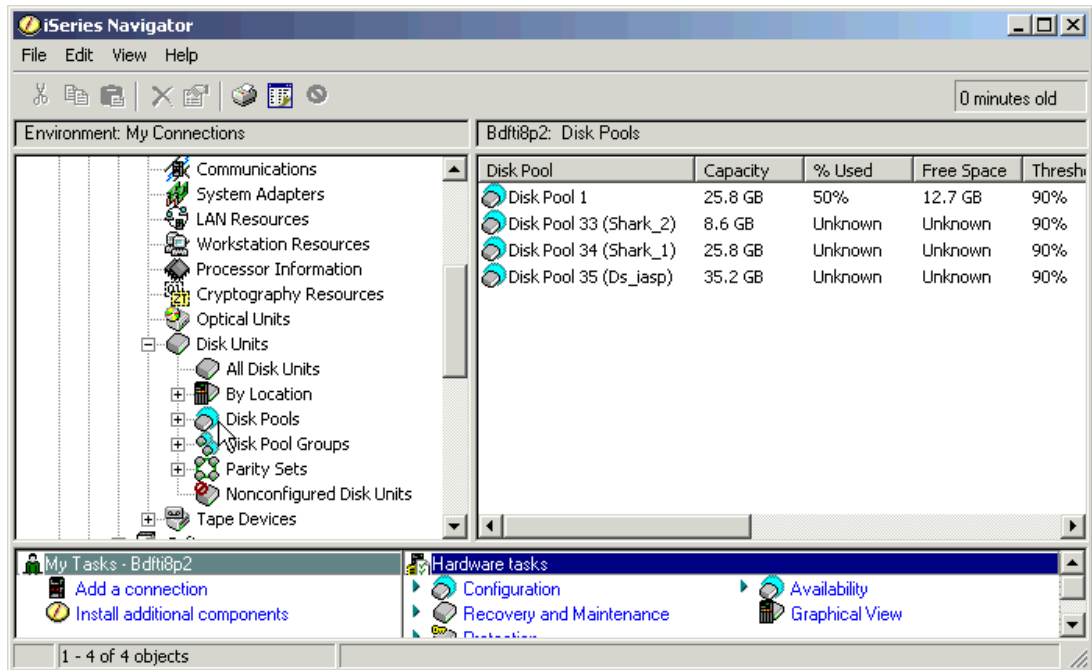


Figure B-28 New Disk Pool shown on iSeries Navigator

To see the logical volume, as shown in Figure B-29, expand **Configuration and Service**, **Hardware**, **Disk Pools** and click the disk pool you just created.

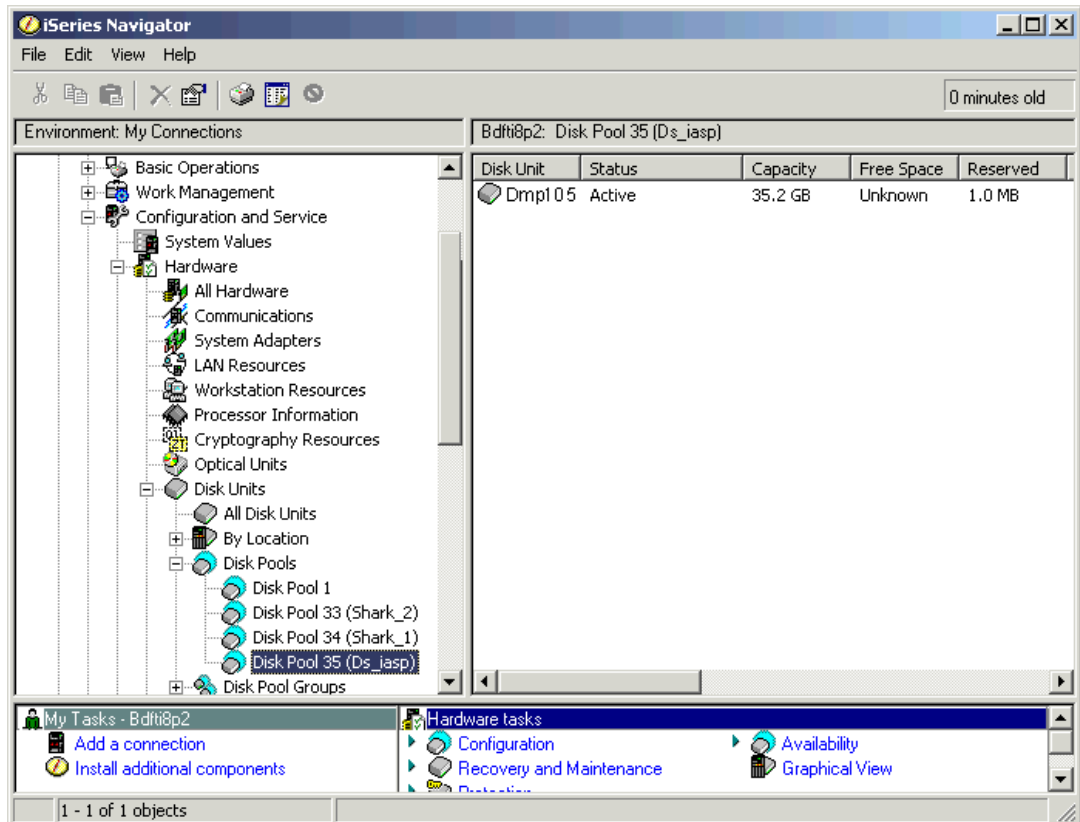


Figure B-29 New logical volume shown on iSeries Navigator

## Managing multipath volumes using iSeries Navigator

All units are initially created with a prefix of DD. As soon as the system detects that there is more than one path to a specific logical unit, it will automatically assign a unique resource name with a prefix of DMP for both the initial path and any additional paths.

When using the standard disk panels in iSeries Navigator, only a single (the initial) path is shown. The following steps show how to see the additional paths.

To see the number of paths available for a logical unit, open iSeries Navigator and expand **Configuration and Service, Hardware**, and **Disk Units** as shown in Figure B-30 and click **All Disk Units**. The number of paths for each unit is shown in column *Number of Connections* visible on the right of the panel. In this example, there are 8 connections for each of the multipath units.

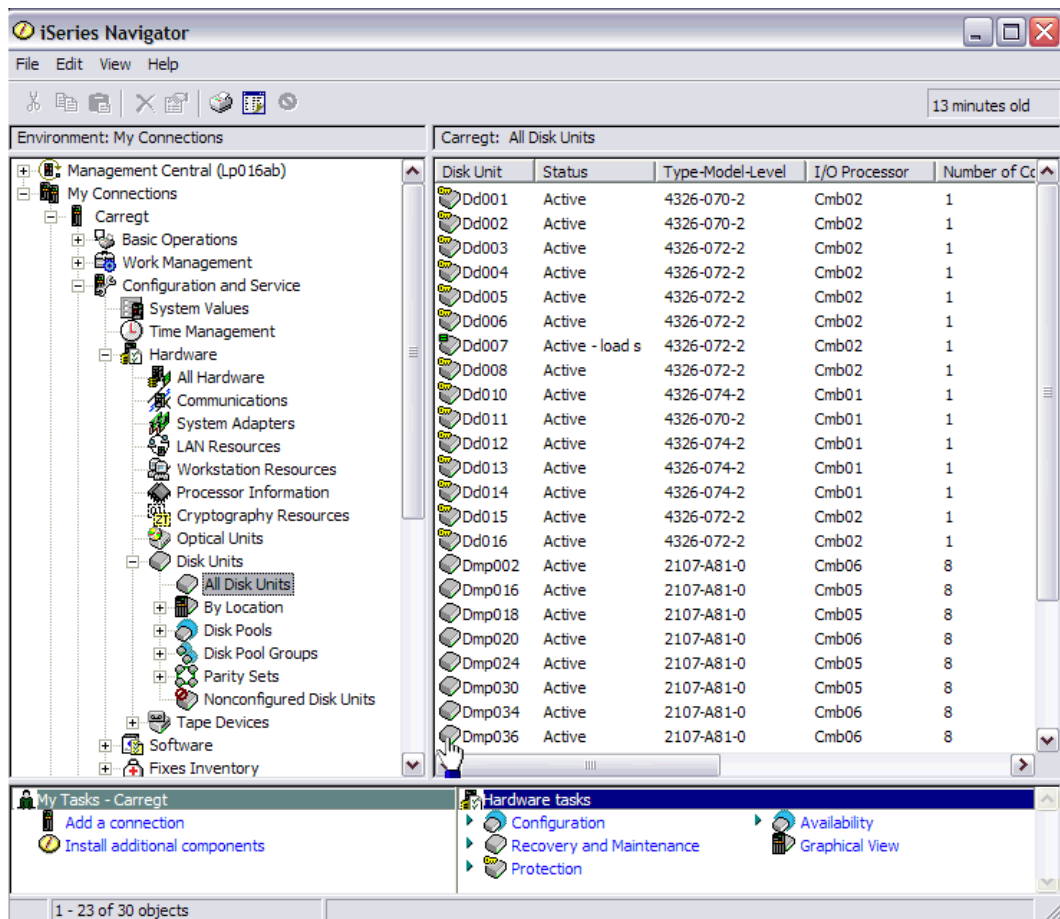


Figure B-30 Example of multipath logical units

To see the other connections to a logical unit, right click on the unit and select **Properties**, as shown in Figure B-31 on page 393.

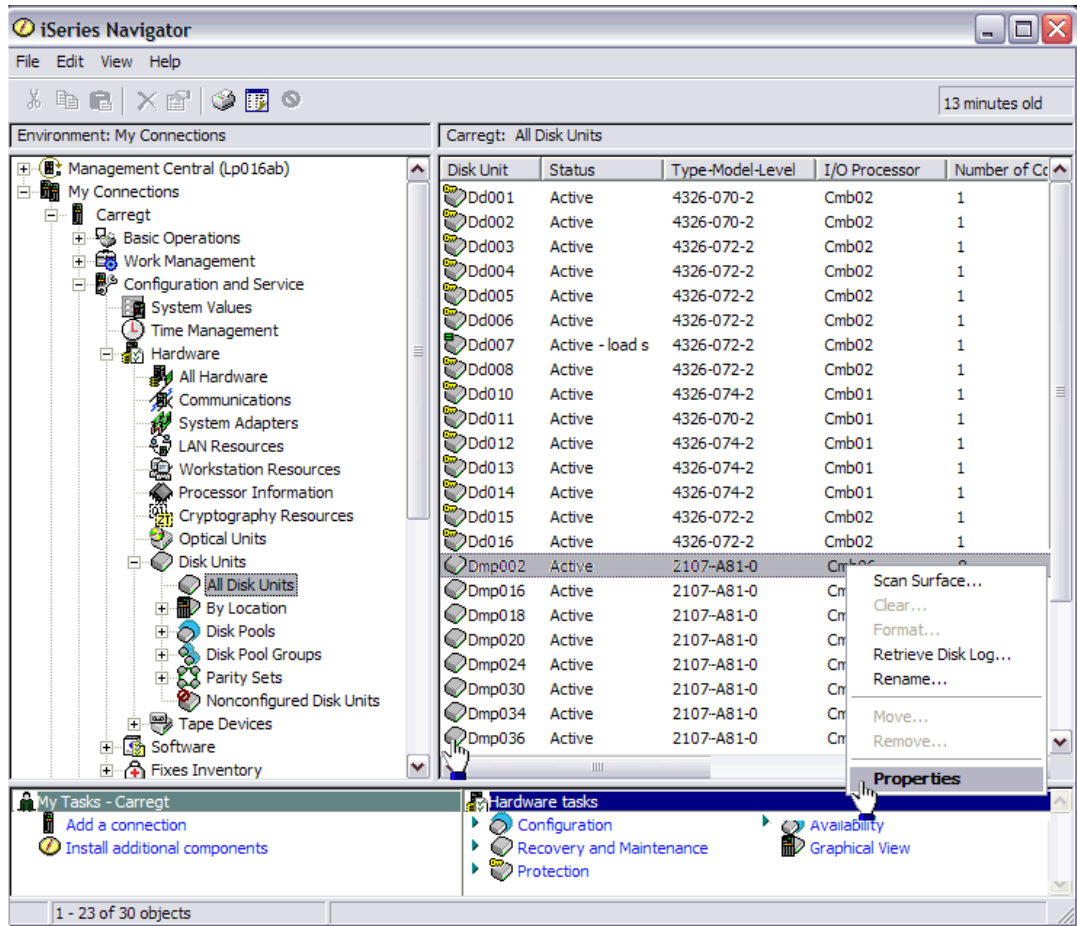


Figure B-31 Selecting properties for a multipath logical unit

You will then see the General Properties tab for the selected unit, as in Figure B-32 on page 394. The first path is shown as **Device 1** in the box labelled *Storage*.



Figure B-32 Multipath logical unit properties

To see the other paths to this unit, click the **Connections** tab, as shown in Figure B-33 on page 395, where you can see the other seven connections for this logical unit.



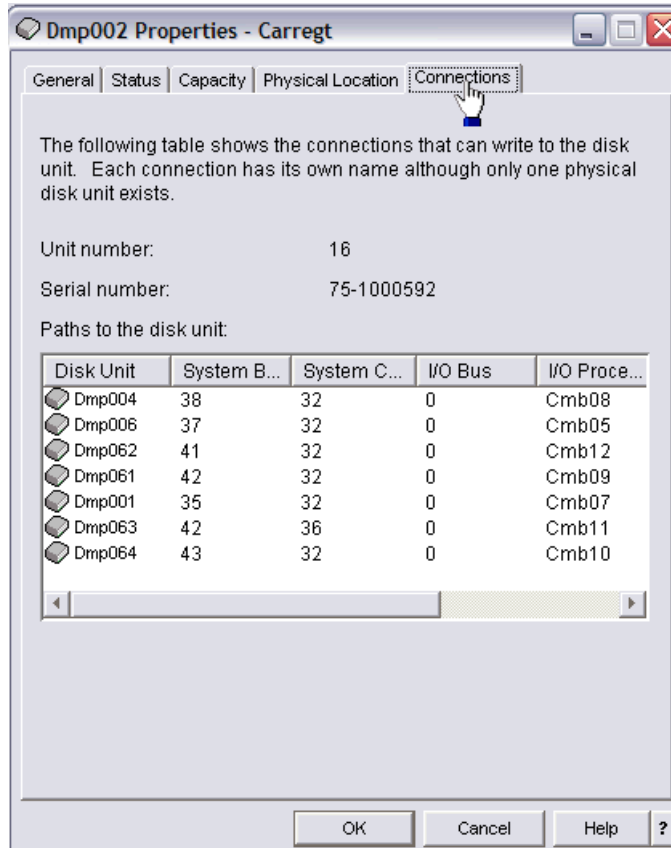


Figure B-33 Multipath connections

## Multipath rules for multiple iSeries systems or partitions

When you use multipath disk units, you must consider the implications of moving IOPs and multipath connections between nodes. You must not split multipath connections between nodes, either by moving IOPs between logical partitions or by switching expansion units between systems. If two different nodes both have connections to the same LUN in the DS8000, both nodes might potentially overwrite data from the other node.

The system enforces the following rules when you use multipath disk units in a multiple-system environment:

- ▶ If you move an IOP with a multipath connection to a different logical partition, you must also move all other IOPs with connections to the same disk unit to the same logical partition.
- ▶ When you make an expansion unit switchable, make sure that all multipath connections to a disk unit will switch with the expansion unit.
- ▶ When you configure a switchable independent disk pool, make sure that all of the required IOPs for multipath disk units will switch with the independent disk pool.

If a multipath configuration rule is violated, the system issues warnings or errors to alert you of the condition. It is important to pay attention when disk unit connections are reported missing. You want to prevent a situation where a node might overwrite data on a LUN that belongs to another node.

Disk unit connections might be missing for a variety of reasons, but especially if one of the preceding rules has been violated. If a connection for a multipath disk unit in any disk pool is found to be missing during an IPL or vary on, a message is sent to the QSYSOPR message queue.

If a connection is missing, and you confirm that the connection has been removed, you can update Hardware Service Manager (HSM) to remove that resource. Hardware service manager is a tool for displaying and working with system hardware from both a logical and a packaging viewpoint, an aid for debugging Input/Output (I/O) processors and devices, and for fixing failing and missing hardware. You can access Hardware Service Manager in System Service Tools (SST) and Dedicated Service Tools (DST) by selecting the option to start a service tool.

## Changing from single path to multipath

If you have a configuration where the logical units were only assigned to one I/O adapter, you can easily change to multipath. Simply assign the logical units in the DS8000 to another I/O adapter and the existing DDxxx drives will change to DMPxxx and new DMPxxx resources will be created for the new path.

## Sizing guidelines

In Figure B-34 on page 397, we show the process you can use to size external storage on iSeries. Ideally, you should have OS/400 Performance Tools reports, which can be used to model an existing workload. If these are not available, you can use workload characteristics from a similar workload to understand the I/O rate per second and the average I/O size. For example, the same application may be running at another site and its characteristics can be adjusted to match the expected workload pattern on your system.

Using this base information, and the rules-of-thumb that follow, you can estimate an approximate configuration which can then be used as input into Disk Magic (DM). This will give an indication of the service and wait time per I/O. If these do not meet your requirements, then you can adjust the hardware configuration in Disk Magic accordingly.

**Note:** Disk Magic is for IBM and IBM Business Partner use only. Customers should contact their IBM or IBM Business Partner representative for assistance with Capacity Planning, which may be a chargeable service.

Once you have this base modelling completed, you should also consider the other influences on the storage subsystem (such as any requirement for Copy Services and other workloads from other systems), and re-assess the hardware configuration and adjust accordingly until your requirements are met.

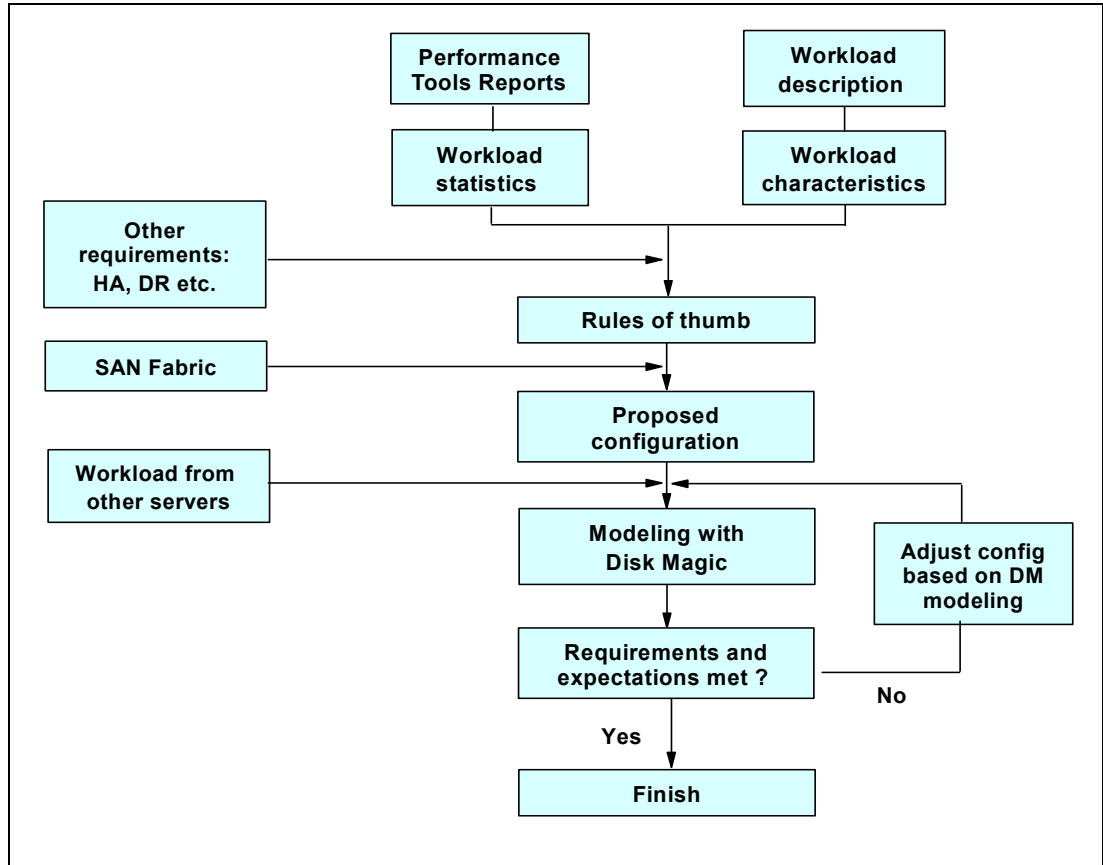


Figure B-34 Process for sizing external storage

## Planning for arrays and DDMs

In general, although it is possible to use 146 GB and 300 GB 10K RPM DDMs, we recommend that you use 73 GB 15K RPM DDMs for iSeries production workloads. The larger, slower drives may be suitable for less I/O intensive work, or for those workloads which do not require critical response times (for example, archived data or data which is high in volume but low in use such as scanned images).

For workloads with critical response times, you may not want to use all the capacity in an array. For 73 GB DDMs you may plan to use about 300 GB capacity per 8 drive array. The remaining capacity could possibly be used for infrequently accessed data. For example, you may have archive data, or some data such as images, which is not accessed regularly, or perhaps FlashCopy target volumes which could use this capacity, but not impact on the I/O /sec on those arrays.

For very high write environments, you may also consider using RAID-10, which offers a higher I/O rate per GB than RAID-5 as shown in Figure B-3 on page 399. However, the majority of iSeries workloads do not require this.

## Cache

In general, iSeries workloads do not benefit from large cache, so in the initial planning for Disk Magic, consider the minimum cache size available. Depending on the workload (as shown in OS/400 Performance Tools System, Component and Resource Interval reports) you may see

some benefit in larger cache sizes. However, in general, with large iSeries main memory sizes, OS/400 Expert Cache can reduce the benefit of external cache.

## Number of iSeries Fibre Channel adapters

The most important factor to take into consideration when calculating the number of Fibre Channel adapters in the iSeries is the throughput capacity of the adapter and IOP combination.

Since this guideline is based only on iSeries adapters and Access Density (AD) of iSeries workload, it doesn't change when using the DS8000. (The same guidelines are valid for ESS 800.)

**Note:** Access Density is the capacity of occupied disk space divided by the average I/O /sec. These values can be obtained from the OS/400 System, Component and Resource Interval performance reports.

Table B-2 shows the approximate capacity which can be supported with various IOA/IOP combinations.

Table B-2 Capacity per I/O Adapter

I/O Adapter	I/O Processor	Capacity per IOA	Rule of thumb
2787	2844	1022/AD	500GB
2766	2844	798/AD	400GB
2766	2843	644/AD	320GB

For most iSeries workloads, Access Density is usually below 2, so if you do not know it, the *Rule of thumb* column is a typical value to use.

## Size and number of LUNs

As discussed in “Logical volume sizes” on page 374, OS/400 can only use fixed logical volume sizes. As a general rule of thumb, we recommend that you should configure more logical volumes than actual DDMs. As a minimum, we recommend 2:1. For example, with 73 GB DDMs, you should use a maximum size of 35.1 GB LUNs. The reason for this is that OS/400 does not support command tag queuing. Using more, smaller LUNs can reduce I/O queues and wait times by allowing OS/400 to support more parallel I/Os.

From the values in Table B-2, you can calculate the number of iSeries Fibre Channel adapters for your required iSeries disk capacity. As each I/O adapter can support a maximum of 32 LUNs, divide the capacity per adapter by 32 to give the approximate average size of each LUN.

For example, assume you require 2 TB capacity and are using 2787 I/O adapters with 2844 I/O processors. If you know the access density, calculate the capacity per I/O adapter, or use the rule-of-thumb. Let's assume the rule-of-thumb of 500 GB per adapter. In this case, we would require four I/O adapters to support the workload. If we were able to have variable LUNs sizes, we could support 32 15.6 GB LUNs per I/O adapter. However, since OS/400 only supports fixed volume sizes, we could support 28 17. 5 GB volumes to give us approximately 492 GB per adapter.

## Recommended number of ranks

As a general guideline, you may consider 1500 disk operations/sec for an *average* RAID rank.

When considering the number of ranks, take into account the maximum disk operations per second per rank as shown in Table B-3. These are measured at 100% DDM Utilization with no cache benefit and with the average I/O being 4KB. Larger transfer sizes will reduce the number of operations per second.

Based on these values you can calculate how many host I/O per second each rank can handle at the recommended utilization of 40%. This is shown for workload read-write ratios of 70% read and 50% read in Table B-3.

Table B-3 Disk operations per second per RAID rank

RAID rank type	Disk ops/sec	Host I/O /sec (70% read)	Host I/O /sec (50% read)
RAID-5 15K RPM (7 + P)	1700	358	272
RAID-5 10K RPM (7 + P)	1100	232	176
RAID-5 15K RPM (6 + P + S)	1458	313	238
RAID-5 10K RPM (6 + P + S)	943	199	151
RAID-10 15K RPM (3 + 3 + 2S)	1275	392	340
RAID-10 10K RPM (3 + 3 + 2S)	825	254	220
RAID-10 15K RPM (4 + 4)	1700	523	453
RAID-10 15K RPM (4 + 4)	1100	338	293

As can be seen in Table B-3, RAID-10 can support higher host I/O rates than RAID-5. However, you must balance this against the reduced effective capacity of a RAID-10 rank when compared to RAID-5.

## Sharing ranks between iSeries and other servers

As a general guideline consider using separate extent pools for iSeries workload and other workloads. This will isolate the I/O for each server.

However, you may consider sharing ranks when the other servers' workloads have a sustained low disk I/O rate compared to the iSeries I/O rate. Generally, iSeries has a relatively high I/O rate while that of other servers may be lower – often below one I/O per GB per second.

As an example, a Windows file server with a large data capacity may normally have a low I/O rate with less peaks and could be shared with iSeries ranks. However, SQL, DB or other application servers may show higher rates with peaks, and we recommend using separate ranks for these servers.

Unlike its predecessor the ESS, capacity used for logical units on the DS8000 can be reused without reformatting the entire array. Now, the decision to mix platforms on an array is only one of performance, since the disruption previously experienced on ESS to reformat the array no longer exists.

## Connecting via SAN switches

When connecting DS8000 systems to iSeries via switches, you should plan that I/O traffic from multiple iSeries adapters can go through one port on a DS8000 and zone the switches accordingly. DS8000 host adapters can be shared between iSeries and other platforms.

Based on available measurements and experiences with the ESS 800 we recommend you should plan no more than four iSeries I/O adapters to one host port in the DS8000.

For a current list of switches supported under OS/400, refer to the iSeries Storage Web site at:

[http://www-1.ibm.com/servers/eserver/iseries/storage/storage\\_hw.html](http://www-1.ibm.com/servers/eserver/iseries/storage/storage_hw.html)

## Migration

For many iSeries customers, migrating to the DS8000 will be best achieved using traditional Save/Restore techniques. However, there are some alternatives you may wish to consider.

### OS/400 mirroring

Although it is possible to use OS/400 to mirror the current disks (either internal or external) to a DS8000 and then remove the older disks from the iSeries configuration, this is not recommended because both the source and target disks must initially be unprotected. If moving from internal drives, these would normally be protected by RAID-5 and this protection would need to be removed before being able to mirror the internal drives to DS8000 logical volumes.

Once an external logical volume has been created, it will always keep its model type and be either protected or unprotected. See Table B-1 on page 374 for DS8000 logical volume models and types. Therefore, once a logical volume has been defined as unprotected to allow it to be the mirror target, it cannot be converted back to a protected model, and therefore will be a candidate for all future OS/400 mirroring, whether you want this or not.

### Metro Mirror and Global Copy

Depending on the existing configuration, it may be possible to use Metro Mirror or Global Copy to migrate from an ESS to a DS8000 (or indeed, any combination of external storage units which support Metro Mirror and Global Copy). For further discussion on Metro Mirror and Global Copy, please see 16.2.2, "Subsystem-based data migration" on page 339. Consider the example shown in Figure B-35 on page 401.

Here, the iSeries has its internal Load Source Unit (LSU) and possibly some other internal drives. The ESS provides additional storage capacity. Using Metro Mirror or Global Copy, it is possible to create copies of the ESS logical volumes in the DS8000.

When ready to migrate from the ESS to the DS8000, you should do a complete shutdown of the iSeries, unassign the ESS LUNs and assign the DS8000 LUNs to the iSeries. After IPLing the iSeries, the new DS8000 LUNs will be recognized by OS/400, even though they are different models and have different serial numbers.

**Note:** It is important to ensure that both the Metro Mirror or Global Copy source and target copies are not assigned to the iSeries at the same time as this is an invalid configuration. Careful planning and implementation is required to ensure this does not happen, otherwise unpredictable results may occur.

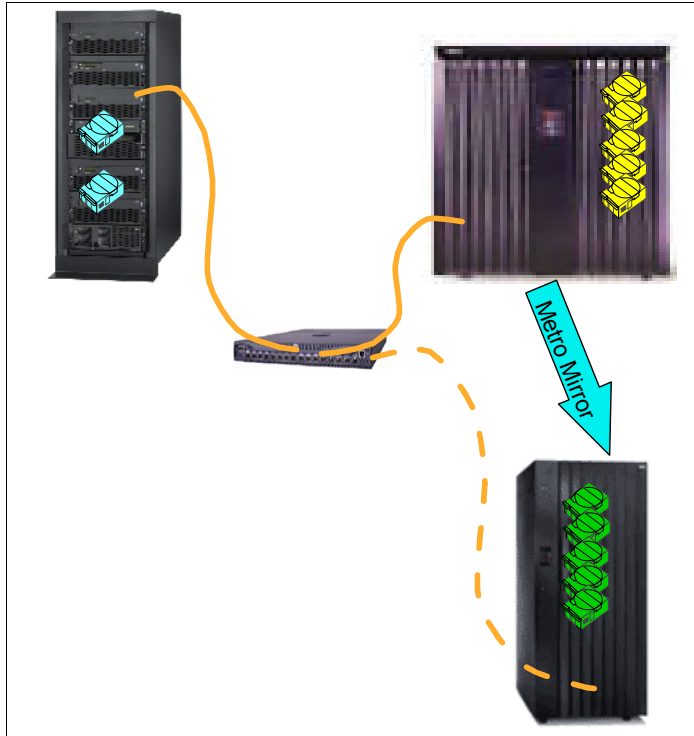


Figure B-35 Using Metro Mirror to migrate from ESS to DS8000

The same setup can also be used if the ESS LUNs are in an IASP, although the iSeries would not require a complete shutdown since varying off the IASP in the ESS, unassigning the ESS LUNs, assigning the DS8000 LUNs and varying on the IASP would have the same effect.

Clearly, you must also take into account the licensing implications for Metro Mirror and Global Copy.

**Note:** This is a special case of using Metro Mirror or Global Copy and will only work if the same iSeries is used, along with the LSU to attach to both the original ESS and the new DS8000. It is not possible to use this technique to a different iSeries.

## OS/400 data migration

It is also possible to use native OS/400 functions to migrate data from existing disks to the DS8000, whether the existing disks are internal or external. When you assign the new DS8000 logical volumes to the iSeries, initially they are non-configured (see “Adding volumes to iSeries configuration” on page 376 for more details). If you add the new units and choose to spread data, OS/400 will automatically migrate data from the existing disks onto the new logical units.

You can then use the OS/400 command STRASPBAL TYPE(\*ENDALC) to mark the units to be removed from the configuration as shown in Figure B-36 on page 402. This can reduce the

down time associated with removing a disk unit. This will keep new allocations away from the marked units.

```

                                Start ASP Balance (STRASPBAL)

Type choices, press Enter.

Balance type . . . . . > *ENDALC          *CAPACITY, *USAGE, *HSM...
Storage unit . . . . .                   1-4094
      + for more values

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

```

Figure B-36 Ending allocation for existing disk units

When you subsequently run the OS/400 command STRASPBAL TYPE(\*MOVDTA) all data will be moved from the marked units to other units in the same ASP, as shown in Figure B-37. Clearly you must have sufficient new capacity to allow the data to be migrated.

```

                                Start ASP Balance (STRASPBAL)

Type choices, press Enter.

Balance type . . . . . > *MOVDTA          *CAPACITY, *USAGE, *HSM...
Time limit . . . . .                   1-9999 minutes, *NOMAX

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

```

Figure B-37 Moving data from units marked \*ENDALC

You can specify a time limit that the function is to run for each ASP being balanced or the balance can be set to run to completion. If the balance function needs to be ended prior to this, use the End ASP Balance (ENDASPBAL) command. A message will be sent to the system history (QHST) log when the balancing function is started for each ASP. A message will also be sent to the QHST log when the balancing function completes or is ended.

If the balance function is run for a few hours and then stopped, it will continue from where it left off when the balance function restarts. This allows the balancing to be run during off hours over several days.

In order to finally remove the old units from the configuration, you will need to use Dedicated Service Tools (DST) and re-IPL the system (or partition).

Using this method allows you to remove the existing storage units over a period of time. However, it does require that both the old and new units are attached to the system at the same time so it may require additional IOPs and IOAs if migrating from an ESS to a DS8000.

It may be possible in your environment to re-allocate logical volumes to other IOAs, but careful planning and implementation will be required.



## Copy Services for iSeries

Due to OS/400 having a single level storage, it is not possible to copy some disk units without copying them all, unless specific steps are taken.

**Attention:** You should not assume that Copy Services with iSeries works the same as with other open systems.

### FlashCopy

When FlashCopy was first available for use with OS/400, it was necessary to copy the entire storage space, including the Load Source Unit (LSU). However, since the LSU must reside on an internal disk, this first had to be mirrored to a LUN in the external storage subsystem. Because it is not possible to IPL from external storage, it was then necessary to *D-Mode IPL* the target system/partition from CD, then Recover Remote LSU. This is sometimes known as *basic FlashCopy*.

In order to ensure the entire single level storage is copied, memory needs to be flushed - preferably with a PWRDWN SYS or perhaps more acceptable, taking the system into a Restricted State using ENDSBS \*ALL.

For most customers, this is not a practical solution.

To avoid this and to make FlashCopy more appropriate to iSeries customers, IBM has developed a service offering to allow Independent Auxiliary Storage Pools (IASPs) to be used with FlashCopy independently from the LSU and other disks which make up \*SYSBAS (ASP1-32). This has three major benefits:

1. Less data is copied.
2. Recover Remote LSU recovery is not necessary.
3. Communication configuration details are not affected.

The target system can be a live system (or partition) used for other functions such as test, development and Lotus® Notes®. When backups are to be done, the FlashCopy target can be attached to the partition without affecting the rest of the users. Or perhaps more likely, the target will be a partition on the same system as production but may have no CPU or memory allocated to it until the backups are taken, when these resources are then reallocated from the production environment (or other) and moved to the backup partition.

Again, like the basic FlashCopy, it is necessary to flush memory for those objects to be saved so that all objects reside in the ESS (where FlashCopy runs). However, unlike basic FlashCopy, this is achieved by varying off the IASP rather than powering off the system or taking it to restricted state. The rest of the system is unaffected.

### Remote Mirror and Copy

Although the same considerations apply to the Load Source Unit as for FlashCopy, unlike FlashCopy, which would likely be done on a daily/nightly basis, Remote Mirror is generally used for DR. The additional steps required to make the target volumes usable (D-IPL, recover remote LSU, abnormal IPL) are more likely to be acceptable due to the infrequent nature of invoking the DR copy. Recovery time may be affected but the recovery point will be to the point of failure. This is sometimes known as *basic Remote Mirroring*.

Using the advantages previously discussed when using IASPs with FlashCopy, we are able to use this technology with Remote Mirror as well. Instead of the entire system (single level

storage) being copied, only the application resides in an IASP and in the event of a disaster, the target copy is attached to the DR server.

Additional considerations must be taken into account such as maintaining user profiles on both systems, but this is no different from using other availability functions such as switched disk between two local iSeries Servers on a High Speed Link (HSL) and Cross Site Mirroring (XSM) to a remote iSeries. However, with Remote Mirror, the distance can be much greater using the synchronous Metro Mirror than the 250 meter limit of HSL, and with the asynchronous Global Mirror, there is little performance impact on the production server.

Whereas with FlashCopy where you would likely only have data in an IASP, for DR with Remote Mirror, you also need the application on the DR system. This can either reside in \*SYSBAS or in an IASP. If it is in an IASP, then the entire application would be copied. If it is in \*SYSBAS, you will need to ensure good change management facilities to ensure both systems have the same level of the application. Also, you must ensure that system objects in \*SYSBAS, such as User Profiles, are synchronized.

Again, the DR server could be a dedicated system, or perhaps more likely, a shared system with development, testing, or other function in other partitions or IASPs.

## iSeries toolkit for Copy Services

Although it is possible to use FlashCopy and Remote Mirror functions with iSeries, for practical reasons, many customers will choose not to use basic FlashCopy and Remote Mirror functions due to the LSU restrictions discussed previously. Instead, IBM recommends that Copy Services should only be used with the iSeries toolkit for Copy Services, which uses OS/400 IASPs and provides control of the clustering environment necessary to use IASPs with Copy Services. More information about the Copy Services toolkit can be found at:

<http://www-1.ibm.com/servers/eserver/series/service/itc/pdf/Copy-Services-ESS.pdf>

Contact your IBM representative if you wish to implement the iSeries toolkit for Copy Services, or contact the iSeries Technology Center directly at:

<mailto:rhc1st@us.ibm.com>

## AIX on IBM iSeries

With the announcement of the IBM iSeries i5, it is now possible to run AIX in a partition on the i5. This can be either AIX 5L V5.2 or V5.3. All supported functions of these operating system levels are supported on i5, including HACMP for high availability and external boot from Fibre Channel devices.

The DS6000 requires the following i5 I/O adapters to attach directly to an i5 AIX partition:

- ▶ 0611 Direct Attach 2 Gigabit Fibre Channel PCI
- ▶ 0625 Direct Attach 2 Gigabit Fibre Channel PCI-X

It is also possible for the AIX partition to have its storage virtualized, whereby a partition running OS/400 hosts the AIX partition's storage requirements. In this case, if using DS8000, they would be attached to the OS/400 partition using either of the following I/O adapters:

- ▶ 2766 2 Gigabit Fibre Channel Disk Controller PCI
- ▶ 2787 2 Gigabit Fibre Channel Disk Controller PCI-X

For more information on running AIX in an i5 partition, refer to the i5 Information Center at:

- ▶ [http://publib.boulder.ibm.com/infocenter/series/v1r2s/en\\_US/index.htm?info/iphatl/iphatlparkickoff.htm](http://publib.boulder.ibm.com/infocenter/series/v1r2s/en_US/index.htm?info/iphatl/iphatlparkickoff.htm)

**Note:** AIX will not run in a partition on earlier 8xx and prior iSeries systems.

## Linux on IBM iSeries

Since OS/400 V5R1, it has been possible to run Linux in an iSeries partition. On iSeries models 270 and 8xx, the primary partition must run OS/400 V5R1 or higher and Linux is run in a secondary partition. For later i5 systems (models i520, i550, i570 and i595), Linux can run in any partition.

On both hardware platforms, the supported versions of Linux are:

- ▶ SUSE Linux Enterprise Server 9 for POWER  
(New 2.6 Kernel based distribution also supports earlier iSeries servers)
- ▶ RedHat Enterprise Linux AS for POWER Version 3  
(Existing 2.4 Kernel based update 3 distribution also supports earlier iSeries servers)

The DS8000 requires the following iSeries I/O adapters to attach directly to an iSeries or i5 Linux partition.

- ▶ 0612 Linux Direct Attach PCI
- ▶ 0626 Linux Direct Attach PCI-X

It is also possible for the Linux partition to have its storage virtualized, whereby a partition running OS/400 hosts the Linux partition's storage requirements. In this case, if using DS8000, they would be attached to the OS/400 partition using either of the following I/O adapters:

- ▶ 2766 2 Gigabit Fibre Channel Disk Controller PCI
- ▶ 2787 2 Gigabit Fibre Channel Disk Controller PCI-X

More information on running Linux in an iSeries partition can be found in the iSeries Information Center at:

- ▶ <http://publib.boulder.ibm.com/series/v5r2/ic2924/index.htm>

More information on running Linux in an i5 partition can be found in the i5 Information Center at:

- ▶ [http://publib.boulder.ibm.com/infocenter/series/v1r2s/en\\_US/info/iphbi/iphbi.pdf](http://publib.boulder.ibm.com/infocenter/series/v1r2s/en_US/info/iphbi/iphbi.pdf)





## Service and support offerings

This appendix provides information about the service offerings which are currently available for the new DS6000 and DS8000 series. It includes a brief description of each offering and where you can find more information on the Web:

- ▶ IBM Implementation Services for TotalStorage disk systems
- ▶ IBM Implementation Services for TotalStorage Copy Functions
- ▶ IBM Implementation Services for TotalStorage Command-Line Interface
- ▶ IBM Migration Services for eServer zSeries data
- ▶ IBM Migration Services for open systems attached to TotalStorage disk systems
- ▶ IBM Geographically Dispersed Parallel Sysplex™ (GDPS®)
- ▶ Enterprise Remote Copy Management Facility (eRCMF)
- ▶ Geographically Dispersed Sites for Microsoft Cluster Services
- ▶ IBM eServer iSeries Copy Services
- ▶ IBM Operational Support Services - Support Line

## IBM Web sites for service offerings

IBM Global Services (IGS) and the IBM Systems Group can offer comprehensive assistance, including planning and design as well as implementation and migration support services. For more information on all of the following service offerings, contact your IBM representative or visit the following Web sites.

The IBM Global Service Web site can be found at:

<http://www.ibm.com/services/us/index.wss/home>

The IBM System Group Web site can be found at:

<http://www.ibm.com/servers/storage/services/>

## IBM service offerings

This section describes the service offerings available from IBM Global Services and IBM Systems Group.

### **IBM Implementation Services for TotalStorage disk systems**

This service includes planning for a new IBM TotalStorage disk system followed by implementation, configuration, and basic skills transfer instruction. For more information visit the following Web site:

<http://www.ibm.com/services/us/index.wss/so/its/a1005008>

### **IBM Implementation Services for TotalStorage Copy Functions**

This service is designed to assist in the planning, implementation, and testing of the IBM TotalStorage advanced copy functions, point-in-time copy and remote mirroring solutions. For more information visit the following Web site:

<http://www.ibm.com/services/us/index.wss/so/its/a1005009>

### **IBM Implementation Services for TotalStorage Command-Line Interface**

IBM provides a service through Global Services to help you with using the Command-Line Interface (CLI) in your system environment. It is designed to provide you with the ability to create and apply configurations online. For more information visit the following Web site:

<http://www.ibm.com/services/us/index.wss/so/its/a1005334>

### **IBM Migration Services for eServer zSeries data**

IBM provides a technical specialist at your site to help plan and assist in the implementation of non-disruptive DASD migration to a new or existing IBM TotalStorage disk system. The migration is accomplished using the following software and hardware that allows DASD volumes to be copied to the new storage devices without interruption to service.

- ▶ Innovation Fast Dump Restore Plug and Swap (FDRPAS)
- ▶ Peer-to-Peer Remote Copy Extended Distance (PPRC-XD)
- ▶ Softek Replicator (TDMF)
- ▶ IBM Piper hardware assisted migration

The IBM Piper hardware assisted migration in the zSeries environment is described in this redbook in “Data migration with Piper for z/OS” on page 299. Additional information about this offering is on the following Web site:

[http://www.ibm.com/servers/storage/services/featured/hardware\\_assist.html](http://www.ibm.com/servers/storage/services/featured/hardware_assist.html)

For more information about IBM Migration Services for eServer zSeries data visit the following Web site:

<http://www.ibm.com/services/us/index.wss/so/its/a1005010>

### **IBM Migration Services for open systems attached to disk systems**

IBM Migration Services for open systems attached to TotalStorage disk systems include planning for and implementation of data migration from an existing UNIX or Windows server to new or existing larger capacity IBM storage with minimal disruption. This service uses the following hardware and software tools:

- ▶ Peer-to-Peer Remote Copy Extended Distance (PPRC-XD)
- ▶ Softek Replicator (TDMF) for Open
- ▶ Native operating system mirroring
- ▶ IBM Piper hardware assisted migration

The IBM hardware-assisted data migration services (IBM Piper), which we already mentioned in regard to the zSeries environment, can also be used in an Open System environment. You can find information about the use of this service in an Open System environment in 16.2.3, “IBM Piper migration” on page 341. For the latest information about this service, please visit the following Web site:

▶ [http://www.ibm.com/servers/storage/services/featured/hardware\\_assist.html](http://www.ibm.com/servers/storage/services/featured/hardware_assist.html)

For more information about IBM Migration Services for open systems attached to TotalStorage disk systems visit the following Web site:

<http://www.ibm.com/services/us/index.wss/so/its/a1005012>

### **IBM Geographically Dispersed Parallel Sysplex (GDPS)**

A Geographically Dispersed Parallel Sysplex (GDPS) is the ultimate disaster recovery and continuous availability solution for a zSeries multi-site enterprise. This solution automatically mirrors critical data and efficiently balances workload between the sites. The GDPS solution also uses automation and Parallel Sysplex technology to help manage multi-site databases, processors, network resources and storage subsystem mirroring. For the latest information on this service, please visit the following Web site:

<http://www.ibm.com/services/us/index.wss/rs/its/a1005497>

### **Enterprise Remote Copy Management Facility (eRCMF)**

IBM Implementation Services for enterprise Remote Copy Management Facility (eRCMF) is intended as a multisite Disaster Recovery solution for Open Systems and provides automation for repairing inconsistent PPRC pairs. This is the software that communicates with the ESS Copy Services server.

It is a scalable, flexible Open Systems solution that protects the business (data) and can be used for both:

- ▶ Planned outages (hardware and software upgrades)

- ▶ Unplanned outages (disaster recovery, testing a disaster)

It simplifies the disaster recovery implementation and concept. Once eRCMF is configured in the customer environment, it will monitor the PPRC states of all specified LUNs/volumes. Visit the following Web site for the latest information:

<http://www.ibm.com/services/us/index.wss/so/its/a1000110>

### **Geographically Dispersed Sites for Microsoft Cluster Services**

IBM TotalStorage Support for Geographically Dispersed Sites for Microsoft Cluster Service (MSCS) is designed to allow Microsoft Cluster installations to span geographically dispersed sites. It helps to protect clients from site disasters or storage subsystem failures. This service is designed to enable a tier 7 disaster recovery solution. It also provides high availability for applications and data running in Windows clustered server environments, by extending the distance that cluster nodes and storage can be separated, mirroring data between two IBM TotalStorage disk subsystems, and providing improved failure detection. You can find the latest information on this service on the following Web site:

[http://www.ibm.com/servers/storage/services/featured/microsoft\\_application\\_environment.html#GDSSolution](http://www.ibm.com/servers/storage/services/featured/microsoft_application_environment.html#GDSSolution)

### **IBM eServer iSeries Copy Services**

For the iSeries environment IBM offers a special toolkit, which allows you to use the advanced Copy Services functions with the iSeries. For more information on this, see , “iSeries toolkit for Copy Services” on page 404.

## **IBM Operational Support Services - Support Line**

IBM offers telephone or electronic access to highly-trained technical support specialists, who can serve as your one source for remote software support services.

Highlights of the offering are the following:

- ▶ High-quality technical support for IBM and select multivendor software including the Linux operating system and Linux clusters
- ▶ A supplement to your internal staff with IBM's skilled services specialists
- ▶ Fast, accurate problem resolution to help keep your IT staff productive
- ▶ Options for enhanced coverage and a single interface for remote support
- ▶ Support for your international environment

For more information on the IBM Support Line visit the following Web site:

<http://www.ibm.com/services/us/index.wss/so/its/a1000030>

In Figure 16-9 on page 411 you can see an example for the Support Product List (SPL) of the IBM Support Line with the DS6800 support and the DS8100 support highlighted. You get the complete SPL on the following Web site:

<http://www.ibm.com/services/sl/products/java3.html>



IBM Support Line Supported Products List (SPL) as of December 3, 2004 (contracts on or after 8/ - Microsoft L...

Address: <http://www.ibm.com/services/sl/products/java3.html>

Country: United States

Support Group: DSKTP - Disk and tape

Name, ID, End Date: [ ] [ ] Year [ ] Month [ ]

30 products supported.

Support Group	Product Name	ID	VRM	End Date
DSKTP	1750 DS6800 Storage Server	6942-63F	1.1.0	2008-12-31
DSKTP	2105 Enterprise Storage Server	6942-63F	1.1.0	
DSKTP	2106 Modular Storage Server	6942-63F	1.1.0	
DSKTP	2107 DS8100 Storage Server	6942-63F	1.1.0	2008-12-31
DSKTP	3494 Tape Library	6942-64F	1.1.0	
DSKTP	3542 FAST200 HA Storage Server	6942-67F	1.1.0	
DSKTP	3542 FAST200 Storage Server	6942-67F	1.1.0	
DSKTP	3552 FAST500 Storage Server	6942-67F	1.1.0	
DSKTP	3560 FastT Express 500	6942-67F	1.1.0	
DSKTP	3570 Tape Subsystem	6942-64F	1.1.0	
DSKTP	3575 Tape Library Dataserver	6942-64F	1.1.0	
DSKTP	3582 Ultrium Tape Library	6942-64F	1.1.0	2007-04-30

Buttons: Search, Print, Save, Select All, Copy Rows

Footer: About IBM | Privacy | Legal | Contact

Figure 16-9 Example of the Supported Product List (SPL) from the IBM Support Line



# Related publications

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

## IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 415. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM TotalStorage Multiple Device Manager Usage Guide*, SG24-7097
- ▶ *IBM TotalStorage Solutions for Disaster Recovery*, SG24-6547
- ▶ *The IBM TotalStorage Solutions Handbook*, SG24-5250
- ▶ *IBM TotalStorage DS6000 Series: Concepts and Architecture*, SG24-6471
- ▶ *IBM TotalStorage Enterprise Storage Server: Implementing ESS Copy Services in Open Environments*, SG24-5757
- ▶ *IBM TotalStorage Enterprise Storage Server: Implementing ESS Copy Services with IBM eServer zSeries*, SG24-5680
- ▶ *DFSMSHsm ABARS and Mainstar Solutions*, SG24-5089
- ▶ *Practical Guide for SAN with pSeries*, SG24-6050
- ▶ *Fault Tolerant Storage - Multipathing and Clustering Solutions for Open Systems for the IBM ESS*, SG24-6295
- ▶ *Implementing Linux with IBM Disk Storage*, SG24-6261
- ▶ *Linux with zSeries and ESS: Essentials*, SG24-7025

## Other publications

These publications are also relevant as further information sources. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM TotalStorage DS8000 Command-Line Interface User's Guide*, SC26-7625
- ▶ *IBM TotalStorage DS8000: Host Systems Attachment Guide*, SC26-7628
- ▶ *IBM TotalStorage DS8000: Introduction and Planning Guide*, GC35-0495
- ▶ *IBM TotalStorage Multipath Subsystem Device Driver User's Guide*, SC30-4096
- ▶ *IBM TotalStorage DS8000: User's Guide*, SC26-7623
- ▶ *IBM TotalStorage DS Open Application Programming Interface Reference*, GC35-0493
- ▶ *IBM TotalStorage DS8000 Messages Reference*, GC26-7659
- ▶ *z/OS DFSMS Advanced Copy Services*, SC35-0248
- ▶ *Device Support Facilities: User's Guide and Reference*, GC35-0033
- ▶ *z/OS Advanced Copy Services*, SC35-0248

## Online resources

These Web sites and URLs are also relevant as further information sources:

- ▶ IBM Disk Storage Feature Activation (DSFA) Web site at  
<http://www.ibm.com/storage/dsfa>
- ▶ The PSP information can be found at:  
<http://www-1.ibm.com/servers/resourceLink/svc03100.nsf?OpenDatabase>
- ▶ Documentation for the DS8000:  
<http://www.ibm.com/servers/storage/support/disk/2107.html>
- ▶ Supported servers for the DS8000:  
<http://www.storage.ibm.com/hardsoft/products/DS8000/supserver.htm>
- ▶ The interoperability matrix:  
<http://www.ibm.com/servers/storage/disk/DS8000/interop.html>
- ▶ Fibre Channel host bus adapter firmware and driver level matrix:  
<http://knowledge.storage.ibm.com/HBA/HBASearchTool>
- ▶ ATTO:  
<http://www.attotech.com/>
- ▶ Emulex:  
<http://www.emulex.com/ts/dds.html>
- ▶ JNI:  
<http://www.jni.com/OEM/oem.cfm?ID=4>
- ▶ QLogic:  
[http://www.qlogic.com/support/oem\\_detail\\_all.asp?oemid=22](http://www.qlogic.com/support/oem_detail_all.asp?oemid=22)
- ▶ IBM:  
<http://www.ibm.com/storage/ibmsan/products/sanfabric.html>
- ▶ CNT (INRANGE):  
<http://www.cnt.com/ibm/>
- ▶ McDATA:  
<http://www.mcdata.com/ibm/>
- ▶ Cisco:  
<http://www.cisco.com/go/ibm/storage>
- ▶ CIENA:  
<http://www.ciena.com/products/transport/shorthaul/cn2000/index.asp>
- ▶ CNT:  
<http://www.cnt.com/ibm/>
- ▶ Nortel:  
<http://www.nortelnetworks.com/>
- ▶ ADVA:  
<http://www.advaoptical.com/>

## 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](http://ibm.com/redbooks)

## Help from IBM

IBM Support and downloads

[ibm.com/support](http://ibm.com/support)

IBM Global Services

[ibm.com/services](http://ibm.com/services)



# Index

## Symbols

151, 298, 335, 368

## A

AAL 36–37, 256  
    benefits 37  
address groups 96, 196  
Advanced Copy Services 162  
AIX  
    boot support 353  
    I/O access methods 352  
    LVM configuration 352  
    managing multiple paths 349  
    monitoring I/O performance  
        filemon 354  
        iostat 354  
    on iSeries 404  
    other publications 348  
    SDD 349  
    WWPN 348  
architecture 22  
array sites 86, 192, 200  
arrays 35, 87, 192, 200  
arrays across loops see AAL  
Asynchronous Cascading PPRC see Metro/Global Copy  
Asynchronous PPRC see Global Mirror

## B

base frame 20  
battery backup assemblies 40  
battery backup unit see BBU  
BBU 79  
Business Continuity 4  
business continuity  
    Resiliency Family 8

## C

cache management 24  
call home 164  
capacity  
    well-balanced configuration 181  
capacity planning 174  
CIM server 162  
CKD volumes 91  
    allocation and deletion 93  
components 19  
configuration planning 158  
Consistency Group 133  
    commands 9  
    data consistency 132  
    what is it? 132  
Consistency Group FlashCopy 121

Control Unit Initiated Reconfiguration see CUIR  
cooling 79  
    disk enclosure 40  
    rack cooling fans 79  
Copy Services 115  
    interfaces 136  
    interoperability with ESS 139  
CUIR 74

## D

DA 30  
    Fibre Channel 260  
data consistency 132–133  
    Consistency Group 132  
data migration  
    basic commands 184  
    migration appliances 185  
    open systems  
        backup and restore 338  
        basic copy commands 336  
        copy raw devices 336  
        Global Copy 339  
        host operating system based 335  
        IBM Migration Services 342  
        Metro Mirror 339  
        other migration applications 342  
        Piper 341  
        rsync 337  
        using volume management software 337  
    operating system mirroring 184  
    planning 183  
    remote copy 184  
    software packages 184  
VSE/ESA 315  
z/OS 185, 293  
    bridge from ESCON to FICON 302  
    combine physical and logical data migration 308,  
    314  
    consolidate logical volumes 295  
    consolidate storage 294  
    DFSMSdss 307  
    DFSMSHsm 308  
    Global Copy 303  
    logical migration 307  
    Metro Mirror 303  
    physical migration 298  
    physical migration with DFSMSdss 298  
    Piper 299  
    system-managed storage 308  
    z/OS Global Mirror 301  
z/VM 315  
data placement 99  
Data Set FlashCopy 9, 119  
DDMs 37, 192

- hot plugable 78
- layer 200
- device adapter see DA
- DFSMSdss 298, 307
- DFSMSHsm 308
- disk drive set 7
- disk enclosure 31
  - power and cooling 40
- Disk Magic 186–187
- disk storage feature activation (DSFA) 173
- disk subsystem 30
- disk virtualization 85
- DS CLI 12, 138, 231, 325
  - changes from ESS CLI 138
  - co-existence with ESS CLI 237
  - command flow 234–235
  - command modes 239
    - interactive mode 240
    - script mode 240
    - single command mode 239
  - DS6000 command flow 236
  - DS8000 split network 236
  - functionality 232
  - installation methods 233
  - introduction 232
  - migration 244
    - example 245
    - tasks 245
  - mixed device environments 244
  - return codes 242
  - supported environments 233
  - syntax conventions 241
  - usage examples 243
  - user assistance 241
  - user security 239
- DS Open API 13, 138
- DS Open application programming interface see DS Open API
- DS Storage Manager 12, 137
  - introduction 203
  - logical configuration 189
  - navigating the GUI 208
  - Welcome panel 203
- DS6000
  - compared to DS8000 12
- DS8000
  - AAL
  - architecture 22
  - arrays 35
  - base frame 20
  - battery backup assemblies 40
  - Business Continuity 4
  - cache management 24
  - capacity planning 174, 344
  - common set of functions 11
  - compared to DS6000 12
  - compared to ESS 11
  - components 19
  - configuration planning 158
  - connecting to 202

- Copy Services 115
- DA
- data migration
  - basic commands 184
  - migration appliances 185
  - operating system mirroring 184
  - planning 183
  - remote copy 184
  - software packages 184
  - z/OS 185
- data placement 265
- DDMs 37
- delivery requirements 144
- differences in operability to DS6000 323
- disk drive set 7
- disk enclosure 31
- Disk Magic 186–187
- disk storage feature activation 173
- disk subsystem 30
- DS CLI 231
- DS8100 13
- DS8100 Model 921 105
- DS8300 13
- DS8300 Model 922 106
- DS8300 Model 9A2 106
- environmental requirements 149
- EPO 21, 80
- ESCON attached 151
- expansion enclosure 35
- expansion frame 21
- Fibre Channel disk drives 6
- FICON 14, 151
- FlashCopy 9, 116
- frames 20
- Global Mirror
- HA
- hardware overview 6
- host adapter 6
- host attachment 150
- I/O enclosure 29
- I/O priority queuing 15
- Information Center 207
- Information Lifecycle Management 4
- Infrastructure Simplification 4
- installation planning 144
- interoperability 10
- Interoperability Matrix 321
- iSeries 373
- licensed functions 167
- logical configuration 174, 189
- Metro Mirror
- Metro/Global Copy 10
- model comparison 108
- model conversion 53
- model naming conventions 104
- model overview 103
- model upgrades 113
- modular expansion 20
- multiple allegiance 15
- network requirements 153



- OEL
- open systems 150, 320
- ordering licensed functions 170
- p5 570 21, 27
- PAV
- performance 14, 186, 253
- placement of data 99
- planning 344
- positioning 11
- power and cooling 39
- power control 80
- power requirements 147
- POWER5 6, 26, 261
- PPS 40
- processor complex 26
- processor memory 28
- rack operator panel 21
- RAS
- remote power control 154
- remote support 154
- Resiliency Family 8
- RIO-G 29
- RPC
- SAN requirements 153–154
- SARC 24
- scalability 13, 103, 109
  - benefits 110, 112
  - for capacity 109
  - for performance 110
- SDD
- series overview 4
- server-based 24
- service 10
- service offerings 407
- service processor 28
- setup 10
- S-HMC 40, 153
- site preparation 145
- SMP 24
- spares 35
- SPCN
- split network 236
- Standby CoD 180
- storage capacity 7
- stripe size 267
- support offerings 407
- supported environment 8
- switched FC-AL 34
- z/OS enhancements 282
- z/OS Global Mirror
- z/OS Metro/Global Mirror 10
- zSeries performance 14
- DS8000 Global Mirror Utility see GMU
- DS8100 13
  - Model 921 105
  - processor memory 28
- DS8300 13
  - Copy Services 55
  - disk subsystem 30
  - FlashCopy 56

- LPAR 48
  - LPAR benefits 56
  - LPAR implementation 49
  - LPAR security 54
  - Model 922 106
  - Model 9A2 106
  - Model 9A2 configuration options 52
  - processor memory 28
  - Remote Mirror and Copy 56
  - storage system LPAR 8

## E

- Enterprise Remote Copy Management Facility see eRCMF
- EPO 21, 80
- eRCMF 331, 409
- ESCON 73, 139, 256
  - architecture 38
  - distances 38
  - Remote Mirror and Copy 38
  - supported servers 38
- ESS
  - compared to DS8000 11
- ESS CLI 138, 237
  - storage management 238
- ESS CS CLI
  - command flow 234
- Ethernet
  - switches 41, 82
- expansion frame 21
- Extended Remote Copy see z/OS Global Mirror
- extent pools 89, 192, 201

## F

- FC-AL
  - non-switched 32
  - overcoming shortcomings 257
  - shortcomings 257
  - switched 6
- Fibre Channel
  - distances 39
  - host adapters 38
- FICON 14, 73, 256
  - host adapters 38
- filemon 354
- fixed block LUNs 91
- FlashCopy 9, 56, 116, 168
  - benefits 118
  - Consistency Group 121
  - data set 119
  - establish on existing RMC primary 121–122
  - inband commands 9, 123
  - Incremental 118
  - Multiple Relationship 120
  - no background copy 118
  - on iSeries 403
  - options 118
  - persistent 123
  - Refresh Target Volume 118

- FlashCopy to a Remote Mirror primary 9
- frames 20
  - base 20
  - expansion 21
- ftp offload option 166

## G

- GDPS 409
- GDS for MSCS 410
- Geographically Dispersed Sites for Microsoft Cluster Services see GDS for MSCS
- Global Copy 9, 116, 124, 131, 168, 303, 339–340, 400
- Global Mirror 10, 116, 125, 131, 168
  - how works 126
- GMU 330

## H

- HA 6, 37
- HBA vendor resources 321
- host adapter
  - four port 260
- host adapter see HA
- host adapters
  - Fibre Channel 38
  - FICON 38
- host attachment 97, 190
- host connection
  - zSeries 74
- hot spot avoidance 188
- HP OpenVMS 368
  - command console LUN 370
  - FC port configuration 368
  - volume configuration 369
  - volume shadowing 370
- Hypervisor 66

## I

- I/O enclosure 29, 40
- I/O priority queuing 15, 187
- IASP 378
- IBM Geographically Dispersed Parallel Sysplex see GDPS
- IBM Migration Services 342, 408–409
- IBM Service Offerings 408
- IBM TotalStorage DS Command-Line Interface see DS CLI
- IBM TotalStorage DS Storage Manager see DS Storage Manager
- IBM TotalStorage Multipath Subsystem Device Driver see SDD
- IBM TotalStorage Multiple Device Manager 327
- IBM TotalStorage Productivity Center see TPC
- IBM Virtualization Engine
  - LPAR 44
- inband commands 123
- Incremental FlashCopy 9, 118
- Independent Auxiliary Storage Pools see IASP
- Information Center 207

- Information Lifecycle Management 4
- Infrastructure Simplification 4
- installation planning 144
  - delivery requirements 144
  - environmental requirements 149
  - floor and space requirements 145
  - host attachment 150
    - ESCON 151
    - FICON 151
    - open systems 150
    - updated info 152
  - network requirements 153
  - remote power control 154
  - remote support 154
  - S-HMC 153
  - power requirements 147
  - SAN requirements 153–154
  - site preparation 145
- iostat 345, 354, 365
- iSeries
  - adding multipath volumes using 5250 interface 388
  - adding volumes 376
  - adding volumes to an IASP
  - adding volumes using iSeries Navigator 390
  - AIX 353
  - AIX on 404
  - avoiding single points of failure 386
  - cache 397
  - changing from single path to multipath 396
  - changing LUN protection 375
  - configuring multipath 387
  - connecting via SAN switches 400
  - Copy Services 403
  - FlashCopy 403
  - Global Copy 400
  - hardware 374
  - Linux 363, 405
  - logical volume sizes 374
  - LUNs 93
  - managing multipath volumes using iSeries Navigator 392
  - Metro Mirror 400
  - migration to DS8000 400
  - multipath 386
  - multipath rules for multiple iSeries systems or partitions 395
  - number of fibre channel adapters 398
  - OS/400 data migration 401
  - OS/400 mirroring 400
  - planning for arrays and DDMs 397
  - protected versus unprotected volumes 375
  - recommended number of ranks 399
  - Remote Mirror and Copy 403
  - sharing ranks with other servers 399
  - size and number of LUNs 398
  - sizing guidelines 396
  - software 374
  - toolkit for Copy Services 404
  - using 5250 interface 376

## L

- LCU 196
- licensed functions 167
  - ordering 170
  - scenarios 174
- Linux 356
  - /proc pseudo file system 364
  - issues 358
  - managing multiple paths 360
  - missing device files 359
  - on iSeries 363, 405
  - performance monitoring with iostat 365
  - reference material 357
  - RH-EL
  - SCSI basics 358
  - SCSI device assignment changes 360
  - support issues 356
- logical configuration 174
  - assigning LUNs to hosts 226
  - creating arrays 219
  - creating CKD LCUs 227
  - creating CKD volumes 227
  - creating FB volumes 222
  - deleting LUNs 226
  - displaying WWNN 228
  - extent pools 221
  - flow 201
  - host systems 216
  - planning 199
  - process 211
  - steps 199
  - storage complex 211
  - storage unit 212
  - terminology 190
  - volume groups 224
- logical migration 307
- logical partition see LPAR
- logical subsystem see LSS
- logical volumes 91, 194, 201
- LPAR 44–45, 62
  - application isolation 47
  - benefits 56
  - consolidation of multiple versions 47
  - Copy Services 55
  - DS8300 48
  - DS8300 implementation 49
  - Hypervisor 66
  - increased flexibility 48
  - increased hardware utilization 47
  - production and test environments 47
  - security through Power Hypervisor 54
  - server consolidation 47
  - storage facility image 48
  - why? 47
- LSS 94, 196
- LUNs
  - allocation and deletion 93
  - fixed block 91
  - iSeries 93
- LVM

- configuration 352
- mirroring 352
- striping 266, 352

## M

- Metro Mirror 9, 116, 123, 130, 168, 303, 339–340, 400
- Metro/Global Copy 10, 302
- microcode updates
  - installation process 81
- Microsoft Windows 2000/2003 366
  - HBA 366
  - SDD 366
- mirroring 352
- modular expansion 20
- multipathing
  - other solutions 325
- multiple allegiance 15
- Multiple Relationship FlashCopy 9, 120

## N

- non-volatile storage see NVS
- NVS 70
- NVS recovery

## O

- OEL 167
- open systems
  - cache size 265
  - data migration 333
  - DS CLI 325
  - eRCMF
  - IBM Migration Services 409
  - operating system specifics 343
  - performance 264
  - SDD 324
  - sizing 264
  - support 320
    - boot support 323
    - differences from ESS 2105 322
    - RPQ 323
  - supported systems 320
- operating environment license see OEL
- OS/400 data migration 401
- OS/400 mirroring 400

## P

- p5 570 21, 27
- panel
  - rack operator 21
- Parallel Access Volumes see PAV
- partitioning
  - concepts 44
- PAV 14, 168, 170, 187
- performance
  - open systems 264
- performance
  - AAL 256
  - challenge 254

- data placement 265
- Disk Magic 186–187
- ESCON 256
- FICON 256
- hot spot avoidance 188
- I/O priority queuing 187
- LVM striping 266
- monitoring 187
- monitoring tools
  - UNIX 345
  - vmstat 347
- number of host ports/channels 187
- open systems
  - determining the connections 267
  - determining the number of paths to a LUN 268
- PAV 187
- planning 186
- PPRC over FCP links 256
- remote copy 187
- size of cache 187
- where to attach the host 268
- workload characteristics 265
- z/OS 269
  - appropriate DS8000 size 271
  - channel consolidation 272
  - configuration recommendations 274
  - connect to zSeries hosts 269
  - disk array sizing 273
  - potential 270
  - processor memory size 271
- Persistent FlashCopy 123
- PFA 78
- physical partitioning 45
- Piper 299, 341
- placement of data 99
- Point-in-time Copy see PTC
- positioning 11
- power 40
  - BBU
    - building power lost 80
    - disk enclosure 40
    - fluctuation protection 80
    - I/O enclosure 40
  - PPS
    - processor enclosure 40
  - RPC 79
- power and cooling 39, 79
  - BBU
  - PPS
    - rack cooling fans 79
  - RPC 79
- Power Hypervisor 54
- POWER5 6, 26, 261
- PPRC Extended Distance see Global Copy
- PPRC over FCP links 256
- PPRC-XD see Global Copy
- PPS 79
- Predictive Failure Analysis see PFA
- primary power supply see PPS
- processor complex 26, 63

- processor enclosure
  - power 40
- PTC 116, 118, 168

## R

- rack operator panel 21
- rack power control cards see RPC
- RAID-10
  - AAL 77
  - drive failure 77
  - implementation 77
  - theory 77
- RAID-5
  - drive failure 76
  - implementation 76
  - theory 76
- ranks 88, 192, 201
- RAS 61
  - CUIR
    - disk scrubbing 79
    - disk subsystem 75
      - disk path redundancy 75
    - EPO 80
    - fault avoidance 64
    - first failure data capture 64
    - host connection availability 71
    - Hypervisor 66
    - I/O enclosure 67
    - metadata 67
    - microcode updates 81
      - installation process 81
    - naming 62
    - NVS recovery
    - permanent monitoring 64
    - PFA
      - power and cooling 79
    - processor complex 63
    - RAID-10 77
    - RAID-5 76
    - RIO-G 67
    - server 67
      - server failover and failback 68
    - S-HMC 82
      - spare creation 77
  - Real-time configuration 204
    - Copy Services 204–205
  - Recovery Point Objective see RPO
  - Redbooks Web site 415
    - Contact us xix
  - RedHat Enterprise Linux see RH-EL
  - reliability, availability, serviceability see RAS
  - remote access 164
  - Remote Mirror and Copy 56, 116
    - ESCON 38
      - on iSeries 403
  - Remote Mirror and Copy see RMC
  - Remote mirror for z/OS see RMZ
  - remote power control 154
  - remote support 154
  - RH-EL 361

- RIO-G 29
- RMC 123, 169
  - comparison of functions 130
  - Global Copy 124
  - Global Mirror 125
  - Metro Mirror 123
- RMF 289
- RMZ 169
- RPC 39
- RPO 131
- rsync 337

**S**

- SAN 73
- SAR 346
- SARC 14, 24
- scalability 13
  - DS8000
    - scalability 264
- SDD 14, 268, 324, 349
  - for Windows 366
- security 165
- Sequential Prefetching in Adaptive Replacement Cache  
see SARC
- server
  - RAS 67
- server consolidation 47
- server-based SMP 24
- service offerings
  - Web sites 408
- service processor 28
- S-HMC 7, 40, 82, 136, 158
  - Advanced Copy Services 162
  - call home 164
  - dial up connection 163
  - DS Storage Manager 160
  - external 159
  - ftp offload option 166
  - network topology 162
  - remote access 164
  - remote services 160
  - secure high speed connection 163
  - security considerations 165
  - service 162
  - software components 160
  - system and partition management 162
  - user management 166
- simulated configuration 205
- sizing
  - open systems 264
  - z/OS 269
- SMI-S 162
- spares 35, 77
  - floating 78
- sparing
  - examples 177
  - rules 176
- SPCN 28
- SSA 256
- Standby Capacity on Demand see Standby CoD

- Standby CoD 7, 180
- storage capacity 7
- storage complex 62, 190
- storage facility image 48, 62
  - addressing capabilities 58
  - hardware components 50
  - I/O resources 51
  - processor and memory allocations 51
  - RIO-G interconnect separation 51–52
- Storage Hardware Management Console see S-HMC
- storage image 190
- storage system logical partitions see storage system LPAR
- storage system LPAR 7
  - DS8300 8
  - future directions 13
- storage unit 62, 190
- stripe
  - size 267
- switched FC-AL 6
  - advantages 33
  - DS8000 implementation 34
- Synchronous PPRC see Metro Mirror
- System Activity Report see SAR
- system power control network see SPCN

**T**

- TPC 326
  - for disk 329
  - for replication 330
- TPF 291

**U**

- UNIX
  - iostat 345
  - SAR
  - vmstat 347

**V**

- VDS 367
- Virtual Disk Service see VDS
- virtualization
  - abstraction layers for disk 85
  - address groups 96
  - array sites 86
  - arrays 87
  - benefits 100
  - concepts 83
  - definition 84
  - extent pools 89
  - hierarchy 98
  - host attachment 97
  - logical volumes 91
  - ranks 88
  - storage system 84
  - volume group 97
- vmstat 347
- volume groups 97, 194

- volumes
  - CKD 91
- VSE
  - ESA 315

## **W**

- Windows Server 2003
  - VDS support
- WWNN 228
- WWPN 348

## **X**

- XRC see z/OS Global Mirror

## **Z**

- z/OS
  - coexistence considerations 290
  - DASD Unit Information Module 284
  - device recognition 284
  - IBM Migration Services 408
  - IOS scalability 282
  - IPL enhancements 284
  - large volume support 283
  - migration considerations 289–290
  - new performance statistics 285
  - read availability mask support 283
  - read control unit 284
  - RMF 289
- z/OS Global Mirror 10, 116, 128, 132, 168, 301
- z/OS Metro/Global Mirror 10, 129
- z/VM 290, 315
- Z/VE 290
- zSeries
  - host connection 74
  - performance 14



## The IBM TotalStorage DS8000 Series: Concepts and Architecture

(0.5" spine)  
0.475" x 0.873"  
250 x 459 pages









# The IBM TotalStorage DS8000 Series: Concepts and Architecture



**Advanced features and performance breakthrough with POWER5 technology**

**Configuration flexibility with LPAR and virtualization**

**Highly scalable solutions for on demand storage**

This IBM Redbook describes the IBM TotalStorage DS8000 series of storage servers, its architecture, logical design, hardware design and components, advanced functions, performance features, and specific characteristics. The information contained in this redbook is useful for those who need a general understanding of this powerful new series of disk enterprise storage servers, as well as for those looking for a more detailed understanding of how the DS8000 series is designed and operates.

The DS8000 series is a follow-on product of the IBM TotalStorage Enterprise Storage Server with new functions related to storage virtualization and flexibility. This book describes the virtualization hierarchy that now includes virtualization of a whole storage subsystem. This is made possible by utilizing IBM's pSeries POWER5-based server technology and its Virtualization Engine LPAR technology. This LPAR technology offers totally new options to configure and manage storage.

In addition to the logical and physical description of the DS8000 series, the fundamentals of the configuration process are also described in this redbook. This is useful information for proper planning and configuration when installing the DS8000 series, as well as for the efficient management of this powerful storage subsystem.

## **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](http://ibm.com/redbooks)

SG24-6452-00

ISBN 0738492205