IBM

# The IBM TotalStorage DS6000 Series: Concepts and Architecture

- Enterprise-class storage functions in a compact and modular design
- On demand scalability and multi-platform connectivity
- Enhanced configuration flexibility with virtualization


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


Redbooks

ibm.com/redbooks

International Technical Support Organization

# The IBM TotalStorage DS6000 Series: Concepts and Architecture

March 2005

SG24-6471-00

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

**First Edition (March 2005)**

This edition applies to the DS6000 series per the October 12, 2004 announcement. Please note that an early version of DS6000 microcode was used for the screen captures and command output, so some details may vary from the currently available microcode.

**Note:** This book contains detailed information about the architecture of IBM's DS6000 product family. We recommend that you consult the product documentation and the Implementation Redbooks for more detailed information on how to implement the DS6000 in your environment.

# Contents



# 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:

iSeries™
i5/OS™
pSeries®
xSeries®
z/OS®
z/VM®
z/VSE™
zSeries®
AIX 5L™
AIX®
AS/400®
BladeCenter®
CICS®
DB2®
DFSMS/MVS®
DFSMS/VM®
DFSMSdss™
DFSMShsm™
DFSORT™
Enterprise Storage Server®
ESCON®
FlashCopy®
FICON®
Geographically Dispersed Parallel Sysplex™
GDPS®
HACMP™
IBM®
IMS™
Lotus Notes®
Lotus®
Multiprise®
MVS™
Netfinity®
Notes®
OS/390®
OS/400®
Parallel Sysplex®
Power PC®
PowerPC®
Predictive Failure Analysis®
POWER™
Redbooks™
RMF™
RS/6000®
S/390®
Seascape®
System/38™
Tivoli®
TotalStorage Proven™
TotalStorage®
VSE/ESA™

The following terms are trademarks of other companies:

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

Intel, Intel Inside (logos), Itanium, 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.

# Summary of changes

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

Summary of Changes
for SG24-6471-00
for DS6000 Series: Concepts and Architecture
as created or updated on December 13, 2005.

## March 2005, First Edition

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

### Changed information (see change bars)

- Fixed errors in table on p. 330-331
- SAN boot is available for the IBM® eServer BladeCenter®
- Updated with August 2005 announcement information
- Updated code load information
- Updated DS CLI user management information
- Updated DS CLI command examples with corrected syntax
- Removed references to pre and post GA restrictions that no longer apply
- Updated bibliography

# Preface

This IBM Redbook describes the IBM TotalStorage® DS6000 storage server series, its architecture, its 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 disk subsystem, as well as for those looking for a more detailed understanding of how the DS6000 series is designed and operates.

The DS6000 series is a follow-on product to the IBM TotalStorage Enterprise Storage Server® with new functions related to storage virtualization and flexibility.

The DS6000 series is a storage product targeted for the midrange market, but it has all the functions and availability features that normally can be found only in high end storage systems. In a very small enclosure, which fits in a standard 19-inch rack, the DS6000 series offers capacity, reliability functions, and performance similar to those of an ESS 800 or comparable high-end storage systems.

For example, zSeries® customers now also have the option to buy a midrange priced storage subsystem for their environment without giving up the functionality of an ESS.

In addition to the logical and physical description of the DS6000 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 DS6000 series, as well as for the efficient management of this powerful storage subsystem.

Characteristics of the DS6000 series described in this redbook also include the DS6000 copy functions: FlashCopy®, Metro Mirror, Global Copy, and Global Mirror. The performance features, particularly the switched FC-AL implementation of the DS6000 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 7 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 IBMs Enterprise Storage Server, the 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 at 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 FAStT. He has worked at IBM for six and a half years. Before joining IBM, Chuck was a hardware CE on UNIX systems for 10 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 6 first years. She has 7 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 transfered 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 IBM 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 responsiblity 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, 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.

*Figure 0-1 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, 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

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

## 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

- Send your comments in an email to:

  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 DS6000 series and its key features. The topics covered include:

- Product overview
- Positioning
- Performance

1

# Introducing the IBM TotalStorage DS6000 series

This chapter provides an overview of the features, functions, and benefits of the IBM TotalStorage DS6000 series of storage servers. The topics covered include:

- Overview of the DS6000 series and its benefits
- Positioning the DS6000 series within the whole family of IBM Disk Storage products
- Performance of the DS6000 series

## 1.1 The DS6000 series, a member of the TotalStorage DS Family

IBM has a wide range of product offerings that are based on open standards and share a common set of tools, interfaces, and innovative features. The IBM TotalStorage DS Family and its new member, the DS6000 series (see Figure 1-1), 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.

*Figure 1-1 DS6000 series*

### 1.1.1 Infrastructure simplification

Consolidation begins with compatibility. The IBM TotalStorage DS Family and the DS6000 support a broad array of IBM and non-IBM server platforms, including IBM z/OS, z/VM®, OS/400®, i5/OS™, and AIX 5L™ operating systems, as well as Linux, HP-UX, Sun Solaris, Novell NetWare, UNIX, and Microsoft® Windows environments. Consequently, you have the freedom to choose preferred vendors and run the applications you require to meet your enterprise's needs while extending your previous IT investments.

Storage asset consolidation can be greatly assisted by virtualization. Virtualization software solutions are designed to logically combine separate physical storage systems into a single, virtual storage pool, thereby offering dramatic opportunities to help reduce the total cost of ownership (TCO), particularly when used in combination with the DS6000 series.

### 1.1.2 Business continuity

The IBM TotalStorage DS Family and the DS6000 series as a member of this family, supports enterprise-class data backup and disaster recovery capabilities. As part of the IBM TotalStorage Resiliency Family of software, IBM TotalStorage FlashCopy point-in-time copy capabilities back up data in the background, while allowing users nearly instant access to information on both the source and target volumes. Metro and Global Mirror and Global Copy capabilities allow the creation of duplicate copies of application data at remote sites.

### 1.1.3 Information lifecycle management

By retaining frequently accessed or high-value data in one storage server and archiving less valuable information in a less costly one, systems like the DS6000 series can help improve the management of information according to its business value—from the moment of its creation to the moment of its disposal. The policy-based management capabilities built into the IBM TotalStorage Open Software Family, IBM DB2® Content Manager and IBM Tivoli®

Storage Manager for Data Retention, are designed to help you automatically preserve critical data, while preventing deletion of that data before its scheduled expiration.

## 1.2 IBM TotalStorage DS6000 series unique benefits

The IBM TotalStorage DS6000 series is a Fibre Channel based storage system that supports a wide range of IBM and non-IBM server platforms and operating environments. This includes open systems, zSeries, and iSeries servers.

In a small 3U footprint, the new storage subsystem provides performance and functions for business continuity, disaster recovery and resiliency, previously only available in expensive high-end storage subsystems. The DS6000 series is compatible regarding Copy Services with the previous Enterprise Storage Server (ESS) Models 800 and 750, as well as with the new DS8000 series.

The DS6000 series offers an entirely new era in price, performance, and scalability. Now for the first time zSeries and iSeries customers have the option for a midrange priced storage subsystem with all the features and functions of an enterprise storage subsystem.

Some clients do not like to put large amounts of storage behind one storage controller. In particular, the controller part of a high-end storage system makes it expensive. Now you have the option of choice: you can build very cost efficient storage systems by adding expansion enclosures to the DS6800 controller, but since the DS6800 controller is not really expensive, you can also grow horizontally by adding other DS6800 controllers. You also have the option to easily grow into the DS8000 series by adding DS8000 systems to your environment or by replacing DS6000 systems (see Figure 1-2).

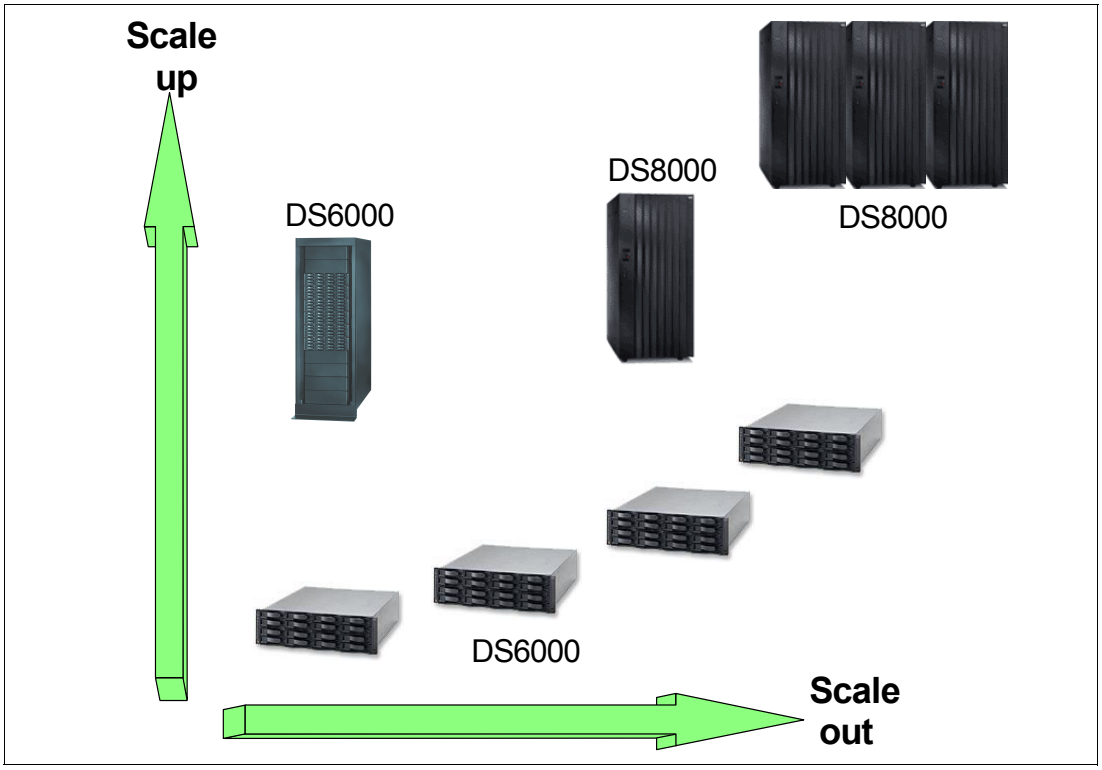



*Figure 1-2 Scaling options of the DS6000 and DS8000 series*

### 1.2.1 Hardware overview

The DS6000 series consists of the DS6800, Model 1750-511, which has dual Fibre Channel RAID controllers with up to 16 disk drives in the enclosure (see Figure 1-1 on page 4). Capacity can be increased by adding up to 7 DS6000 expansion enclosures, Model 1750-EX1, each with up to 16 disk drives, as shown Figure 1-3.

*Figure 1-3 DS6800 with five DS6000 expansion enclosures in a rack*

#### DS6800 controller enclosure (Model 1750-511)

IBM TotalStorage systems are based on a server architecture. At the core of the DS6800 controller unit are two active/active RAID controllers based on IBM's industry leading PowerPC® architecture. By employing a server architecture with standard hardware components, IBM's storage division can always take advantage of the best of breed components developed by other IBM divisions. The customer gets the benefit of a very cost efficient and high performing storage system.

##### *The processors*

The DS6800 utilizes two 64-bit PowerPC 750GX 1 GHz processors for the storage server and the host adapters, respectively, and another PowerPC 750FX 500 MHz processor for the device adapter on each controller card. The DS6800 is equipped with 2 GB memory in each controller card, adding up to 4 GB. Some part of the memory is used for the operating system and another part in each controller card acts as nonvolatile storage (NVS), but most of the memory is used as cache. This design to use processor memory makes cache accesses very fast.

When data is written to the DS6800, it is placed in cache and a copy of the write data is also copied to the NVS of the other controller card, so there are always two copies of write data until the updates have been destaged to the disks. On zSeries, this mirroring of write data can be disabled by application programs, for example, when writing temporary data (Cache Fast Write). The NVS is battery backed up and the battery can keep the data for at least 72 hours if power is lost.

The DS6000 series controller's Licensed Internal Code (LIC) is based on the DS8000 series software, a greatly enhanced extension of the ESS software. Since 97% of the functional code of the DS6000 is identical to the DS8000 series, the DS6000 has a very good base to be a stable system.

### *The disk drives*

The DS6800 controller unit can be equipped with up to 16 internal FC-AL disk drive modules, offering up to 4.8 TB of physical storage capacity in only 3U (5.25”) of standard 19” rack space.

### *Dense packaging*

Calibrated Vectored Cooling technology used in xSeries® and BladeCenter to achieve dense space saving packaging is also used in the DS6800. The DS6800 weighs only 49.6 kg (109 lbs.) with 16 drives. It connects to normal power outlets with its two power supplies in each DS6800 or DS6000 expansion enclosure. All this provides savings in space, cooling, and power consumption.

### *Host adapters*

The DS6800 has eight 2 Gbps Fibre Channel ports that can be equipped with two or up to eight shortwave or longwave Small Formfactor Plugables (SFP). You order SFPs in pairs. The 2 Gbps Fibre Channel host ports (when equipped with SFPs) can also auto-negotiate to 1 Gbps for legacy SAN components that support only 1 Gbps. Each port can be configured individually to operate in Fibre Channel or FICON® mode, but you should always have pairs. Host servers should have paths to each of the two RAID controllers of the DS6800.

### *Switched FC-AL subsystem*

The disk drives in the DS6800 or DS6000 expansion enclosure have a dual ported FC-AL interface. Instead of forming an FC-AL loop, each disk drive is connected to two Fibre Channel switches within each enclosure. With this switching technology there is a point-to-point connection to each disk drive. This allows maximum bandwidth for data movement, eliminates the bottlenecks of loop designs, and allows for specific disk drive fault indication.

There are four paths from the DS6800 controllers to each disk drive to provide greater data availability in the event of multiple failures along the data path. The DS6000 series systems provide preferred path I/O steering and can automatically switch the data path used to improve overall performance.

Here is a summary of the DS6800 major features:

- Two RAID controller cards.
- Two PowerPC 750GX 1 GHz processors and one PowerPC 750FX processor on each RAID controller card.
- 4 GB of cache.
- Two battery backup units (one for each controller card).
- Two AC/DC power supplies with imbedded enclosure cooling units.
- Eight 2 Gbps device ports - for additional DS6000 expansion enclosures connectivity.
- Two Fibre Channel switches for disk drive connectivity in each DS6000 series enclosure.
- Eight Fibre Channel host ports that can be configured as pairs of FCP or FICON host ports. The host ports auto-negotiate to either 2 Gbps or 1 Gbps link speeds.
- Attachment to up to seven (7) DS6000 expansion enclosures.
- Very small size, weight, and power consumption. All DS6000 series enclosures are 3U in height and mountable in a standard 19-inch rack.

### DS6000 expansion enclosure (Model 1750-EX1)

The size and the front look of the DS6000 expansion enclosure (1750-EX1) is the same as the DS6800 controller enclosure. In the front you can have up to 16 disk drives.

Aside from the drives, the DS6000 expansion enclosure contains two Fibre Channel switches to connect to the drives and two power supplies with integrated fans.

Up to 7 DS6000 expansion enclosures can be added to a DS6800 controller enclosure. The DS6800 supports two dual redundant switched loops. The first loop is for the DS6800 and up to three (3) DS6000 expansion enclosures. The second switched loop is for up to four (4) expansion enclosures. For connections to the previous and next enclosure, four inbound and four outbound 2 Gbps Fibre Channel ports are available.

## 1.2.2 Storage capacity

The DS6000 series offers outstanding scalability with physical capacities ranging from 584 GB up to 38.4 TB, while maintaining excellent performance. Physical capacity for the DS6800 and DS6000 expansion enclosure is purchased via disk drive sets. A disk drive set contains four identical disk drives (same capacity and revolutions per minute (RPM)). Currently, a minimum of eight drives (two disk drive sets) are required for the DS6800. You can increase the capacity of your DS6000 by adding one or more disk drive sets to the DS6800 or DS6000 expansion enclosure. Within the controller model DS6800, you can install up to four disk drive sets (16 disk drive modules (DDMs)). Up to 7 DS6000 expansion enclosures can be added non-disruptively, on demand, as your storage needs grow.

According to your performance needs you can select from four different disk drive types: fast 73 GB and 146 GB drives rotating at 15,000 RPM, good performing and cost efficient 146 GB drives operating at 10,000 RPM, and high capacity 300 GB drives running at 10,000 RPM.

The minimum storage capability with eight 73 GB DDMs is 584 GB. The maximum storage capability with 16 300 GB DDMs for the DS6800 controller enclosure is 4.8 TB. If you want to connect more than 16 disks, you can use the optional DS6000 expansion enclosures that allow a maximum of 128 DDMs per storage system and provide a maximum storage capability of 38.4 TB.

Every four or eight drives form a RAID array and you can choose between RAID-5 and RAID-10. The configuration process enforces that at least two spare drives are defined on each loop. In case of a disk drive failure or even when the DS6000's predictive failure analysis comes to the conclusion that a disk drive might fail soon, the data of the failing disk is reconstructed on the spare disk. More spare drives might be assigned if you have drives of mixed capacity and speed. The mix of different capacities and speeds will not be available at general availability, but at a later time.

## 1.2.3 DS management console

The DS management console consists of the DS Storage Manager software, shipped with every DS6000 series system, and a computer system on which the software can run. The DS6000 management console running the DS Storage Manager software is used to configure and manage DS6000 series systems. The software runs on a Windows or Linux system that the client can provide.

### 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 to the DS management console 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 new or un-configured storage unit at a later time.

- **Real-time (online) configuration**

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

The DS6000 series' Express Configuration Wizards guide you through the configuration process and help get the system operational in minimal time. The DS Storage Manager's GUI is intuitive and very easy to understand.

### IBM TotalStorage DS Command-Line Interface

The DS Command-Line Interface (CLI) is a single CLI that has the ability to perform a full set of commands for logical configuration, Copy Services activities, or both. The DS CLI can also issue Copy Services commands to an ESS Model 750, ESS Model 800 (at LIC level 2.4.3 and above), or DS8000 series system. It is possible to combine the DS CLI commands into a script. This can help enhance your productivity since it eliminates the previous (on ESS) requirement for you to create and save a task using the GUI.

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

### 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.

## 1.2.4 Supported environment

The DS6000 system can be connected across a broad range of server environments, including IBM eServer, zSeries, iSeries, xSeries, BladeCenter, and pSeries servers, as well as servers from Sun Microsystems, Hewlett-Packard, and other providers. You can easily split up the DS6000 system storage capacity among the attached environments. This makes it an ideal system for storage consolidation in a dynamic and changing on demand environment.

Particularly for zSeries and iSeries customers, the DS6000 series will be an exciting product, since for the first time it gives them the choice to buy a midrange priced storage system for their environment with a performance that is similar to or exceeds that of an IBM ESS.

## 1.2.5 Business continuance functions

As data and storage capacity are growing faster year by year most customers can no longer afford to stop their systems to back up terabytes of data, it just takes too long. Therefore, IBM has developed fast replication techniques that can provide a point-in-time copy of the customer's data in a few seconds or even less. This function is called FlashCopy on the DS6000 series, as well as on the ESS models and DS8000 series.

As data becomes more and more important for an enterprise, losing data or access to data, even only for a few days, might be fatal for the enterprise. Therefore, many customers, particularly those with high end systems like the ESS and the DS8000 series, have implemented Remote Mirroring and Copy techniques previously called Peer-to-Peer Remote Copy (PPRC) and now called Metro Mirror, Global Mirror, or Global Copy. These functions are also available on the DS6800 and are fully interoperable with ESS 800 and 750 models and the DS8000 series.

### Point-in-time Copy feature

The Point-in-time Copy feature consists of the FlashCopy function. The primary objective of FlashCopy is to create very quickly a point-in-time copy of a source volume on the target volume. When you initiate a FlashCopy operation, a FlashCopy relationship is created between the source volume and target volume. A FlashCopy relationship is a mapping of a FlashCopy source volume and a FlashCopy target volume. The FlashCopy relationship exists between the volume pair from the time that you initiate a FlashCopy operation until the storage unit copies all data from the source volume to the target volume, or until you delete the FlashCopy relationship if it is a persistent FlashCopy. The benefits of FlashCopy are that the point-in-time copy is immediately available for use for backups and the source volume is immediately released so that applications can be restarted, with minimal application downtime. The target volume is available for read and write processing so it can be used for testing or backup purposes. You can choose to leave the copy as a logical copy or choose to physically copy the data. If you choose to physically copy the data, a background process copies tracks from the source volume to the target volume.

FlashCopy is an additional charged feature. You have to order the Point-in-time Copy feature, which includes FlashCopy. Then you have to follow a procedure to get the key from the internet and install it on your DS6800.

To make a FlashCopy of a LUN or a z/OS CKD volume you need a target LUN or z/OS CKD volume of the same size as the source within the same DS6000 system (some operating systems also support a copy to a larger volume). z/OS customers can even do FlashCopy on a data set level basis when using DFSMSdss™. Of course the DS6000 also supports Concurrent Copy.

The DS Storage Manager's GUI provides an easy way to set up FlashCopy or Remote Mirror and Copy functions. Not all functions are available via the GUI; instead, we recommend that you use the new DS Command-Line Interface (DS CLI), which is much more flexible.

#### *Data Set FlashCopy*

In a z/OS environment when DFSMSdss is used to copy a data set, by default FlashCopy is used to do the copy. In this environment FlashCopy can operate at a data set level.

#### *Multiple Relationship FlashCopy*

Multiple Relationship FlashCopy allows a source to have FlashCopy relationships with multiple targets simultaneously. This flexibility allows you to initiate up to 12 FlashCopy relationships on a given logical unit number (LUN), volume, or data set, without needing to first wait for or cause previous relationships to end.

### *Incremental FlashCopy*

Incremental FlashCopy provides the capability to *refresh* a LUN or volume involved in a FlashCopy relationship. When a subsequent FlashCopy is initiated, only the data required to bring the target current to the source's newly established point-in-time is copied. This unburdens the backend and the disk drives are not so busy and can do more production I/Os. The direction of the refresh can also be reversed, in which case the LUN or volume previously defined as the target becomes the source for the LUN or volume previously defined as the source (and now the target).

### *FlashCopy to a remote mirror primary*

When we mention *remote mirror primary*, we mean any primary volume of a Metro Mirror or Global Copy pair. FlashCopy to a remote mirror primary lets you establish a FlashCopy relationship where the target is a remote mirror primary volume. This enables you to create full or incremental point-in-time copies at a local site and let remote mirroring operations copy the data to the remote site. Many clients that utilize a remote copy function use it for all their volumes. The DS6800 allows a FlashCopy onto a remote mirror primary volume.

Previous ESS clients faced the issue that they could not use FlashCopy since a FlashCopy onto a volume that was mirrored was not possible. This restriction particularly affected z/OS clients using data set level FlashCopy for copy operations within a mirrored pool of production volumes.

### *Consistency Group commands*

The Consistency Group function of FlashCopy allows the DS6800 to hold off I/O activity to a LUN or volume until all LUNs/volumes within the group have established a FlashCopy relationship to their targets and the FlashCopy **`Consistency Group Created`** command is issued or a timer expires. Consistency Groups can be used to help create a consistent point-in-time copy across multiple LUNs or volumes, and even across multiple DS6800 systems, as well as across DS8000 series, ESS 800, and ESS 750 systems.

### *Inband commands over remote mirror link*

In a remote mirror environment where you want to do a FlashCopy of the remote volumes, instead of sending FlashCopy commands across an Ethernet connection to the remote DS6800, Inband FlashCopy allows commands to be issued from the local or intermediate site, and transmitted over the remote mirror Fibre Channel links for execution on the remote DS6800. This eliminates the need for a network connection to the remote site solely for the management of FlashCopy.

## Remote Mirror and Copy feature

Remote Mirror and Copy is another separately orderable priced feature; it includes Metro Mirror, Global Copy, and Global Mirror. The local and remote storage systems must have a Fibre Channel connection between them. Remote Mirror and Copy functions can also be established between DS6800 and ESS 800/750 systems, but not between a DS6800 and older ESS models like the F20, because these models do not support remote mirror or copy across Fibre Channel (they only support ESCON®) and the DS6800 does not support ESCON.

### *Metro Mirror*

Metro Mirror was previously called Synchronous Peer-to-Peer Remote Copy (PPRC) on the ESS. It provides a synchronous copy of LUNs or zSeries CKD volumes. A write I/O to the source volume is not complete until it is acknowledged by the remote system. Metro Mirror supports distances of up to 300km.

### *Global Copy*

This is a non-synchronous long distance copy option for data migration and backup.

Global Copy was previously called PPRC-XD on the ESS. It is an asynchronous copy of LUNs or zSeries CKD volumes. An I/O is signaled complete to the server as soon as the data is in cache and mirrored to the other controller cache. The data is then sent to the remote storage system. Global Copy allows for copying data to far away remote sites. However, if you have more than one volume, there is no mechanism that guarantees that the data of different volumes at the remote site is consistent in time.

### *Global Mirror*

Global Mirror is similar to Global Copy but it provides data consistency.

Global Mirror is a long distance remote copy solution across two sites using asynchronous technology. It is designed to provide the following:

- Support for virtually unlimited distances between the local and remote sites, with the distance typically limited only by the capabilities of the network and channel extension technology being used. This can better enable 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 little impact to applications at the local site.
- Data currency, where for many environments the remote site lags behind the local site an average of three to five seconds, helps to minimize the amount of data exposure in the event of an unplanned outage. The actual lag in data currency experienced will depend upon a number of factors, including specific workload characteristics and bandwidth between the local and remote sites.
- Efficient synchronization of the local and remote sites, with support for failover and failback modes, which helps to reduce the time required to switch back to the local site after a planned or unplanned outage.

### *Three sites 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 and it 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. Metro/Global Copy is *not* available on the DS6800, but the following General Statement of Direction from IBM was included in the October 12, 2004 Hardware Announcement:

*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. The solution may include any mix of DS6000 series, ESS 750, ESS 800, or the DS8000 series.*

### *Remote Mirror connections*

All of the remote mirroring solutions described here use Fibre Channel as the communications link between the primary and secondary systems. The Fibre Channel ports used for Remote Mirror and Copy can be configured either as dedicated remote mirror links or as shared ports between remote mirroring and Fibre Channel Protocol (FCP) data traffic.

### *z/OS Global Mirror*

z/OS Global Mirror (previously called XRC) offers a specific set of very high scalability and high performance asynchronous mirroring capabilities designed to match very demanding,

large zSeries resiliency requirements. The DS6000 series systems can only be used as a target system in z/OS Global Mirror operations.

## 1.2.6 Resiliency

The DS6000 series has built in resiliency features that are not generally found in small storage devices. The DS6000 series is designed and implemented with component redundancy to help reduce and avoid many potential single points of failure.

Within a DS6000 series controller unit, there are redundant RAID controller cards, power supplies, fans, Fibre Channel switches, and Battery Backup Units (BBUs).

There are four paths to each disk drive. Using Predictive Failure Analysis®, the DS6000 can identify a failing drive and replace it with a spare drive without customer interaction.

### *Four path switched drive subsystem*

Most vendors feature Fibre Channel Arbitrated Loops, which can make it difficult to identify failing disks and is more susceptible to losing access to storage. The IBM DS6000 series provides dual active/active design architecture, including dual Fibre Channel switched disk drive subsystems which provide four paths to each disk drive. This ensures high data availability even in the event of multiple failures along the data path. Fibre Channel switches are used for each 16 disk drive enclosure unit, so that if a connection to one drive is lost, the remaining drives can continue to function, unlike disk drives configured in a loop design.

### *Spare drives*

The configuration process when forming RAID-5 or RAID-10 arrays will require that two global spares are defined in the DS6800 controller enclosure. If you have expansion enclosures, the first enclosure will have another two global spares. More spares could be assigned when drive groups with larger capacity drives are added.

### *Predictive Failure Analysis*

The DS6800 uses the well known (xSeries and ESS) Predictive Failure Analysis (PFA) to monitor the operations of its drives. PFA takes pre-emptive and automatic actions before critical drive failures occur. This functionality is based on a policy-based disk responsiveness threshold and takes the disk drive offline. The content of the failing drive is reconstructed from data and parity information of the other RAID array drives on the global spare disk drive. At the same time, service alerts are invoked, the failed disk is identified with Light Path indicators, and an alert Message Popup occurs on the management server.

## 1.2.7 Interoperability

The DS6800 features unsurpassed enterprise interoperability for a modular storage subsystem because it uses the same software as the DS8000 series, which is an extension of the proven IBM ESS code. This allows for cross-DS6000/8000 management and common software function interoperability, for example, Metro Mirror between a DS6000 and an ESS Model 800, while maintaining a Global Mirror between the same DS6000 and a DS8000 for some other volumes.

## 1.2.8 Service and setup

DS6000 series systems are designed to be easy to install and maintain; they are customer setup products. The DS Storage Manager's intuitive Web-based GUI makes the configuration process easy. For most common configuration tasks, Express Configuration Wizards are available to guide you through the installation process.

Light Path Diagnostics and controls are available for easy failure determination, component identification, and repair if a failure does occur. The DS6000 series can also be remotely configured and maintained when it is installed in a remote location.

The DS6800 consists of only five types of customer replaceable units (CRU). Light Path indicators will tell you when you can replace a failing unit without having to shut down your whole environment. If a concurrent maintenance is not possible, which might be the case for some double failures, the DS Storage Manager's GUI will guide you on what to do.

Of course a customer can also sign a service contract with IBM or an IBM Business Partner for extended service.

The DS6800 can be configured for a call home in the event of a failure and it can do event notification messaging. In this case an Ethernet connection to the external network is necessary. The DS6800 can use this link to place a call to IBM or to another service provider when it requires service. With access to the machine, service personnel can perform service tasks, such as viewing error and problem logs or initiating trace and dump retrievals.

At regular intervals the DS6000 sends out a heartbeat. The service provider uses this report to monitor the health of the call home process.

Configuration changes like adding disk drives or expansion enclosures are a non-disruptive process. Most maintenance actions are non-disruptive, including downloading and activating new Licensed Internal Code.

The DS6000 comes with a four year warranty. This is outstanding in the industry and shows IBM's confidence in this product. Once again, this makes the DS6800 a product with a low total cost of ownership (TCO).

## 1.2.9 Configuration flexibility

The DS6000 series uses virtualization techniques to separate the logical view of hosts onto LUNs from the underlying physical layer. This provides high configuration flexibility (see Chapter 4, "Virtualization concepts" on page 65).

### Dynamic LUN/volume creation and deletion

The DS6800 gives you a high degree of flexibility in managing your storage. LUNs can be created and also deleted without having to reformat a whole array.

LUN and volume creation and deletion is non-disruptive. When you delete a LUN or volume, the capacity can be reused, for example, to form a LUN of a different size.

### Large LUN and large CKD volume support

You can configure LUNs and volumes to span arrays, which allows for large LUN sizes.

LUNs can be as large as 2 TB.

The maximum volume size has also been increased for CKD volumes. Volumes with up to 65520 cylinders can now be defined, which corresponds to about 55.6 GB. This can greatly reduce the number of volumes that have to be managed.

### Flexible LUN to LSS association

A Logical Subsystem (LSS) is constructed to address up to 256 devices.

On an ESS there was a predefined association of arrays to Logical Subsystems. This caused some inconveniences, particularly for zSeries customers. Since in zSeries one works with relatively small volume sizes, the available address range for an LSS often was not sufficient to address the whole capacity available in the arrays associated with the LSS. On the other hand, in large capacity configurations where large CKD volumes were used, clients had to define several address ranges, even when there were only a few addresses used in an LSS. Several customers were confronted with a zSeries addressing limit of 64K addresses.

There is no predefined association of arrays to LSSs on the DS6000 series. Clients are free to put LUNs or CKD volumes into LSSs and make best use of the 256 address range of an LSS.

### Simplified LUN masking

The access to LUNs by host systems is controlled via volume groups. Hosts or disks in the same volume group share access to data. This is a new form of LUN masking. Instead of doing LUN masking for individual World Wide Port Names (WWPN), as implemented on the ESS, you can now do LUN masking at the host level by grouping all or some WWPNs of a host into a so-called Host Attachment and associating the Host Attachment to a Volume Group.

This new LUN masking process simplifies storage management because you no longer have to deal with individual Host Bus Adapters (HBAs) and volumes, but instead with groups.

### Summary

In summary, the DS6000 series allows for:

- Up to 32 logical subsystems
- Up to 8192 logical volumes
- Up to 1040 volume groups
- Up to 2 TB LUNs
- Large z/OS volumes with up to 65520 Cylinders

## 1.3 Positioning the IBM TotalStorage DS6000 series

The IBM TotalStorage DS6000 series is designed to provide exceptional performance, scalability, and flexibility while supporting 24 x 7 business operations to help provide the access and protection demanded by today's business environments. It is also designed to deliver 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 it is the disk product of choice for environments that require the highest levels of reliability, scalability, and performance available from IBM for mission-critical workloads. Clients who currently have IBM TotalStorage ESS models in their enterprise should also consider the IBM TotalStorage DS6000 series when they plan to replace or buy additional storage.

The IBM TotalStorage DS6000 series is designed for the cost, performance, and high capacity requirements of today's on demand business environments.

The DS6000 series is ideally suited for storage consolidation because it offers extensive connectivity options. A broad range of server environments, including IBM eServer zSeries, iSeries, xSeries, BladeCenter, and pSeries servers, as well as servers from Sun, HP, and other server providers are supported. This rich support of heterogeneous environments and

attachments, along with the flexibility to easily partition the DS6000 series storage capacity among the attached environments, makes the DS6000 series system a very attractive product in dynamic, changing environments.

In the midrange market it competes with many other products. The DS6000 series overlaps in price and performance with the DS4500 of the IBM DS4000 series of storage products. But the DS6000 series offers enterprise capabilities not found in other midrange offerings. A zSeries or iSeries customer might ponder whether to buy a DS6000 series or a DS8000 series class storage subsystem or a competitive subsystem. When comparing the DS6000 series with the DS4000 series and DS8000 series of storage products it is important to have a closer look at advanced functions and the management interfaces.

### 1.3.1 Common set of functions

The DS6000 series, the DS8000 series, and even the ESS storage subsystems share a common set of advanced functions, including FlashCopy, Metro Mirror, Global Copy, and Global Mirror.

There is also a set of common functions for storage management, including the IBM TotalStorage DS Command-Line Interface (DS CLI) and the IBM TotalStorage DS open application programming interface (API).

It is unique in the industry that IBM can offer, with the DS6000 series and the DS8000 series of products, storage systems with common management and copy functions for the whole DS family. This allows customers to easily expand their storage server farm to high end or midrange systems without the need for the storage administrators to learn about a new product. This will reduce your management costs and the total cost of ownership.

### 1.3.2 Common management functions

Within the DS6000 series and DS8000 series of storage systems, the provisioning tools like the DS Storage Manager's configuration GUI or CLI are very similar. Scripts written for one series member of storage servers will also work for the other series. Given this, it is easy for a storage administrator to work with either of the products. This reduces management costs since no training on a new product is required when adding a product of another series.

### 1.3.3 DS6000 series compared to other members of the TotalStorage DS Family

Here we compare the DS6000 series to other members of the IBM TotalStorage DS Family.

#### DS6000 series compared to ESS

The ESS clients will find it very easy to replace their old systems with a DS6800. All functions (with the exception of cascading Metro/Global Copy and z/OS Global Mirror), are the same as on the ESS and are also available on a DS6800.

If you want to keep your ESS and if it is a model 800 or 750 with Fibre Channel adapters, you can use your old ESS, for example, as a secondary for remote copy. With the ESS at the appropriate LIC level, scripts or CLI commands written for Copy Services will work for both the ESS and the DS6800.

For most environments the DS6800 performs much better than an ESS. You might even replace two ESS 800s with one DS6800. The sequential performance of the DS6800 is excellent. However, when you plan to replace an ESS with a large cache (let's say more than 16 GB) with a DS6800 (which comes with 4 GB cache) and you currently get the benefit of a

high cache hit rate, your cache hit rate on the DS6800 will drop down. This is because of the smaller cache. z/OS benefits from large cache, so for transaction-oriented workloads with high read cache hits, careful planning is required.

## DS6000 series compared to DS8000 series

You can think of the DS6000 series as the small brother or sister of the DS8000 series. All Copy Services (with the exception of z/OS Global Mirror) are available on both systems. You can do Metro Mirror, Global Mirror, and Global Copy between the two series. The CLI commands and the DS Storage Manager GUI look the same for both systems.

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

It is very easy to have a mixed environment, with DS8000 series systems where you need them and DS6000 series systems where you need a very cost efficient solution.

Logical partitioning with some DS8000 models is not available on the DS6000. For more information about the DS8000 refer to *The IBM TotalStorage DS8000 Series: Concepts and Architecture,* SG24-6452*.*

## DS6000 series compared to DS4000 series

Previous DS4000 series (formerly called FAStT) clients will find more of a difference between the DS4000 series and DS6000 series of products.

Both product families have about the same size and capacity, but their functions differ. With respect to performance, the DS4000 series range is below the DS6000 series. You have the option to choose from different DS4000 models, among them very low cost entry models. DS4000 series storage systems can also be equipped with cost efficient, high capacity Serial ATA drives.

The DS4000 series products allow you to grow with a granularity of a single disk drive, while with the DS6000 series you have to order at least four drives. Currently the DS4000 series also is more flexible with respect to changing RAID arrays on the fly and changing LUN sizes.

The implementation of FlashCopy on the DS4000 series is different as compared to the DS6000 series. While on a DS4000 series you need space only for the changed data, you will need the full LUN size for the copy LUN on a DS6000 series. However, while the target LUN on a DS4000 series cannot be used for production, it can on the DS6000 series. If a real copy of a LUN is needed on a DS4000 series, there is the option to do a volume copy. However, this is a two step process and it can take a long time until the copy is available for use. On a DS6000 series the copy is available for production after a few seconds.

Some of the differences in functions will disappear in the future. For the DS6000 series there is a General Statement of Direction from IBM (from the October 12, 2004 Hardware Announcement):

*Extension of IBM's dynamic provisioning technology within the DS6000 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.*

While the DS4000 series also offers remote copy solutions, these functions are not compatible with the DS6000 series.

### 1.3.4 Use with other virtualization products

IBM TotalStorage SAN Volume Controller is designed to increase the flexibility of your storage infrastructure by introducing a new layer between the hosts and the storage systems. The SAN Volume Controller can enable a tiered storage environment to increased flexibility in storage management. The SAN Volume Controller combines the capacity from multiple disk storage systems into a single storage pool, which can be managed from a central point. This is simpler to manage and helps increase utilization. It also allows you to apply advanced Copy Services across storage systems from many different vendors to help further simplify operations.

Currently, the DS4000 series product family (FAStT) is a popular choice for many customers who buy the SAN Volume Controller. With the DS6000 series, they now have an attractive alternative. Since the SAN Volume Controller already has a rich set of advanced copy functions, clients were looking for a cost efficient but reliable storage system. The DS6000 series fits perfectly into this environment since it offers good performance for the price while still delivering all the reliability functions needed to protect your data.

The SAN File System provides a common file system for UNIX, Windows, and Linux servers, with a single global namespace to help provide data sharing across servers. It is designed as a highly scalable solution supporting both very large files and very large numbers of files, without the limitations normally associated with Network File System (NFS) or Common Internet File System (CIFS) implementations.

To be able to share data in a heterogeneous environment the storage system must support the sharing of LUNs. The DS6000 series can do this and thus is an ideal candidate for your SAN File System data.

## 1.4 Performance

At the time this redbook was written, no officially published performance benchmarks were available.

With its fast six processors on the controller cards and the switched FC-AL disk subsystem, the DS6000 series is a high-performance modular storage system.

Some other performance relevant features are discussed in the following sections.

### 1.4.1 Tagged Command Queuing

Tagged Command Queuing allows Multiple AIX/UNIX I/O commands to be queued to the DS6800, which improves performance through autonomic storage management versus the server queuing one I/O request at a time. The DS6800 can reorder the queue to optimize disk I/O.

### 1.4.2 Self-learning cache algorithms - SARC

Cache algorithms determine what data is stored in cache and what data is removed. Read ahead caching will not store recently used data in cache, but will pre-fetch data and load it into cache. This is based on the idea that the application will want the next chunks of data in addition to the data it just received.

Most vendors use a cache algorithm based on what is commonly known as Last Recently Used (LRU), which places data to cache based on server access patterns. IBM's patent

pending *Sequential prefetching in Adaptive Replacement Cache (SARC)* places data in cache based not only on server access patterns, but also on frequency of data utilization.

### 1.4.3 IBM multipathing software

IBM Multi-path Subsystem Device Driver (SDD) provides load balancing and enhanced data availability capability in configurations with more than one I/O path between the host server and the DS6800. The data path from the host to the RAID controller is pre-determined by the LUN. Below the RAID controller, load balancing algorithms are designed to direct the data to the path that will have the best throughput.

Most vendors' priced multipathing software selects the preferred path at the time of initial request. IBM's free of charge *preferred path* multipathing software offers performance beyond this, by dynamically selecting the most efficient and optimum path to use at each data interchange during read and write operations.

### 1.4.4 Performance for zSeries

In this section we discuss some z/OS relevant performance features available on the DS6000 series.

#### Parallel Access Volumes (PAV)

PAV is an optional feature for zSeries environments which enables a single zSeries server to simultaneously process multiple I/O operations to the same logical volume, which can help to significantly improve throughput. 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. To utilize dynamic PAV, the Workload Manager must be used in Goal Mode.

#### Multiple Allegiance

Multiple Allegiance is a standard DS6800 feature which expands simultaneous logical volume access capability across multiple zSeries servers. This function, along with the software function PAV, enables the DS6800 to process more I/Os in parallel, helping to dramatically improve performance and enabling greater use of large volumes.

#### Priority I/O Queuing

Priority I/O Queuing improves performance in z/OS environments with several z/OS images. You can, for example, favor I/O from production systems compared to I/O from test systems. Storage administrator productivity can also be improved and Service Level Agreements better managed due to this capability.

# Part 2

# Architecture

In this part we describe various aspects of the DS6000 series architecture. These include:

- Hardware components
- RAS - reliability, availability, and serviceability
- Virtualization concepts
- Overview of the models
- Copy Services

2

# Components

This chapter details the hardware components of the DS6000. Here you can read about the DS6000 hardware platform and its components:

- Server enclosure
- Expansion enclosure
- Controller architecture
- Disk subsystem
- Server enclosure RAID controller card
- Expansion enclosure SBOD controller card
- Front panel
- Rear panel
- Power subsystem
- System service card
- Storage Manager console
- Cables

## 2.1 Server enclosure

The entire DS6800 including disks, controllers, and power supplies, is contained in a single 3U chassis which is called a server enclosure. If additional capacity is needed, it can be added by using a DS6000 expansion enclosure.

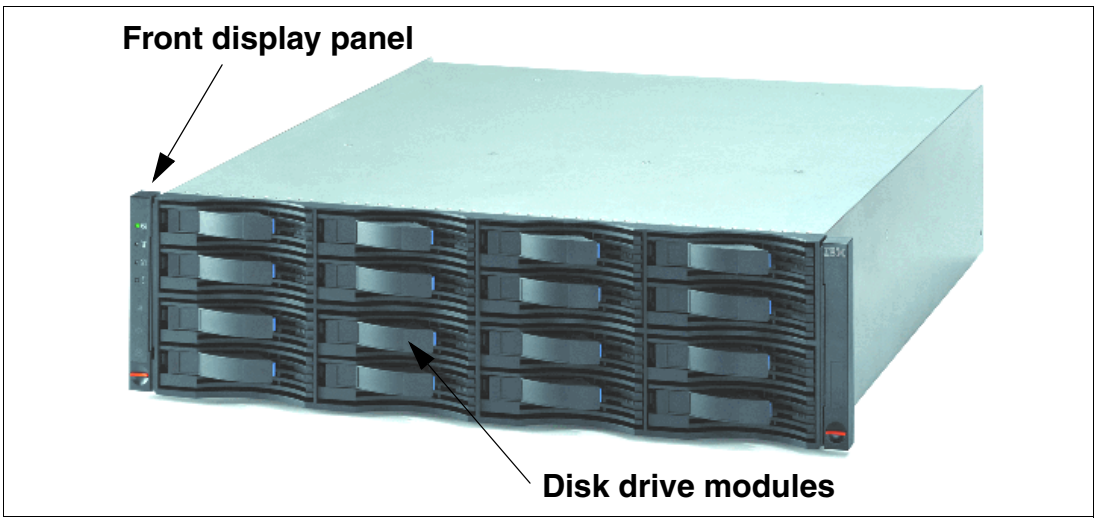



*Figure 2-1 DS6800 front view*

The front view of the DS6800 server enclosure is shown in Figure 2-1. On the left is the front display panel that provides status indicators. You can also see the disk drive modules or DDMs. Each enclosure can hold up to 16 DDMs.

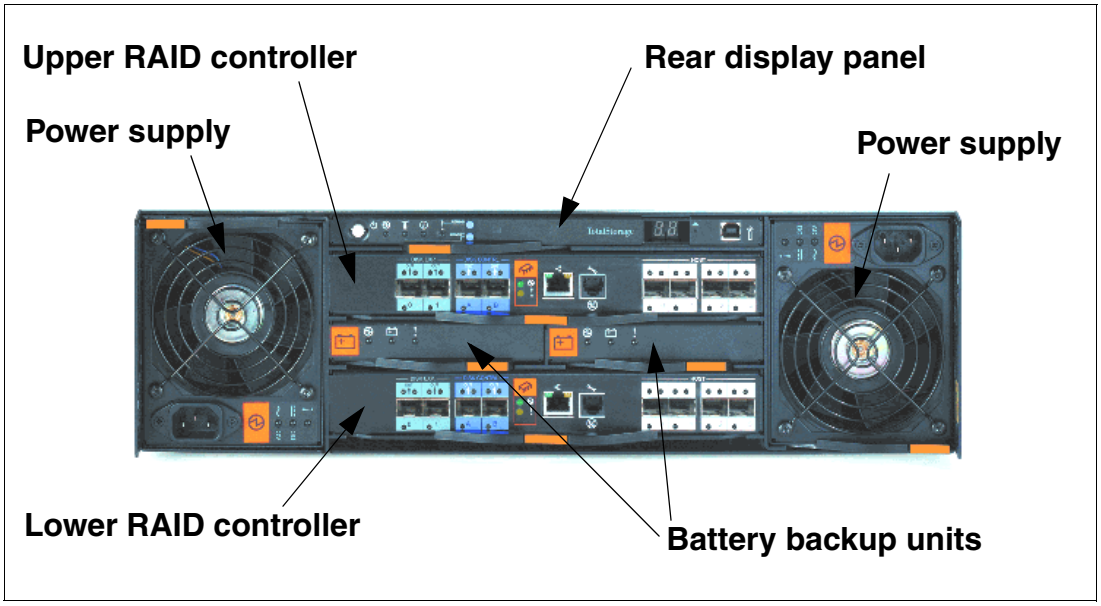



*Figure 2-2 DS6800 rear view*

The rear view of the DS6800 server enclosure is shown in Figure 2-2. You can see the left and right power supplies, the rear display panel, the upper and lower RAID controllers and the battery backup units. Each of these components is described separately later in this chapter.

## 2.2 Expansion enclosure

The DS6000 expansion enclosure is used to add capacity to an existing DS6800 server enclosure. From the front view, it is effectively identical to the server enclosure (so it is not pictured). The rear view is shown in Figure 2-3. You can see the left and right power supplies, the rear display panel, and the upper and lower SBOD (Switched Bunch Of Disks) controllers. The power supplies and rear display panel used in the expansion enclosure are identical to the server enclosure.

The rear view shows two small but important differences. First, the RAID controller cards are replaced with SBOD controller cards. Second, there are no batteries (since there is no persistent memory in the expansion enclosure). Instead the expansion enclosure has blockouts where the battery backup units were, to ensure correct airflow within the enclosure.

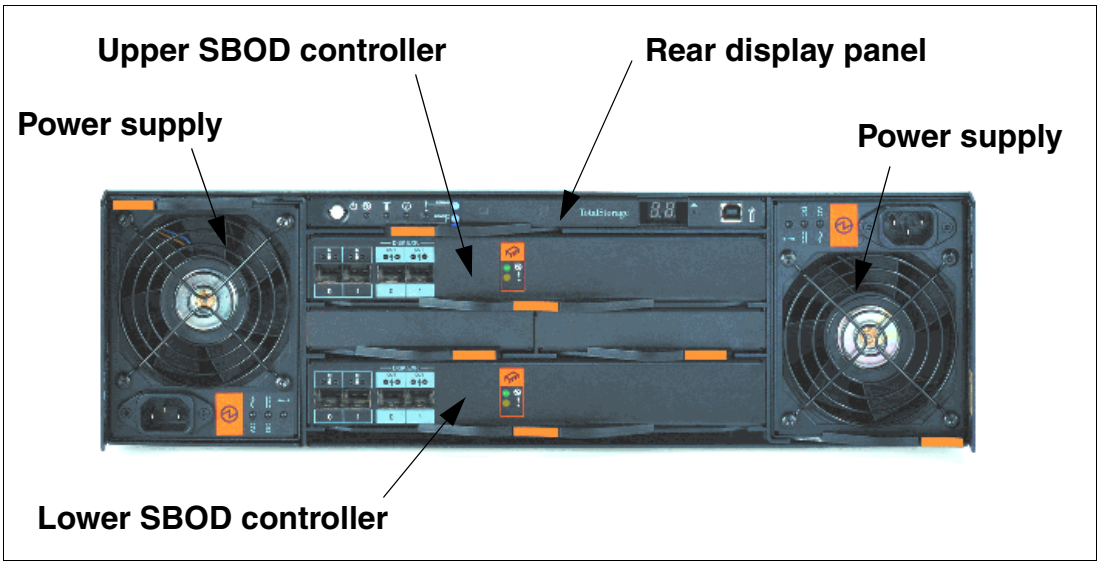



*Figure 2-3 DS6000 expansion enclosure rear view*

## 2.3 Controller architecture

Having described the enclosures themselves, the rest of the chapter explores the technical details of each of the components. The architecture that connects these components is pictured in Figure 2-4 on page 26.

Effectively the DS6800 consists of two controller cards. Each controller card contains an integrated four port host adapter to connect Fibre Channel and FICON hosts. For the disk subsystem, each controller card has an integrated four port FC-AL (Fibre Channel Arbitrated Loop) device adapter that connects the controller card to two separate Fibre Channel loops. Each switched loop attaches disk enclosures that each contain up to 16 disks. Each enclosure contains two 22 port Fibre Channel switches. Of these 22 ports, 16 are used to attach to the 16 disks in the enclosure and four are used to interconnect with other enclosures. The remaining two are reserved for internal use. 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 microcode running on a Power PC® chipset to access data on logical volumes. The microcode manages all read and write requests to the logical volumes on the disk arrays. For write I/O operations, the controllers use fast-write, whereby

the data is written to volatile memory on one controller and persistent memory on the other controller. The DS6800 then reports to the host that the write is complete before it has actually been written to disk. This provides much faster write performance. Persistent memory is also called NVS or non-volatile storage.

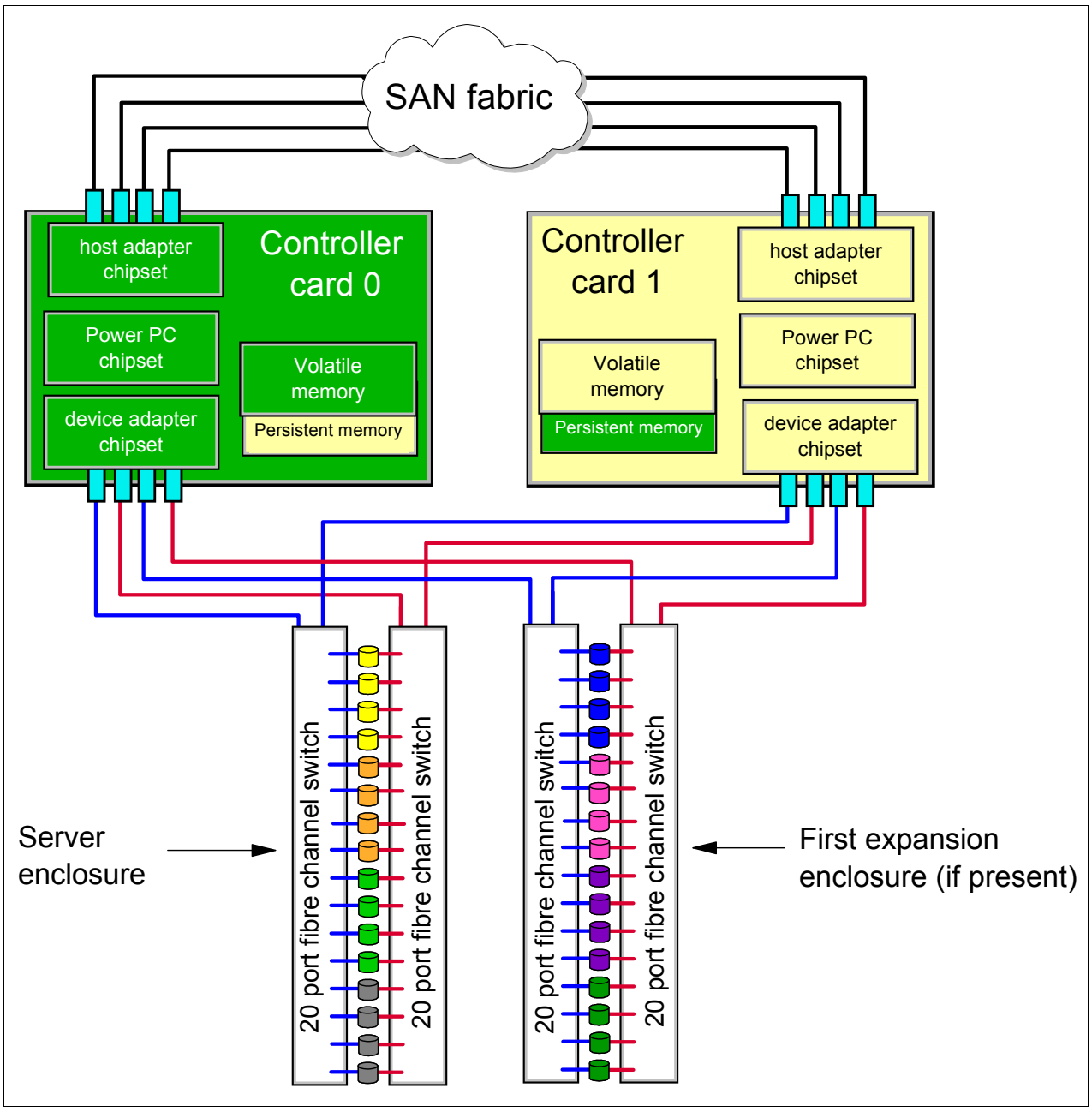



*Figure 2-4 DS6000 architecture*

When a host performs a read I/O, the controllers 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 controllers 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 read from the disks.

If you can view Figure 2-4 on page 26 in color, you can use the colors as indicators of how the DS6000 hardware is shared between the controllers (in black and white, the dark color is green and the light color is yellow). On the left side is the green controller. The green controller records its write data and caches its read data in its volatile memory area (in green). For fast-write data it has a persistent memory area on the right controller. It uses its device adapter chipset to access the disk arrays under its management. The yellow controller on the right operates in an identical fashion.

### 2.3.1 Server-based design

The DS6800 benefits from a fully assembled, leading edge processor and memory system. Using the PowerPC architecture as the primary processing engine sets the DS6800 apart from other disk storage systems on the market.

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

With the DS6800, the cache access has been accelerated further by making the non-volatile storage (NVS) a part of the main memory.

### 2.3.2 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 DS6800 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 decision to copy some amount of data into the DS6000/DS8000 cache can be triggered from two policies: demand paging and prefetching. *Demand paging* means that disk blocks 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. 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 one of the following:

- 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 2-5). 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 to not keep the same pages twice in memory.

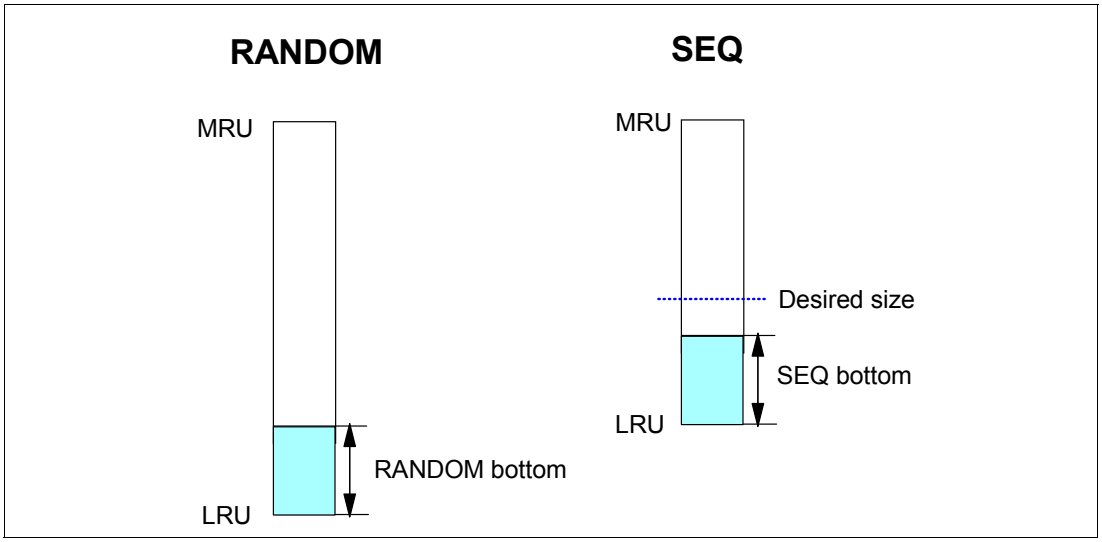



*Figure 2-5 Cache lists of the SARC algorithm for random and sequential data*

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 the 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 destaging, and the cooperation with Copy Services. In this manner, the DS6800 and DS8000 cache management goes far beyond the usual variants of the LRU/LFU (Least Recently Used / Least Frequently Used) approaches.

## 2.4 Disk subsystem

Each DS6000 storage or expansion enclosure can contain 16 DDMs or dummy carriers. A dummy carrier looks very similar to a DDM in appearance but contains no electronics. As discussed earlier, from the front, the server enclosure and the expansion enclosure appear almost identical. When identifying the DDMs, they are numbered 1 to 16 from front top left to front bottom right as depicted in Figure 2-6.

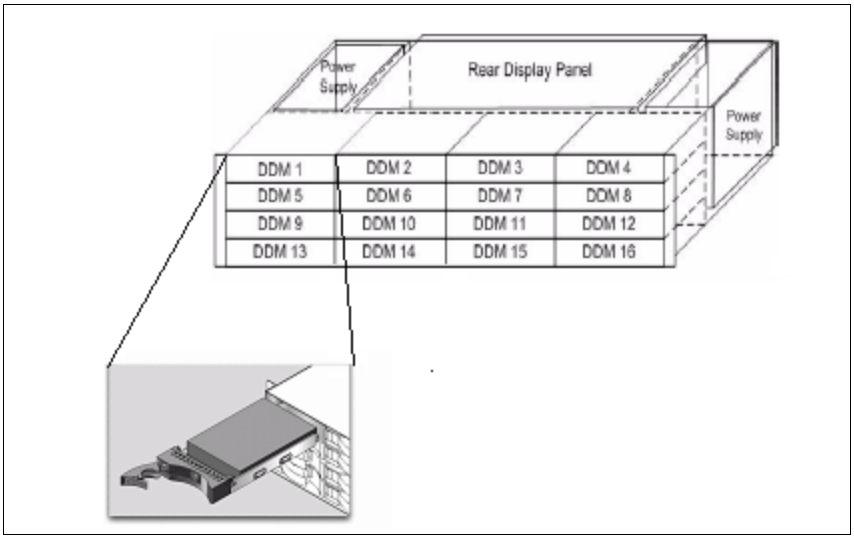



*Figure 2-6 DS6000 DDMs*

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

Each DDM is an industry standard FC-AL disk. Each disk plugs into the DS6000 midplane. The midplane is the electronic and physical backbone of the DS6000.

### 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 2-7 on page 30. This loop-based architecture means that data flows through all disks before arriving at either end of the RAID controller (shown here as *Storage Server*).

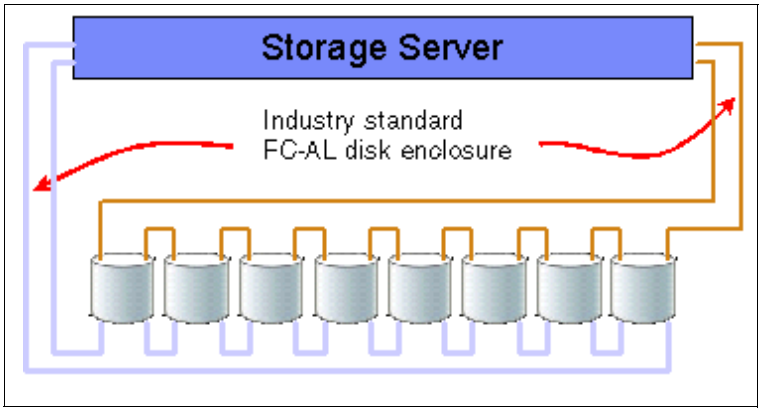



*Figure 2-7 Industry standard FC-AL disk enclosure*

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 drop off 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 DS6000.

## Switched FC-AL advantages

The DS6000 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, which provides easy problem determination
- Predictive failure statistics
- Simplified expansion: no cable rerouting required when adding another disk enclosure

The DS6000 architecture employs dual redundant switched FC-AL access to each of the disk enclosures. The key benefits of doing this are:

- Two independent switched 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 the DS6000, the switch chipset is completely integrated into the controllers. Each controller contains one switch. Note, however, that the switch chipset itself is completely separate from the controller chipset. In Figure 2-8 on page 31 each DDM is depicted as being attached to two separate Fibre Channel switches. This means that with two RAID controllers, we have four effective data paths to each disk, each path operating at 2Gb/sec.

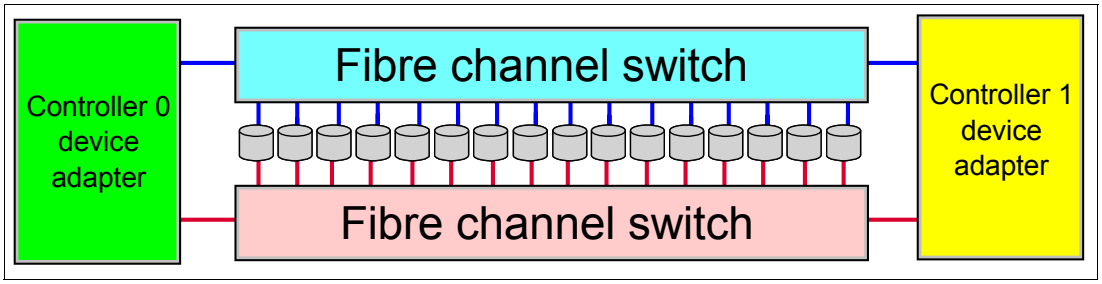



*Figure 2-8 Disk enclosure*

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 2-9 depicts four simultaneous and independent connections, one from each device adapter port.

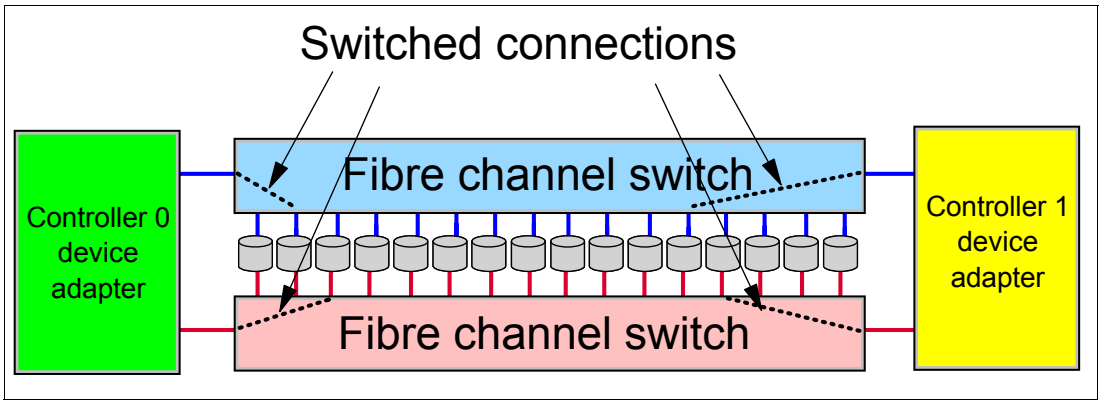



*Figure 2-9 Disk enclosure switched connections*

### DS6000 switched FC-AL implementation

For a more detailed look at how the switched disk architecture expands in the DS6000, refer to Figure 2-10 on page 32. It depicts how the DS6000 is divided into two disk loops. The server enclosure (which contains the first 16 DDMs) is on loop 0. The first expansion enclosure is placed on loop 1. This allows for the best performance since we are now using all four ports on the device adapter chipset. Expansion is achieved by adding expansion enclosures onto each loop, until each loop has four enclosures (for a total of 128 DDMs). The server enclosure is the first enclosure on loop 0, which is why we can only add a total of seven expansion enclosures.

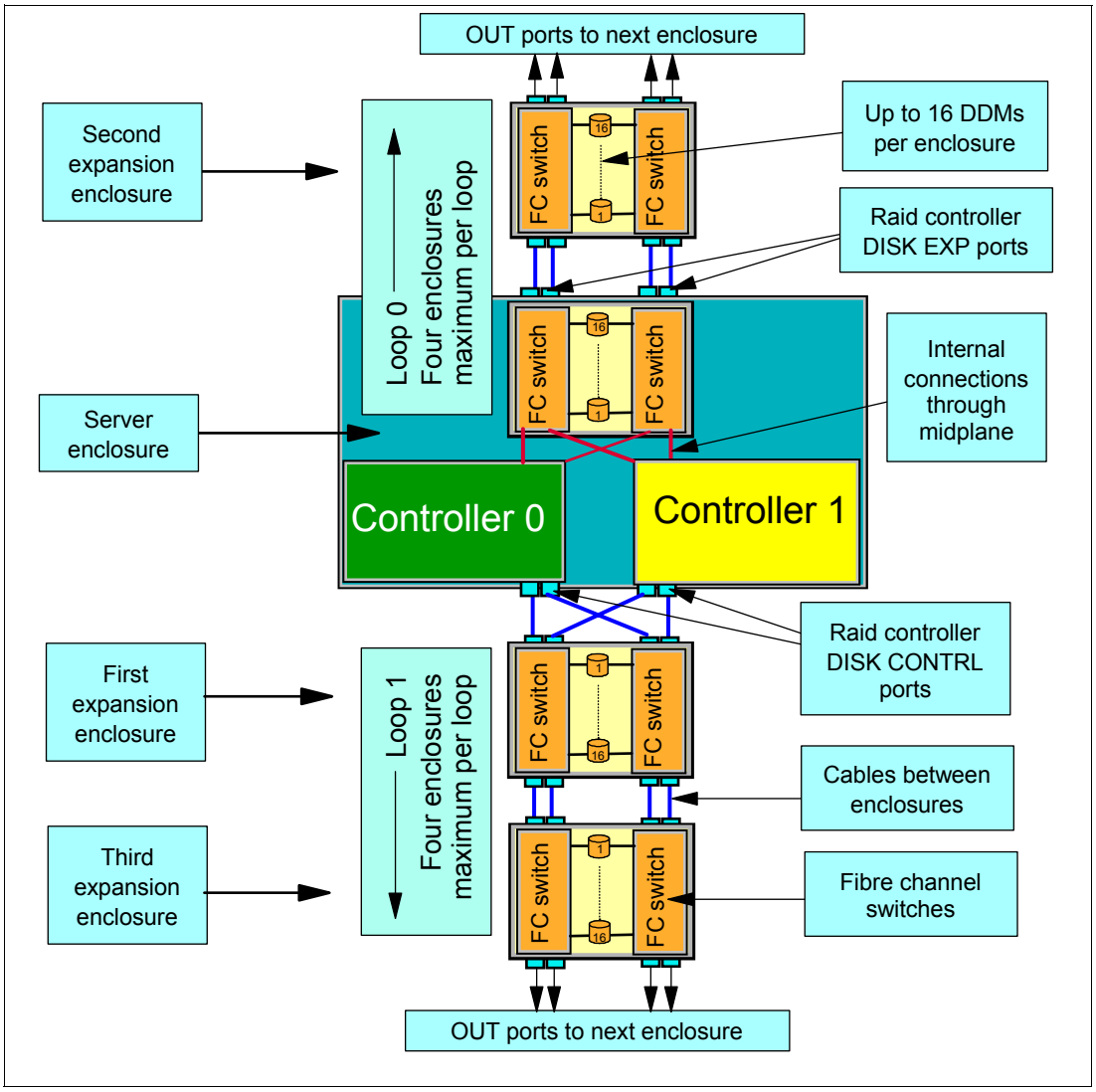



*Figure 2-10 Switched disk expansion*

## DDMs

Each DDM is hot plugable and has two indicators. The green indicator shows disk activity while the amber indicator is used with light path diagnostics to allow the customer to identify and replace a failed DDM.

At present, the DS6000 allows the choice of four different DDM types:

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

All DDMs exist in what are called array sites. Array sites containing four DDMs are created as DDMs are installed. During configuration discussed in Chapter 10, “DS CLI” on page 195, you have the choice of creating a RAID-5 or RAID-10 array by choosing one or two array sites. The first and the third array sites created on each loop each contribute one DDM to be a

spare. So at least two spares are created per loop, which will serve up to four enclosures, depending on the disk intermix.

## 2.5 Server enclosure RAID controller card

The RAID controller cards are the heart and soul of the system. Each card is the equivalent of a cluster node in an ESS. IBM has leveraged its extensive development of the ESS host adapter and device adapter function to create a total repackaging. It actually uses DS8000 host adapter and device adapter logic, which allows almost complete commonality of function and code between the two series (DS6000 and DS8000).

### 2.5.1 Technical details

From a technical point of view the controller card is powered by an IBM PowerPC 750GX 1GHz processor. The controllers do not have an internal hard drive, but instead contain a compact flash memory card to act as a boot device and to store microcode and log data. Each controller contains 2 GB of server memory, giving the DS6800 a total of 4 GB. A certain portion of that memory is reserved as persistent memory or non-volatile storage (NVS). The NVS memory is not located on a separate battery protected card like you would find in an ESS 800. It instead shares the same memory DIMMs with all the other functions. To protect the NVS memory area, a battery backup unit preserves the entire cache memory in the event of an unexpected power failure. If the DS6800 were to power off with un-destaged writes in NVS, then after reboot, the controller would read this reserved area and destage the writes. For more details on the batteries themselves and controller failover see Chapter 3, “RAS” on page 45.


*Figure 2-11 DS6800 controller card*

### 2.5.2 Device adapter ports

The DS6800 controller card is pictured in Figure 2-11. On the left-hand side, surrounded by light and dark blue boxes (for readers seeing this in black and white, they appear to be light and dark grey), are the disk expansion and disk control ports respectively. These ports are used to attach up to a total of seven expansion enclosures to the server enclosure.

The device adapter ports provided in each controller are effectively the chipset from one DS8000 device adapter. This provides remarkable performance thanks to a new high function/high performance ASIC. To ensure maximum data integrity it supports metadata creation and checking. Each controller provides four 2 Gb/sec device adapter ports, giving the machine a total of 8 device adapter ports. These ports must be short wave and use multimode cables with LC connectors.

The disks in the server enclosure are on the first disk loop (loop 0). When you attach the first expansion enclosure you attach it to the *DISK CONTRL* ports to start the second disk loop (loop 1). The *DISK EXP* ports are used to attach the second expansion enclosure. It joins the same switched loop as the disks in the server enclosure (loop 0). The two loops are depicted

in Figure 2-10 on page 32 (with loop 0 going upwards and loop 1 going in the downwards direction).

You add one expansion enclosure to each loop until both loops are populated with four enclosures each (remembering the server enclosure represents the first enclosure on the first loop). Note that while we use the term disk loops, and the disks themselves are FC-AL disks, each disk is actually attached to two separate Fibre Channel switches.

### Device adapter port indicators

For each device adapter port, there are two indicators. The top left-hand indicator is green and is used to indicate port status. The top right-hand indicator is amber and is used to show port activity.

## 2.5.3 Host adapter ports

From a host connectivity point of view, each DS6800 controller comes with four Fibre Channel/FICON host ports, giving the machine a total of eight host ports. You can see these on the right-hand side of Figure 2-11 on page 33. These host ports auto-negotiate to either 2 Gbps or 1 Gbps link speeds. These ports can be either short wave or long wave, which use multimode or single mode cables respectively, all with LC connectors.

The ports in each controller are effectively on a PCI-X 64 Bit 133 MHz card, the same card used in the DS8000. The chipset is driven by a new high function/high performance ASIC. To ensure maximum data integrity it supports metadata creation and checking.

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.

It is important to understand that an attached host must have connectivity to both controllers. This is better detailed in Chapter 3, “RAS” on page 45.

### Host adapter port indicators

For each host attachment port, there are three indicators. The top left-hand indicator is green and is used to indicate port status. The top right-hand indicator is amber and is used to show a faulty port. The bottom left-hand indicator is green and is used to indicate activity.

## 2.5.4 SFPs

The disk expansion and host attachment ports both use SFPs (which stands for small form factor plugable). These SFPs are 2 Gbps. The RAID controller card pictured in Figure 2-11 on page 33 does not have these SFPs inserted (which is why you can’t see them). These SFPs are hot plugable and are supplied as a priced feature of the DS6000.

*Figure 2-12 SFP hot-plugable fibre port with LC connector fiber cable*

### Ethernet and serial ports

Each controller card has a 10/100 copper Ethernet port to attach to a customer-supplied LAN. Both controllers must be attached to the same LAN and have connectivity to the customer-supplied PC that has the DS Storage Manager software installed on it. This port has both a status and an activity light. In addition, there is a serial port provided for each controller. This is not a modem port and is not intended to have a modem attached to it. Its main purpose is for maintenance by an IBM System Service Representative (SSR), and possibly for some initial setup tasks.

### Health indicators

Contained in an orange box, each controller card has two status indicators located below a *chip* symbol. The upper indicator is green and indicates that the controller card is powered on. The lower indicator is amber and indicates that this controller requires service.

## 2.6 Expansion enclosure SBOD controller card

The DS6000 SBOD controller card is only found in the expansion enclosure. Each SBOD controller card contains an independent 22 port Fibre Channel switch. Of these 22 ports, 16 are used to attach to the 16 disks in the expansion enclosure. Four more are used to interconnect with other enclosures, with the remaining two ports reserved for internal use. Figure 2-13 on page 36 shows the connectors on the SBOD controller card. The two *in* ports on the left are the switch ports that connect either to the server enclosure (to either the *disk exp* loop or the *disk contrl* loop) or to the *out* ports of a previous expansion enclosure. The two *out* ports on the right (in light blue boxes) are the switch ports that attach to the *in* ports of the next expansion enclosure. If there are no extra expansion enclosures then they are not used.


*Figure 2-13 DS6000 expansion enclosure SBOD controller card*

## Indicators

On the right-hand side, contained in an orange box, each SBOD controller card has two status indicators located below a *chip* symbol. The upper indicator is green and indicates that the SBOD controller card is powered on. The lower indicator is amber and indicates that this SBOD controller requires service.

## Cabling

Examples of how the expansion enclosures are cabled are shown in Figure 2-14 and Figure 2-15 on page 37.

In Figure 2-14, the server enclosure has two expansion enclosures attached to the *disk exp* loop (loop 0). The server enclosure itself is the first enclosure on loop 0. The upper controller in the server enclosure is cabled to the upper SBOD card in the expansion enclosure. The lower controller is cabled to the lower SBOD card. In each case cables run from the *disk exp* ports to the *in* ports of the SBOD card. A second expansion enclosure has been added by running cables from the *out* ports on the first expansion enclosure to the *in* ports on the second expansion enclosure. At the bottom of the diagram, dotted lines indicate the potential cabling to add more expansion enclosures to that loop.

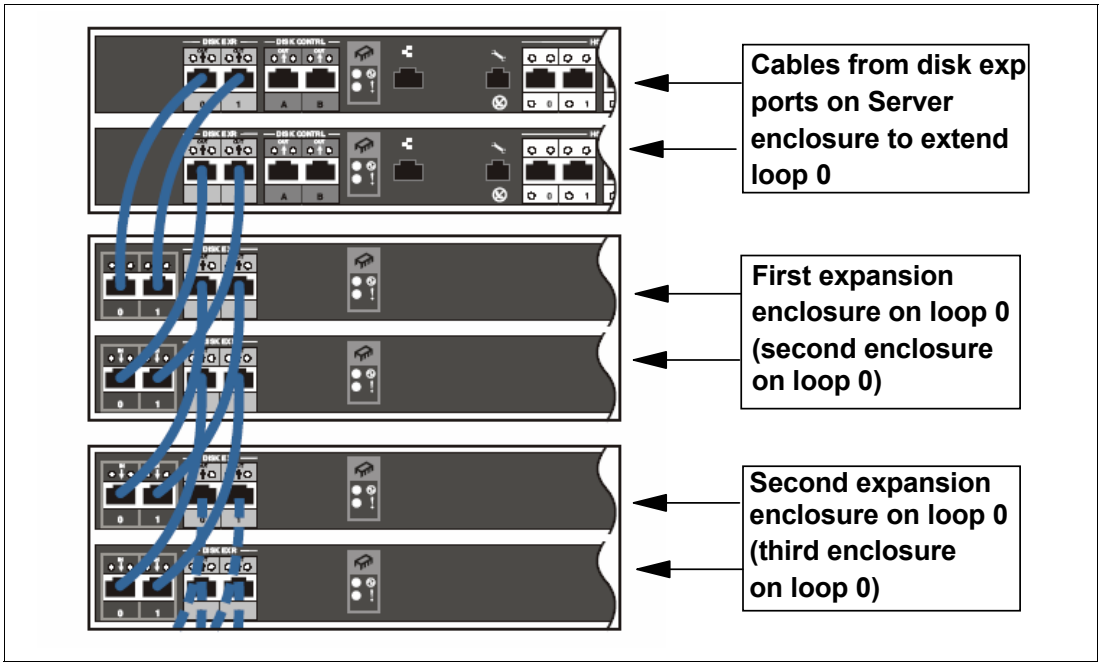



*Figure 2-14 Expansion enclosure cabling (Disk Exp ports on loop 0)*

In Figure 2-15, the server enclosure has two expansion enclosures attached to the *disk contrl* loop (loop 1). The first expansion enclosure plugged into the disk contrl ports is the first enclosure on loop 1. The cabling from the server enclosure to the first expansion enclosure on this loop is slightly different. Each controller attaches to both SBOD cards, which is why

these cables are pictured in orange and green (which appear darker if viewed in black and white). In each case cables run from the *disk contrl* ports to the *in* ports of the SBOD card. A second expansion enclosure has been added by running cables from the *out* ports on the first expansion enclosure to the *in* ports on the second expansion enclosure. At the bottom of the diagram, dotted lines indicate the potential cabling to add more expansion enclosures to that loop

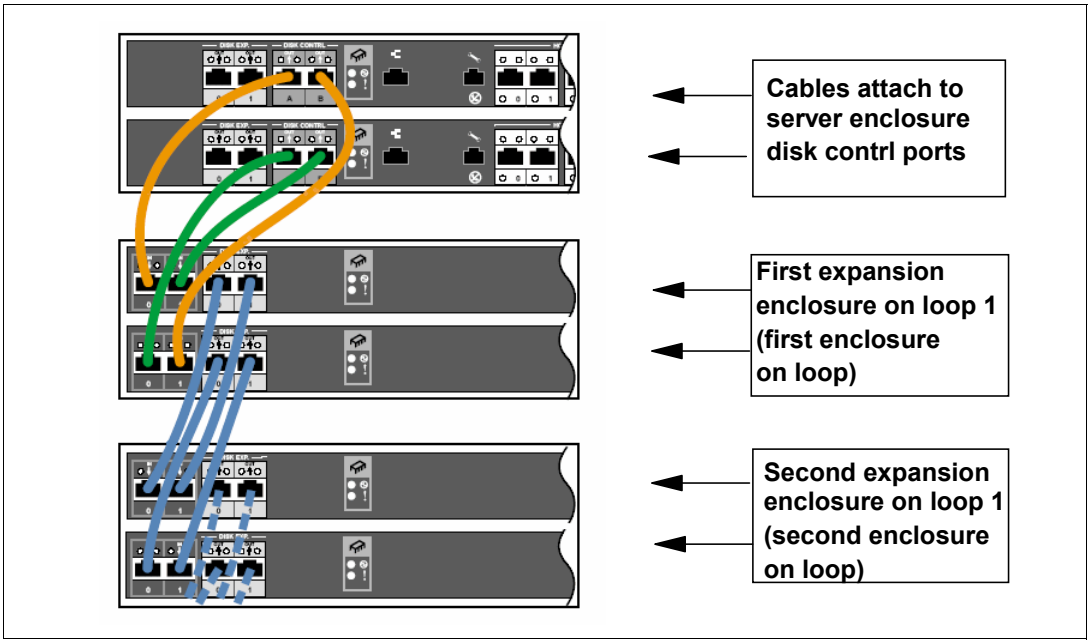



*Figure 2-15 Expansion enclosure cabling (Disk Contrl ports on loop 1)*

## 2.7 Front panel

The DS6000 front operator panel allows you to perform a health check with a single glance. There are 7 indicators present on the panel; they are depicted in Figure 2-16.

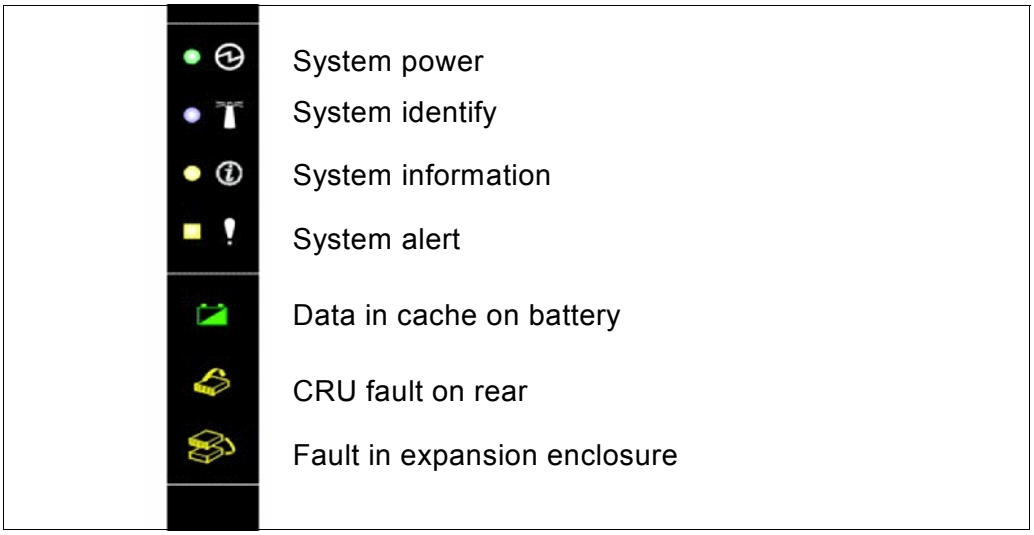



*Figure 2-16 DS6000 front panel*

Table 2-1 summarizes the purpose of each indicator.

*Table 2-1 DS6000 front panel indicators*

| Indicator | Symbol | Purpose |
|---|---|---|
| System Power (green) | Lightning bolt | If this indicator is on solid then DC power is present and the system is powered on. If it is blinking then AC Power is present but the DS6000 is not powered on. If this indicator is off then AC power is not present. |
| System Identify (blue) | Lighthouse | This indicator is normally off. It can be made to blink by pressing the lightpath identify button. It is used to identify all enclosures that are grouped together in one system. |
| System Information (amber) | Circled letter *i* | This indicator is normally off. If it is on solid, then an error has occurred that cannot be fixed by light path diagnostics. To turn this light off you need to use the GUI to correct the error condition. This may be as little as to just view the error log. |
| System Alert (amber) | Exclamation mark | This indicator is normally off. This indicator turns on solid when a fault has been detected and will remain on until the fault condition has been corrected. |
| Data Cache On Battery (green) | Battery | This indicator is normally off. If it is blinking then the battery is charging. If it is on solid then AC power has been lost and the DS6000 has data in cache being protected by battery. |
| CRU Fault on Rear (amber) | Box with arrow pointing to rear | This indicator is normally off. If it is on solid then a fault has occurred within a CRU in the rear of the enclosure and can be repaired using the light path indicators. |
| Fault in External Enclosure (amber) | Two boxes with arrow pointing to lower box. | This indicator is normally off. If it is on solid then a fault has occurred within an attached expansion enclosure. |

## 2.8 Rear panel

All of the indicators on the DS6000 front panel are mirrored to the rear panel. The rear panel is pictured in Figure 2-17 on page 39. The same colors and symbols are used for both the front and rear displays. The rear panel also has several push buttons which are detailed in Table 2-2 on page 39.

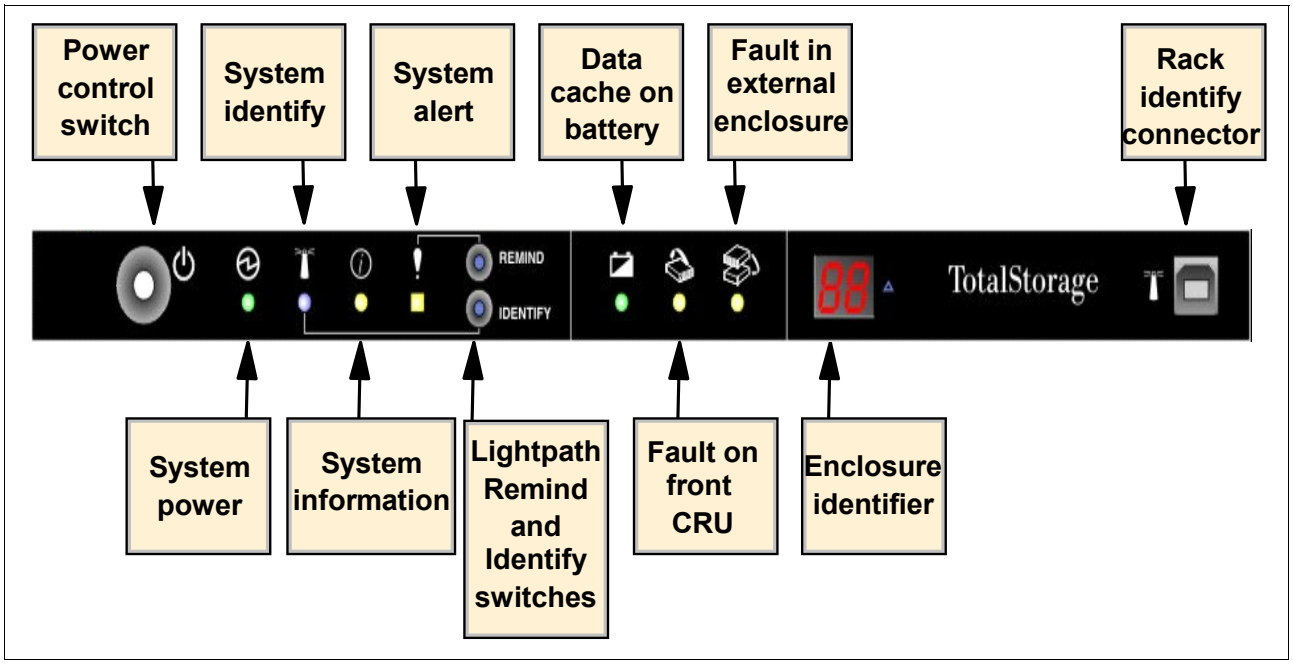



*Figure 2-17 DS6000 rear panel*

*Table 2-2 DS6000 rear panel push buttons*

| Button | Purpose |
|---|---|
| Power Control Switch (white) | This button can be seen on the left-hand side of the rear panel. You press it once to begin the power on or power off sequence. While the button is present on the expansion enclosure, only the power button on the server enclosure can power off the entire complex. |
| Lightpath *REMIND* Switch (blue) | The upper of the two blue buttons, you push this button to re-activate the light path remind. This will allow you to identify a failed component that requires replacement. |
| Lightpath *IDENTIFY* Switch (blue) | The lower of the two blue buttons, you push this button to activate the system identify indicator |

### *Enclosure ID indicator*

The rear display also has an enclosure identifier indicator. This uses two seven-segment LEDs to display the enclosure identifier number. The left-hand digit displays the device adapter switched loop the enclosure resides on. This will be 0 or 1 depending on whether the expansion enclosure is attached to the *disk exp* ports or the *disk contrl* ports respectively. On the server enclosure it will always be 0. The right-hand digit displays the enclosure base address. It will range from 0 to 3. This address will be set automatically after the enclosure is powered on and joins the loop.

### *Rack identify connector*

On the far right-hand end of the rear panel is a connector known as the rack identify connector. The intention is to allow a user to attach the enclosure to eServer rack identifier hardware. This allows you to identify in which rack a particular DS6000 storage or expansion enclosure is located.

## 2.9 Power subsystem

The power subsystem of the DS6800 consists of two redundant power supplies and two battery backup units (BBUs). DS6000 expansion enclosures contain power supplies but not BBUs. The power supplies convert input AC power to 3.3V, 5V, and 12V DC power. The battery units provide DC power, but only to the controller card memory cache in the event of a total loss of all AC power input. The DS6000 power supplies are hot swapable and a single power supply is able to support the power requirements of an entire enclosure. The second power supply is supplied by default, meaning you do not need to request redundant power as a feature.

Each power supply has two integrated fans for cooling the entire enclosure. If a power supply fails it should not be removed from the enclosure until a replacement one is available. Both power supplies must be physically present to ensure that cooling air flows correctly through the enclosure, even if one has failed. If during replacement of the failed supply, you take more than five minutes to install the new supply, the DS6000 may power off. Replacement, however, can be accomplished in less than 30 seconds.

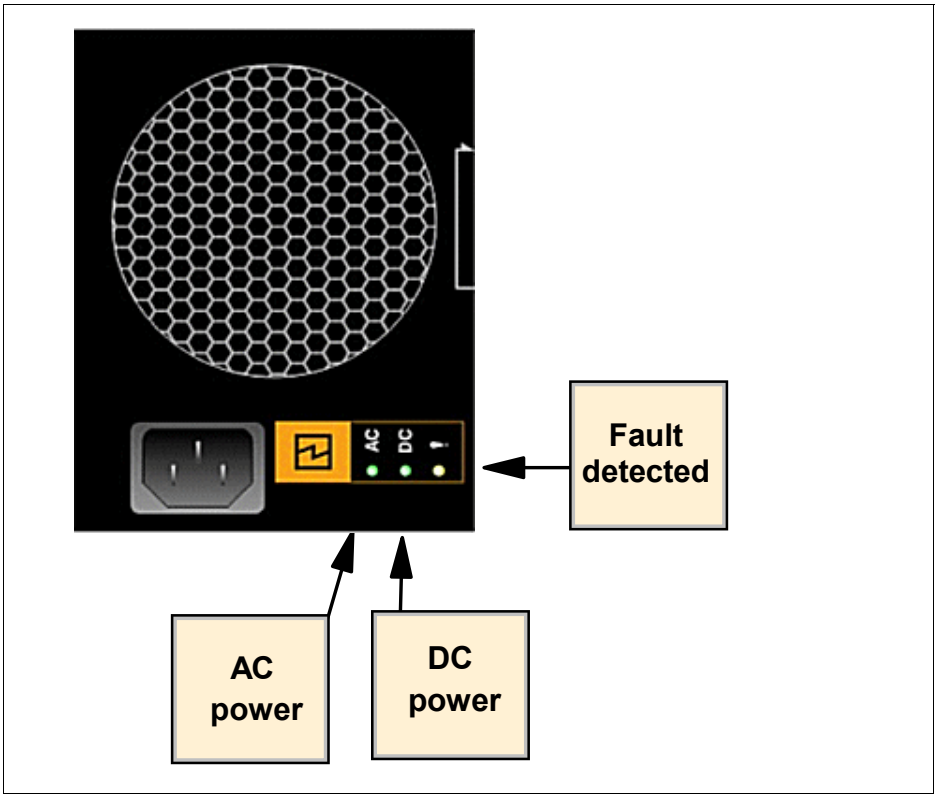



*Figure 2-18 DS6000 power supplies*

The DS6000 power supply has three indicators; they are defined in Table 2-3.

*Table 2-3 DS6000 power supply indicators*

| Indicator | Symbol | Purpose |
|---|---|---|
| AC Power (green) | AC | This indicator shows that main AC power is present. If it is off then no AC power is being supplied to the power supply, or the power supply is faulty. |

| Indicator | Symbol | Purpose |
|---|---|---|
| DC Power (green) | DC | If this indicator is on solid then the power supply is producing correct DC power. If it is blinking then the DS6000 is not powered on. If it is off then either the enclosure is powered off or the power supply is faulty. |
| Fault detected (yellow) | Exclamation mark | If this indicator is on solid then the DS6000 has identified this power supply as being faulty and it requires replacement. |

### 2.9.1 Battery backup units

Each DS6800 RAID controller has a battery backup unit (BBU) to provide DC power to that controller in the event of a complete loss of power. There are thus two BBUs present in the DS6800 server enclosure. If you compare their function to that of the different batteries in the ESS, they are the NVS batteries. They allow un-destaged cache writes in the NVS area of controller memory to be protected in the event of a sudden loss of AC power to both power supplies. The BBUs will protect the contents of NVS for at least 72 hours.

From the rear of the unit, the left-hand BBU supports the upper controller, while the right-hand BBU supports the lower controller.

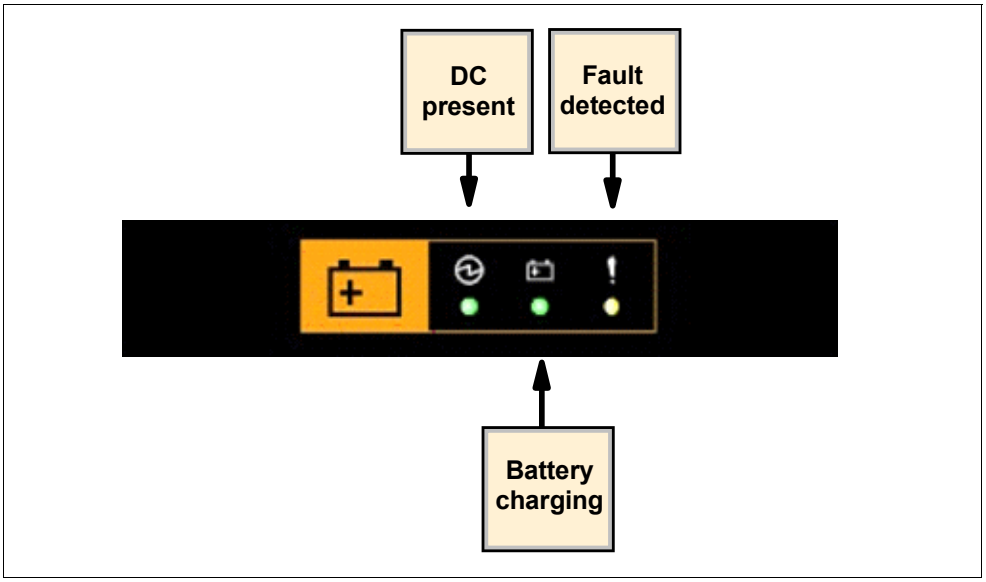



*Figure 2-19 DS6000 battery backup unit indicators*

The DS6000 battery units have three indicators, detailed in Table 2-4.

*Table 2-4 DS6800 battery backup unit indicators*

| Indicator | Symbol | Purpose |
|---|---|---|
| DC present (green) | Lightning bolt | If this indicator is on solid then DC power is present. If this indicator is off then DC power is not available from this battery. |

| Indicator | Symbol | Purpose |
|---|---|---|
| Battery charging (green) | Battery symbol | If this indicator is on solid then the battery backup unit is fully charged. If this indicator is blinking then the battery is charging. As the battery approaches full charge the blink speed will slow. If the indicator is off then the battery is not operational. |
| Fault detected (yellow) | Exclamation mark | If this indicator is on solid then a fault has been detected and this battery requires service. |

## 2.10 System service card

The system service card which ships with the DS6800 can be placed in a cavity below the lower RAID controller card. It is a plastic card that contains important information on how to maintain your DS6800.

## 2.11 Storage Manager console

The DS6800 requires a separate PC to act as the DS Storage Manager console. This PC is not a feature of the DS6800 and must be ordered separately. The hardware and software requirements for this PC can be found in Chapter 8, “Configuration planning” on page 125.

## 2.12 Cables

The DS6800 ships with three different sorts of cables:

- **Power cords**

  Each DS6000 enclosure ships with six power cords. These should allow attachment to both rack power outlets and standard power outlets. The rack power outlet cords come in two lengths. The *standard* power cords are specified by feature code at time of purchase (based on the outlet used in your country). Only two of the six cords are used.

- **Ethernet cables**

  Each DS6800 server enclosure ships with one Ethernet cross-over cable. This cable is used during initial configuration of the controllers to set IP addresses. After this, it is not used. You do not use it to connect the two controllers together for normal operation or during initial power on (this connectivity is supplied by an Ethernet switch).

  Each DS6800 server enclosure ships with two standard Ethernet cables. These should be used in conjunction with an Ethernet hub or switch (that will need to be supplied or ordered separately) so that the controllers are able to communicate with each other and the DS Storage Management Console. This connectivity is for configuration and regular maintenance.

- **Service cable**

  Each DS6800 server enclosure ships with a special service cable and DB9 converter. This cable looks very similar to a telephone cable. These components should be kept aside for use by an IBM System Service Representative and will normally be used only for problem debug.

## 2.13 Summary

This chapter has described the various components that make up a DS6000. For additional information, there is documentation available on the Web at:

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

3

# RAS

This chapter describes the RAS (reliability, availability and serviceability) characteristics of the DS6000 series. Specific topics covered are:

- Controller RAS
- Host connection availability
- Disk subsystem RAS
- Power subsystem RAS
- Light path guidance strategy
- Microcode updates

# 3.1 Controller RAS

The DS6800 design is built upon IBM's highly redundant storage architecture. It has the benefit of more than five years of ESS 2105 development. The DS6800, therefore, employs similar methodology to the ESS to provide data integrity when performing fast write operations and controller failover.

## 3.1.1 Failover and failback

To understand the process of controller failover and failback, we have to understand the logical construction of the DS6800. To better understand the contents of this section you may want to refer to Chapter 10, "DS CLI" on page 195. In short, to create logical volumes on the DS6000, 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 controller 0 or controller 1.

Within each extent pool we create logical volumes (which for open systems are called LUNs and for zSeries, 3390s). These logical volumes belong 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 z/Series disk control unit which the DS6800 emulates). What is important, is that LSSs that have an even identifying number have an affinity with controller 0, while LSSs that have an odd identifying number have an affinity with controller 1.

When a host operating system issues a write to a logical volume, it is preferable that it is issued to the controller that *owns* the LSS of which that logical volume is a member. Understanding this controller affinity is important for achieving the best performance and it is also very important when we look at host pathing. More details are in 3.2, "Host connection availability" on page 49.

### Data flow

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

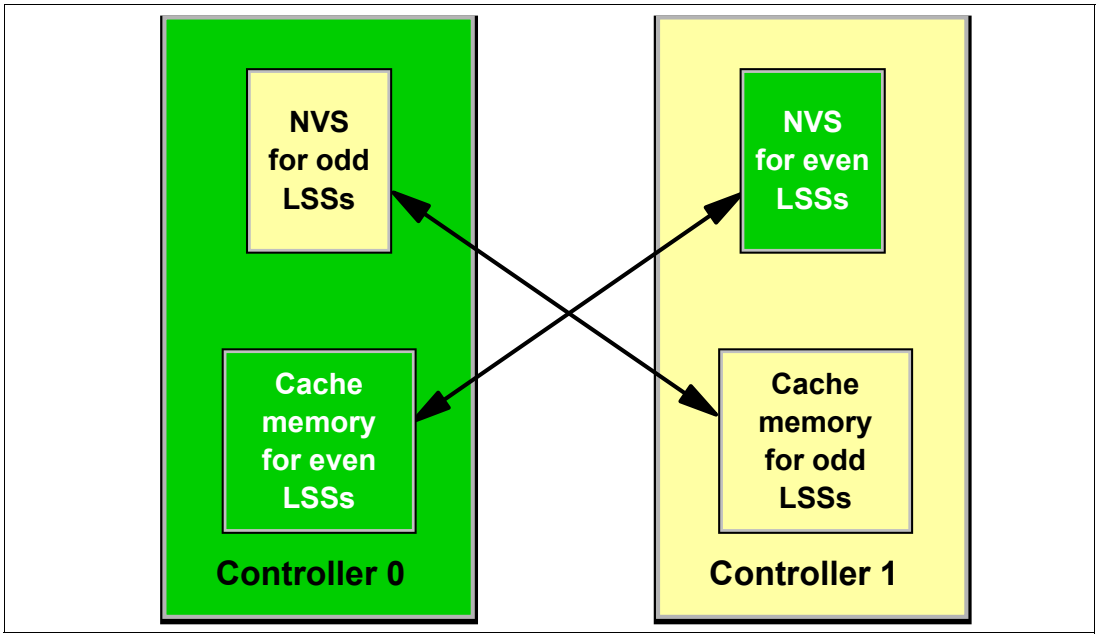



*Figure 3-1 DS6800 normal data flow*

Figure 3-1 illustrates how the cache memory of controller 0 is used for all logical volumes that are members of the even LSSs. Likewise, the cache memory of controller 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 opposite controller. So the normal flow of data for a write is:

1. Data is written to cache memory in the owning controller.
2. Data is written to NVS memory of the alternate controller.
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 then discarded from the NVS memory of the alternate controller.

Under normal operation, both DS6800 controllers are actively processing I/O requests. This section describes the failover and failback procedures that occur between the DS6800 controllers when an abnormal condition has affected one of them.

## Failover

In the example depicted in Figure 3-2 on page 48, controller 0 in the DS6800 has failed. The remaining controller has to take over all of its functions. The host adapters located in controller 0 are now no longer available. All the RAID arrays in the DS6800 will be accessed from the device adapter in controller 1. First, controller 1 has to process the data it is holding in NVS. It then starts operating the entire machine in single controller mode. The steps it takes are:

1. It de-stages the contents of its NVS to disk.
2. The NVS and cache of controller 1 are divided in two, half for the odd LSSs and half for the even LSSs.
3. Controller 1 now begins processing the writes (and reads) for *all* the LSSs.

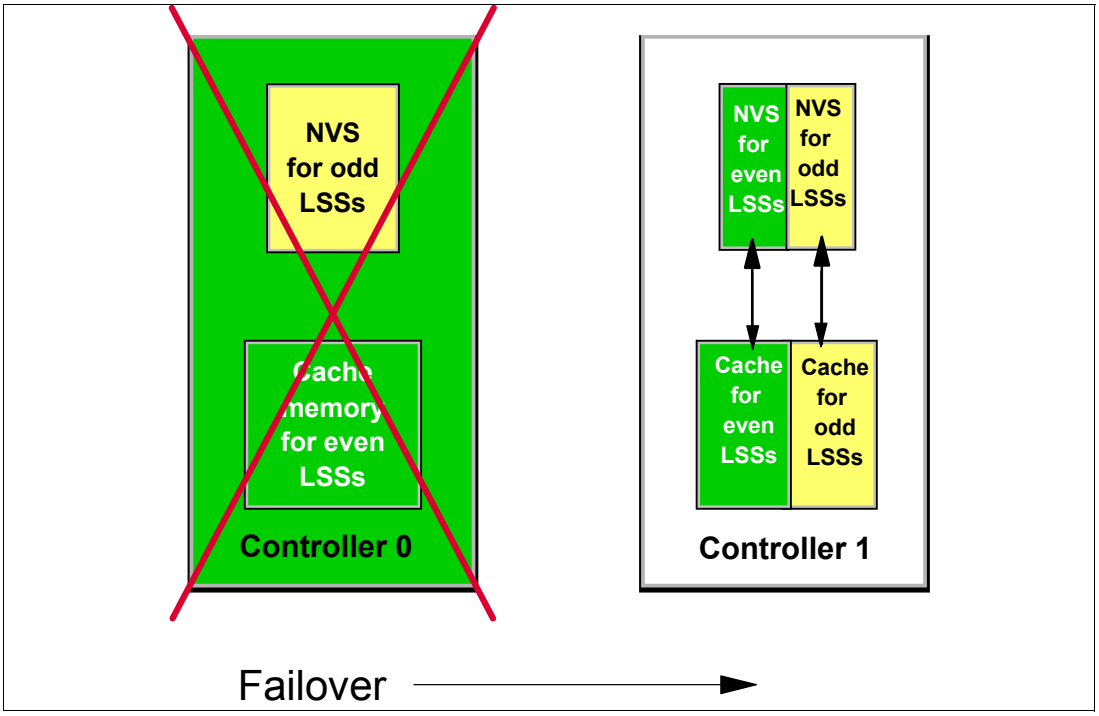



*Figure 3-2 Controller failover*

This entire process is known as a failover. After failover, controller 1 now owns all the LSSs, which means all reads and writes will be serviced by controller 1. The NVS inside controller 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 controller has been repaired and restarted, the failback process is activated. Controller 1 starts using the NVS in controller 0 again, and the ownership of the even LSSs is transferred back to controller 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 DS6000 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.

## 3.1.2 NVS recovery after complete power loss

During normal operation, the DS6800 preserves un-destaged write data using the NVS copy in the alternate controller. To ensure that these writes are not lost, each controller has a dedicated battery backup unit (BBU). If this BBU were to fail, the controller would lose this protection and consequently that controller would remove itself from service. If power is lost to a single power supply, this does not affect the ability of the other power supply to keep both BBUs charged, so both controllers would remain online. In other words, there is an affinity between controllers and BBUs, but not between power supplies and BBUs.

The single purpose of the BBUs is to preserve the NVS area of controller memory in the event of a complete loss of input power to the DS6800. If both power supplies were to stop receiving

input power, the DS6800 controller cards would detect that they were now running on batteries and immediately shut down. The BBUs are not sufficient to keep the disks spinning so there is nowhere to put the modified data. All that the BBUs will do is preserve all data in memory while input power is not available. When power becomes available again, the DS6800 controllers begin the bootup process, but leave the NVS portion of controller memory untouched. During the initialization process, the NVS data area is examined and if any un-destaged write data is found, it is destaged to disk prior to the controllers coming online.

The BBUs are capable of preserving the contents of controller memory for at least 72 hours, and possibly much longer. If the DS6800 unexpectedly powers off while processing host I/Os, and is then left without power for more than 72 hours, then any un-destaged writes in NVS may be permanently lost. Since we do not know which tracks were in NVS at the time of the power failure, all data on the DS6000 would have to be considered as suspect, and data integrity checking at the least—and data recovery at worst—may be necessary.

The DS6800 BBUs are designed to be replaced every four years.

From the rear of the server enclosure, the left-hand BBU supports the upper controller, while the right hand BBU supports the lower controller.

### 3.1.3 Metadata checks

When application data enters the DS6000, 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 DS6800. 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 DS6800 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.

## 3.2 Host connection availability

Each DS6800 controller card contains four Fibre Channel ports, for connection either directly to a host or to a Fibre Channel SAN switch. This gives the DS6800 a total of eight ports for host connections.

#### Single and preferred path

Unlike the DS8000 or the ESS 800, the DS6800 uses the concept of *preferred path*, since the host adapters are integrated into the controller hardware rather than in separate I/O bays. What this means is that the attached host systems must be aware that it is preferential to direct I/O to a particular controller. If an I/O request for a particular LUN is delivered to a host adapter located in the non-owning controller, that controller will use an internal data bus to route the request to the owning controller. This re-route of the I/O request has a performance cost but does not affect the reliability or availability of the DS6800.

If a host were to only have a single path to a DS6800, as depicted in Figure 3-3 on page 50, then it would still be able to access volumes belonging to all LSSs, but I/O for odd LSS volumes would use the internal data path between the controllers. However, if controller 0 were to have a hardware failure, then all connectivity would be lost. Within the figure itself, an HP is a host port (a fibre port located in the DS6800 controller card), a DA is two device adapter ports (also located on the DS6800 controller card) and an HBA is a host bus adapter (a Fibre Channel card located in the attached host).

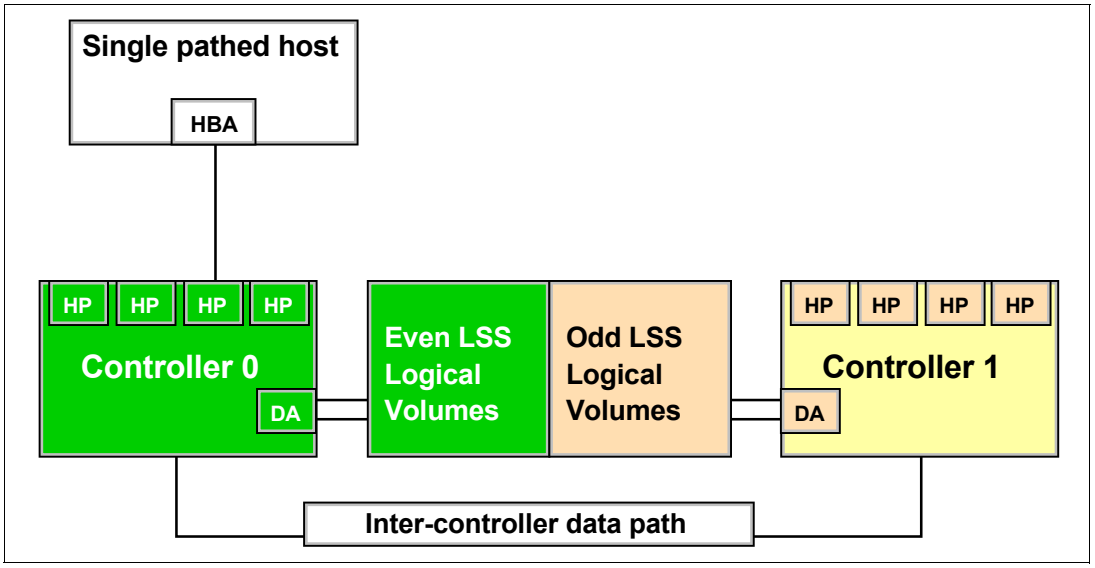



*Figure 3-3 A host with a single path to the DS6800*

For best reliability and performance, it is recommended that each attached host has two connections, one to each controller as depicted in Figure 3-4. This allows it to maintain connection to the DS6800 through both controller failure and HBA or HA (host adapter) failure.

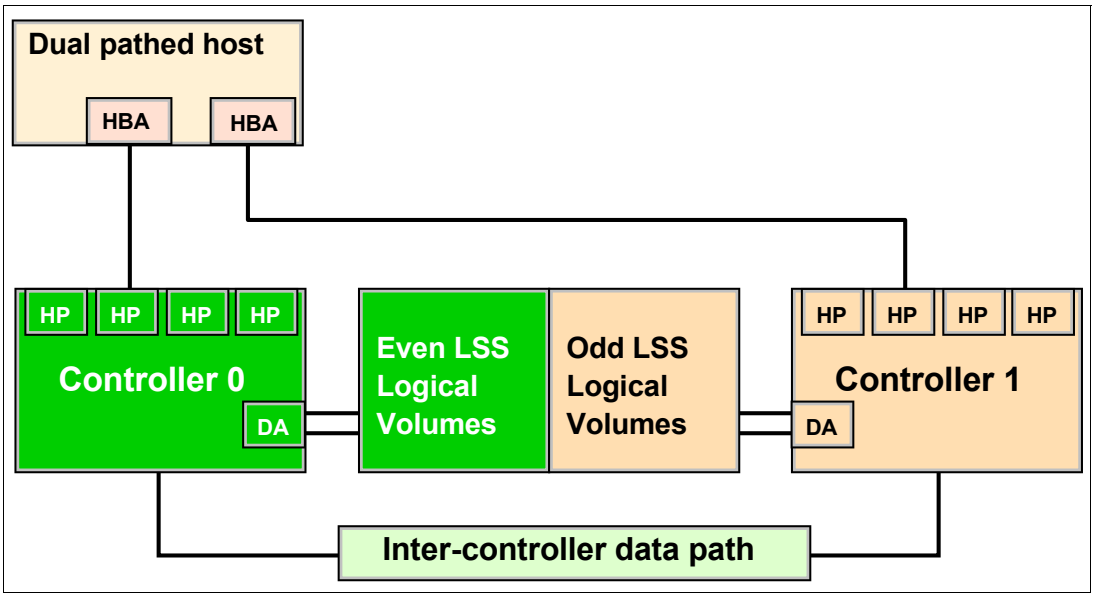



*Figure 3-4 A host with two paths to the DS6800*

## SAN/FICON switches

Because a large number of hosts may be connected to the DS6800, each using multiple paths, the eight host adapter ports that are available in the DS6800 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.

A logic or power failure in a SAN switch can interrupt communication between hosts and the DS6800. We recommend that more than one SAN switch be provided to ensure continued availability. For example, four of the eight fibre ports in a DS6800 could be configured to go through each of two directors. The complete failure of either director leaves half of the paths still operating.

### Multipathing software

For each attached host we now require a mechanism to allow the attached operating system to manage multiple paths to the same device, and to also show a preference in this routing so that I/O requests for each LUN go to the preferred controller. Also, when a controller failover occurs, attached hosts that were routing all I/O for a particular group of LUNs (LUNs on either even or odd LSSs) to a particular controller (because it was the preferred controller) must have a mechanism to allow them to detect that the preferred path is gone. It should then be able to re-route all I/O for those LUNs to the alternate, previously non-preferred controller. Finally, it should be able to detect when a controller comes back online so that I/O can now be directed back to the preferred controller on a LUN by LUN basis (determined by which LSS a LUN is a member of). The mechanism that will be used varies by the attached host operating system, as detailed in the next two sections.

## 3.2.1 Open systems host connection

In the majority of open systems environments, IBM recommends the use of the Subsystem Device Driver (SDD) to manage both path failover and preferred path determination. SDD is supplied free of charge to all IBM customers who use ESS 2105, SAN Volume Controller (SVC), DS6800, or DS8000. A new version of SDD (Version 1.6) will also allow SDD to manage pathing to the DS6800 and DS8000.

SDD provides availability through automatic I/O path failover. If a failure occurs in the data path between the host and the DS6800, SDD automatically switches the I/O to another path. SDD will also 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.

SDD is not available for all supported operating systems, so attention should be directed to the *IBM TotalStorage DS6000 Host Systems Attachment Guide*, GC26-7680, and the interoperability Web site for direction as to which multi-pathing software will 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 and preferred path. The interoperability Web site is located at:

http://www.ibm.com/servers/storage/disk/ds6000/interop.html

## 3.2.2 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 DS6800 are defined in the HCD (or IOCDS) 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 DS6800 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 DS6800.

A physical FICON path is established when the DS6800 port sees light on the FICON fiber (for example, a cable is plugged in to a DS6800 host adapter, or a processor, or the DS6800 is powered on, or a path is configured online by OS/390). At this time, logical paths are established through the FICON port between the host and some or all of the LCUs in the DS6800, controlled by the HCD definition for that host. This happens for each physical path between a zSeries CPU and the DS6800. There may be multiple system images in a CPU. Logical paths are established for each system image. The DS6800 then knows which FICON paths can be used to communicate between each LCU and each host.

Provided you have the correct maintenance level, all major zSeries operating systems should support preferred path (z/OS, z/VM, VSE/ESA™, TPF).

### 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 DS6800 service actions.

Control Unit Initiated Reconfiguration is available for the DS6800 when operating 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 DS6800 service actions.

CUIR allows the DS6800 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 DS6800 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 DS6800.

**Note:** CUIR support will be included in a future release of microcode.

## 3.3 Disk subsystem RAS

The DS6000 currently supports only RAID-5 and RAID-10. It does not support non-RAID configurations of disks (JBOD - just a bunch of disks).

### 3.3.1 RAID-5 overview

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

#### RAID-5 theory

The DS6000 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 DS6000

In a DS6000, a RAID-5 array built on one array site will contain either three disks or four disks, depending on whether the array site chosen had a pre-allocated spare. A three disk array effectively uses 1 disk for parity, so it is referred to as a 2+P array (where the P stands for parity). The reason only three disks are available to a 2+P array is that the fourth disk in the array site used to build the array, was used as a spare. This can be referred to as a 2+P+S

array site (where the S stands for spare). A four disk array also effectively uses 1 disk for parity, so it is referred to as a 3+P array.

In a DS6000, a RAID-5 array built on two array sites will contain either seven disks or eight disks, again depending on whether the array sites chosen had pre-allocated spares. A seven disk array effectively uses one disk for parity, so it is referred to as a 6+P array. The reason only 7 disks are available to a 6+P array is that the eighth disk in the two array sites used to build an array, was already a spare. This is referred to as a 6+P+S array site. 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 (DDM) 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 is 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 controller and switched network resources are being used to do the reconstruction. Due to the switched architecture, this effect will be minimal. Additionally, any read requests for data on the failed drive require data to be read from the other drives in the array to reconstruct the data. The remaining requests are satisfied by reading the drive containing the data in the normal way.

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 on the workload on the array and the controller.

## 3.3.2 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-10 array. The same data is also striped across the other half of the array, creating a mirror. Access to data is usually preserved, even if multiple disks fail. 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 DS6000

In the DS6000 the RAID-10 implementation is achieved by using one or two array sites (either four or eight DDMs). If a single array site array is created and that site includes one spare, then only two DDMs will be available for this array. This makes the array a 1+1 array that is effectively just RAID-1. The other two DDMs will both be spares. If an array site with no spares is selected then the array will be 2+2.

If two array sites are used to make a RAID-10 array and the array sites contain spares, then six DDMs are used to make two RAID-0 arrays which are mirrored. If spares do not exist on the array sites then eight DDMs are used to make two RAID-0 arrays which are 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 spare drives. The spare that is used is 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 controller 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 controller and switched network resources are being used to do the reconstruction. Due to the switched architecture of the DS6000, 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-0 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 on the workload on the array and the controller.

## 3.3.3 Spare creation

There are four array sites in each enclosure of the DS6000. The first and third array sites created on each loop are used to supply spares. This normally means that two spares will be created in the server enclosure and two spares in the first expansion enclosure. Spares are created as the array sites are created, which occurs when the DDMs are installed. After four spares have been created for the entire storage unit, no more spares are normally needed.

On the ESS 800 the spare creation policy was to have four DDMs on each SSA (Serial Storage Architecture) 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 types. With the DS6000 the intention is to not do this. Where DDMs with different sizes, but the same RPM, exist in the complex, the spares will be taken from the array sites with the larger sized DDMs. This means in most cases the DS6000 will continue to have only four spares for the entire complex regardless of DDM size intermix.

### Floating spares

The DS6000 implements a smart floating technique for spare DDMs. When a spare *floats*, this means that when a DDM fails and the data it contained is rebuilt onto a spare, then the disk is replaced, the replacement disk becomes the spare. The data is not copied back to the original position which the failed DDM occupied. The DS6000 microcode may choose to allow the hot spare to remain where it has been *moved*, but it may instead choose to move the spare to a more optimum position. This will be done to better balance the spares across the DA pairs and 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 DS6000 until the migration process is complete.

A smart process may be used to ensure that the larger or higher RPM DDMs act as spares. This is preferable because if we were to rebuild the contents of a 73 GB DDM onto a 146 GB

DDM, then approximately half of the 146 GB DDM would be wasted since that space is not needed. The problem here is that the failed 73 GB DDM will be replaced with a new 73 GB DDM. So the DS6000 microcode will most likely migrate the data on the 146 GB DDM onto the recently replaced 73 GB DDM. When this process completes, the 73 GB DDM will rejoin the array and the 146 GB will become the spare again. Another example would be if we fail a 10k RPM DDM onto a 15k RPM DDM. While this means that the data has now moved to a faster DDM, the replacement DDM will be the same as the failed DDM. This means the spare will now be a 10k RPM DDM. This could result in a 15k RPM DDM being spared onto a 10k RPM DDM. This is not desirable. Again a smart failback of the spare will be performed once a suitable replacement DDM has been made available.

#### Hot plugable DDMs

Replacement of a failed drive does not affect the operation of the DS6000 because the drives are fully hot plugable. 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.

### 3.3.4 Predictive Failure Analysis (PFA)

The drives used in the DS6000 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-5 or RAID-10 recovery to reconstruct all of the data onto the spare drive. The DS6000 will alert the user and can also use call home e-mail notification.

### 3.3.5 Disk scrubbing

The DS6000 will periodically read all sectors on a disk. This is designed to occur without any interference to application performance. If ECC-correctable bad bits are identified, the bits are corrected immediately by the DS6000. 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 on the disk. The scrubbing process applies to both array members and spare DDMs.

### 3.3.6 Disk path redundancy

Each DDM in the DS6000 is attached to two 22 port SAN switches. These switches are built into the RAID or SBOD controller cards. Figure 3-5 on page 56 depicts the redundancy features of the DS6000 switched disk architecture. Each disk has two separate connections to the midplane. This allows it to be simultaneously attached to both switches. If either a RAID or SBOD controller card is removed from an enclosure, the switch that is included in that controller is also removed. However, the remaining controller retains the ability to communicate with all the disks via the remaining switch.

Figure 3-5 also shows the connection paths to the expansion enclosures. To the left and right you can see paths from the switches and Fibre Channel chipset that travel to the device adapter ports at top left and top right. These ports are depicted in Figure 2-2 on page 24. From each controller we have two paths to each expansion enclosure. This means that we can easily survive the loss of a single path (which would mean the loss of one out of four paths) due to the failure of, for instance, a cable or an optical port. We can also survive the loss of an entire RAID controller or SBOD controller (which would remove two out of four

paths) since two paths to the expansion controller would be available for the remaining controller.

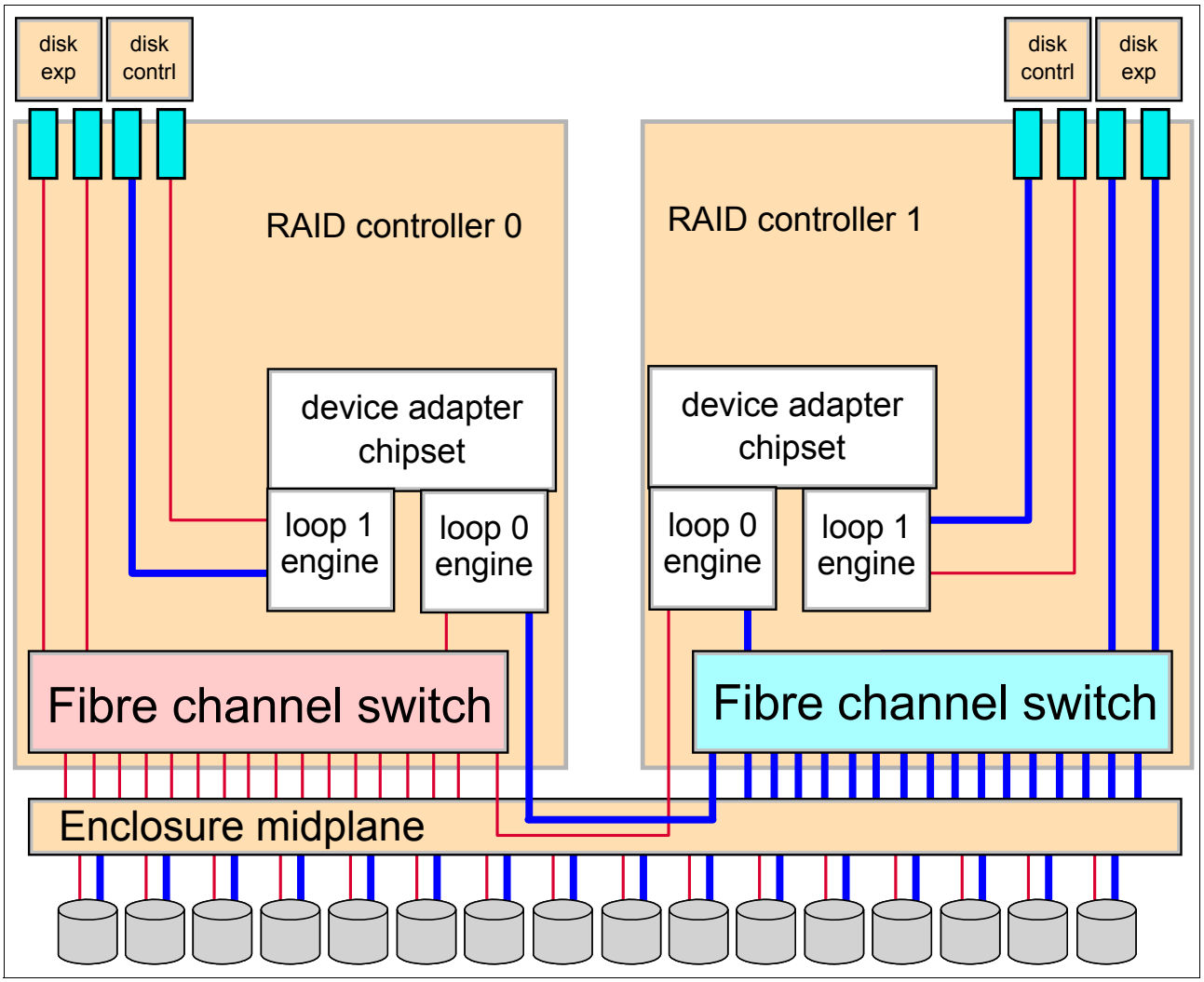



*Figure 3-5 DS6000 switched disk connections*

## 3.4 Power subsystem RAS

As discussed in Chapter 2, “Components” on page 23, the DS6000 is equipped with two BBUs and two power supplies. This provides redundancy in case of either an external power failure or an internal power subsystem failure. The DS6000 is able to control the state of the power supplies in the expansion enclosures via in-band commands sent through the device adapter Fibre Channel connections.

In the event of a BBU failure, the RAID controller that relies on that BBU for data protection will remove itself from service and go offline until its BBU is fully charged. If both BBUs were to fail, then the entire system would have to go offline until the problem is corrected. This possibility is highly unlikely.

All power components are hot plugable and can usually be replaced without employing the DS Storage Manager GUI. If more information is needed, however, the GUI could be employed, as described in “Example 2: Using the GUI to replace a power supply” on page 57.

**Important:** If you install the DS6000 so that both power supplies are attached to the same power strip, or where two power strips are used but they connect to the same circuit breaker or the same switch-board, then the DS6000 will not be well protected from external power failures. This is a very common cause of unplanned outages.

### Redundant cooling

The DS6000 gets its cooling from the two fan assemblies in each power supply. Provided one power supply is working and the other power supply is physically inserted, sufficient cooling will be available for that DS6000 enclosure. The DS6000 microcode can modify the speed of each fan in response to environmental conditions or the failure of a single power supply.

**Important:** If any component fails, it should not be removed from the enclosure until a replacement part is available. Power supplies in particular must be physically present to ensure that cooling air flows correctly through the enclosure. Replacement of a failed supply can be accomplished in less than 30 seconds. If you remove the failed power supply and do not insert a replacement within five minutes, the DS6000 may overheat and power off.

## 3.5 System service

The DS6000 uses a light path guidance strategy that allows the user in many cases to both detect and repair a failure without using a GUI. However, if desired, guided maintenance in the form of a GUI with animation is also available. This is done by using the DS Storage Manager GUI. Most parts can be replaced without this GUI, though this may not always be the case depending on what parts have failed and the failure mode of those parts.

### 3.5.1 Example 1: Using light path indicators to replace a DDM

An example of the use of the light path guided repair is a disk failure. The user sees that the System Alert Indicator is on, and that a DDM fault indicator is also lit. They refer to the Service Card shipped with the DS6000 and using the simple replacement instructions detailed there, they remove and replace the failed DDM with a new one. After replacing the DDM the System Alert Indicator will be turned off automatically.

### 3.5.2 Example 2: Using the GUI to replace a power supply

As an alternative to using light path guidance, the user also has the alternative of using the GUI. After either receiving an alert or determining a fault exists via the system alert indicator, the user could start the DS Storage Manager and switch to the *component view* to confirm system status.

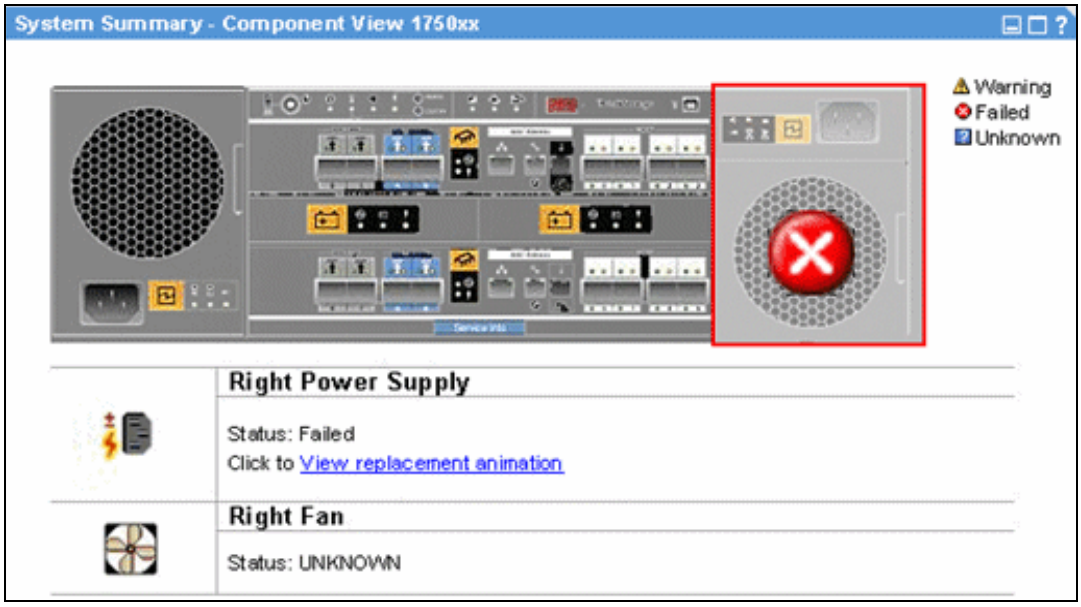



*Figure 3-6 Failed power supply*

If a power supply failure is indicated, the user could then follow this procedure:

1. Review the online component removal instructions. Figure 3-7 on page 59 shows an example of the screen the user may see. On this screen, users are given the ability to do things like:
   a. View an animation of the removal and replacement procedures.
   b. View an informational screen to determine what affect this repair procedure will have upon the DS6000.
   c. Order a replacement part from IBM via an internet connection.

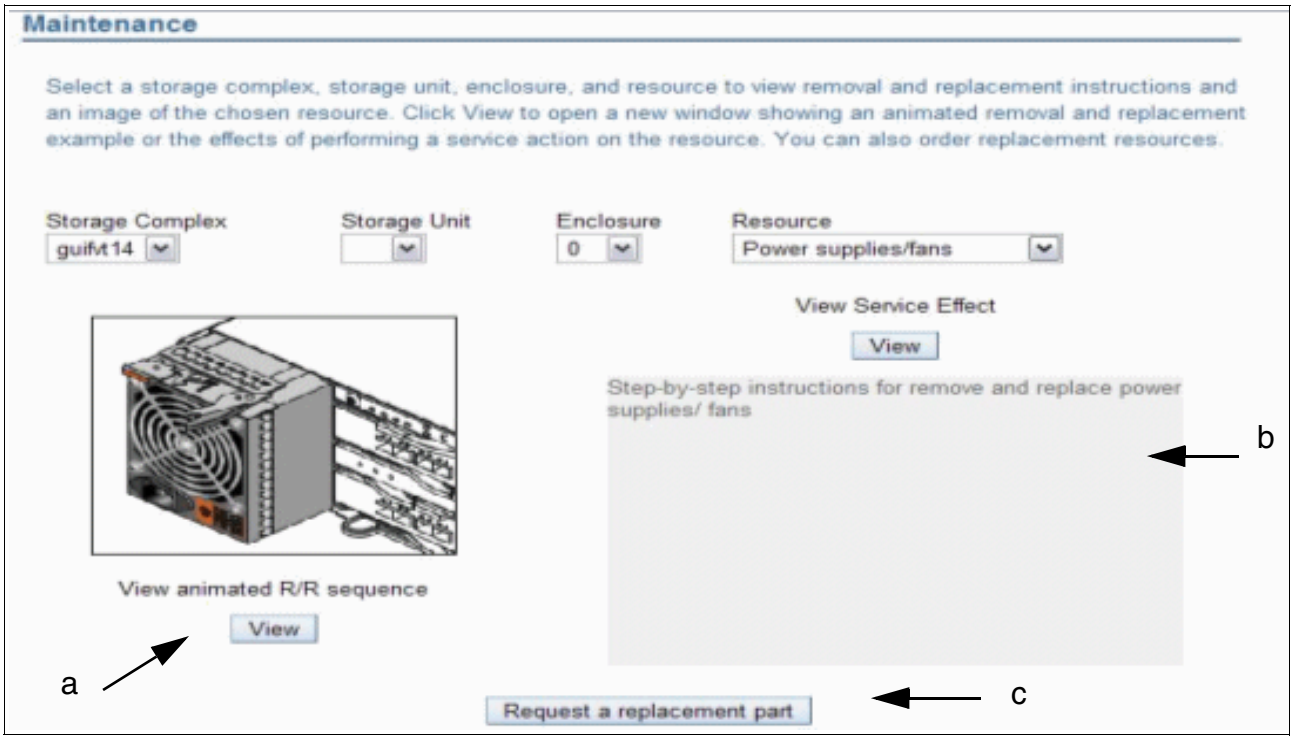



*Figure 3-7 Power supply replacement via the GUI*

2. Upon arrival of the replacement supply, the user physically removes the faulty power supply and then installs the replacement power supply.
3. Finally, the user checks the component view to review system health after the repair. An example of this is shown in Figure 3-8. In this example we can see that all the components displayed are *normal*.

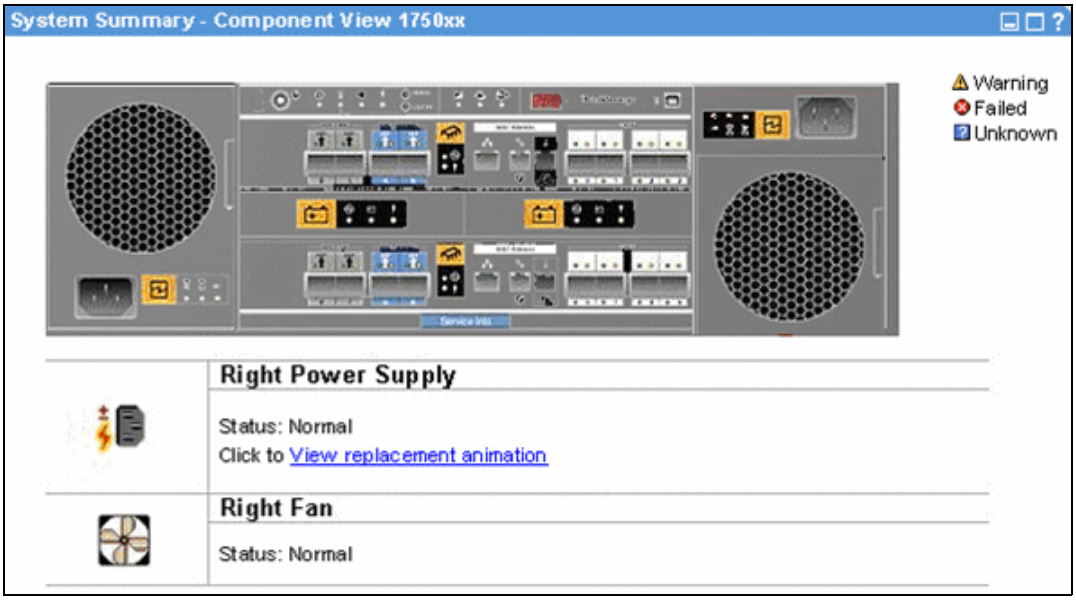



*Figure 3-8 Power supply replaced*

All parts are very easy to remove and replace. There is comprehensive online documentation and help.

## 3.5.3 System indicators

The DS6000 uses several simple indicators to allow a user to quickly determine the health of the DS6000. These indicators were previewed in Chapter 2, “Components” on page 23.

### System Identify Indicator (blue light)

In addition to the light path diagnostics discussed here, the DS6000 implements a method for the person servicing the system to identify all of the enclosures associated with a given system. This is a blue LED visible on the front and rear of an enclosure. The blue light can be turned on or off by using the system identify button on the storage or any expansion enclosure.

This indicator will begin blinking when the Light Path Identify push button is pressed. At this time a request is generated to all other enclosures in the subsystem to turn their System Identify indicators on. When all enclosures have their indicators on solid, the indicator in this enclosure changes from blinking to solid on.

When the Light Path Identify button of the enclosure with its System Identify Indicator active (solid) is pressed, this indicator is changed to blinking. In turn, a request is generated to all enclosures in the subsystem to turn off their System Identify Indicators. When all other Identify indicators are turned off, the Identify indicator in this enclosure will change from blinking to solid off.

#### *System Alert Indicator (amber light)*

This indicator is similar in function to the xSeries Fault Light. It is present on both the storage and expansion enclosures and is used in problem determination. This indicator will be turned on solid when a fault is detected in the system. The indicator will remain on solid until the fault condition is corrected or the Light Path Remind button is pressed and activated. If the Light Path Remind button is pressed, the Alert indicator will go from solid to a 2 second blip (short blip on for 250 msec every 2 seconds). The System Alert Indicator and the System Information Indicator can be on at the same time, if you have two independent error conditions, one minor, one major. When the last fault is corrected, the indicator will be turned off.

#### *System Information Indicator (amber light)*

This indicator is similar in function to the xSeries Information indicator. It is present on both the server and expansion enclosures and is used in problem determination. This indicator will be on solid when a minor error condition exists in the system. For example, a log entry has been written that the user needs to look at. It will remain on solid until the fault condition is corrected (for example, by the user viewing the log). The System Alert Indicator and the System Information Indicator can be on at the same time if you have two independent error conditions, one minor, one major.

#### *CRU Endpoint Indicator (amber light)*

A CRU is a customer replaceable unit. A CRU is a part of the machine that can be replaced safely and easily by an end user. This indicator is present on both the controller and expansion enclosures and is used in problem determination. Each CRU has an amber indicator that, when lit, indicates to the user that a fault exists and that the CRU should be replaced. If a fault indicator is on, it is not necessary to prepare that part for replacement. This means it is not necessary to *quiesce* a resource prior to replacement, as is the case on an ESS 2105. Any time a CRU fault light is turned on, the System Alert Indicator will also be turned on. If there are multiple CRU failures then the CRUs can be replaced in any order. The DS6000 can light more than one CRU indicator. For example, if a power supply and a disk drive both fail, both CRU lights will be turned on and either CRU can be replaced in any order. The CRU Endpoint Indicator will be blocked from being illuminated in any case where

additional guidance on the CRU replacement procedure is required. This includes a situation in which it is unclear which CRU has failed. This will prevent an incorrect maintenance procedure from taking place. After the defective CRU has been replaced, the CRU fault indicator will turn off. Pressing the Remind button will have no effect on the state of the CRU Endpoint Indicator.

## 3.5.4 Parts installation and repairs

The DS6000 has been designed for ease of maintenance. This allows the user to perform the vast majority of service tasks.

### Parts replacement

With the DS6000, an IBM Systems Service Representative (SSR) is not needed to perform the majority of service tasks required during normal operations. Using light path diagnostics it is possible for the user to perform problem determination, parts ordering, and parts replacement.

#### *CRU parts versus FRU parts*

Within all IBM machines, spare parts are divided into two categories: CRU parts (customer replaceable units) and FRU parts (field replaceable units). If a part is designated a CRU, this implies that it can be safely and easily replaced by an end user with few or no tools. If a part is designated a FRU, then this implies that the spare part needs to be replaced by an IBM Service Representative. Within CRU parts, there are currently two tiers: Tier 1 CRUs are relatively easy to replace, while Tier 2 CRUs are generally more expensive parts or parts that require more skill to replace.

#### *Tier 1 CRU parts:*

- Battery backup units
- Cables - Ethernet, serial, fibre optic, and power
- Disk drive modules
- Operator panels - front and rear display
- Power supplies
- RAID controller and SBOD controller cards
- SFPs (2Gbps small form factor plugable fibre optic units)

Currently the only Tier 2 CRU is the entire chassis for either the storage or expansion enclosures. There are currently no FRU parts for the DS6000.

While the DS6000 is under warranty, IBM will ship a replacement CRU to the machine location free of charge, provided the machine is located in a metro area that is serviced by IBM. You should check with your IBM sales representative or IBM Business Partner for details. Installation of Tier 1 CRUs is the customer's responsibility. If an IBM SSR installs a Tier 1 CRU at the customer's request, there will be a charge for the installation. However, for machines with an on-site same-day response service agreement, IBM will replace a Tier 1 CRU at the customer's request, at no additional charge. The customer may choose to install a Tier 2 CRU themselves, or request IBM to install it, at no additional charge. When the customer calls in for service and the problem can be fixed by a Tier 2 part, the customer is given the choice to decide at that point if they have the skills on hand to replace the part.

## 3.6 Microcode updates

The DS6000 contains several discrete redundant components. Most of these components have firmware that can be updated. This includes the controllers, device adapters, host adapters, and network adapters. Each DS6800 controller also has microcode that can be updated. All of these code releases come as a single package installed all at once. As IBM continues to develop and improve the DS6800, new releases of firmware and microcode will become available which offer improvements in both function and reliability.

The architecture of the DS6800 allows for concurrent code updates. This is achieved by using the redundant design of the DS6800. In general, redundancy is lost for a short period as each component in a redundant pair is updated. This also depends on the attached hosts having a separate path to each controller.

Each DS6800 controller card maintains a copy of the previous code version and the active code version. When a code update is performed, the new code version is written to the controller and then activated.

### Maintaining code

It is the responsibility of the user to ensure that their DS6800 is running on the currently available version of microcode. They can do this by monitoring the code level currently available at:

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

The currently available version of code can also be downloaded from this Web site.

Users can also register for support e-mails at:
https://www-1.ibm.com/support/mysupport/us/en/

### Installation process

The concurrent installation process proceeds as follows:

1. Copy the new DS6800 code version onto the DS Management Console and from there to the internal storage of each controller.
2. The user selects to active the code and from then the process is automated. First, controller 1 is quiesced.
3. The active microcode version in controller 1 is updated to the new level. This may include firmware updates to the controller card, host adapter, device adapter, and network adapter.
4. Controller 1 is rebooted and then resumes operation.
5. Controller 0 is quiesced.
6. The active microcode version in controller 0 is updated to the new level. This may include firmware updates to the controller card, host adapter, device adapter and network adapter.
7. Controller 0 is rebooted and then resumes operation.
8. Now the user must manually update the DS Management Console to the matching level of DS System Manager GUI. You also need to update the version of DS CLI being used to match the SM GUI version.

As noted, after step two, the installation process described above should not require any user intervention. The user should be able to simply start the process and then monitor its

progress using the DS Management Console GUI. Clearly a multipathing driver (such as SDD) is required for this process to be concurrent.

There is also the alternative to load code non-concurrently. This means that both controllers are unavailable for a short period of time. This method can be performed in a smaller window of time.

## 3.7 Summary

This chapter has described the RAS characteristics of the DS6000. These characteristics combine to make the DS6000 a world leader in reliability, availability, and serviceability.

4

# Virtualization concepts

This chapter describes the virtualization concepts for the DS6000 and the abstraction layers for disk virtualization. The topics covered are:

- Array sites
- Arrays
- Ranks
- Extent pools
- Logical volumes
- Logical storage subsystems
- Address groups
- Volume groups
- Host attachments

## 4.1 Virtualization definition

In our fast changing 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.

Our 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.

## 4.2 The abstraction layers for disk virtualization

In this chapter, when we talk 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 DS6000 is populated with switched FC-AL disk drives that are mounted in storage enclosures. You order disk drives in disk drive sets. A disk drive set is a group of 4 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). In normal operation, however, disk drives are typically accessed by one device adapter and one server. Each path on each device adapter can be active concurrently, but the set of eight paths on the two device adapters can all be concurrently accessing independent disk drives. This avoids any contention between the two device adapters for access to the same disk, such that all eight ports on the two device adapters can be concurrently communicating with independent disk drives.

Figure 4-1 on page 67 shows the physical layer on which virtualization is based.

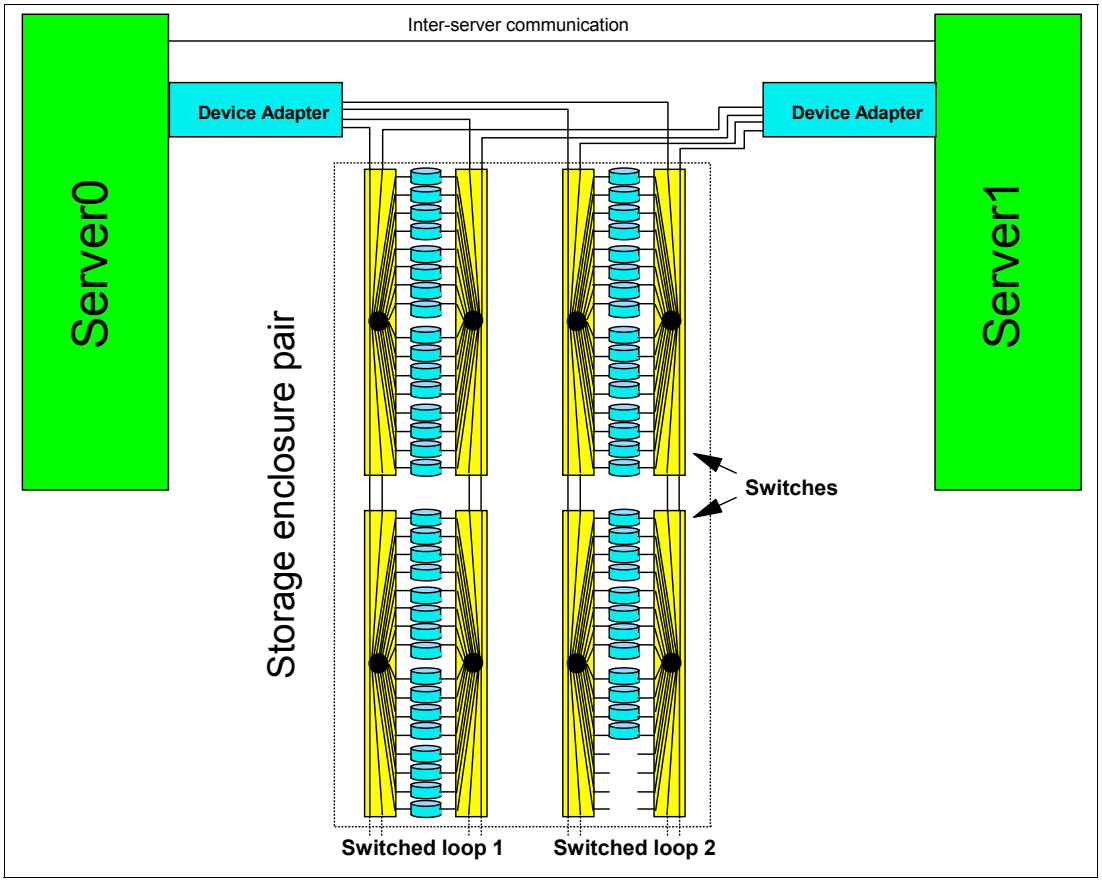



*Figure 4-1 Physical layer as the base for virtualization*

When you 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 DS6000. 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).

## 4.2.1 Array sites

An array site is a group of four DDMs. What DDMs make up an array site is pre-determined by the DS6000, but note, that there is no pre-determined server affinity for array sites. The DDMs selected for an array site are chosen from the same disk enclosure string (see Figure 4-2 on page 68). All DDMs in an array site are of the same type (capacity and RPM).

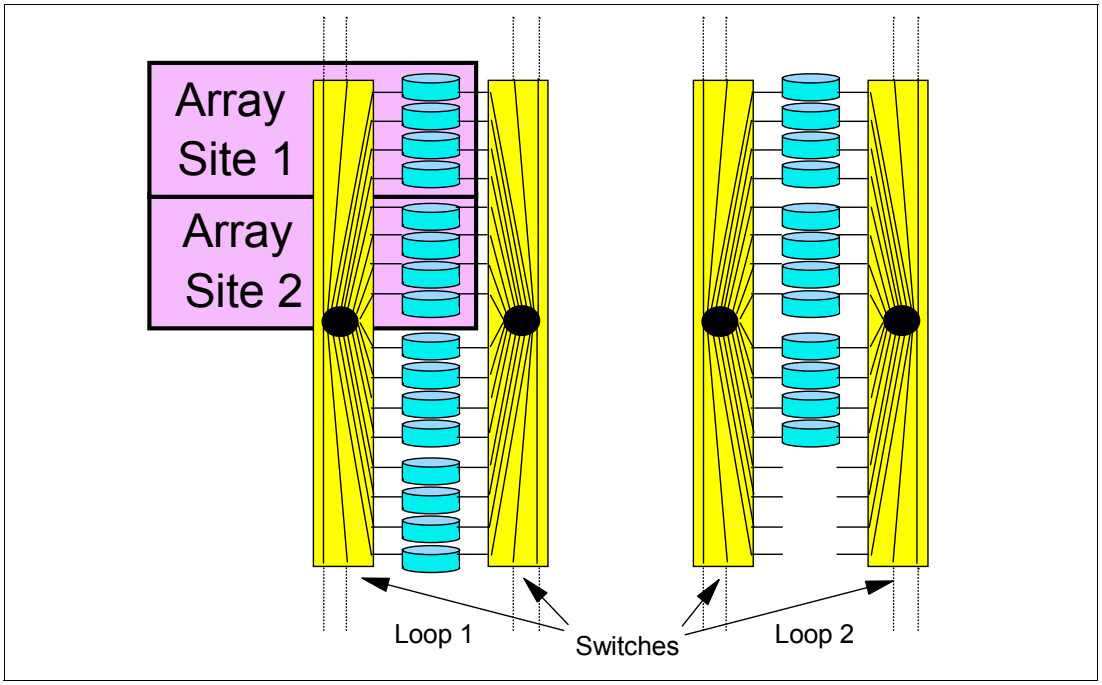



*Figure 4-2 Array sites*

*Array sites* are the building blocks used to define *arrays*.

### 4.2.2 Arrays

*Arrays* are created from one or two *array sites*. Forming an array means defining it for a specific RAID type.The supported RAID types are RAID-5 and RAID-10 (see 3.3.1, “RAID-5 overview” on page 52 and 3.3.2, “RAID-10 overview” on page 53). For each array site or for a group of two array sites you can select a RAID type. The process of selecting the RAID type for an array is also called *defining* an array.

According to the DS6000 sparing algorithm, up to two spares may be taken from the array sites used to construct the array on each device interface (loop). See Chapter 5, “IBM TotalStorage DS6000 model overview” on page 83 for more details.

Figure 4-3 on page 69 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) from two array sites. According to the RAID-5 rules, parity is distributed across all seven drives in this example.

Also referring to Figure 4-3, on the right side 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.

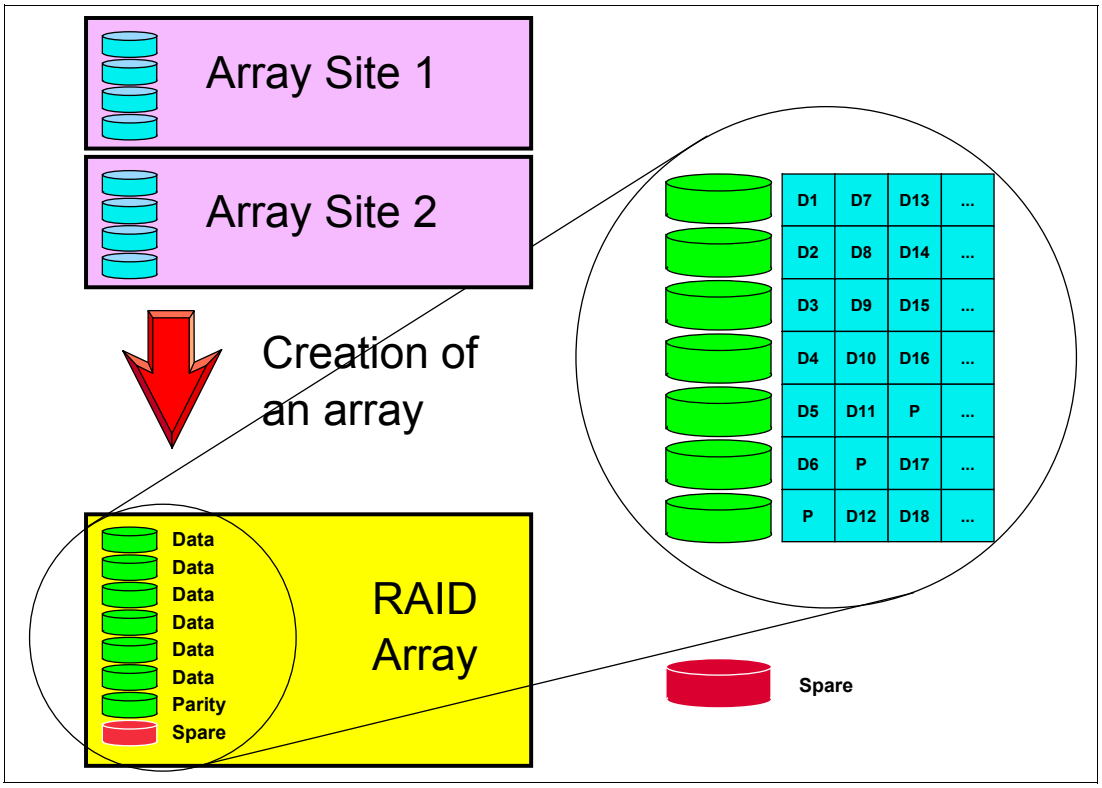



*Figure 4-3 Creation of an array*

So, an array is formed using one or two array sites, 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.

### 4.2.3 Ranks

In the DS6000 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 DS6000 implementation, a rank is built using just one array.

The available space on each rank is 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 4-4 on page 70 and indicated by the small squares in Figure 4-5 on page 71.

The process of forming a rank does two things:

- The array is defined for either fixed block (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 but think of storage in metrics of the old 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 4-4 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).

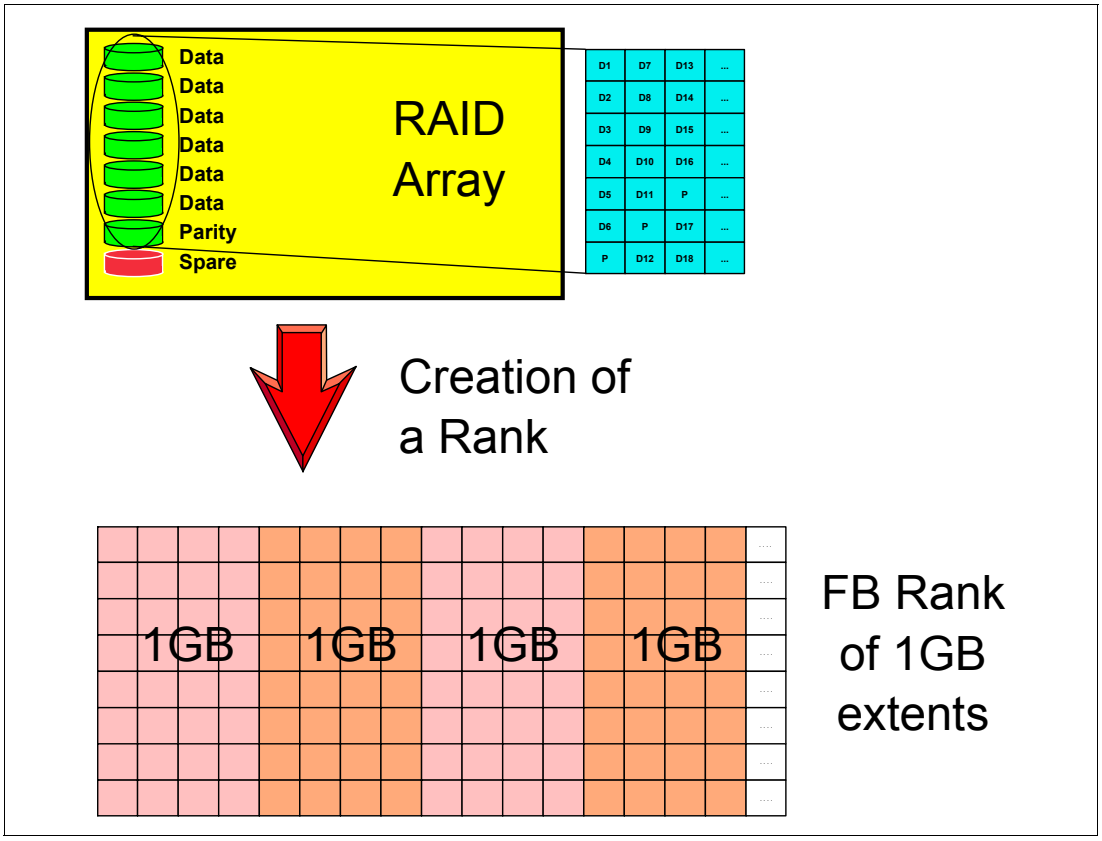



*Figure 4-4 Forming an FB rank with 1 GB extents*

## 4.2.4 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 one assigned to server 1 so that both servers are active. In an environment where FB and CKD are to go onto the DS6000 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 4-5 is an example of a mixed environment with CKD and FB extent pools.

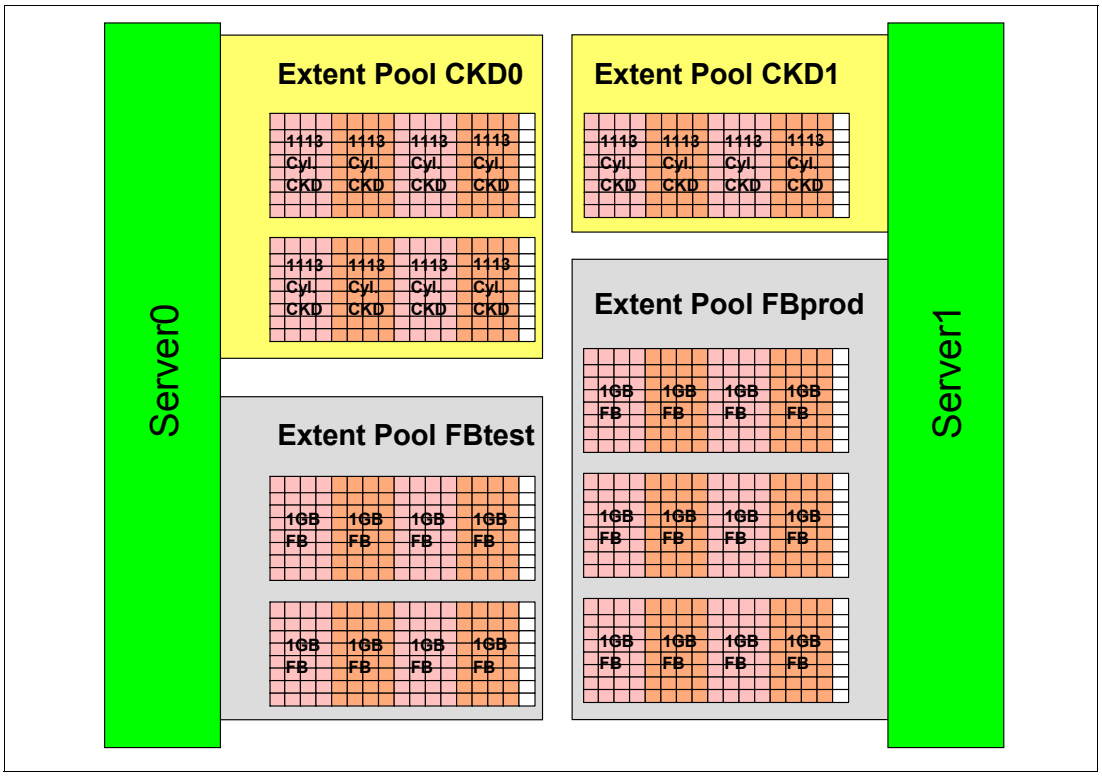



*Figure 4-5 Extent pools*

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

## 4.2.5 Logical volumes

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

On a DS6000 up to 8192 (8K) volumes can be created (8K CKD, or 8K FB volumes, or a mix of both types (4K CKD plus 4K FB).

### 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 ($2^{40}$) in any integral multiple of 64K bytes. The capacity allocated to a LUN is always a multiple of the 1 GB extent, so any LUN size that is not a multiple of 1 GB wastes some space in the last extent allocated to the LUN. LUNs can be allocated in binary GB ($2^{30}$ bytes), decimal GB ($10^{9}$ bytes), or 512 or 520 byte blocks. However, when you define a LUN that is not a multiple of 1 GB, the capacity up to the next multiple of 1 GB 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, your should consider allocating volumes that are exact multiples of 1113 cylinders. In fact, integral multiples of 3339 cylinders should be consider 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).

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

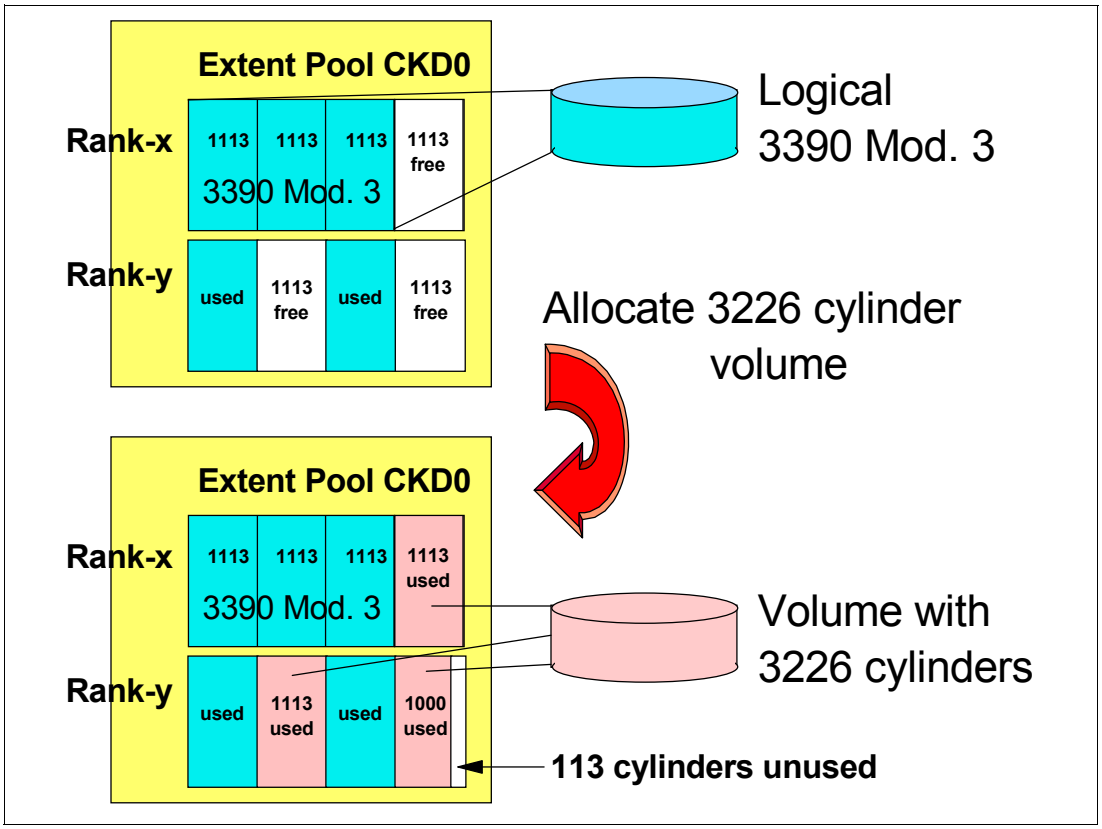



*Figure 4-6 Allocation of a CKD logical volume*

Figure 4-6 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 4-7 on page 74.

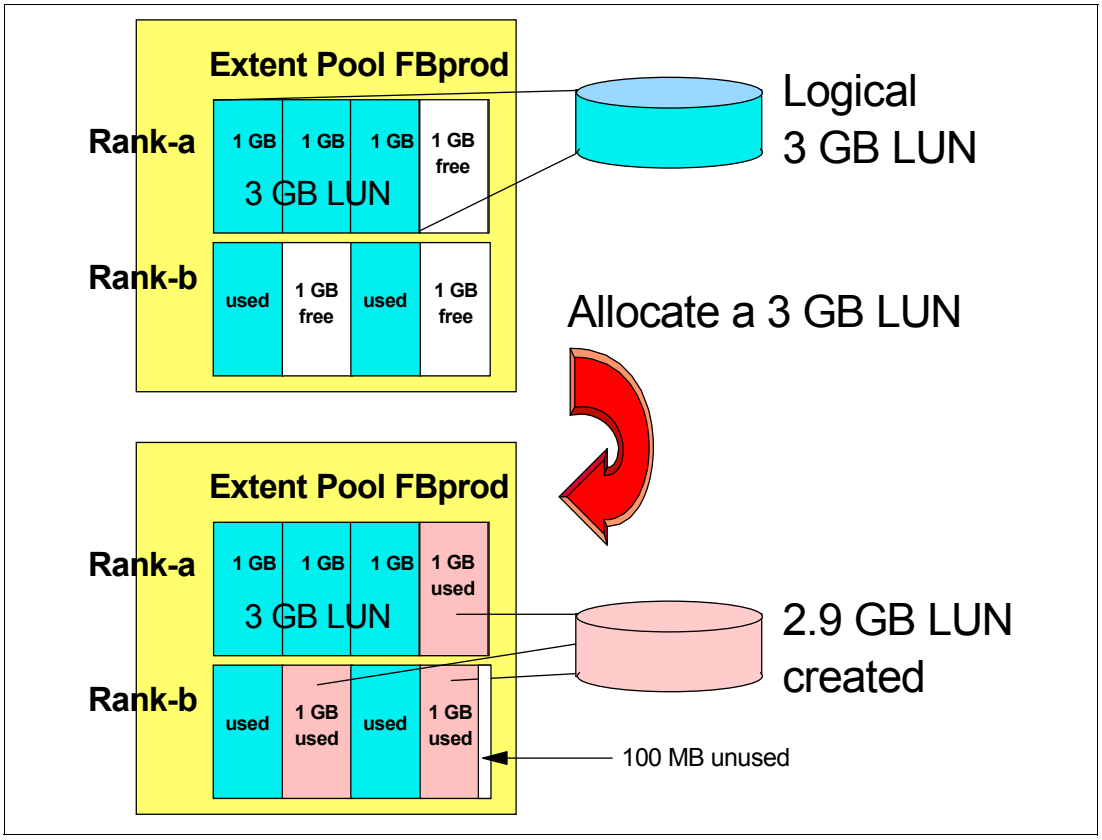



*Figure 4-7 Creation of an FB LUN*

## 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 DS6000 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 DS6000 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 which are 1 GB ($2^{30}$). For more information on this topic see Appendix B, “Using the DS6000 with iSeries” on page 329.

## 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 DS6000 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.

This construction method of using fixed extents to form a logical volume in the DS6000 allows flexibility in the management of the logical volumes. We can now delete LUNs and reuse the extents of that LUN to create another LUN, maybe of a different size. 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 DS6000 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 DS6000 announcement letter the following Statement of General Direction:

*Extension of IBM's dynamic provisioning technology within the DS6000 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.*

### 4.2.6 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 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 DS6000, 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 DS6000. Also, the number of LSSs has changed. You can now define up to 32 LSSs for the DS6000. 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’1E’) belong to server 0 and all odd numbered LSSs (X’01’, X’03’, X’05’, up to X’1F’) belong to server 1.

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 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, “DS CLI” on page 195.

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 4-8). 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; from a performance standpoint this was not desirable. Now on the DS6000 you can group your volumes in one or a few LSSs, but still have the volumes in many arrays or ranks.

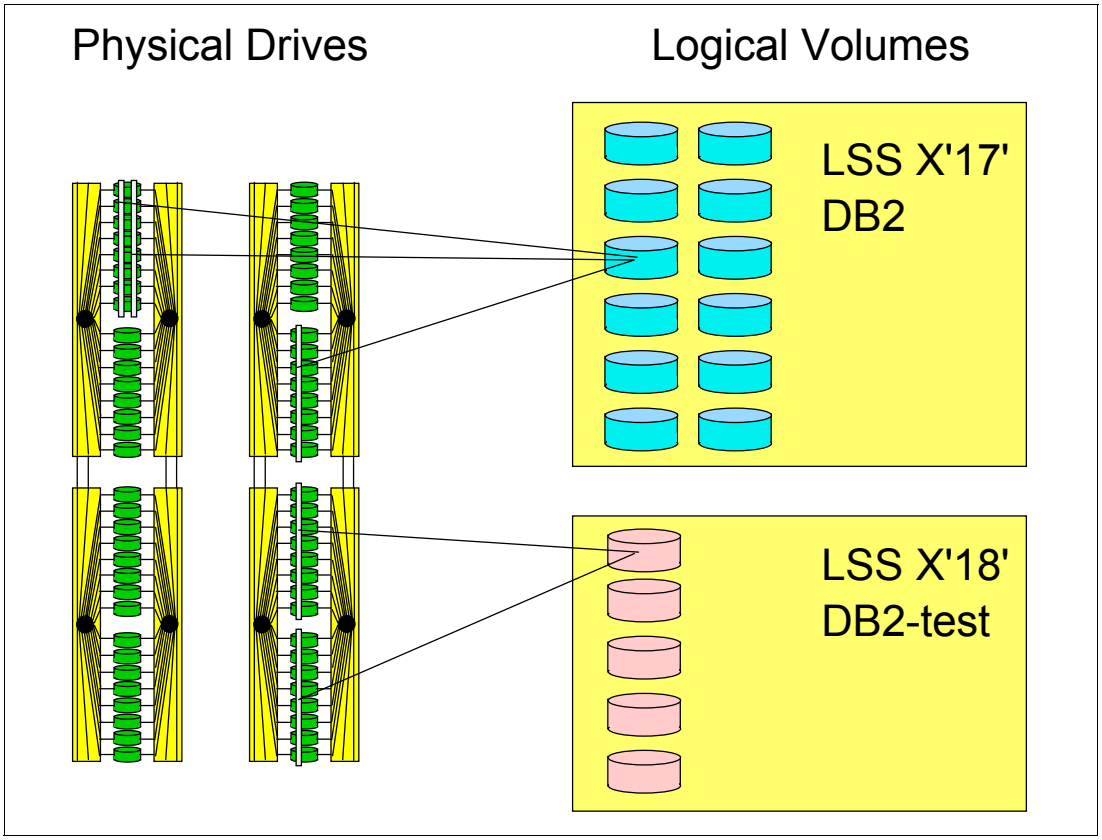



*Figure 4-8 Grouping of volumes in LSSs*

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, must be created before the first CKD logical volume can be created on the LSS, and must be deleted manually after the last CKD logical volume on the LSS is deleted.

## 4.2.7 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.

Figure 4-9 illustrates the concept of LSSs and address groups.

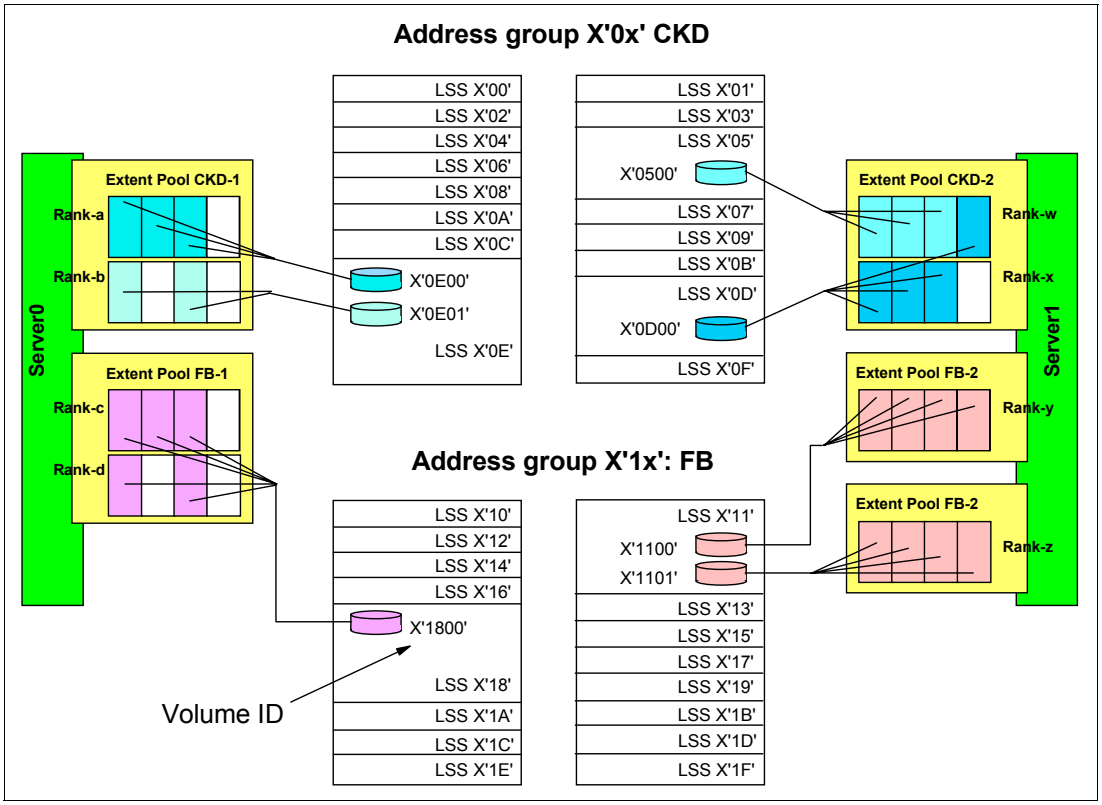



*Figure 4-9 Logical subsystems*

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'1101' denotes the second (X'01') LUN in LSS X'11' of address group 1.

## 4.2.8 Volume access

A DS6000 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 DS6000 introduced the concept of host attachments and volume groups.

## Host attachment

HBAs are identified to the DS6000 in a host attachment construct that specifies the HBA's 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 DS6000 I/O ports that 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 DS6000 is 1024.

## 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 bit map that specifies which LUNs are accessible. For this volume group type, the logical volume number X'abcd' is mapped to LUN_ID X'40ab40cd00000'. 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 is 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 on a DS6000 is 1040.

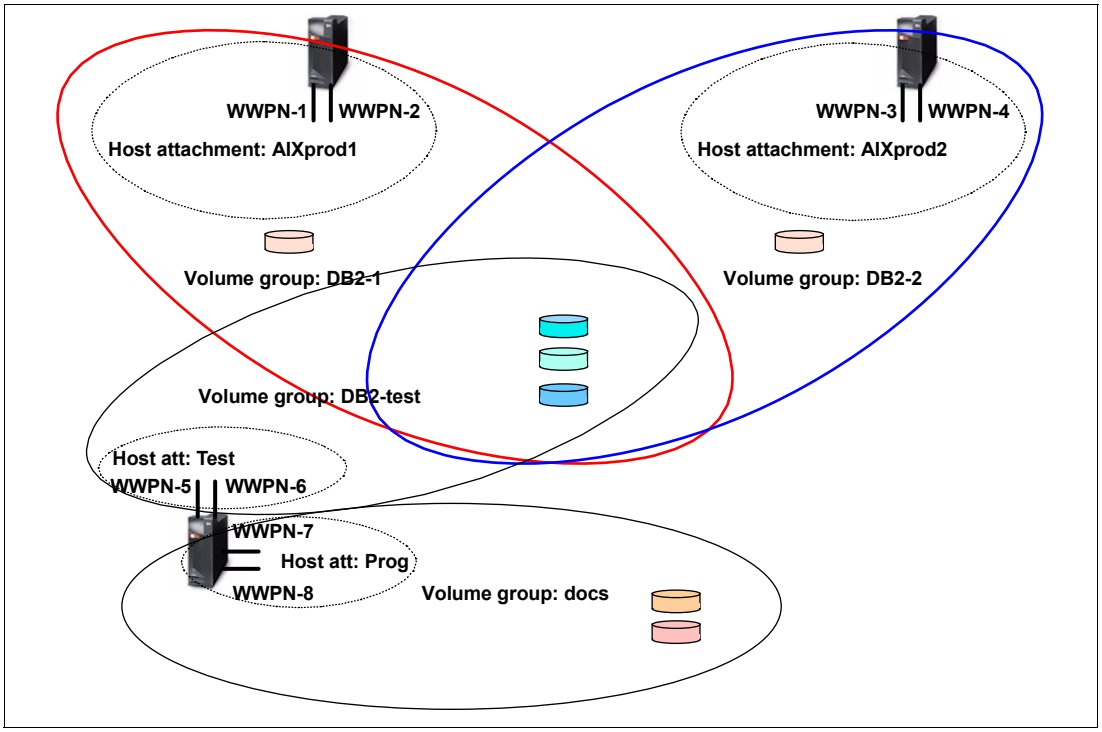



*Figure 4-10 Host attachments and volume groups*

Figure 4-10 shows the relationships between host attachments and volume groups. Host AIXprod1 has two HBAs, that 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 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.

### 4.2.9 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 which 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 efforts.

Figure 4-11 on page 80 summarizes the virtualization hierarchy.

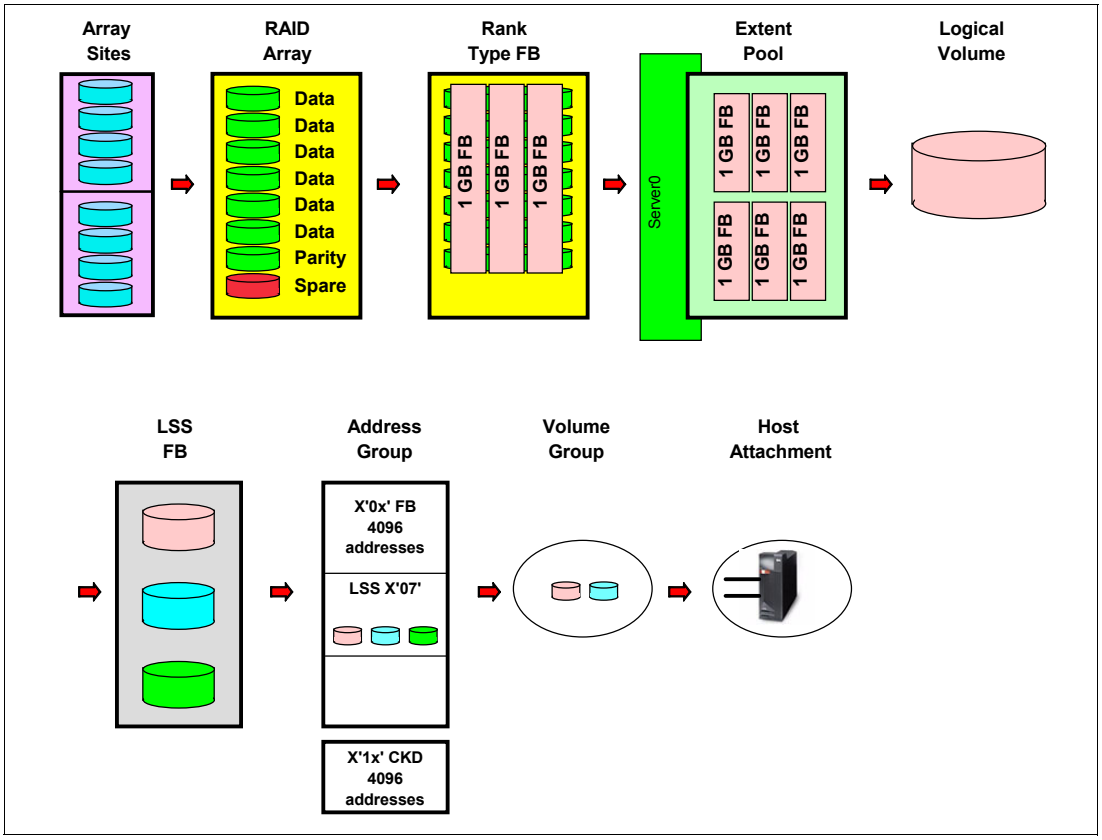



*Figure 4-11 Virtualization hierarchy*

## 4.2.10 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 loops or from the same loop. Figure 4-12 on page 81 shows an optimal distribution of four logical volumes within a DS6000. 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 and across the two 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 DS6000 logical volumes (LUNs). You can select LUNs from different DS6000 servers and loops as shown in Figure 4-12. By striping your host logical volume across the LUNs, you will get the best performance for this LVM volume.

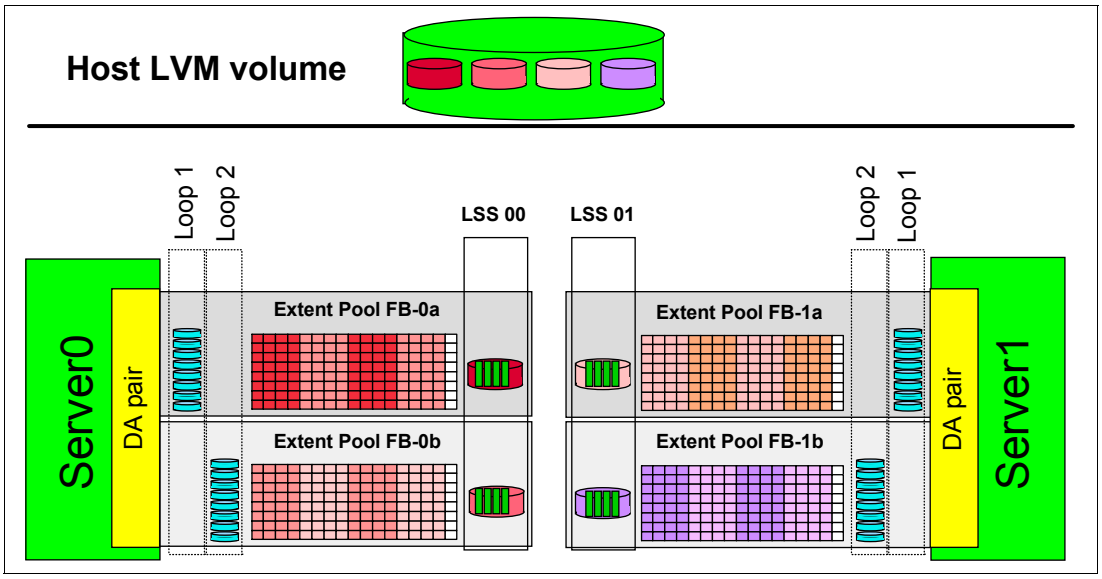



*Figure 4-12 Optimal distribution of data*

## 4.3 Benefits of virtualization

The DS6000 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 the production systems.
- Increased number of logical volumes:
  - Up to 8192 (CKD)
  - Up to 8192 (FB)
  - Up to 4096 CKD and up to 4096 FB
- 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.

5

# IBM TotalStorage DS6000 model overview

This chapter provides an overview of the IBM TotalStorage DS6000 storage server which is from here on referred to as the DS6000. While the DS6000 is physically small, it is a highly scalable and powerfully performing storage server. Topics covered in this chapter are:

- DS6000 highlights
- DS6800 Model 1750-511
- DS6000 Model 1750-EX1
- Design to scale for capacity

# 5.1 DS6000 highlights

The DS6000 is a member of the DS product family that offers high reliability and enterprise class performance for mid-range storage solutions with the DS6800 model 1750-511.

It is built upon 2 Gbps fibre technology and offers:

- RAID protected storage
- Advanced functionality
- Extensive scalability
- Increased addressing capabilities
- The ability to connect to all relevant host server platforms

## 5.1.1 DS6800 Model 1750-511

The 1750-511 model contains control unit functions as well as a rich set of advanced functions, and holds up to 16 disk drive modules (DDMs). It provides a minimum capacity of 584 GB with 8 DDMs and 73 GB per DDM.

As of this writing, the maximum storage capacity with 300 GB DDMs is 4.8 TB with 16 DDMs in a 1750-511 model.

It measures 5.25 inches high and is available in a 19 inch rack-mountable package.

*Figure 5-1 DS6800 Model 1750-511 and Model 1750-EX1 front view*

The 1750-511 model offers the following features:

- Two Fibre Channel control cards.
- PowerPC 750GX 1 GHz processors.
- Dual active controllers to provide continuous operations and back up the other controller in case of controller maintenance or an unplanned outage of a controller.
- 4 GB of cache.
- Battery backed mirrored cache.
- Two battery backup units - one per controller card.
- Two AC/DC power supplies with imbedded enclosure cooling units.
- Disk subsystem connectivity with eight 2 Gbps device ports.
- Selection of 2 Gbps Fibre Channel disk drives including 73 GB, 146 GB, and 300 GB DDM size with speeds currently of 10K RPM or 15K RPM.

- Front-end connectivity with two to eight Fibre Channel host ports which auto negotiate to either 2 Gbps or 1 Gbps link speeds. Each port, long-wave or short-wave, can be either configured for:
  - FCP to connect to open system hosts or PPRC FCP links, or both
  - FICON host connectivity

The DS6800 storage system can connect to a broad range of servers through its intermix of FCP and FICON front-end I/O adapters. This includes the following servers:

- IBM eServer zSeries
- IBM eServer iSeries
- IBM eServer pSeries
- IBM eServer xSeries
- Servers from Sun Microsystems
- Servers from Hewlett-Packard
- Servers from other Intel®-based platforms

For an up-to-date and complete interoperability matrix refer to:

http://www.ibm.com/servers/storage/disk/ds6000/interop.html

*Figure 5-2 DS6800 Model 1750-511 rear view*

## 5.1.2 DS6000 Model 1750-EX1

To configure more than 4.8 TB, the DS6800 can expand with the expansion enclosure model 1750-EX1 to connect a maximum of 128 DDMs per storage system. This brings the maximum storage capacity to a total of 38.4 TB with 300 GB per DDM, when you expand the DS6800 Model 1750-511 by up to 7 DS6000 Models 1750-EX1.

Each expansion enclosure contains the following features:

- Two expansion controller cards. Each controller card provides the following:
  - Two 2 Gbps inbound ports
  - Two 2 Gbps outbound ports
  - One Fibre Channel switch
- Disk enclosure which holds up to 16 Fibre Channel DDMs.
- Two AC/DC power supplies with imbedded enclosure cooling units.
- Support for attachment to DS6800 Model 1750-511.

The DS6800 Model 1750-EX1 is also a 3 Electrical Industries Association (EIA) self-contained unit, as is the 1750-511, and it can also be mounted in a standard 19 inch rack.

*Figure 5-3 DS6800 Model 1750-EX1 rear view*

Controller model 1750-511 and expansion model 1750-EX1 have the same front appearance. Figure 5-3 displays the rear view of the expansion enclosure, which is a bit different compared to the rear view of the 1750-511 model.

Figure 5-4 shows a 1750-511 model with two expansion 1750-EX1 models.

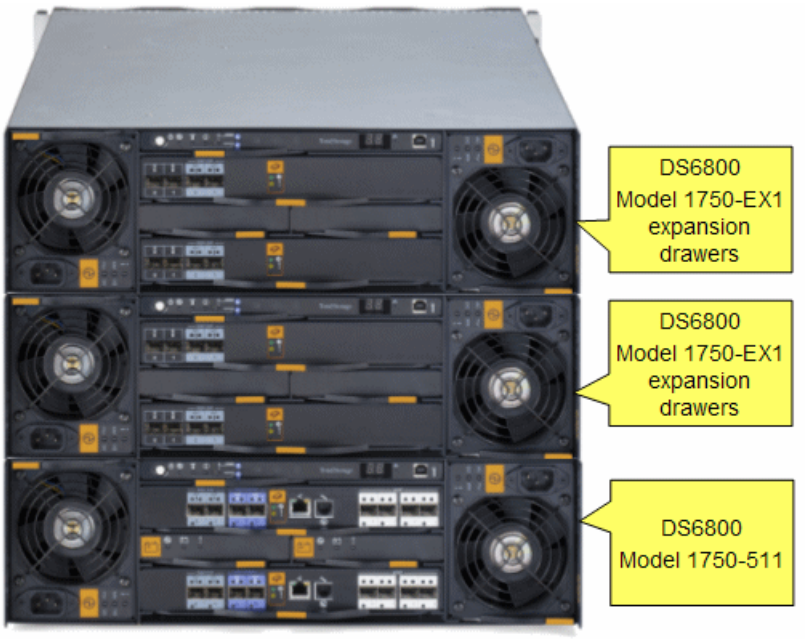



*Figure 5-4 DS6000 with 2 expansion drawer 1750-EX1 models attached to 1750-511 controller*

Up to 7 expansion drawers build up to the maximum configuration of 128 DDMs, which is comprised of 8 drawers x 16 DDMs.

## 5.2 Designed to scale for capacity

The DS6800 has outstanding scalability for capacity ranging from 584 GB up to 38.4TB without disruptive maintenance.

The DS6800 server enclosure can have from 8 up to 16 DDMs and can connect 7 expansion enclosures. Each expansion enclosure also can have 16 DDMs. Therefore, in total a DS6800 storage unit can have 16 + 16 x 7 = 128 DDMs.

You can select from four types of DDMs:

- 73 GB 15k RPM
- 146 GB 10k RPM
- 146 GB 15k RPM
- 300 GB 10k RPM

Therefore, a DS6800 can have from 584 GB (73 GB x 4 DDMs) up to 38.4TB (300 GB x 128 DDMs).

Table 5-1 describes the capacity of the DS6800 with expansion enclosures.

*Table 5-1 DS6800 physical capacity examples*

| Model | with 73 GB DDMs | with 146 GB DDMs | with 300 GB DDMs |
|---|---|---|---|
| 1750-511 (16 DDMs) | 1.17 TB | 2.34 TB | 4.80 TB |
| 1750-511 + 6 Exp (112 DDMs) | 8.18 TB | 16.35 TB | 33.60 TB |

In addition, the DS6800 and expansion enclosures can have different types of DDMs in each enclosure (an intermix configuration).

## The DS6800 capacity upgrade

The DS6800 has two enclosure groups for attaching expansion enclosures. One group can have a server enclosure and up to three expansion enclosures (we call this group *Loop 0*) and the other group can have up to four expansion enclosures (we call this group *Loop 1*). You can attach additional expansion enclosures to the two groups for well-balanced capacity.

Figure 5-5 on page 88 illustrates the connectivity of the server and expansion enclosures. Each DS6800 controller has four FC-AL ports and two of them connect to the dual redundant loops of the first group (Loop 0), and the others to the second group (Loop 1). The FC-AL port in Loop 0 is called the *disk exp* port, and the port in the Loop 1 is called the *disk control* port.

These groups are independent and there is no restriction on the connection sequence of expansion enclosures. Figure 5-5 is an example of how to connect an expansion enclosure to each group alternately.

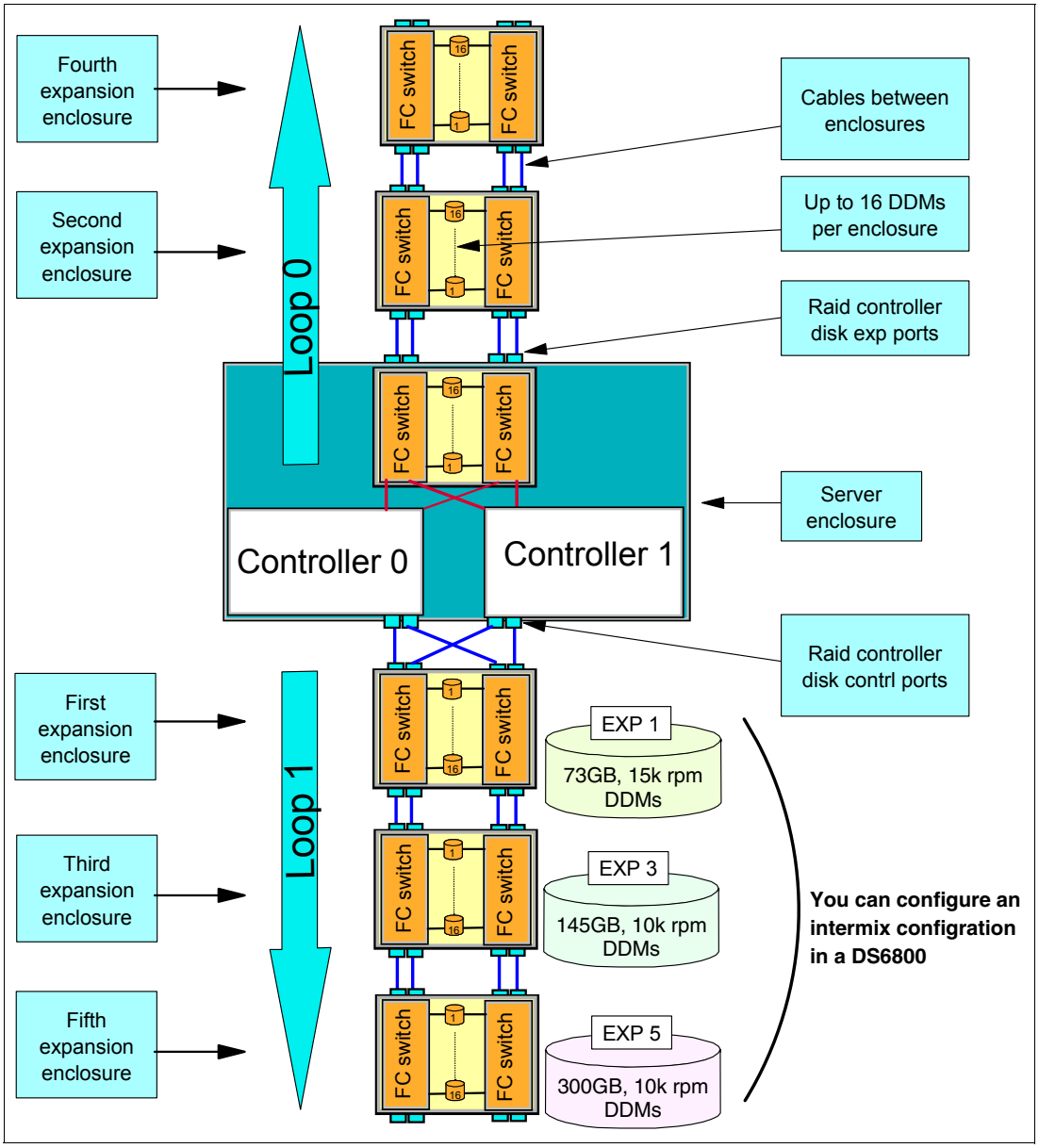



*Figure 5-5 DS6800 switched disk expansion*

When you add new DDMs into an enclosure or attach additional enclosures, you don't need disruptive maintenance. To add new DDMs into an existing enclosure, you have only to pull out the dummy carriers and replace them with the new DDMs. To attach additional enclosures, you only have to attach the new enclosures to the existing enclosures with Fibre Channel cables.

6

# Copy Services

In this chapter, we describe the architecture and functions of Copy Services for the DS6000. Copy Services is a collection of functions that provide disaster recovery, data migration, and data duplication functions. Copy Services run on the DS6000 server enclosure 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
- Future plans

# 6.1 Introduction to Copy Services

Copy Services is a collection of functions that provides 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 a remote site for disaster recovery.

Copy Services run on the DS6000 server enclosure 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 DS6000 and data copying and mirroring capabilities of the 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

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 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 6.3, “Interfaces for Copy Services” on page 108.

# 6.2 Copy Services functions

We describe each function and the architecture of the Copy Services in this section.

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

The Point-in-Time Copy feature, which includes FlashCopy, enables you to create full volume copies of data. 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 section, *track* means a piece of data in the DS6000; the DS6000 uses the logical tracks to manage the Copy Services functions.

See Figure 6-1 on page 91 for an illustration of FlashCopy concepts.

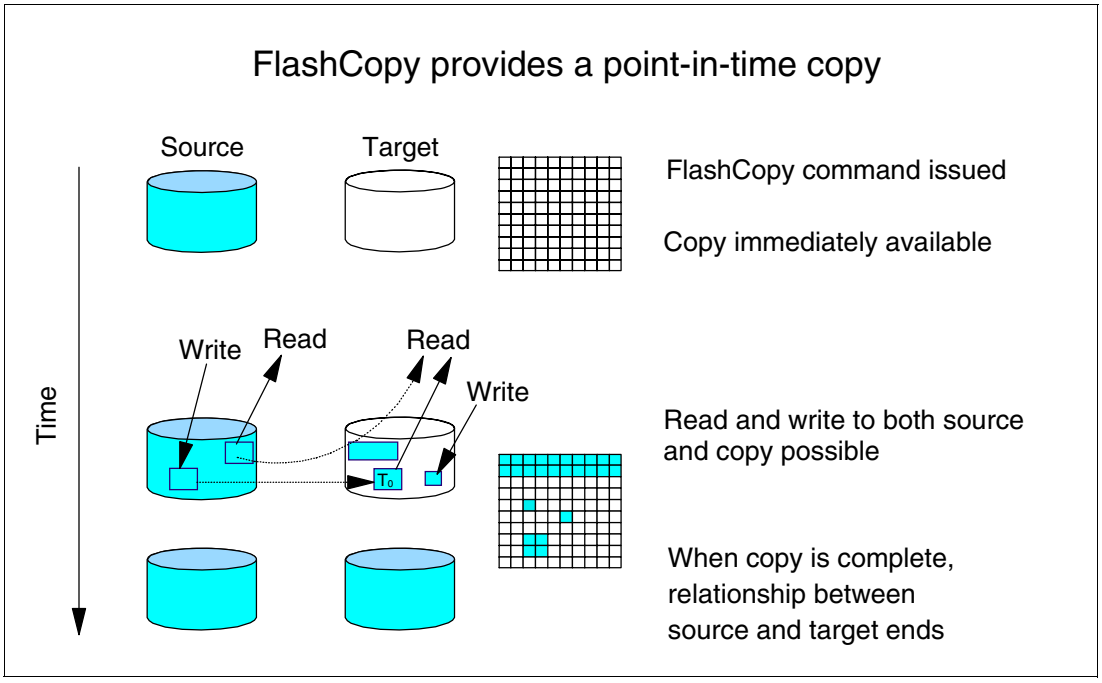



*Figure 6-1 FlashCopy concepts*

When a FlashCopy operation is invoked, the process of establishing the FlashCopy pair and creating the necessary control bitmaps takes only a few seconds to complete. 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 the target volumes.

After creating the bitmaps, 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 bitmaps 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 bitmaps 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 then 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 on your application because the real-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 ESS 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.

## 6.2.2 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 6-2 on page 93 explains the architecture for Incremental FlashCopy.

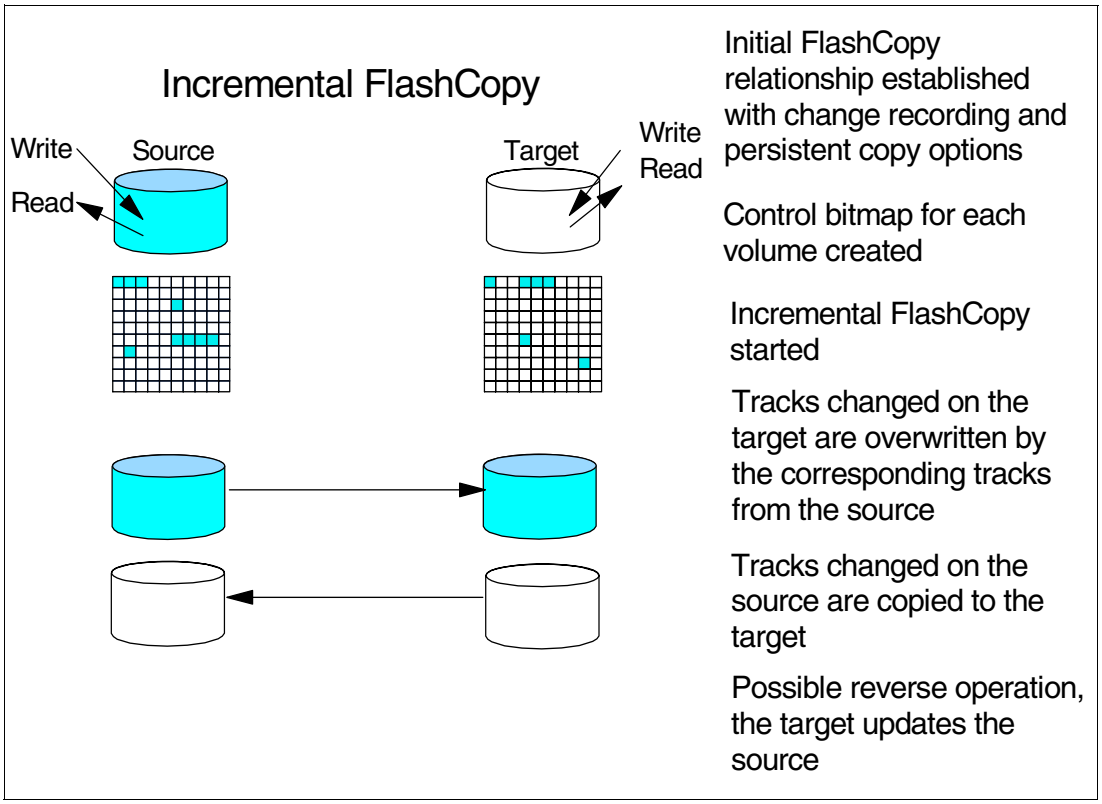



*Figure 6-2 Incremental FlashCopy*

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 server enclosure. 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.

If you want, 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.

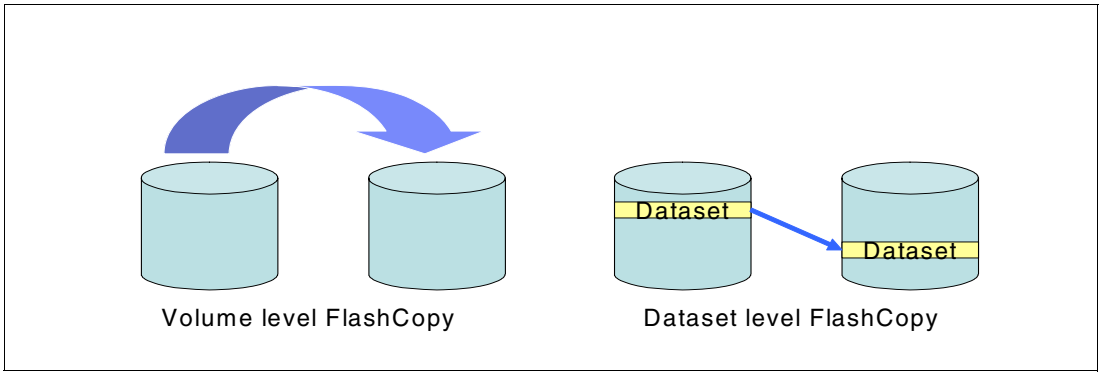



*Figure 6-3 Data Set FlashCopy*

## 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 6-4.

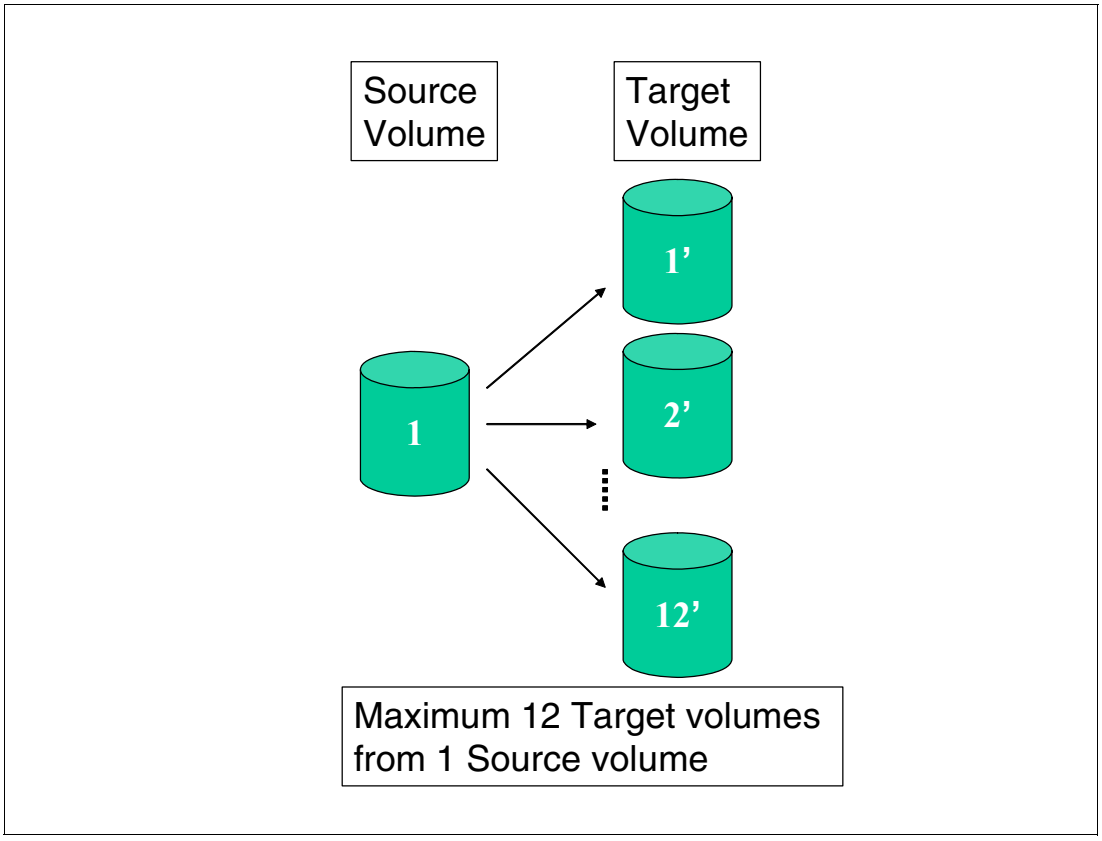



*Figure 6-4 Multiple Relationship FlashCopy*

**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 can 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 issues 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 6-5.

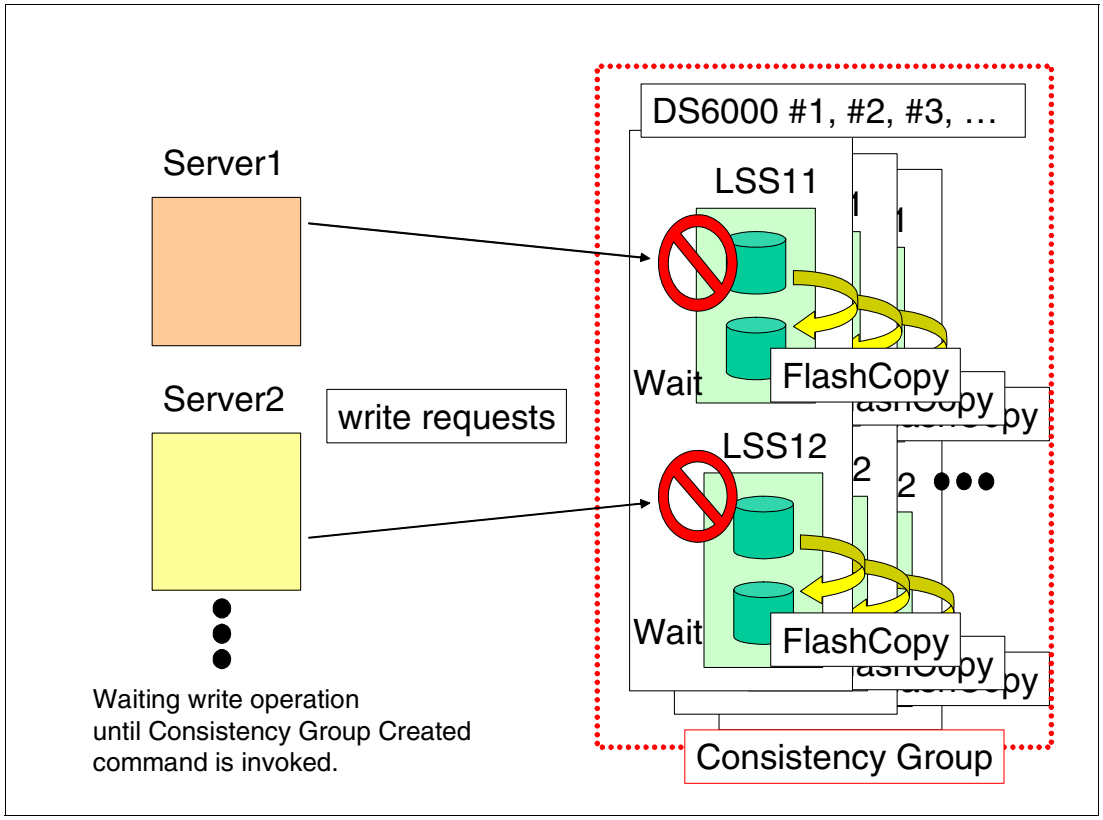



*Figure 6-5 Consistency Group FlashCopy*

A more detailed discussion of the concept of *data consistency* and how to manage the Consistency Group operation is in 6.2.5, “What is Consistency Group?” on page 105.

**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 only available through the use of CLI commands and not 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 6.2.3, “Remote Mirror and Copy (Peer-to-Peer Remote Copy)” on page 97.

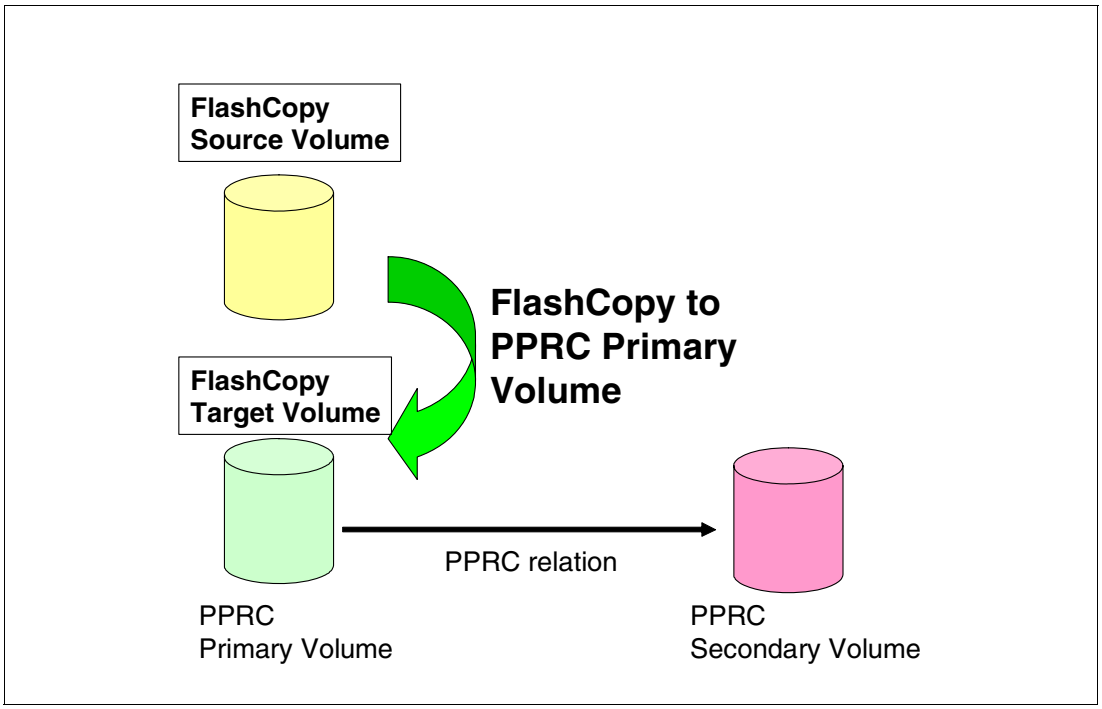



*Figure 6-6 Establish FlashCopy on existing Remote Mirror and Copy primary*

**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 only available through the use of CLI commands and not the DS Storage Manager GUI at the current time.

## 6.2.3 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.

DS6000 server enclosures can participate in Remote Mirror and Copy solutions with the ESS Model 750, ESS Model 800, and DS6000 and DS8000 server enclosures. To establish a PPRC relationship between the DS6000 and 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 DS6000s 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.

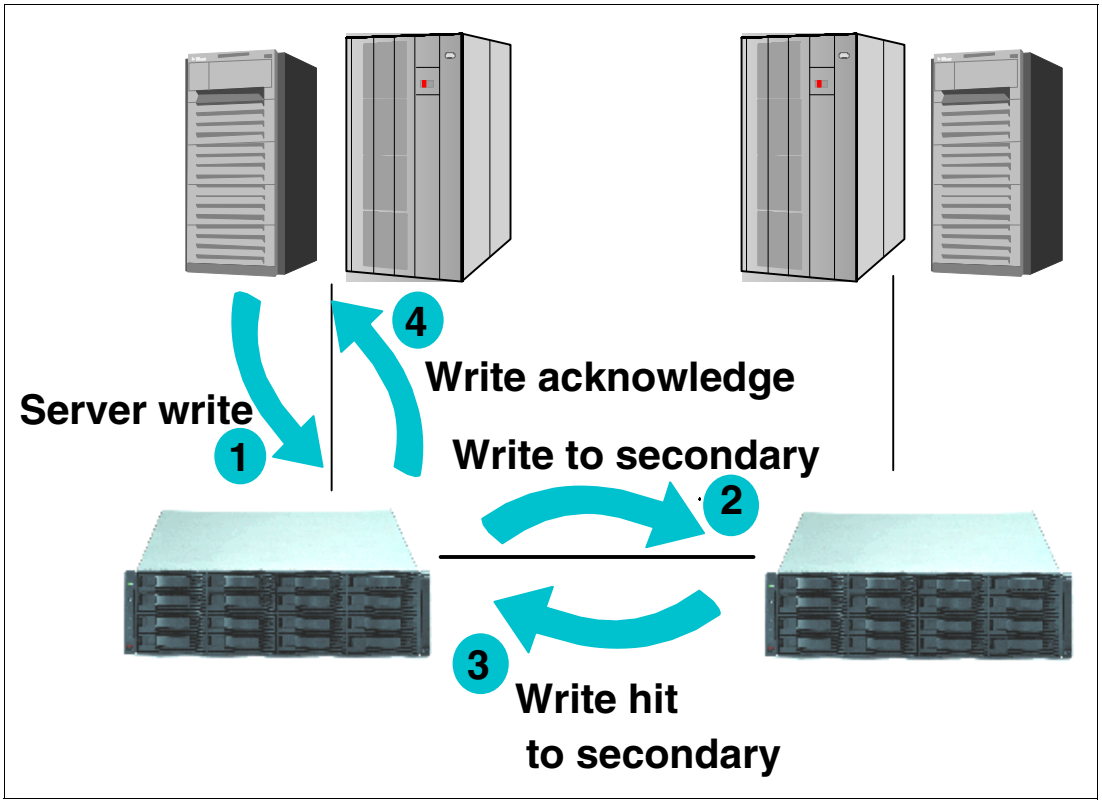



*Figure 6-7 Metro Mirror*

## 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 you establish 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 to make a consistent copy with FlashCopy, you must purchase a Point-in-Time Copy function authorization (2244 Model PTC) for the secondary storage unit.

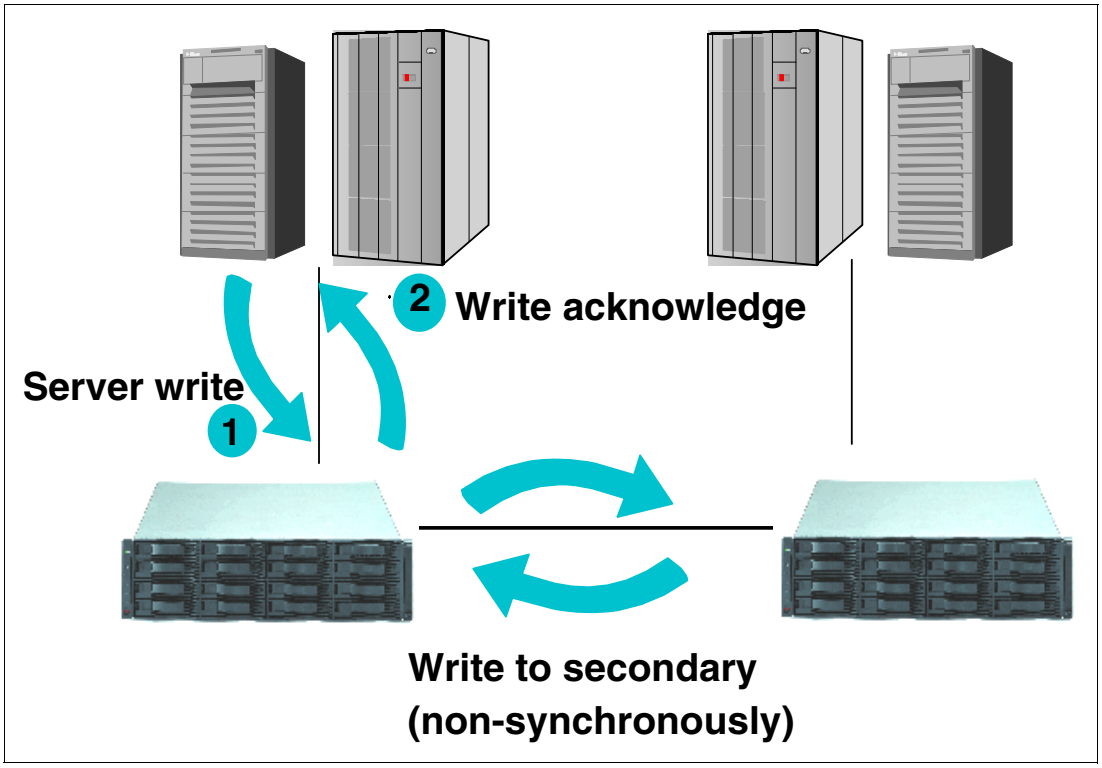



*Figure 6-8 Global Copy*

## 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 server enclosure at the local site is asynchronously shadowed to the server enclosure at the remote site. A consistent copy of the data is then automatically maintained on the server enclosure 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.

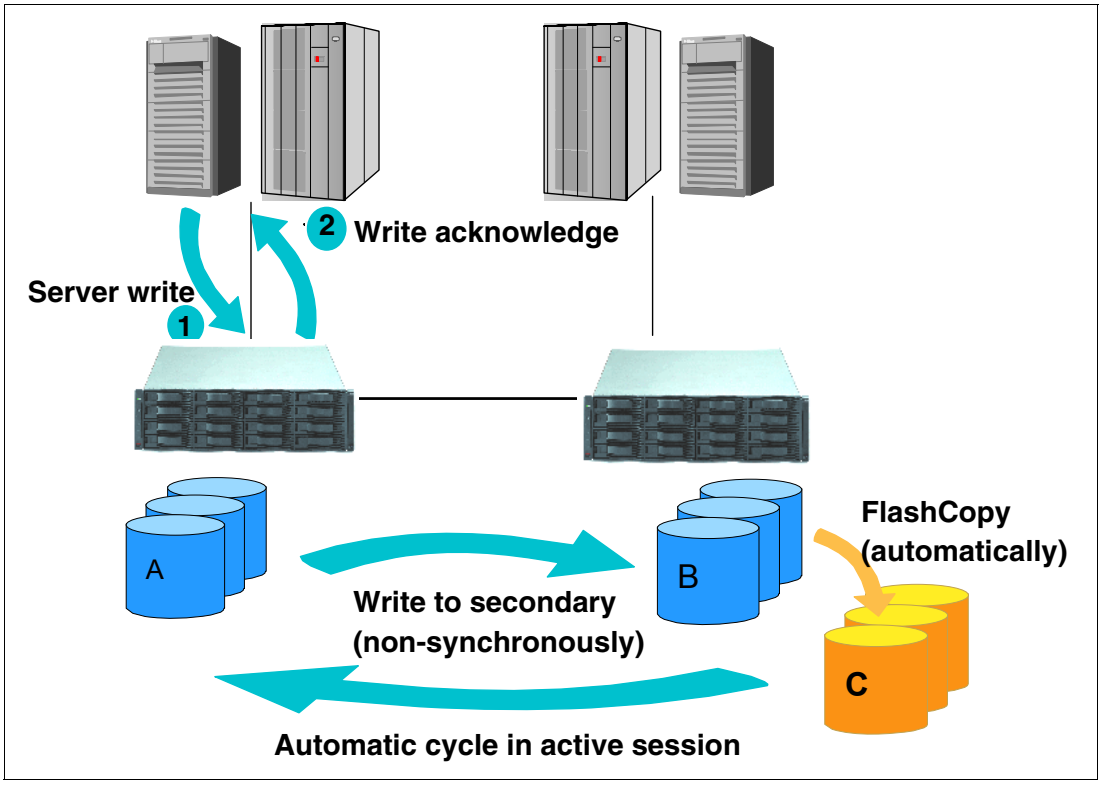



*Figure 6-9 Global Mirror*

### *How Global Mirror works*

We explain how Global Mirror works in Figure 6-10 on page 101.

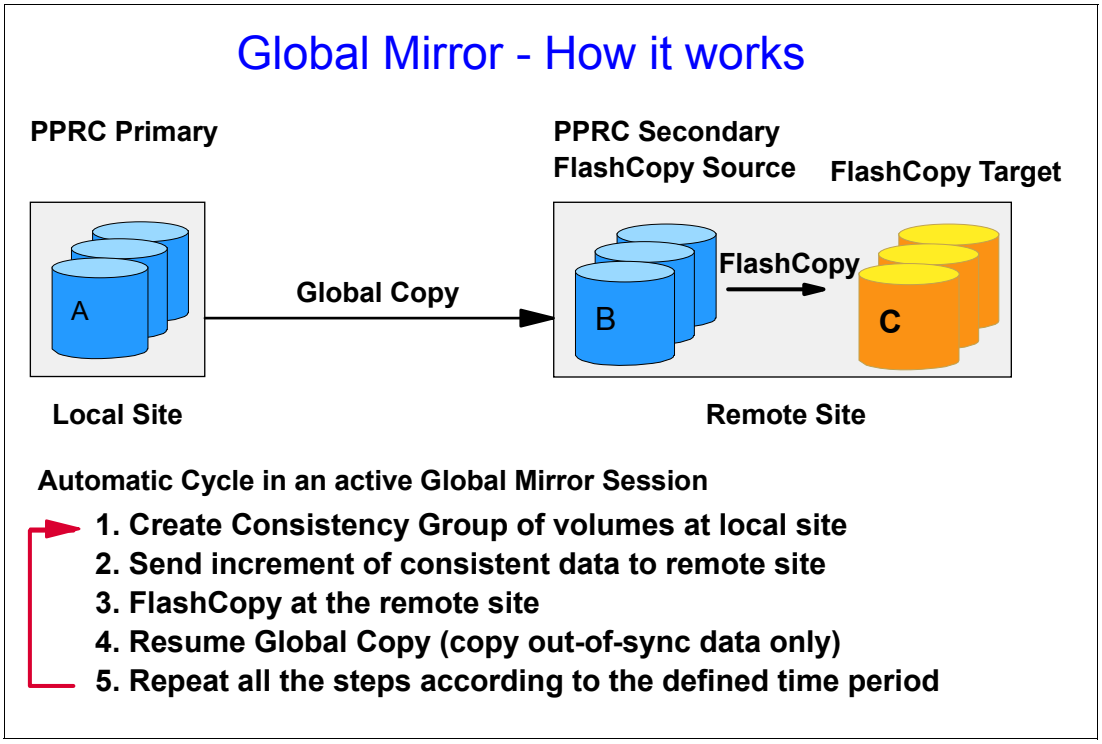



*Figure 6-10 How Global Mirror works*

The A volumes at the local site are the production volumes; they 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 Storage Units. 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 setup 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 Utility), 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)

z/OS Global Mirror is an asynchronous copy function for the z/Series environment. This function has a different architecture than Global Mirror. The DS6000 can only be used as a secondary system for z/OS Global Mirror (it cannot be used as primary system).

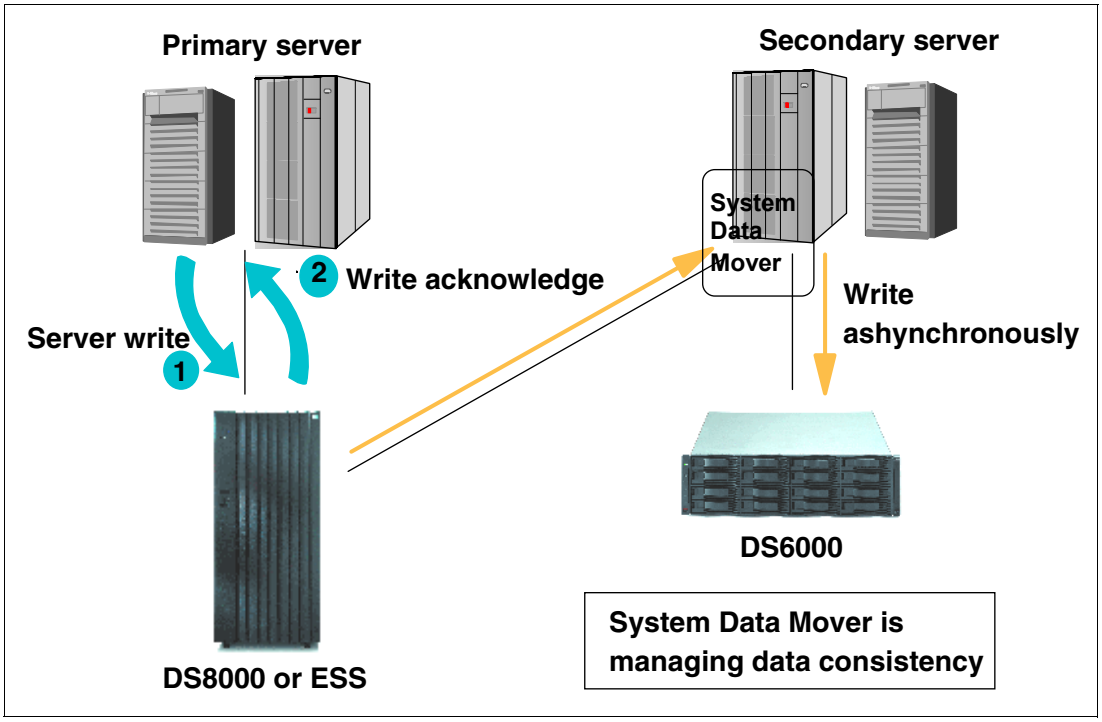



*Figure 6-11 z/OS Global Mirror (DS6000 is used as secondary system)*

## 6.2.4 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 PPRC for mirroring, you need to compare its function with OS mirroring.

Generally speaking, you will have some disruption to recover your system with PPRC secondary volumes in an open systems environment because PPRC secondary volumes are not online to the application servers during the PPRC relationship.

You may 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 PPRC for the mirroring in the local site only, 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 operate to 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 recreate should the system need to be recovered.

- Advantages

  Global Mirror can copy over nearly an unlimited distance. It is scalable across the server enclosures. It can realize low RPO with enough link bandwidth. Global Mirror causes little or no 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.

## 6.2.5 What is 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 the data contains any of these 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 operations, 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 withdrawal 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 an *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 6-12. In this case, three write operations (1st, 2nd, and 3rd) are dependent writes. This means that these operations must be completed sequentially.

.

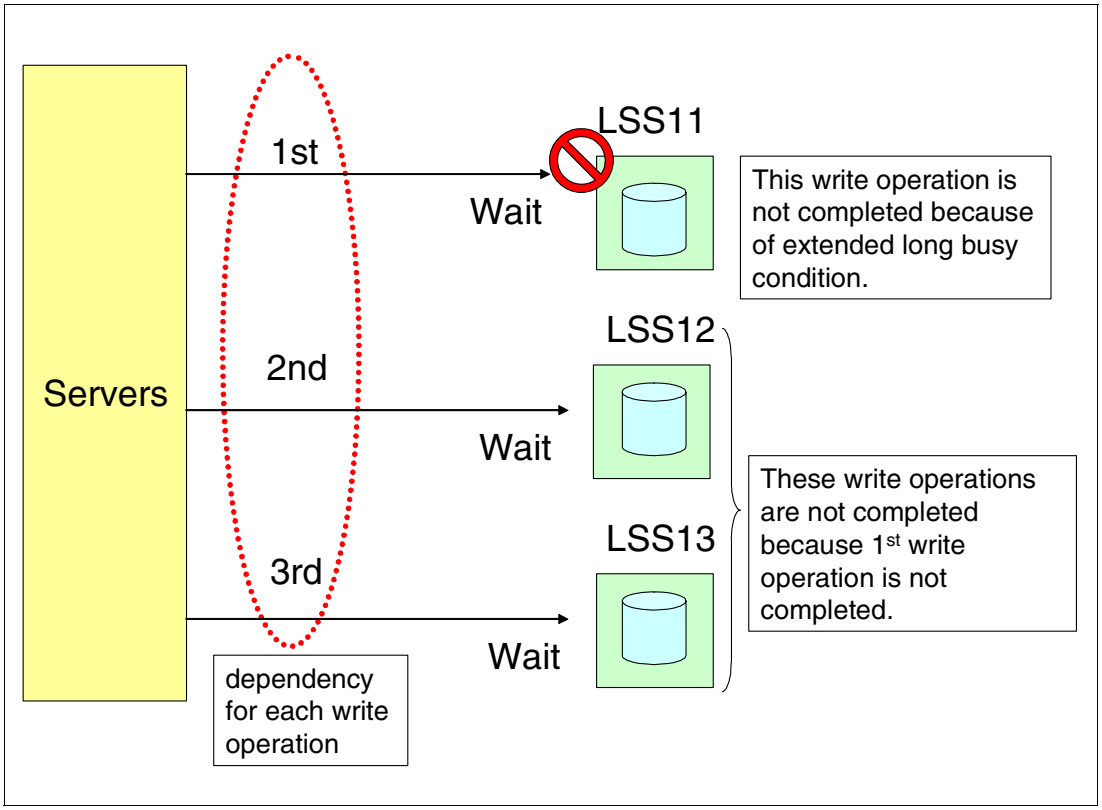



*Figure 6-12 Consistency Group: Example 1*

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 6-12, 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 6-13. 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 an 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.

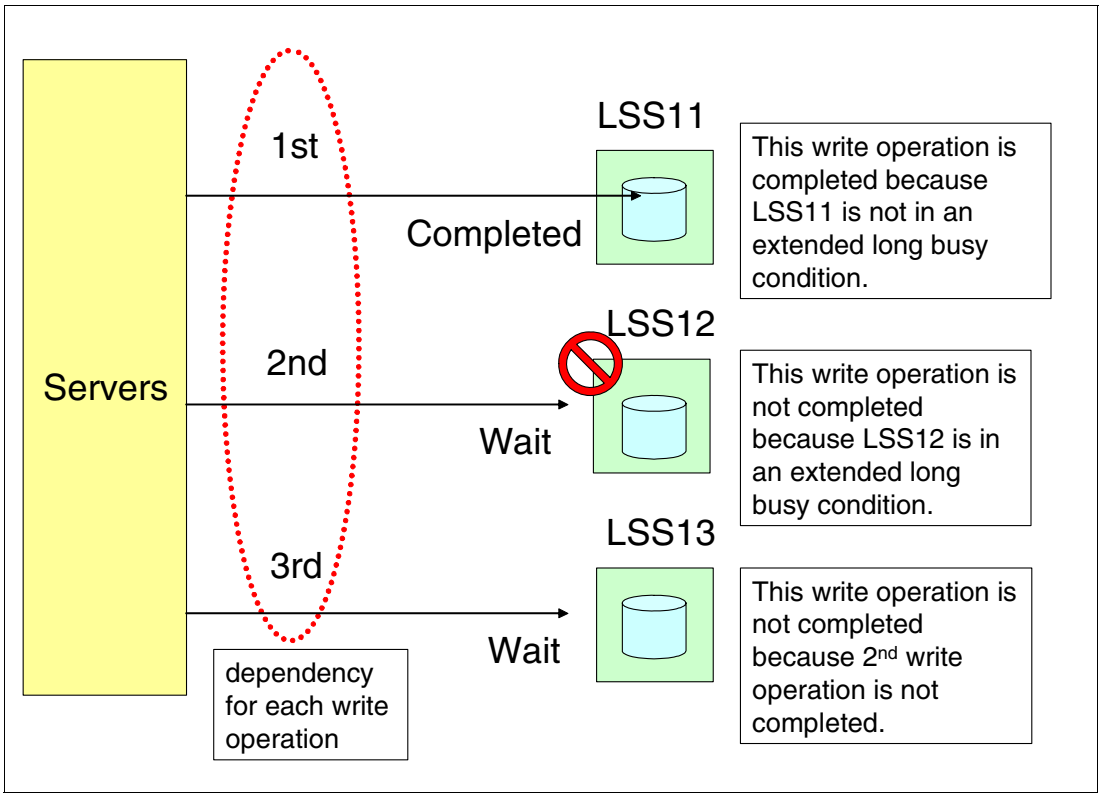



*Figure 6-13 Consistency Group: Example 2*

In all cases, if each write operation is dependent, Consistency Group operations can keep 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 operations. See Figure 6-14 on page 108. 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.

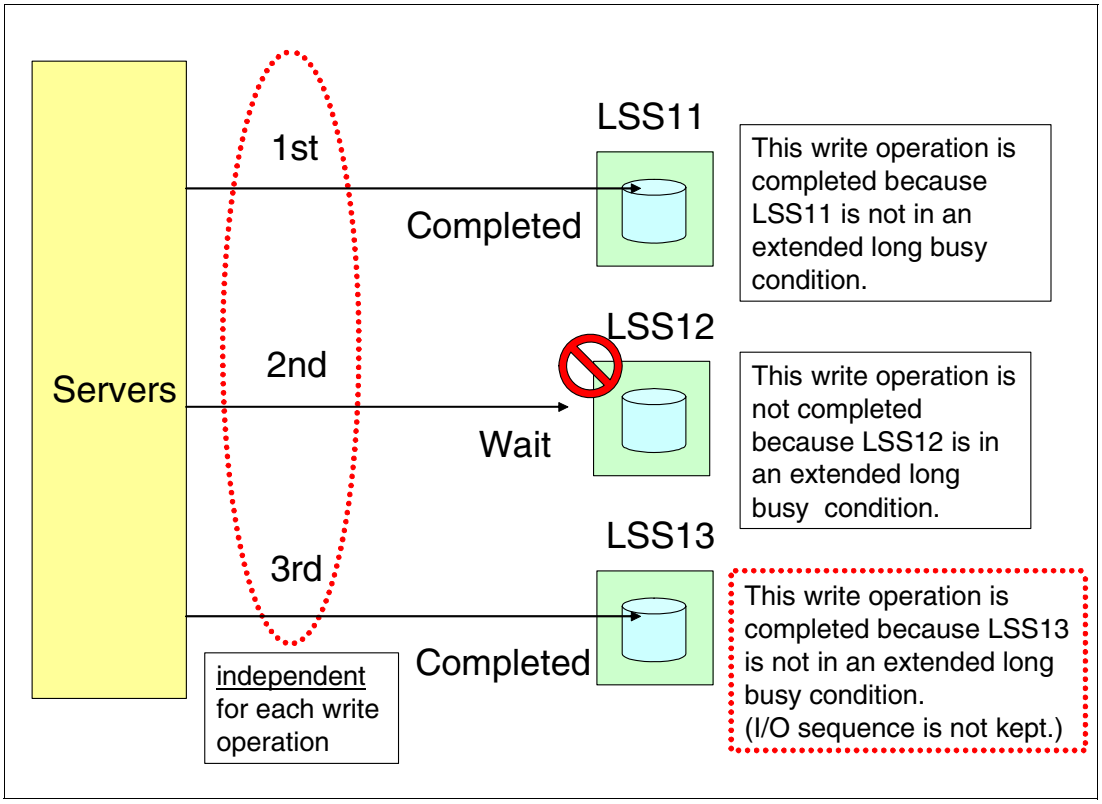



*Figure 6-14 Consistency Group: Example 3*

In this case, the copy created by Consistency Group operation reflects only the 1st and 3rd write operation, not including the 2nd operation.

If you accept this result, you can use Consistency Group operation with your applications. But, if you cannot accept it, you should consider other procedures without Consistency Group operations. For example, you could stop your applications for a slight interval for the backup operations.

## 6.3 Interfaces for Copy Services

There are multiple interfaces for invoking Copy Services. We describe them in this section.

### 6.3.1 DS Management Console

Copy Service 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 DS Management Console (MC). When the MC has the command requests, including those for Copy Services, from these interfaces, the

MC communicates with each server in the storage units via the Ethernet network. Therefore, the MC is a key component to configure and manage the DS6000.

The client must provide a computer to use as the MC. If they want, they can order a computer from IBM as the MC. An additional MC can be provided for redundancy.

For further information about the Management Console, see Chapter 8, “Configuration planning” on page 125.

### 6.3.2 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.

The client will install the DS Storage Manager on the Management Console. You can also install it in other computers that you prepare. When you manage Copy Services with DS Storage Manager in your computers, DS Storage Manager issues its commands to the Management Console 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

### 6.3.3 DS Command-Line Interface (CLI)

The IBM TotalStorage DS Command-Line Interface (CLI) helps enable open systems hosts to invoke and manage FlashCopy and Metro and Global Mirror functions through batch processes and scripts. The CLI provides a full-function command set that allows you to check your storage 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 configuration.
- Check the current Copy Services configuration that is used by the server enclosure.
- Create new logical volumes and Copy Services configuration settings.
- Modify or delete logical volumes and Copy Services configuration settings.

**Tip:** ***What is changed from the ESS CLI?***

There are differences between the ESS CLI and DS CLI.

The ESS CLI needs two steps to issue Copy Service functions:

1. Register 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 10, “DS CLI” on page 195.

### 6.3.4 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 the DS6000 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 DS6000.

IBM will support IBM TotalStorage Multiple Device Manager (MDM) for the DS6000 under the IBM TotalStorage Productivity Center in the future. MDM consists of software components that 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 14.5, “IBM TotalStorage Productivity Center” on page 282.

## 6.4 Interoperability with ESS

Copy Services also supports the IBM Enterprise Storage Server Model 800 (ESS 800) and the ESS 750. To manage the ESS 800 from the Copy Services for DS6000, you need to install licensed internal code version 2.4.2 or later on the ESS 800.

The DS CLI supports the DS6000, DS8000, and ESS 800 at the same time. The DS Storage Manager does not support the ESS 800.

**Note:** The DS6000 does not support PPRC via an ESCON link. If you want to configure a PPRC relationship between a DS6000 and ESS 800, you have to use an FCP link.

## 6.5 Future Plan

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.*

# Part 3

# Planning and configuration

In this part we present an overview of the planning and configuration necessary before installing your DS6000. The topics include:

- Installation planning
- Configuration planning
- Logical configuration
- Command-Line Interface
- Performance

7

# Installation planning

This chapter discusses planning for the physical installation of a new DS6000 in your environment. Refer to the latest version of the *IBM TotalStorage DS6000 Introduction and Planning Guide*, GC26-7679, for further details.

In this chapter we cover the following topics:

- General considerations
- Installation site preparation
- Management interfaces
- Network settings
- SAN requirements and considerations
- Software requirements

## 7.1 General considerations

The successful installation of a DS6000 requires careful planning. The main considerations when planning for the physical installation of a new DS6000 are the following:

- Floor loading
- Floor space
- Electrical power
- Operating environment
- Cooling
- Management console
- Host attachment and cabling
- Network and SAN considerations

Always refer to the most recent information for physical planning in the *IBM TotalStorage DS6000 Introduction and Planning Guide*, GC26-7679.

## 7.2 Installation site preparation

Before you begin to install a new DS6000, you must ensure that the location where you plan to install your DS6000 storage units meets all requirements.

You can install the DS6000 series in a 2101-200 system rack or in any other 19” rack that is compliant with the Electronic Industries Association (EIA) 310-D Type A standard.

You have to use these feature codes when you order a system rack from IBM for your DS6000 series:

- The feature code 0800 is used to indicate that the DS6000 series ordered will be assembled into an IBM TotalStorage 2101-200 System Rack by IBM manufacturing.
- The feature code 0801 is used to indicate that the DS6000 series ordered will be shipped as an assembled enclosure for field integration into a supported rack enclosure. Supported rack enclosures include the IBM 7014 RS/6000® Rack and the IBM 9308 Netfinity® Enterprise Rack. Field integration of the DS6000 series is customer setup, unless the DS6000 Installation Services are utilized.

The next topics in this section discuss how you prepare the installation site to meet all of these requirements.

### 7.2.1 Floor and space requirements

When you are planning the location of your storage units, you need to perform the following steps to ensure that your planned installation location meets space and floor load requirements:

1. Determine the number of server enclosures and expansion enclosures that are included in your order.
2. Decide whether the storage units will come within a 2101-200 rack or will be installed in an existing rack.
   a. If the storage units come in a rack, plan where the floor tiles must be cut to accommodate the cables.
   b. If the storage units will be installed in an existing rack, ensure that there is enough space for cable exits and routing.

3. Ensure that the floor area provides enough stability to support the weight of the fully configured DS6000 series and associated components.
4. Ensure that you have adequate rack space for your hardware. Calculate the amount of space that the storage units will use. Don't forget to leave enough space for future upgrades.
5. Ensure that your racks have space for the clearances that the DS6000 series requires for service and the requirements for floor loading strength.

## Floor load requirements

It is very important that your location meets the floor load requirements. This section provides information you need to ensure that your physical site meets the installation requirements for the DS6000 series.

Table 7-1 gives the dimensions and weight of the DS6800. The DS6000 expansion enclosure is of the same size.

*Table 7-1 DS6800 dimensions and weight*

| Height | Width | Depth | Maximum weight (fully configured) |
|---|---|---|---|
| 5.25 inches (0.134 meters) | 18.80 inches (0.478 meters) | 24.00 inches (0.610 meters) | 109 lbs. (49.6 kg) |

Table 7-2 describes the DS6800 dimensions and weight within a 2101-200 rack.

*Table 7-2 DS6800 dimensions and weight within a 2101-200 rack*

| Height | Width | Depth | Maximum weight (fully configured) |
|---|---|---|---|
| 71.00 inches (1.804 meters) | 25.40 inches (0.644 meters) | 43.30 inches (1.908 meters) | 2034 lbs. (904 kg)) |

1. Find out the floor load rating of the location where you plan to install the storage units.

> **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.

2. Determine whether the floor load rating of the location meets the following requirements:
   – The minimum floor load rating used by IBM is 342 kg per sq m (70 lb per sq ft).

## Service clearance requirements

This section describes the clearances that the DS6000 series requires for service. We include the clearances that are required on the front and to the rear of the rack.

Table 7-3 shows the clearances for the DS6000 series.

*Table 7-3 DS6800 or the DS6000 expansion enclosure clearance*

| Front | Rear |
|---|---|
| 12.00 inches<br>0.305 m | 18.00 inches<br>0.457 m |

Use the following steps to calculate the required space for your storage units:

1. Determine the dimensions of each model configuration in your storage units.
2. Determine the total space that is needed for the storage units by planning where you will place each storage unit in the rack.
3. Verify that the planned space and layout also meets the service clearance requirements for each unit.

## 7.2.2 Power requirements

We describe here the input voltage requirements for the DS6000 series.

### Input voltage requirements

The DS6000 series has built-in redundant, auto-sensing, auto-ranging power supplies. The power supplies are designed for operation in a voltage range of 90-257 V AC, 50-60 Hz.

Table 7-4 lists the input voltages and frequencies that the DS6000 series power line cords support. The values apply to both of the primary line cords to any storage or expansion enclosure in a DS6000 series. The DS6000 series power inputs are single phase.

*Table 7-4 DS6000 series input voltage requirements*

| Characteristic | Value |
|---|---|
| Nominal input voltages | 100-127 RMS V AC<br>200-240 RMS V AC |
| Minimum input voltage | 90 RMS V AC |
| Maximum input voltage | 264 RMS V AC |
| Input frequencies | 50 ± 3.0 Hz<br>60 ± 3.0 Hz |

Two independent power outlets for the two DS6000 power line cords are needed by each base model and expansion model.

**Important:** 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.

## 7.2.3 Environmental requirements

To properly maintain your DS6000 storage unit, you must install your storage unit in a location that meets the operating environment requirements.

Table 7-5 describes the environment operating requirements for the DS6000 series.

*Table 7-5 Operating environment*

| | |
|---|---|
| Powered on temperature limit | 10 - 40° C (50 - 104° F) |
| Powered off temperature limit | 10 - 52° C (50 - 126° F) |

| | |
|---|---|
| Maximum wet bulb temperature | 27° C (80° F) Note: The upper limit of wet bulb temperature must be lowered 1.0° degree C for every 274 meters of elevation above 305 meters. |
| Relative humidity | 8 - 80 percent |
| Typical heat load | 550 watts or 1880 Btu/hr |
| Noise level | 5.9 bels |

The DS6000 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).

### 7.2.4 Preparing the rack

The DS6000 series requires an Electronic Industries Association (EIA) 310-D Type A 19-inch rack cabinet. The distance between EIA rails, from the front to the rear of the rack, is 60.96 centimeters (24 inches) minimum and 81.28 centimeters (32 inches) maximum. This rack conforms to the EIA standard. Where you place the support rails in the rack depends on where you intend to position the server or storage enclosure.

Before you install the DS6000 series in a rack, keep in mind the following considerations:

- Review the documentation that comes with your rack enclosure for safety and cabling considerations.
- Install the DS6000 series in a recommended 22° C (72° F) environment.
- To ensure rack stability, load the rack starting at the bottom.
- If you install multiple components in the rack, do not overload the power outlets.
- Always connect the storage server to a properly grounded outlet.
- Always connect the rack power to at least two different power circuits or sources. This reduces the chance of a simultaneous loss of both AC power sources.

## 7.3 System management interfaces

The DS6000 series provides the following management interfaces:

- IBM TotalStorage DS Storage Manager
- DS Open application programming interface
- DS Command-Line Interface (CLI)

You have to decide where you want to install this software.

### 7.3.1 IBM TotalStorage DS Storage Manager

The IBM TotalStorage DS Storage Manager is an interface that is used to perform logical configurations, service, copy services management, and firmware upgrades.

The DS Storage Manager software must be installed on a user-provided computer. We refer to this computer as the DS Management Console. For more information on the supported operating systems and the minimum hardware requirements see Chapter 8, “Configuration planning” on page 125.

The DS Storage Manager can be accessed from any location that has network access to the DS management console using a Web browser.

Since the DS Storage Manager is required to manage the DS6000, to perform Copy Services operations, or to call home to a service provider, the DS management console should always be on.

For more information about the DS Storage Manager refer to the *DS6000 Installation, Troubleshooting, and Recovery Guide*, GC26-7678.

**Note:** The management console hardware is not part of the DS6000, it has to be ordered separately.

### 7.3.2 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 arrays. The DS Open API also enables Copy Services functions such as FlashCopy and Remote Mirror and Copy (formally known as Peer-to-Peer Remote Copy).

The DS Open API helps integrate DS configuration management support into storage resource management (SRM) applications, which allow our 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.

If you have an application that needs the DS Open API, you have to install the API and you have to decide where you want to install it.

For more information on the DS Open API refer to *DS Open Application Programming Interface Reference, GC35-0493.*

### 7.3.3 DS Command-Line Interface

The IBM TotalStorage DS Command-Line Interface (DS CLI) provides a full-function command set that allows you to check your storage unit configuration, configure the storage unit, manage (create and delete) logical volumes, and perform other specific application functions when necessary.

The DS CLI can be used to manage FlashCopy, Metro Mirror, Global Copy, and Global Mirror functions through batch processes and scripts.

Part of the installation planning is to decide where to install the DS Command-Line Interface. The DS CLI is supported on several operating systems. For more information on the DS CLI and the supported environments see Chapter 10, “DS CLI” on page 195.

For detailed information refer to the *TotalStorage DS6000 Command-Line Interface User’s Guide*, GC26-7681

## 7.4 Network settings

To install a DS6000 in your environment you have to plan for the Ethernet infrastructure that the DS6000 has to be connected to. You have to provide some TCP/IP addresses and you need an Ethernet switch or some free ports on a existing switch (see Figure 7-1).

This following settings are required connect the DS6000 series to a network:

- Controller card IP address

  You must provide a dotted decimal address that you will assign to each storage server controller card in the DS6800. Since there are two controllers, you need two TCP/IP addresses.

- DS Storage Manager IP address

  You need another TCP/IP address for the computer where you run the DS Storage Manager.

- Gateway

  Provide the dotted decimal or symbolic name address of the gateway (for example, 9.123.123.123 or sanjosegate).

- Subnet mask

  Provide the dotted decimal address of the subnet (network) mask.

- Primary domain name server (DNS)

  Provide the Host name and IP address if you are using a domain name server.

- Alternate domain name server (DNS)

  You can optionally provide an alternate DNS.

- Simple Network Management Protocol (SNMP) destination

  Provide the host names or the dotted decimal addresses of the destinations that are to receive SNMP traps (for example, host.com or 9.123.123.123).

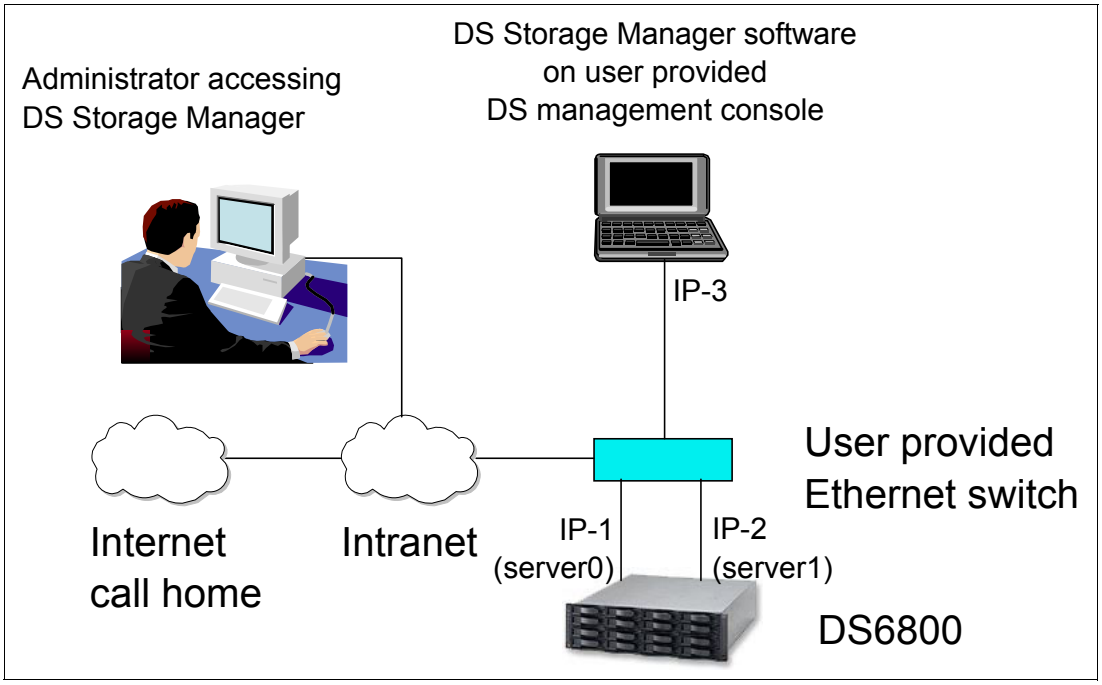



*Figure 7-1 Network requirements for a DS6000 implementation*

## 7.5 SAN requirements and considerations

The DS6000 series provides a variety of host attachments so that you can consolidate storage capacity and workloads for open systems hosts and eServer zSeries hosts. The Fibre Channel adapters of the storage system can be configured to support the Fibre Channel Protocol (FCP) and fibre connection (FICON) protocol.

You should also keep the following consideration 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.

Single Fibre Channel host adapters can have physical access to multiple Fibre Channel ports on the storage unit.

Part of the installation planning process is the planning for your SAN. You may already have a SAN infrastructure with enough unused ports to connect a DS6000, or you might need additional Fibre Channel switches. If you have switches that operate at 1 Gbps, you might consider replacing them with 2 Gbps switches. For high availability, you should connect the DS6000 at least to two different switches (see Figure 7-2).

The same considerations mentioned also apply to HBAs on the host. You might want to replace 1 Gbps HBAs by 2 Gbps HBAs. For fault tolerance, each server should be equipped with two HBAs connected to different Fibre Channel switches.

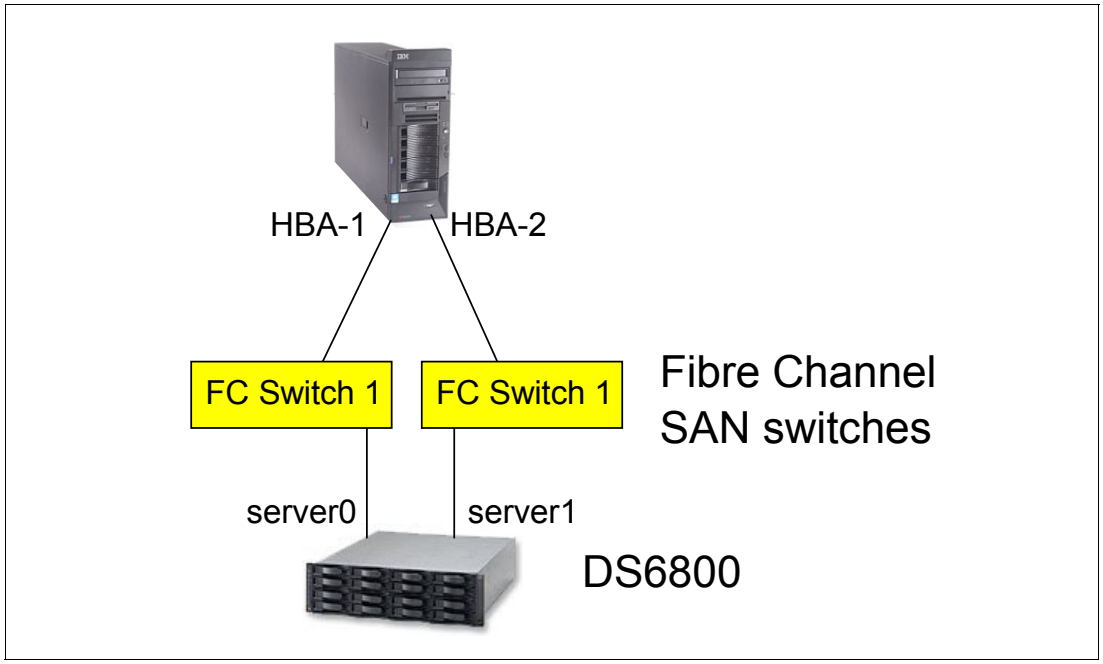



*Figure 7-2 SAN environment*

For 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.

Part of the installation planning is to decide what SAN zones are needed or have to be modified.

### 7.5.1 Attaching to an Open System host

Fibre Channel technology supports increased performance, scalability, availability, and distance for attaching storage subsystems to servers as compared to SCSI attachments. Fibre Channel technology supports sharing of large amounts of disk storage between many servers.

Fibre Channel ports on the DS6000 can be shared between different Host Bus Adapters (HBAs) and operating systems. For more information about supported host systems, operating system levels, host bus adapters, and fabric support, see the *Interoperability Matrix* at:

http://www.ibm.com/servers/storage/disk/ds6000/interop.html

**Attention:** SCSI adapters are not available on the DS6000.

### 7.5.2 FICON-attached zSeries Host

The DS6000 series supports FICON-attached zSeries hosts.

The storage unit attaches to S/390 and zSeries host systems with FICON channels. Each of the four Fibre Channel ports on each controller can be configured to operate in FICON mode. When configured for FICON, the Fibre Channel port supports connections to a maximum of 128 FICON hosts. When you configure a FICON port on one of the two controller cards, you should also configure a FICON port on the second controller card to allow failover from one server to the other to work. A FICON port on the DS6000 can be connected directly to a zSeries host or to a SAN switch supporting FICON connections. With Fibre Channel adapters configured for FICON, the storage unit supports the following:

- Either fabric or point-to-point topologies
- A maximum of 128 host connections per Fibre Channel port
- A maximum of 32 logical subsystems
- A maximum of 8192 logical volumes
- A maximum of 1024 logical paths per FICON port

The storage unit supports the following operating systems for zSeries hosts:

- Transaction Processing Facility (TPF)
- Virtual Storage Extended/Enterprise Storage Architecture (VSE/ESA)
- z/OS
- z/VM

Note, that Linux on zSeries currently supports only FCP-attached disks and not FICON-attached disks.

FICON is an optional feature on the DS6800 system, and is available with the FICON Attachment feature number 5915.

For details about models, versions of operating systems, and releases that the storage unit supports for these host systems, see the *Interoperability Matrix* at:

http://www.ibm.com/servers/storage/disk/ds6000/interop.html

**Attention:** The DS6000 series does not support ESCON attachment.

## 7.6 Software requirements

To see current information on servers, operating systems, host adapters, and connectivity products supported by the DS6000 series, review the Interoperability Matrix at the following DS6000 series Web site:

http://www.ibm.com/servers/storage/disk/ds6000/interop.html

### 7.6.1 Licensed features

Before you can configure your DS6000 series you must activate your licensed features to enable the functions purchased on your machine.

The DS6000 series is licensed at the following levels:

- **Machine licensing** uses licensed machine code (LMC) to activate base functions on your machine. When you receive a DS6800 or DS6000 expansion enclosure, or both, you receive an LMC agreement. The use of the machine constitutes acceptance of the license terms outlined in the LMC agreement. Some DS6000 series features and functions may not be available or supported in all environments. Current information on supported environments, prerequisites, and minimum operating systems levels is available at:

  http://www.ibm.com/servers/storage/disk/ds6000/interop.html

- **Operating environment licensing** (OEL) manages the machine operating environment and is required on every DS6800 system. The extent of IBM authorization acquired through the DS6800 feature numbers (for example: feature code 5001 = OEL 5 TB unit) must cover the physical capacity of the DS6800 system, where system is defined as the base enclosure and all attached expansion enclosures.

  **Note:** If the operating environment license has not been acquired and activated on the machine, disk drives installed within the DS6800 system cannot be logically configured for use. Upon activation, disk drives can be logically configured up to the extent of authorization.

  As additional disk drives are installed, the extent of IBM authorization must be increased by acquiring additional DS6800 feature numbers. Otherwise, the additional disk drives cannot be logically configured for use.

- **Feature licensing** controls the licenses of features of each DS6800. Each DS6800 licensed function feature number enables the use of, and establishes the extent of, IBM authorization for licensed functions acquired for a DS6800 system.

Each licensed function feature number is applicable only for the specific DS6800 (by serial number) for which it was acquired and is not transferable to another serial numbered DS6800.

For feature code information, refer to the *IBM TotalStorage DS6000 Introduction and Planning Guide*, GC26-7679.

To activate the feature licenses for your DS6000 series, you must access the Disk Storage Feature Activation (DSFA) application from the IBM Web site:

http://www.ibm.com/storage/dsfa

8

# Configuration planning

This chapter discusses configuration planning considerations when implementing the DS6000 series in your environment. The topics covered are:

- Configuration planning
- DS6000 Management Console
- DS6000 license functions
- Data migration planning
- Planning for performance

## 8.1 Configuration planning considerations

When installing a DS6000 disk system, various physical requirements need to be addressed to successfully integrate the DS6000 into your existing environment. These requirements include:

- The DS6800 requires Licensed Machine Code (LMC) level 5.0.0.0 or later.
- Appropriate operating environment characteristics such as temperature, humidity, electrical power, and noise levels.
- A PC, the DS Management Console, to host the DS Management Console (MC). This will be the focal point for configuration, Copy Services management, and maintenance for the DS6000 series.
- Input voltages and frequencies that the DS6000 series power line cords support.
- An Electronic Industries Association (EIA) 310-D Type A 19-inch rack cabinet.
- Processor memory refers to the amount of memory you want for the processors on your model. The DS6800 has 4 GB processor memory.
- Licensed functions include both required and optional features such as the operating environment, Point-in-Time Copy, Remote Mirror and Copy and Parallel Access Volumes. See 8.3, “DS6000 licensed functions” on page 131, for a in-depth discussion of the licensed functions.
- Delivery requirements to receive delivery of the DS6000 at your location.
- Adequate floor space, power, environmentals, network, and communication.

For a complete discussion of these requirements, see the *IBM TotalStorage DS6000 Introduction and Planning Guide,* GC26-7679, or visit:

`http://www-1.ibm.com/servers/storage/disk/ds6000/index.html`

## 8.2 DS6000 Management Console

A management console is the configuration, service, and management portal for the DS6000 series. A management console is a requirement for a DS6800. The user must provide a computer to use as the management console or the user can optionally order a computer from IBM. This computer must meet the following minimum set of hardware and operating system compatibility 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 managed Integrated RAID Controller (IRC)
- IP Network connectivity to each RAID controller card

You may also want to have:

- IP network connectivity to an external network to enable call home and remote support
- Serial connectivity to the DS6000

The Management Console is supported on the following operating systems:

- Microsoft Windows 2000
- Microsoft Windows 2000 Server
- Microsoft Windows XP Professional
- Microsoft Windows 2003 Server
- Linux (Red Hat AS 2.1)

**Note:** If the user wants to order the management console, consider the IBM 8141 ThinkCentre M51 Model 23U (8141-23U) Desktop system with a 3.0 GHz/800 MHz Intel Pentium 4 Processor. If a monitor is also required, IBM suggests the IBM 6737 ThinkVison C170 (6737-66x or 6737-K6x) 17-inch full flat shadow mask CRT color monitor with a flat screen (16-inch viewable image size).

The DS Storage Manager software is provided with the DS6000 series at no additional charge. The DS Management Console gives the user the capability to perform the following tasks:

- Configuration management of the DS6000 series
- DS Management Console connectivity
- Local maintenance
- Copy Services management
- Remote service support and call home
- Event notification messaging

### 8.2.1 Configuration management of DS6000 system

The DS Storage Manager is a single integrated Web-based (HTML) graphical user interface (GUI) that offers:

- Offline configuration, which provides the user with the capability to create and save logical configurations while not connected to a DS6000. These logical configurations can then be applied to a DS6000 when the user connects to the appropriate DS6000 series.
- Online configuration, which provides real-time configuration management support.
- The ability to execute Copy Services functions.
- The capability for configuration, problem reporting, and maintenance.
- The ability to support multiple DS6000s.
- The capability to perform maintenance tasks such as downloading of code to the DS Management Console and activating licensed functions and features.

The DS6000 series can be configured by:

- Offline configuration
- Online configuration
- Express configuration

#### Offline configuration

Offline configuration allows the user to manipulate an existing configuration or start configuring a new configuration without being connected to the DS6000 series. In this way, the user can complete the configuration in disconnected mode and when ready, connect to a

new or un-configured DS6000 and apply the configuration. The following tasks may be performed when operating in offline configuration mode:

- Create or import a new simulated instance of a physical storage unit.
- Apply logical configurations to a new or fully deconfigured storage unit.
- View and change communication settings for the DS6000.
- From a single interface, work with new and view existing multiple DS6000s.
- Create, save, and open configuration documents for later reference and retention purposes.
- Print configuration reports.
- Export configuration data in a spreadsheet ready format.

### Online configuration

In online configuration mode, the user updates or views configurations in real time. In this mode, the user is connected to a DS6000 by way of direct or switched Ethernet connections. The following tasks may be performed when operating in online mode:

- Manage physical and logical configuration from one or more existing DS6000s across the network.
- Construct and apply valid logical configuration actions on new or fully deconfigured DS6000s at the time each action is initiated.
- Complete and apply valid logical configuration actions on existing storage plexes and storage units at the time each action is initiated.
- View and change communication settings for DS6000s.

### Express configuration

Express configuration provides the simplest and fastest method to configure a DS6000. The configuration methods other than Express configuration require more time to configure. This is due to the complex functions available to the user when configuring the DS6000. The user is required to make several decisions during the configuration process. Express configuration is ideal for:

- A novice user with little knowledge of storage concepts who wants to quickly and easily set up and begin using the storage.
- An expert user who wants to quickly configure a storage plex, allowing the storage server to make decisions for the best storage allocation.

The Express configuration allows you to:

- Configure fixed block and CKD volumes
- Create a volume group
- Create a host
- Map a volume group to a host attachment

The online configuration and Copy Services are available by way of a Web browser interface installed on the DS Management Console. All storage configuration operations are accomplished either through the DS6000 Storage Manager, through the DS6000 Command-Line Interface (CLI), or through any SMI-S compliant storage management agent, such as IBM's TotalStorage Productivity Center (TPC).

## 8.2.2 DS Management Console connectivity

Connectivity to the DS6000 series from the DS Management Console is needed to perform configuration updates to the DS6000 series. Connectivity to both DS6800 controllers is required to activate updates.

Figure 8-1 illustrates the DS Management console, optionally connected to to the DS6000 by redundant Ethernet switches, but redundant switches are not a requirement. The switches must be part of the same subnet so that the controllers in the DS6000 can communicate with each other through the Ethernet.

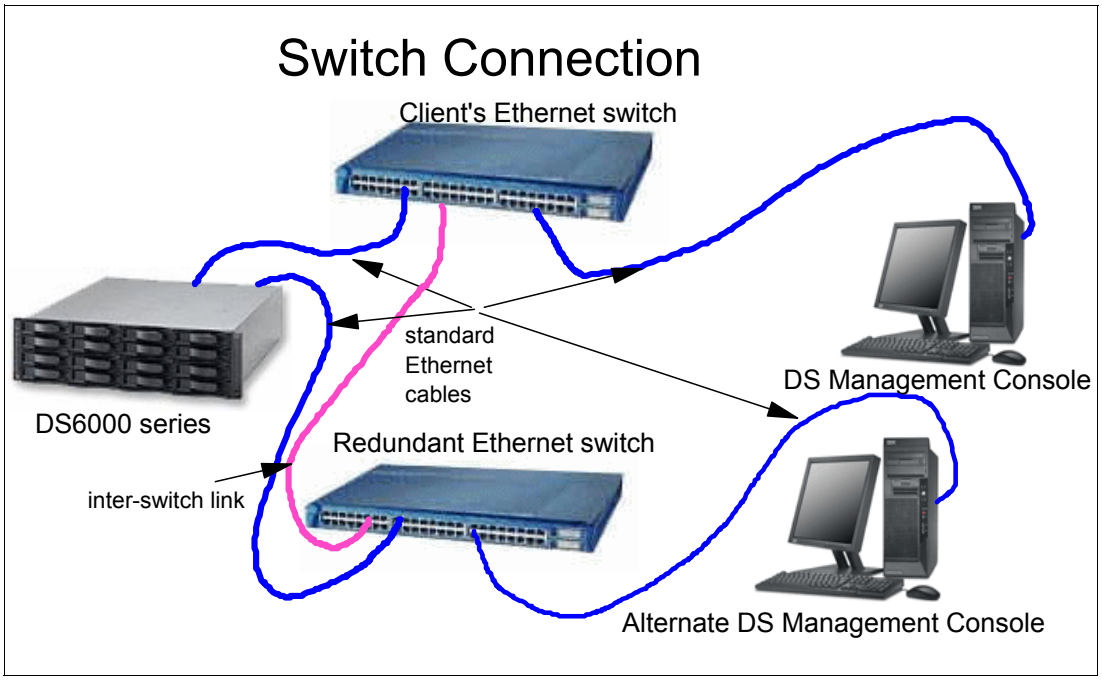



*Figure 8-1 Network configuration for DS Management Console*

## 8.2.3 Local maintenance

The DS6000 series is a customer maintained system. No IBM SSR (Service Support Representative) is involved, unless you have a service contract with IBM. The customer does problem determination, replacement parts ordering, and the actual replacement of the component. Most components are customer replaceable units (CRUs).

In the event of a failure, the following sequence of events takes place:

- A fault conditions occurs, for example, a power supply failure.
- The user is alerted about the failure.
- The user hyperlinks to the component view and identifies the failing component. In this case, it is a power supply.
- The user reviews component removal instructions.
- The user reviews the component installation instructions for the new power supply.
- Before removing a faulty component, ensure that you have a replacement component available on site.
- Remove the faulty power supply from the DS6000 as per the instructions.
- Install the new power supply.

- The DS6000 Storage Manager indicates the condition of the system after service.

Should the user require additional maintenance services, excluding standard maintenance services, IBM and IBM Business Partners have fee-based service offerings to fulfill such maintenance requirements.

## 8.2.4 Copy Services management

The DS6000 series may be part of a Remote Mirror and Copy configuration. In Metro Mirror and Global Mirror configurations the DS6000 series may be either primary or secondary. The DS6000 series cannot be a primary in a z/OS Global Mirror, but may be a secondary. The DS Management Console must be connected to both the primary and secondary to be able to perform Copy Services operations.

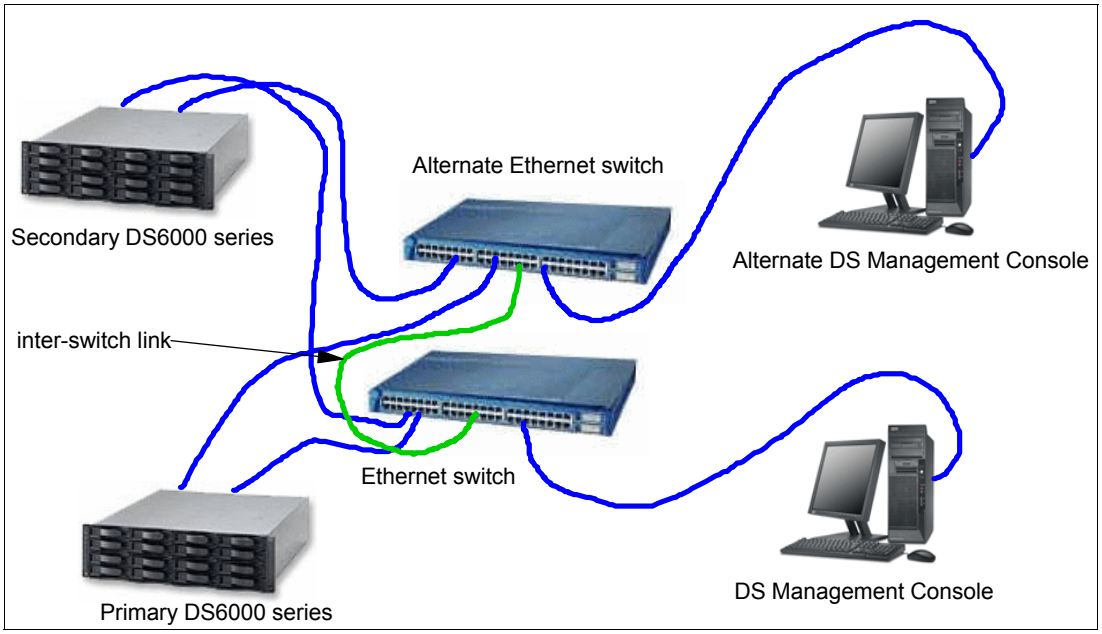



*Figure 8-2 DS Management Console in Copy Services environment*

The DS CLI will communicate with the DS Management Console when performing copy services commands, so the network connecting the management console to the DS6000 should be redundant.

## 8.2.5 Remote service support

In order to use remote service support, the user must allow a VPN connection. For example, Figure 8-3 on page 131 depicts an error condition occurring on the DS6000. The user's support services receives the alert over their local area network. If the user decides to ask IBM for help, and IBM needs to access the DS6000 remotely, then the user must provide connectivity from the DS Management Console to an IBM service support center. The DS Management Console must be connected to the DS6000 to enable IBM to access the DS6000 over this network to analyze the error condition.

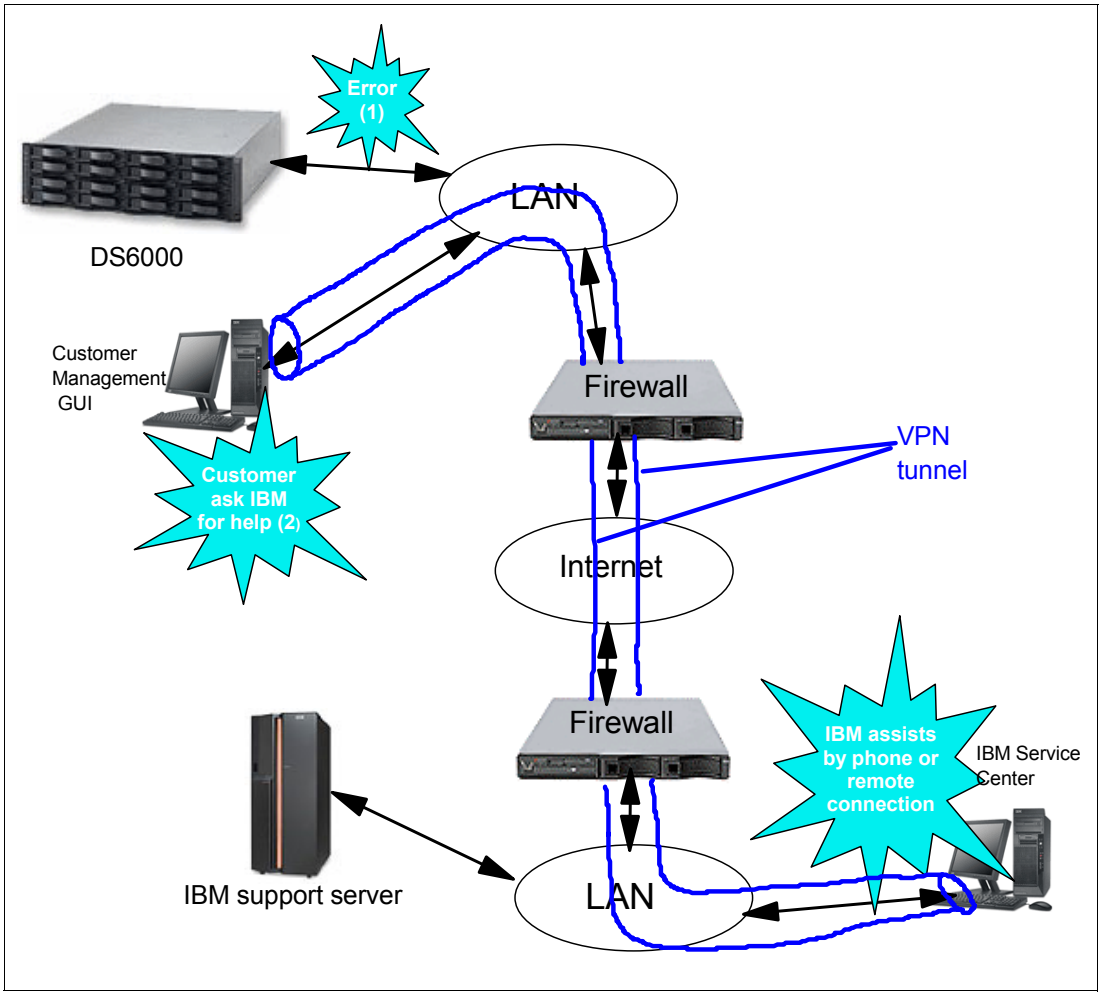



*Figure 8-3 DS6000 series configured for remote support*

### 8.2.6 Call home

The DS6000 has the capability to enable a call home facility. The call home facility is only available by way of e-mail. The DS Management Console must be connected to your local area network to be able to receive e-mail notification. On error identification, the user may decide to ask IBM for assistance.

### 8.2.7 Simple Network Management Protocol (SNMP)

The storage unit generates SNMP traps and supports a read-only management information base (MIB) to allow monitoring by your network. You have to decide which server is going to receive the SNMP traps.

## 8.3 DS6000 licensed functions

Licensed functions include the storage unit's operating system and functions. These include both required features and optional features.

### 8.3.1 Operating Environment License (OEL) - required feature

The user must order an operating environment license (OEL) feature, the IBM TotalStorage DS Operating Environment, for every DS6000 series. The operating environment model and features establish the extent of IBM authorization for the use of the IBM TotalStorage DS Operating Environment. The OEL licenses the operating environment and is based on the total physical capacity of the storage unit (DS6800 plus any DS6000 expansion enclosures). It authorizes you to use the model configuration at a given capacity level.

Once the OEL has been 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 DS Storage Manager. The feature activation process is discussed in more detail in 8.3.7, “Disk storage feature activation” on page 137.

Table 8-1 provides the feature codes for the operating environment licenses for the DS6000 series.

*Table 8-1 Operating environment license feature codes*

| Feature code | Description |
|---|---|
| 5000 | OEL-1TB |
| 5001 | OEL-5TB |
| 5002 | OEL-10TB |
| 5003 | OEL-25TB |
| 5004 | OEL-50TB |

Licensed functions are activated and enforced within a defined license scope. License scope refers to the following types of storage, and types of servers with which the function can be used:

- Fixed block (FB) - The function can only be used with data from Fibre Channel-attached servers.
- Count key data (CKD) - The function can only be used with data from FICON-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 a single license scope.

Table 8-2 provides the license scope options for each licensed function.

*Table 8-2 License scope for each DS6000 licensed function*

| Licensed function | License scope options |
|---|---|
| Operating environment | ALL |
| Point-in-Time Copy (PTC) | FB, CKD, or ALL |
| Remote Mirror and Copy (RMC) | FB, CKD, or ALL |
| PAV | CKD |
| FICON attachment | CKD |

The deactivation of an activated licensed function, or a lateral change or reduction in the licensed scope, is a disruptive activity and requires an initial microcode load (IML). A lateral change occurs when the user changes license scope from CKD to FB or FB to CKD. A reduction in license scope also occurs when the user decides to change from a license scope of ALL to CKD or FB.

### Optional features

The following optional features are available for the DS6000:

- 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
- Parallel Access Volumes
- FICON attachment

## 8.3.2 Point-in-Time Copy function (PTC)

The Point-in-Time Copy licensed function model and features establish the extent of IBM authorization for the use of 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 8-3 provides the feature codes for the Point-in-Time Copy function.

*Table 8-3 Point-in-Time Copy (PTC) feature codes*

| Feature code | Description |
|---|---|
| 5200 | PTC-1TB unit |
| 5201 | PTC-5TB unit |
| 5202 | PTC-10TB unit |
| 5203 | PTC-25TB unit |
| 5204 | PTC-50TB unit |

You can combine feature codes to order the exact capacity that you need. For example, if you determine that you need 23 TB of point-in-time capacity, you can order two 5202 features (this will give you 20TB) and three 5200 features (this will give you 3TB), giving you a total of 23TB.

## 8.3.3 Remote Mirror and Copy functions (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 8-4 provides the feature codes for the Remote Mirror and Copy functions.

*Table 8-4 Remote Mirror and Copy (RMC) feature codes*

| Feature code | Description |
|---|---|
| 5300 | RMC-1TB unit |
| 5301 | RMC-5TB unit |

| Feature code | Description |
|---|---|
| 5302 | RMC-10TB unit |
| 5303 | RMC-25TB unit |
| 5304 | RMC-50TB unit |

### 8.3.4 Parallel Access Volumes (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 8-5 provides the feature codes for the PAV function.

*Table 8-5 Parallel Access Volume (PAV) feature codes*

| Feature code | Description |
|---|---|
| 5100 | PAV-1TB unit |
| 5101 | PAV-5TB unit |
| 5102 | PAV-10TB unit |
| 5103 | PAV-25TB unit |
| 5104 | PAV-50TB unit |

PAV requires the purchase of the FICON attachment feature number 5915.

### 8.3.5 Server attachment license

If you order the PAV function, you must also order a server attachment license for your FICON attachment.

Table 8-6 provides the feature code for a FICON server attachment license.

*Table 8-6 Server attachment license (FICON)*

| Feature code | Description |
|---|---|
| 5915 | FICON attachment |

### 8.3.6 Ordering license functions

An OEL is required for every DS6000 series. This license is for the total physical capacity of the entire DS6000 series, including the DS6800 and DS6000 expansion units. For example, if the total capacity of the DS6000 series is 30 TB, then you will purchase an OEL feature for 30 TB.

Should you increase your capacity in the future, for example, from 30 TB to 40 TB, by adding additional disk drives, you then must purchase an additional 10 TB of OEL features, to be able to use the additional capacity. This activation will be non-disruptive.

The other optional features, such as PTC, RMC and PAV, can use any combination of the features to cover the required capacity for the DS6000. For example, if the DS6000 has a capacity of 20 TB of data and only 5 TB of the data will be used for CKD storage, and only the

CKD volumes will be FlashCopied, then you only need to purchase a license for 5 TB PTC and not the full 20 TB. The 5 TB of PTC would be set to CKD only.

Figure 8-4, 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.

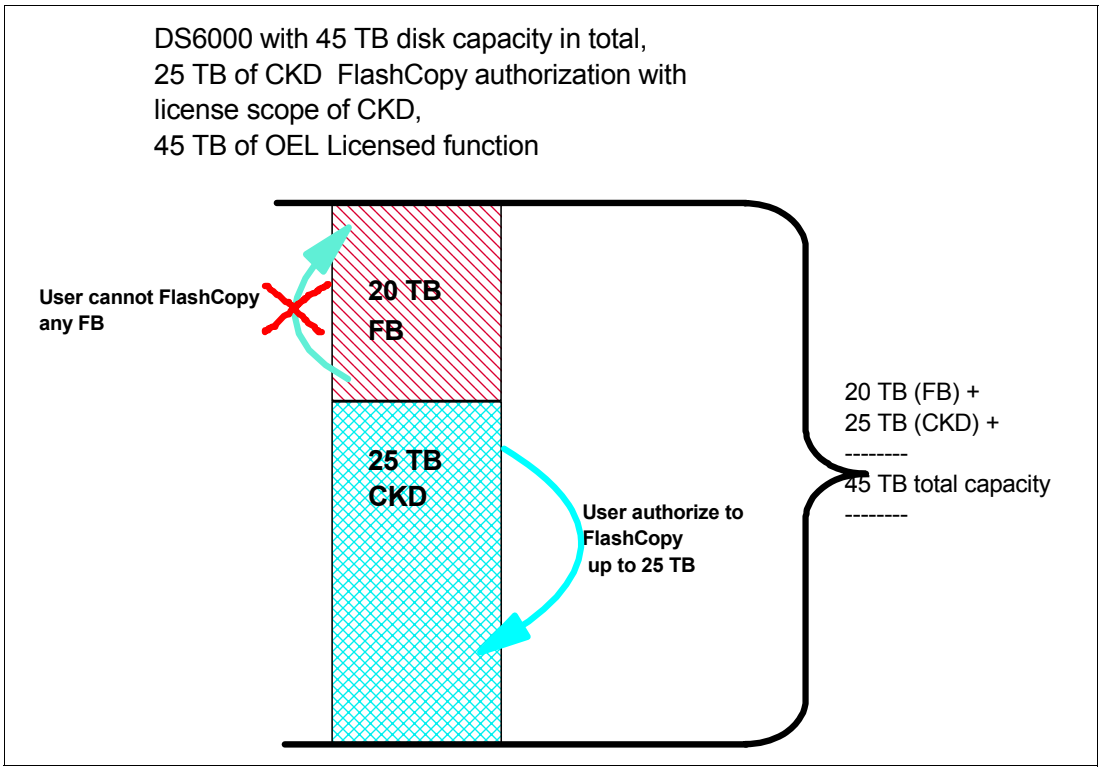



*Figure 8-4 User authorized to FlashCopy 25 TB of CKD data*

In Figure 8-5 on page 136, 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 user has to purchase authorization for the total FB and CKD capacity. In this case the total is the sum of 20 TB of FB and 25 TB of CKD, which equals 45 TB.

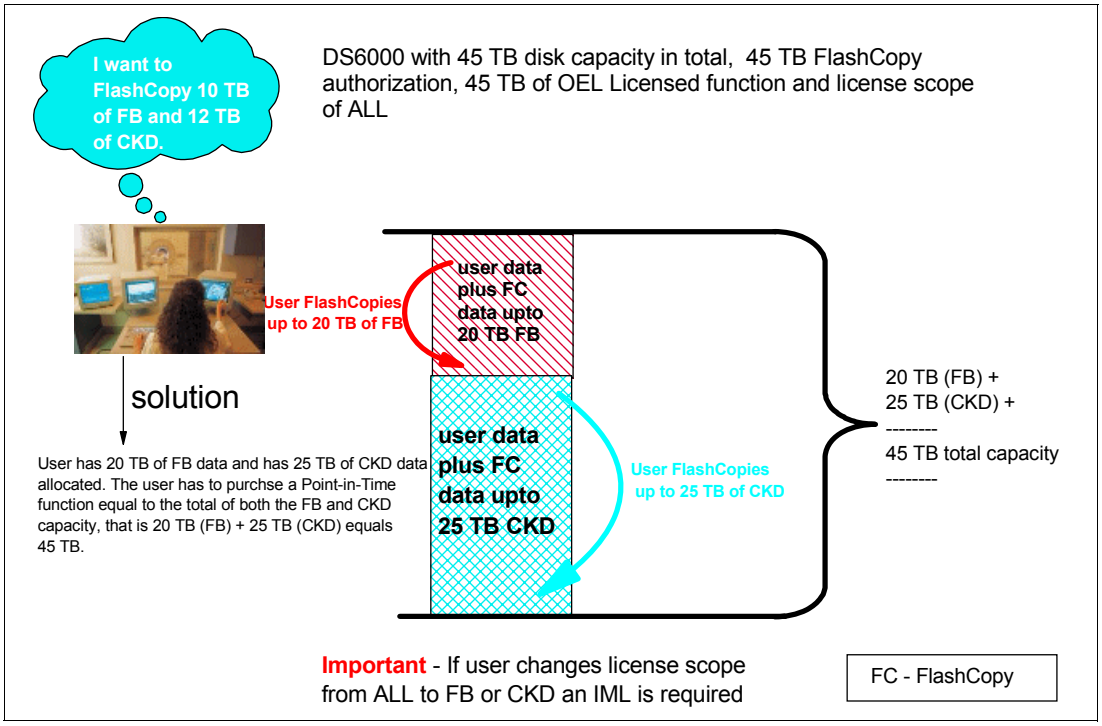



*Figure 8-5 User authorized to FlashCopy 45 TB of data with a license scope of ALL*

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 licensed feature for both the primary storage unit and the secondary storage unit.

Figure 8-6 on page 137 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 DS6000 and secondary DS6000.

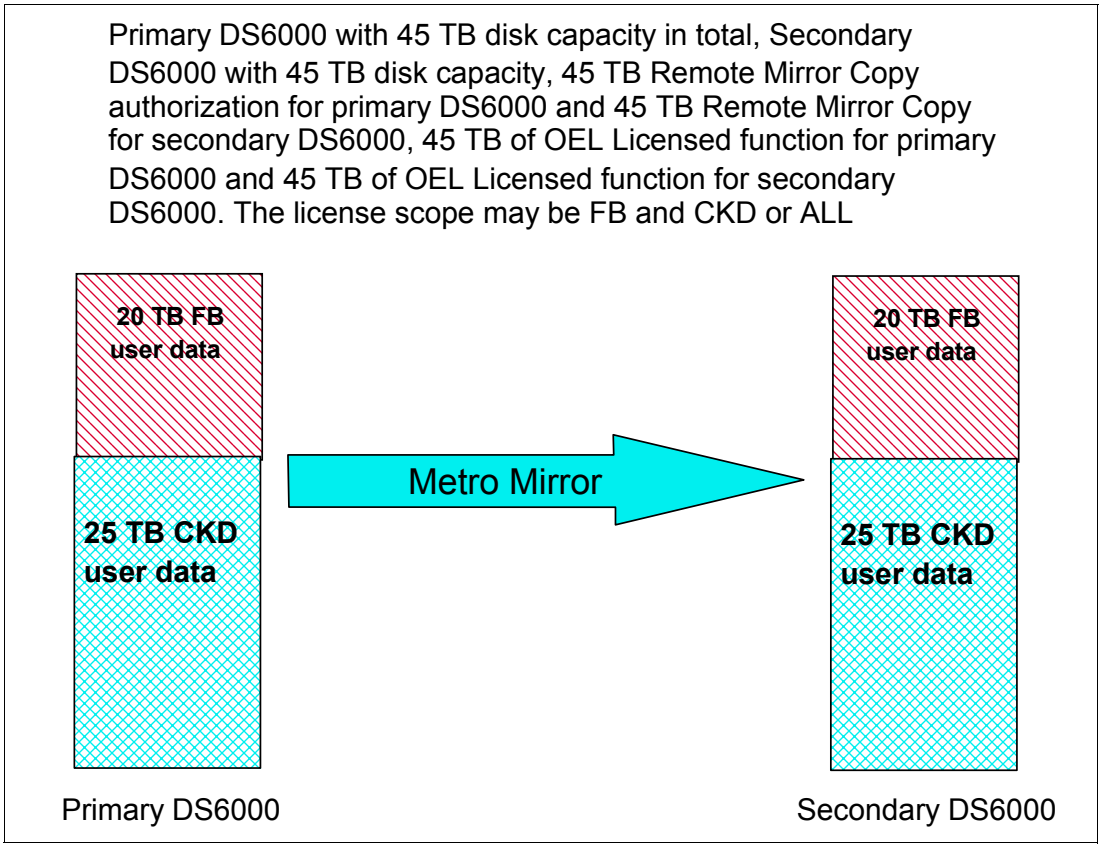



*Figure 8-6 Remote Mirror and Copy*

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 on the primary system.

PAV is only applicable to zSeries, so you will only have CKD data.

The initial enablement of any optional DS6000 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 DS6000 licensed function to deactivate the function is a disruptive activity and requires an IML.

## 8.3.7 Disk storage feature activation

Managing and activating licensed functions is the client's responsibility. The client can 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. The client performs these activities and then activates the function.

Activation refers to the retrieval and installation of the feature activation code into the DS6000 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; that is, model, serial number, and machine signature. This information is obtained from the DS Storage Manager.
- Log on to 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 on to your DS Storage Manager application and applying them to the storage unit.

## 8.4 Capacity planning

The DS6000 can have several kinds of disk drive modules (DDMs), from 8 up to 124 DDMs, and can use different types of RAID protection. The effective capacity depends on these factors. In this section, we explain how to estimate the effective capacity.

### 8.4.1 Physical configurations

The DS6000 series offers high scalability while maintaining excellent performance. With the DS6800 (Model 1750-511), you can install up to 16 DDMs in the server enclosure. The minimum storage capability with 8 DDMs is 584 GB (73 GB x 8 DDMs). The maximum storage capability with 16 DDMs for the DS6800 model is 4.8 TB (300 GB x16 DDMs).

If you want to connect more than 16 DDMs, you can use the optional DS6000 expansion enclosures (Model 1750-EX1). Each expansion enclosure can have up to 16 DDMs and a DS6800 can have up to 7 expansion enclosures. Therefore, the DS6800 allows a maximum of 124 DDMs per storage system and provides a maximum storage capability of 38.4 TB (300 GB x 128DDMs).

#### How to connect the expansion enclosures to the server enclosure

The DS6800 has two loops for connecting disk enclosures, one loop (accessed via the *disk exp* ports) can be connected with the server enclosure and three expansion enclosures, and the other loop (accessed via the *disk control* ports) can be connected with four expansion enclosures. You would normally add one expansion enclosure to each loop until both loops are populated with four enclosures each (realizing the server enclosure represents the first enclosure on the first loop).

Figure 8-7 on page 139 is an example for connecting several expansion enclosures.

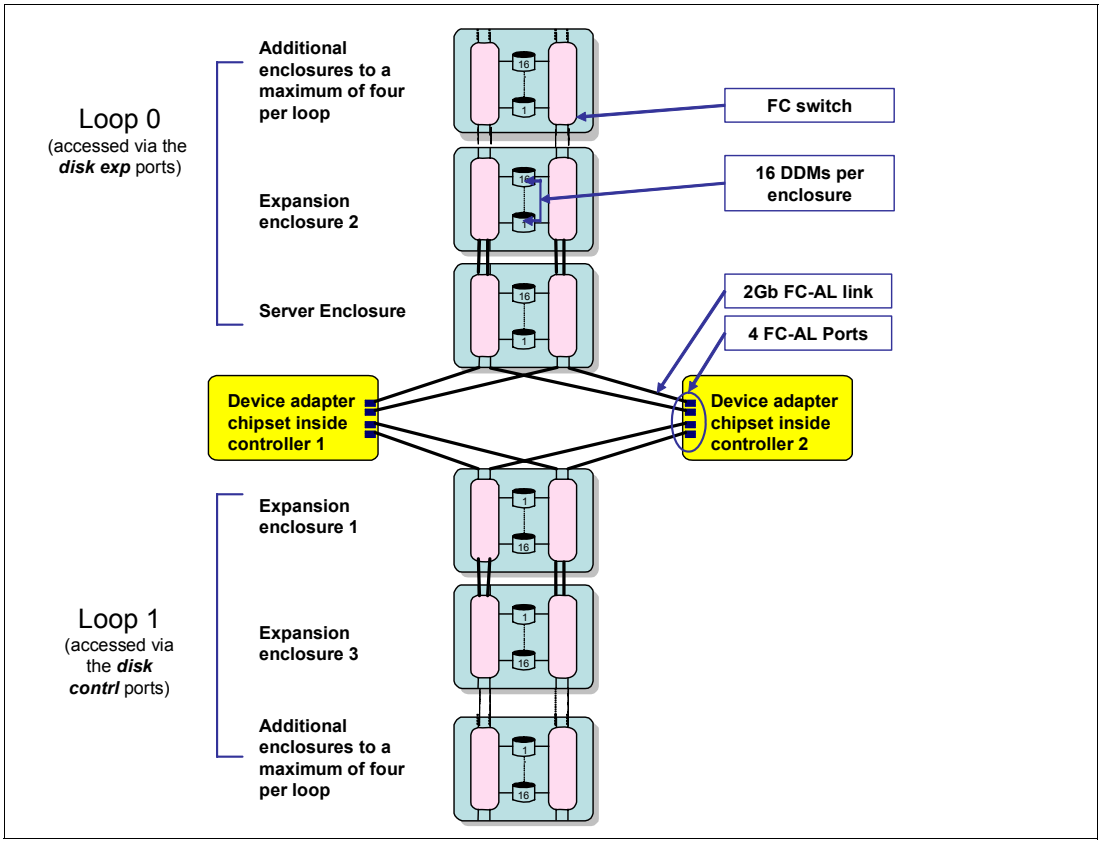



*Figure 8-7 Example of connecting several expansion enclosures*

## 8.4.2 Logical configurations

The capacity of the physical disk drives is not equal to the capacity you can use in the DS6000 series. Among the reasons for this are that the user will configure either RAID-5 or RAID-10, and the DS6000 series exploits virtualization concepts, which use some of the available physical capacity. See Chapter 4, “Virtualization concepts” on page 65 for more details.

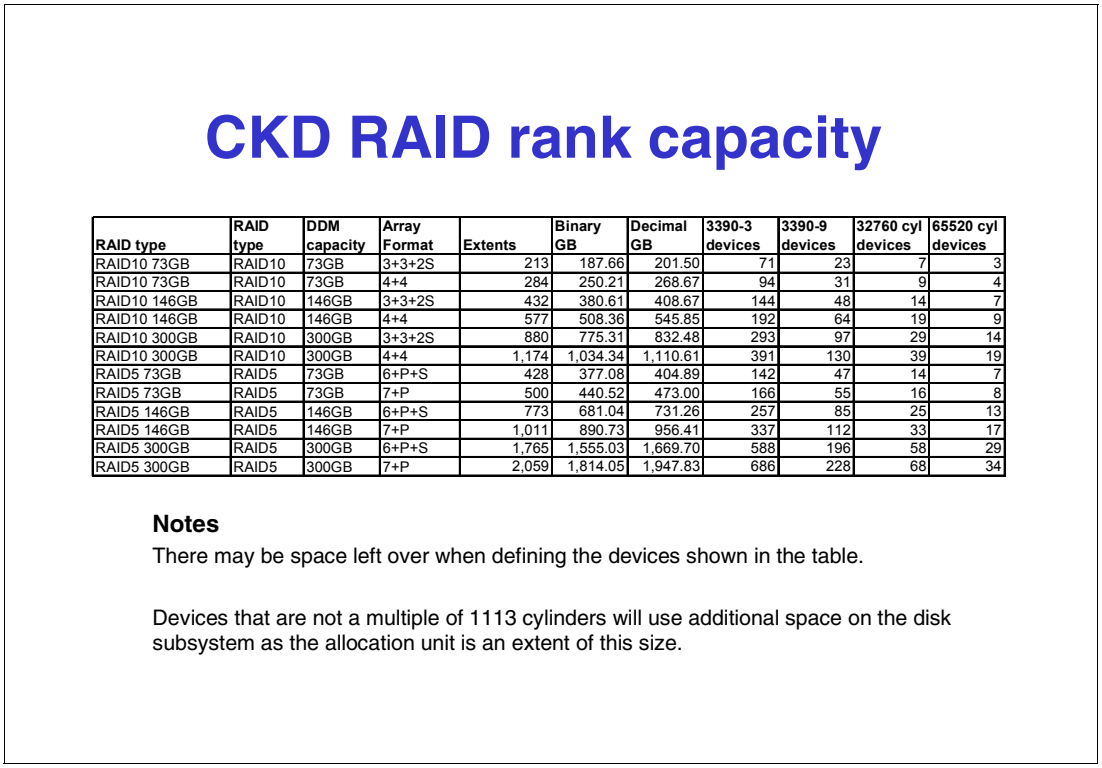
The logical capacity depends on not only the capacity of the DDMs but also the RAID type and the number of spare disks in a rank. We describe the capacity in the following figures.

**Attention:**

- The figures used in this chapter are still subject to change. Specifically, the exact number of extents per rank may be slightly different.
- 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 unexpected situations, a certain amount of margin is required for important systems.

**Note:** IBM plans to offer *Capacity Magic* for DS6000 in the future. Capacity Magic calculates the physical and effective storage capacity of a DS6000. IBM sales representatives and IBM Business Partners can download this tool from the IBM Web site. The values presented here are intended to be used only until Capacity Magic is available.

The following figures are for an 8 DDMs array. The user can also select a 4 DDMs array configuration.



## CKD RAID rank capacity

| RAID type | RAID type | DDM capacity | Array Format | Extents | Binary GB | Decimal GB | 3390-3 devices | 3390-9 devices | 32760 cyl devices | 65520 cyl devices |
|---|---|---|---|---|---|---|---|---|---|---|
| RAID10 73GB | RAID10 | 73GB | 3+3+2S | 213 | 187.66 | 201.50 | 71 | 23 | 7 | 3 |
| RAID10 73GB | RAID10 | 73GB | 4+4 | 284 | 250.21 | 268.67 | 94 | 31 | 9 | 4 |
| RAID10 146GB | RAID10 | 146GB | 3+3+2S | 432 | 380.61 | 408.67 | 144 | 48 | 14 | 7 |
| RAID10 146GB | RAID10 | 146GB | 4+4 | 577 | 508.36 | 545.85 | 192 | 64 | 19 | 9 |
| RAID10 300GB | RAID10 | 300GB | 3+3+2S | 880 | 775.31 | 832.48 | 293 | 97 | 29 | 14 |
| RAID10 300GB | RAID10 | 300GB | 4+4 | 1,174 | 1,034.34 | 1,110.61 | 391 | 130 | 39 | 19 |
| RAID5 73GB | RAID5 | 73GB | 6+P+S | 428 | 377.08 | 404.89 | 142 | 47 | 14 | 7 |
| RAID5 73GB | RAID5 | 73GB | 7+P | 500 | 440.52 | 473.00 | 166 | 55 | 16 | 8 |
| RAID5 146GB | RAID5 | 146GB | 6+P+S | 773 | 681.04 | 731.26 | 257 | 85 | 25 | 13 |
| RAID5 146GB | RAID5 | 146GB | 7+P | 1,011 | 890.73 | 956.41 | 337 | 112 | 33 | 17 |
| RAID5 300GB | RAID5 | 300GB | 6+P+S | 1,765 | 1,555.03 | 1,669.70 | 588 | 196 | 58 | 29 |
| RAID5 300GB | RAID5 | 300GB | 7+P | 2,059 | 1,814.05 | 1,947.83 | 686 | 228 | 68 | 34 |

**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 8-8 CKD RAID rank capacity (8 DDMs array)*

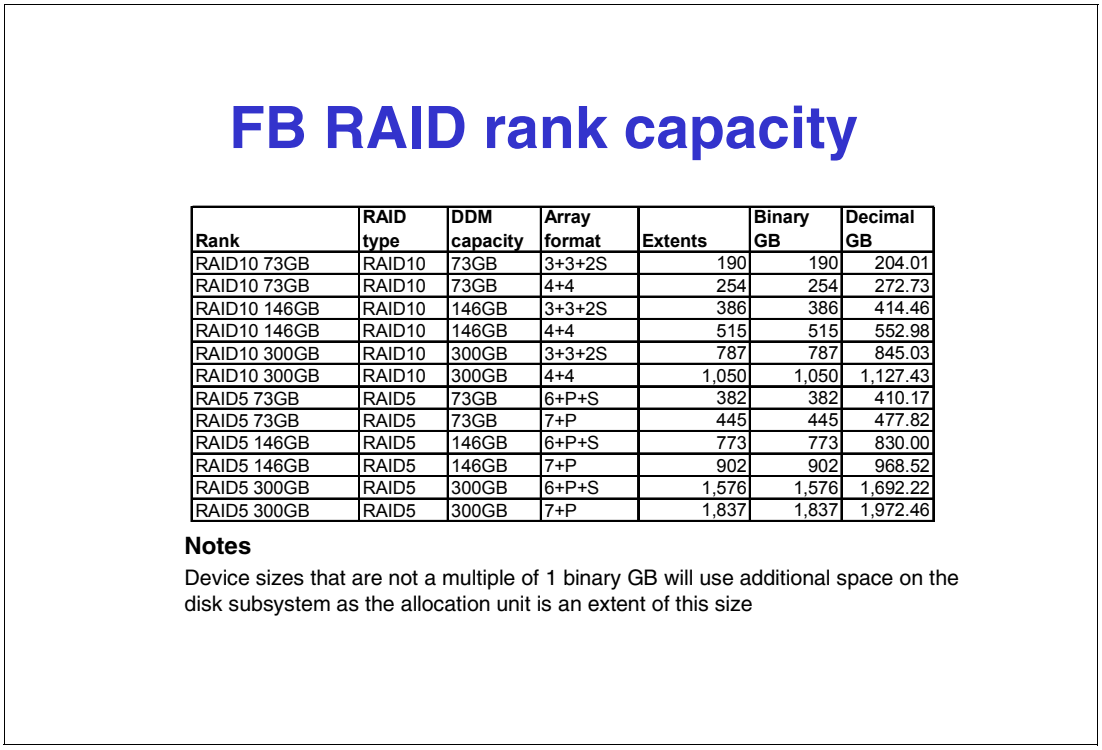



*Figure 8-9 FB RAID rank capacity (8 DDMs array)*

For example, if you configure a RAID-5 8 DDM 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 DS6000 series, the extent size is expressed in binary, not decimal. For example, 1 GB in binary is calculated as 1024 x 1024 x 1024, and 1 GB in decimal is calculated as 1000 x 1000 x 1000. Operating systems adopt binary format for their storage.

## 8.4.3 Sparing rules

To estimate your usable storage capacity, it is helpful to understand the rule of sparing disks. We explain the sparing rules for the DS6000 as follows:

- On a DS6000, you will have up to two spares on each loop (switched FC-AL).
- In the base frame we will have one or two spares: one spare in 8 DDMs configuration and two spares otherwise. There may be three spares if you configure RAID-5 on 8 drives and then RAID-10 on 8 drives in the first enclosure for the loop.
- In the first expansion frame we will have one or two spares. The first expansion frame is on the second loop.

### Configuration with the same type of DDMs

For example, if you configure a DS6000 with all the same type of 32 DDMs (a server enclosure with an expansion enclosure) attached to two switched FC-AL loops, you can configure four RAID arrays of 8 drives and the DS6000 will automatically assign four spare disks (two spare disks for each loop). After that, if you add more expansion enclosures, you can configure RAID arrays and the DS6000 will not assign spare disks.

See Figure 8-10 and Figure 8-11 on page 143. These figures are simple examples for 64 DDM configurations with RAID 5 or RAID 10 arrays.

RAID arrays are configured with spare disks in the server enclosure and the first expansion enclosure (each enclosure has two spare disks). Expansion enclosures 2 and 3 don't have spare disks.

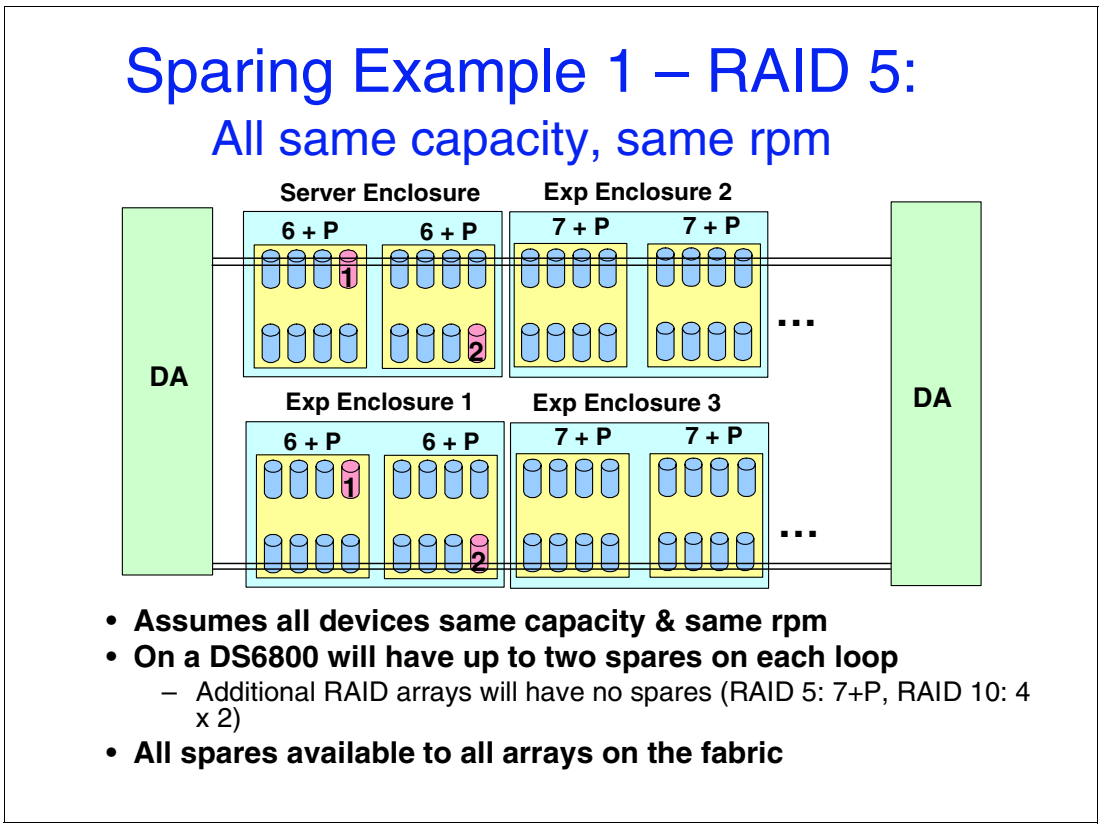



*Figure 8-10 Sparing example 1 (RAID-5)*

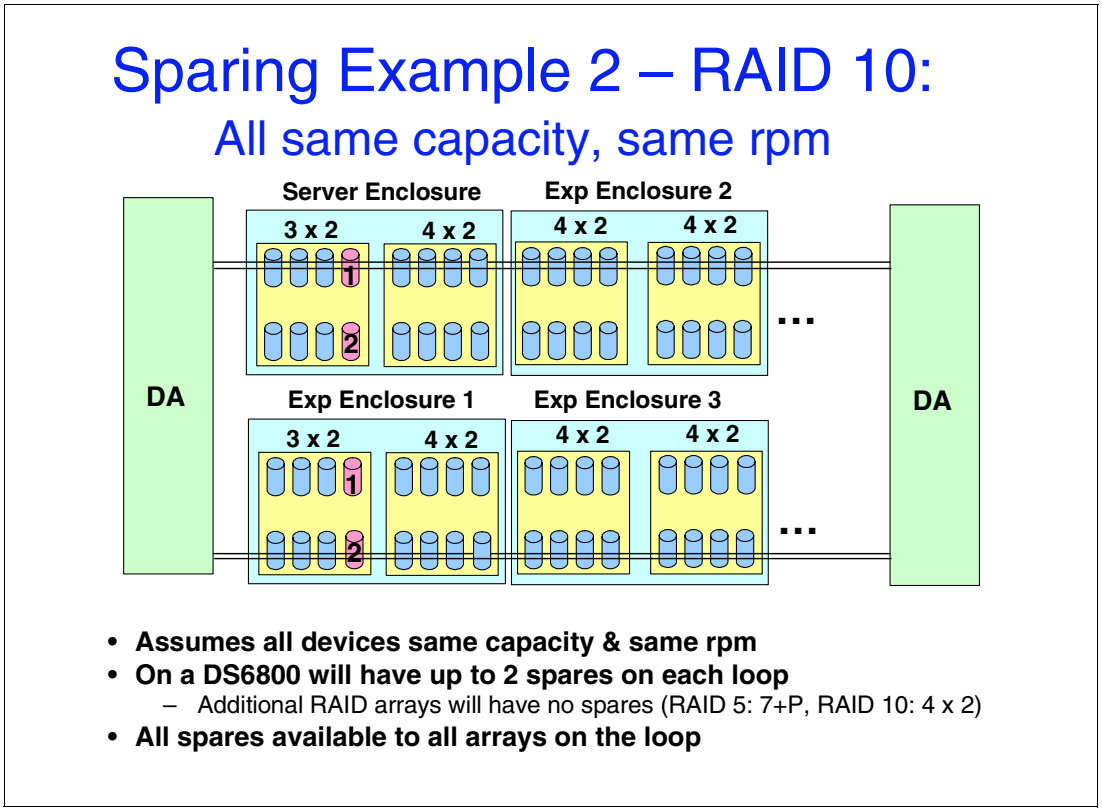



*Figure 8-11 Sparing example 2 (RAID-10)*

## Configuration with different types of DDMs

If you attach other types of DDMs in the loop, the DS6000 series has the following sparing rules:

- Minimum of two spares per the largest capacity array site on the loop.
- Minimum two spares of capacity and RPM greater than or equal to the fastest array site of any given capacity on the DA pair.

See Figure 8-12 on page 144. This is an example of adding larger capacity DDMs. At first, 32 DDMs of 146 GB are installed in the server enclosure and expansion enclosure 1, and after that 32 DDMs of 300 GB are installed in expansion enclosures 2 and 3.

In this case, in the first installation of disks, you configure four spare disks in the server enclosure and expansion enclosure 1. And, in the next installation, you also configure four spare disks in expansion enclosures 2 and 3. A 300 GB DDM cannot be spared onto a 146 GB DDM; therefore, 300 GB arrays need 300 GB spare disks for RAID configuration.

If you want to add additional arrays with 146 GB or 300 GB DDMs in this DA pair, you don't need spare disk in the array.

If the installation sequence is reversed (300 GB DDMs are installed into the first two enclosures, and 146 GB DDMs are installed into the next two expansion enclosures), you don't need spare disks in expansion enclosures 2 and 3 because 146 GB DDM can be spared onto existing 300 GB spare disks.

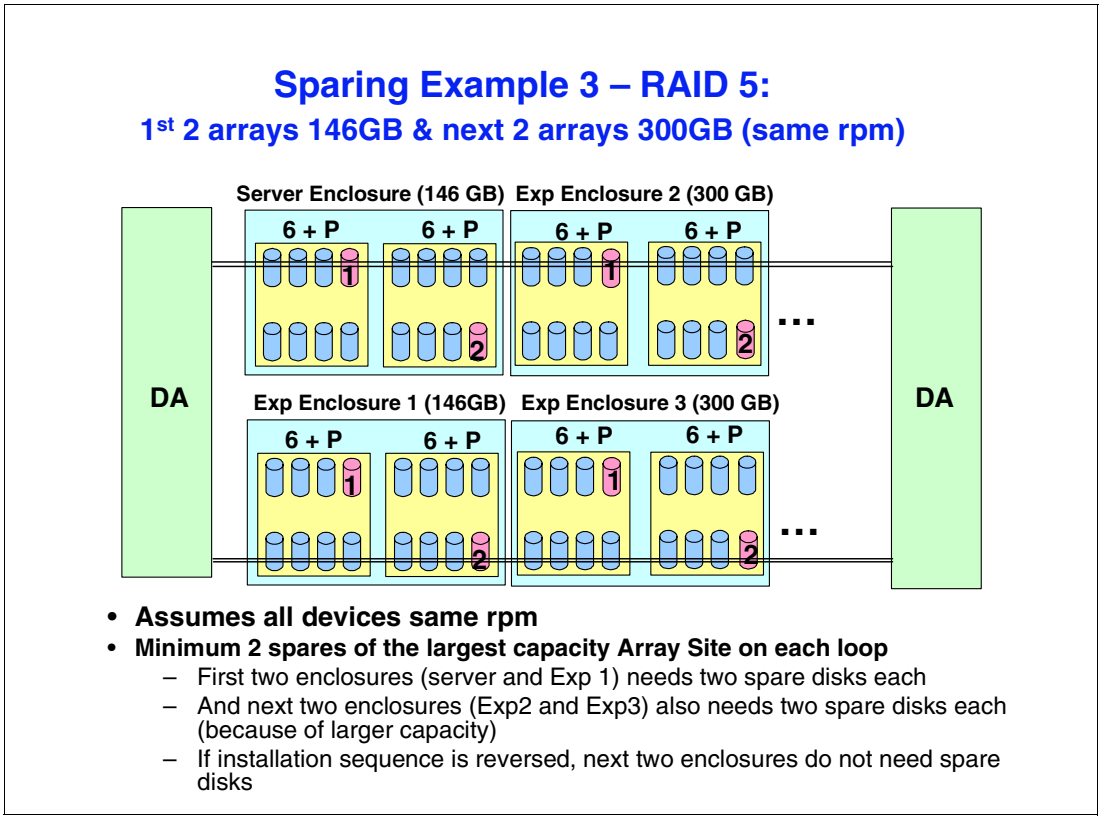



*Figure 8-12 Sparing example 3 (add larger capacity DDMs)*

See Figure 8-13 on page 145. This is an example of adding faster DDMs. At first, 32 DDMs of 146 GB and 10,000 RPM are installed in the server enclosure and expansion enclosure 1, and after that, 32 DDMs of 73 GB and 15,000 RPM are installed in expansion enclosures 2 and 3.

In this case, in the first installation, you configure four spare disks in the server enclosure and expansion enclosure 1. And, in the next installation, you also configure four spare disks in expansion enclosures 2 and 3.

If you want to add additional arrays with 73 GB or 146 GB DDMs in this DA pair, you don't need spare disks in the array.

And even if the installation sequence is reversed (146 GB DDMs are installed into the first two enclosures, and 73 GB DDMs are installed into the next two expansion enclosures), you need spare disks in the expansion enclosures 2 and 3, because 146 GB DDMs cannot be spared onto existing 73 GB spare disks.

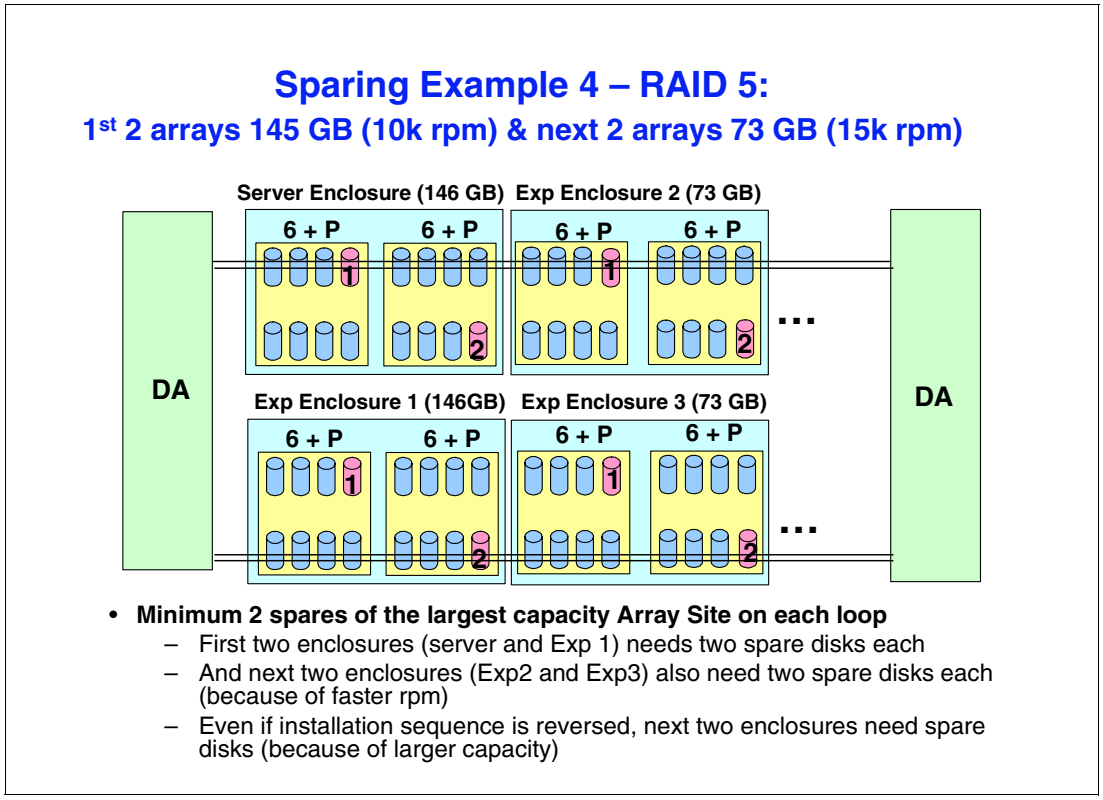



*Figure 8-13 Sparing example 4 (add faster DDMs)*

## 8.5 Data migration planning

When migrating data, the migration objectives should be clearly defined. The following are some key questions to use to define your generic migration environment:

- Why is the data migrating?
- What operating systems are involved?
- 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 DS6000?
- What resources are available for the migration?

Having answered 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—depending on your environment and your data migration objectives.

### 8.5.1 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.
- Application delays are possible due to dual writes occurring.
- This method does not allow for point-in-time copy or easy backout once the first copy is removed.

### 8.5.2 Basic commands

Basic commands such as **`cpio`** (in a UNIX environment) or **`copy`** (in a Windows environment) are common and easy to use. The disadvantages of basic commands are:

- Length of service disruption varies by commands used.
- They are tedious to handle large amount of LUNs.
- Command scripting is prone to human error.
- Security file level access issues can arise, depending on which commands used.

### 8.5.3 Software packages

The user can have data migration, backup or restore packages, and database tools to migrate data. An example of a data migration package is BRMS (Backup Recovery and Media Services). Tivoli Storage Manager (TSM) can also be used in a backup and restore fashion to achieve data migrations. DB2 utilities and tools such as unload, load and copy may 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 offloaded to another server.
- Standard database utilities.

The disadvantages of using a software package are:

- Cost of data migration package.
- Bigger impact and or disruption with large databases due to lack of checkpoint-restart abilities.
- Possibly lengthy application outage to backup and or restore environment.
- Application service interruption.

### 8.5.4 Remote copy technologies

Remote copy technologies include synchronous and asynchronous mirroring. They include Metro Mirror, Global Mirror and Global Copy.

The advantages of remote copy technologies are:

- Other than z/OS Global Mirror, they are operating system independent.
- Minimal host application outages.

The disadvantages of remote copy technologies include:

- Same storage device types are required. For example, in a Metro Mirror configuration you need ESS 800 mirroring to a DS6000 (or an IBM approved configuration), but cannot have non-IBM disk systems mirroring to a DS6000.
- Physical volume ID (PVID) and device name are not maintained, if not under LVM.

### 8.5.5 Migration appliances

The following IBM migration appliances are available:

- IBM Piper Services Offering
- IBM SAN Volume Controller. This offering cannot be decommissioned after the data migration is completed at this stage. Therefore, the user should be aware of the implication of opting for this offering.

These data migration appliances provide smooth, risk-averse data migration to your DS6000. The advantages of these migration appliances are:

- Facilitated by custom migration tools
- Minimal customer involvement by IT staff, except planning
- Minimal disruption or outages to IT operations
- On-line 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
- May fall back to the original data and storage device
- Large number of operating systems and storage devices supported

The disadvantages of migration appliances are:

- Cost of migration appliance or service
- Application disruption to install and remove the appliance

See Appendix C, “Service and support offerings” on page 363 for storage services offerings.

### 8.5.6 z/OS data migration methods

Figure 8-14 on page 148 lists a number of data migration methods available to migrate data from existing disk systems to the DS6000 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 |
| | Softek 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 8-14 Different data migration methods*

See Chapter 13, “Data Migration in zSeries environments” on page 251 for a complete discussion of the different methods available for data migration from any other disk system to the DS6000 series.

## 8.6 Planning for performance

The IBM TotalStorage DS6000 is a cost effective, 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
- Number of ports
- Remote Copy
- Parallel Access Volumes (z/OS only)
- I/O priority queuing (z/OS only)
- Monitoring performance
- Hot spot avoidance
- Preferred paths

### 8.6.1 Disk Magic

An IBM representative or an IBM Business Partner may model your workload using Disk Magic before migrating to the DS6000. 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/sec throughput. Disk Magic will be enhanced to support the DS6000. Disk Magic provides insight when you are considering deploying remote technologies such as Metro Mirror. Confer with your sales representative for any assistance with Disk Magic.

**Note:** Disk Magic is available to IBM sales representatives and IBM Business Partners only.

### 8.6.2 Number of host ports

Plan to have an adequate number of host ports and channels to provide the required bandwidth to support your workload. The ports must also be balanced across the entire DS6000.

### 8.6.3 Remote copy

If the DS6000 is a primary disk system in a remote copy configuration, it will consume more resources, such as fibre channels ports, compared to a standalone disk system. Plan for the number of fibre ports.

### 8.6.4 Parallel Access Volumes (z/OS only)

Configuring the DS6000 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, since 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. Plan for the number of alias devices that your configuration will need.

### 8.6.5 I/O priority queuing (z/OS only)

I/O priority queuing allows the DS6000 series to use I/O priority information provided by the z/OS Workload Manager to manage the processing sequence of I/O operations. This enables the DS6000 to prioritize I/O workload if, for whatever reason, I/Os are queued in the DS6000.

### 8.6.6 Monitoring performance

A number of monitoring tools are available to measure the performance of your DS6000 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 DS6000 is already installed in your environment. New counters will be reported by RMF to 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 can also be used to take a quick snapshot or overview of the system performance.

### 8.6.7 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 DS6000s will enable you to optimally exploit the DS6000 resources, thus providing better performance. I/O activity attributes to consider are to spread I/O activity across:

- The two servers of the DS6000
- The two loops of the DS6000
- All available RAID ranks
- Multiple DS6000s

### 8.6.8 Preferred paths

In the DS6000, host ports have a fixed assignment to a server (or controller card). In other words, preferred paths to a server avoid having to cross the PCI-X connection. Therefore, there is a performance penalty if data from a logical volume managed by one server is accessed from a port that is located on the other server. The request for the logical volume and the data would have to be transferred across the bridge interface that connects both servers. These transfers add some latency to the response time. Furthermore, this interface is also used to mirror the persistent memory and for other inter-server communication. It could become a bottleneck if too many normal I/O requests ran across it, although it is a high bandwidth, low latency, PCI-X connection.

When assigning host ports, always consider preferred pathing because the use of non-preferred paths will have a performance impact on your DS6000.

9

# 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

# 9.1 Configuration hierarchy, terminology, and concepts

The DS Storage Manager provides a powerful, flexible, and easy to use application for the logical configuration of the DS6000. 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. For further assistance, help and guidance can be obtained from an IBM FTSS or an IBM Business Partner.

## 9.1.1 Storage configuration terminology

An understanding of the following concepts and terminology can help you to use and configure the DS Storage Manager configurator.

### Storage complex

A storage complex consists of one or more physical storage units that can be managed from a central management point. DS6000 units can be placed together to form a complex. Multiple DS6000s can be supported by a single Management Console.

### Storage unit

A storage unit, also known as a *storage facility*, is a single physical storage subsystem (DS6000).

### Host attachment

A host attachment is one or more host ports that are grouped together. 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 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 9-1 on page 153.

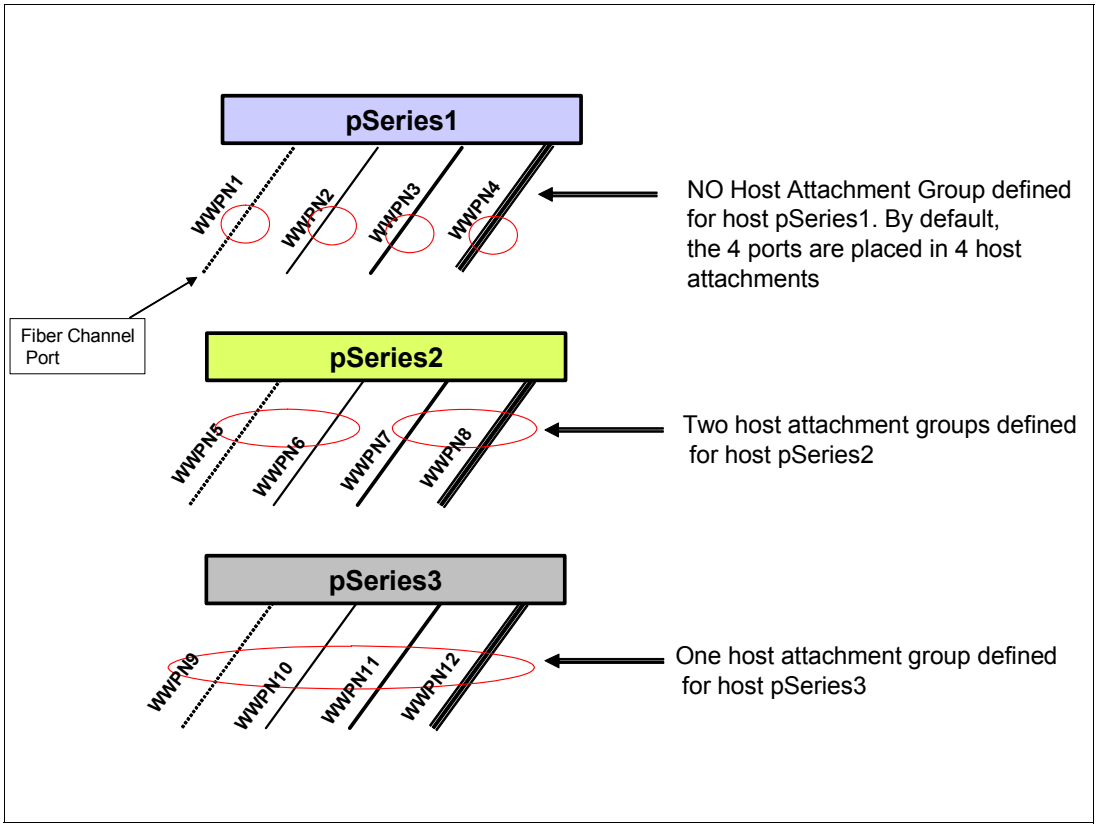



*Figure 9-1 Diagram of hosts and host attachment groups*

- In Figure 9-1 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.
- However, multiple host attachments, even different open system host types, with the same blocksize 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 9-3 on page 157.
- For FICON attachment, access is controlled by zSeries HCD/IOGEN definitions. Some FICON attachment considerations:
  - One storage subsystem FICON host definition will be required as the trigger to set the I/O adapter that will be accessed by the FICON protocol.
  - Default volume groups are automatically created, allowing anonymous access for FICON.
- One set of host definitions may be used for multiple storage, storage units, and storage complexes.

### DDM

A Disk Drive Module (DDM) is a customer-replaceable unit that consists of a single disk drive and its associated packaging.

### Array sites

An array site is a predetermined grouping of four individual DDMs of the same speed and capacity.

### Arrays

Arrays consist of DDMs from one to two array sites, 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 creating them to consist of the same RAID type, speed, and capacity. Also, we recommend that only half of the total of number of ranks be configured 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 1 GB

The storage in an extent pool (the extents from each rank in the extent pool) is used to create logical volumes. We recommend creating 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.

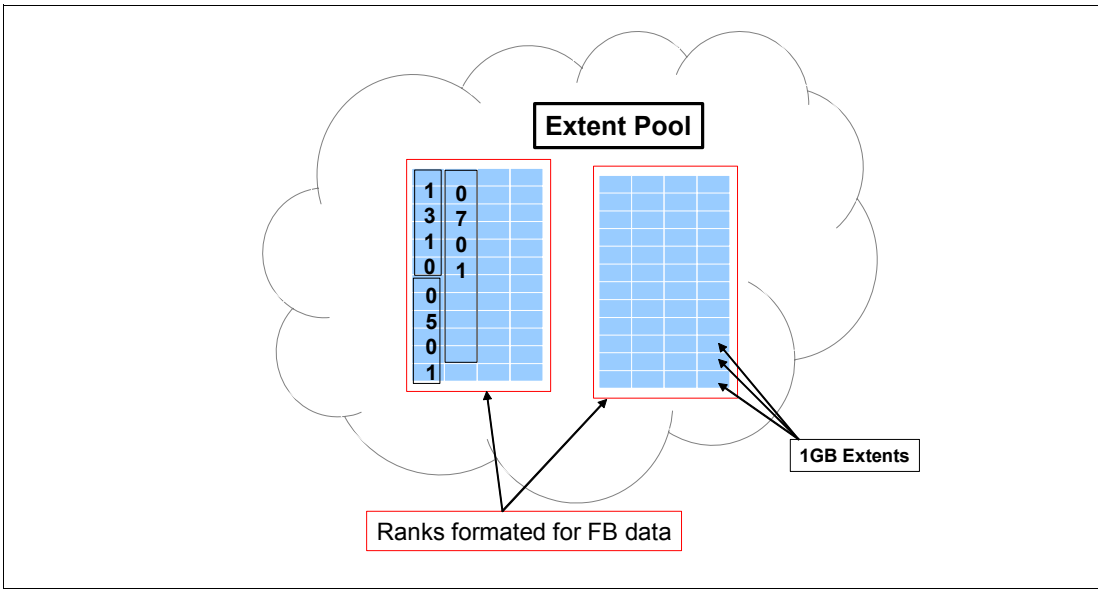



*Figure 9-2 Diagram of an extent pool containing 3 volumes*

In Figure 9-2, there is an example of one extent pool composed of ranks formatted for FB data. Three logical volumes are defined (volumes 1310, 0501, 0701). Two logical volumes (1310 and 0501) are made up of 6 extents at 1 GB each; this makes each volume 6 GBs. Logical Volume 0701 is built with 11 extents making an 11 GB Volume. The numbering sequence of a LUN (for example, 1310 as 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(volid). For example, LUN 1310 has an address group number of 1, is located in LSS 13 and has a volid 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 the client 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.

When creating extent pools, certain rules apply as follows:

- A minimum of two extent pools must be configured to utilize server 0 and server 1.
- More than one rank can reside in an extent pool, but two extent pools can not be made out of only one rank. We recommend that one extent pool be created 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 pools can be in the same LSS, 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 term volume and LUN interchangeably throughout the rest of this chapter when referring to logical volumes.

Some limitations are as follows:

- 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 9-6 on page 160.
- Multiple volumes in different extent pools and on different ranks can be in the same LSS, as shown in Figure 9-6.
- The minimum volume/LUN size is one extent.
  - For CKD the minimum size is a 3390 Mod1.
  - For FB the minimum size is 1 GB.

  **Note:** The user is allowed to specify volume sizes in binary, decimal, or block sizes. In binary form, for example, 1 GB is equal to 1073741924 bytes (1GB=1073741924), instead of the decimal size, as in the ESS, as (1GB=1000000000 bytes).

- The maximum volume or LUN size is equal to the size of the extent pool with the following limitation: 56 GB for CKD, with the 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 DS6000 is 8K (8192). It could be either 8192 FB Volumes (LUNs), or 8192 CKD Volumes, or a mix of 4096 LUNs and 4096 CKD Volumes.
- Logical 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 DS6000 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 hosts systems, even running different operating system types, is allowed in the same volume group.

**Note:** We recommend configuring one host per volume group.

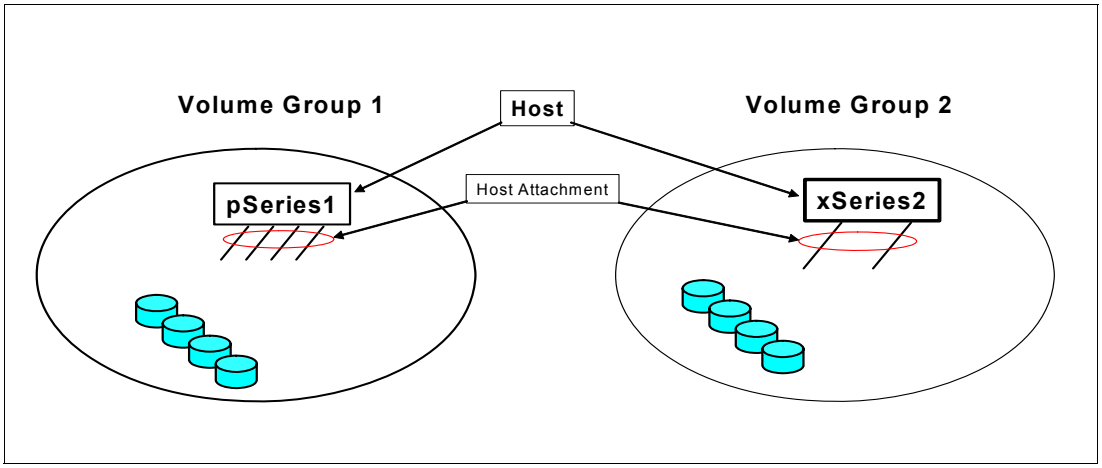



*Figure 9-3 Diagram of the relationship of a host attachment to a volume group*

In Figure 9-3, 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 Group1. 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 9-4 on page 158.

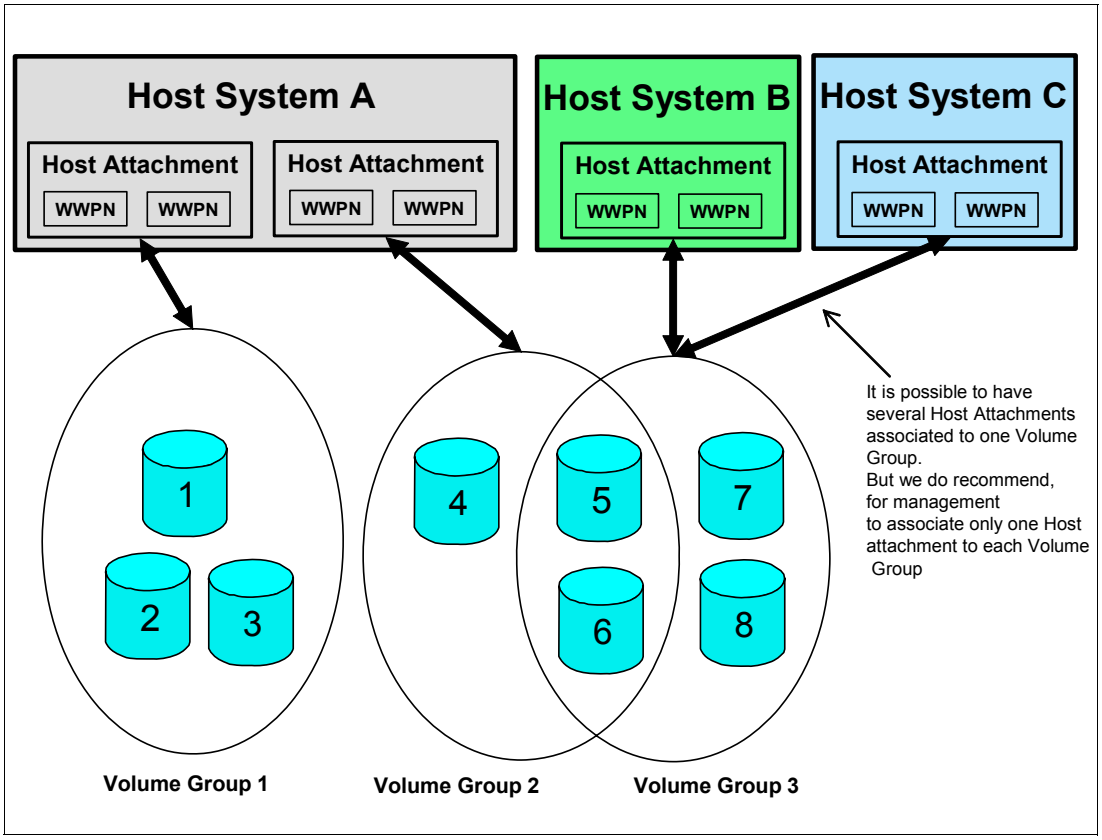



*Figure 9-4 Example of Volume Group, LUNs and host attachment definition*

In Figure 9-4 we show three hosts (Host A, Host B, and Host C). The three hosts are 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 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 DS6000 is 1040.

## Address groups

An address group is a group of FB or CKD LSSs. An address group has up to 16 LSSs. The DS6000 supports two address groups: address group 0 and address group 1.

## LSS/LCU

A Logical Subsystem (LSS) is a topological construct that consists of a group of up to 256 logical volumes. A DS6000 can have up to 32 LSSs. The DS6000 supports a mix of CKD-formatted logical subsystems and FB logical subsystems. There is a one-to-one mapping between a CKD logical subsystem and a zSeries control unit.

For zSeries hosts, a logical subsystem represents a logical control unit (LCU). Each control unit is associated with only one logical subsystem.

Figure 9-5 shows an example of the relationship between LSSs, extent pools, and volume groups: Extent Pool 4, consisting of two LUNs, LUNs 0210 and 1401, and Extent Pool 5, consisting of three LUNs, LUNs 0313, 0512, and 1515.

Here are some considerations regarding the relationship between LSSs, extent pools, and volume groups:

- Volumes from different LSSs and different extent pools can be in one volume group as shown in Figure 9-5. Volume Group 1 consists of Extent Pool 4, LUN 0210 and Extent Pool 5 LUN 0512.
- Volumes from the same LSS or the same extent pool can be in different volume groups. For example, in Figure 9-5, LUN 0512 in Extent Pool 5 and LUN 0313 in Extent Pool 5 both reside in the same extent pool, but LUN 0512 is in Volume Group 1 and LUN 0313 is in Volume Group 2.

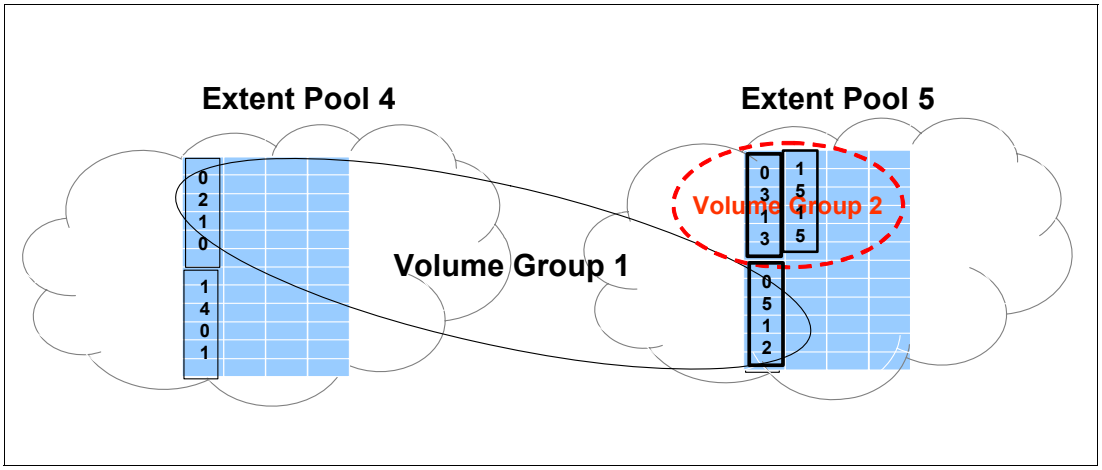



*Figure 9-5 Example of relationship between LSS, extent pools and Volume Group*

The LSS and LCU provide the logical grouping of volumes/LUNs for Copy Services and various other purposes. Some of these purposes are explained here along with general considerations about LSSs in the DS6000, 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 (volid). For example, LUN 1401 has an address group number of 1, is located in LSS 14 and has a volid of 01.
  - 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.
  - Any given PAV can only be used within one LCU.
- LSSs are used for Copy Services, PPRC paths, and consistency group properties/time-outs.
- The DS6000 supports a maximum of 32 LSSs and supports intermix of CKD LSSs and FB 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 9-6.
  - One LCU can contain CKD volumes/LUNs of different types. For example type 3390 Model 3 and Model 9s.
  - LSSs can have a many to many Copy Services relationship as shown by the arrows in Figure 9-6,

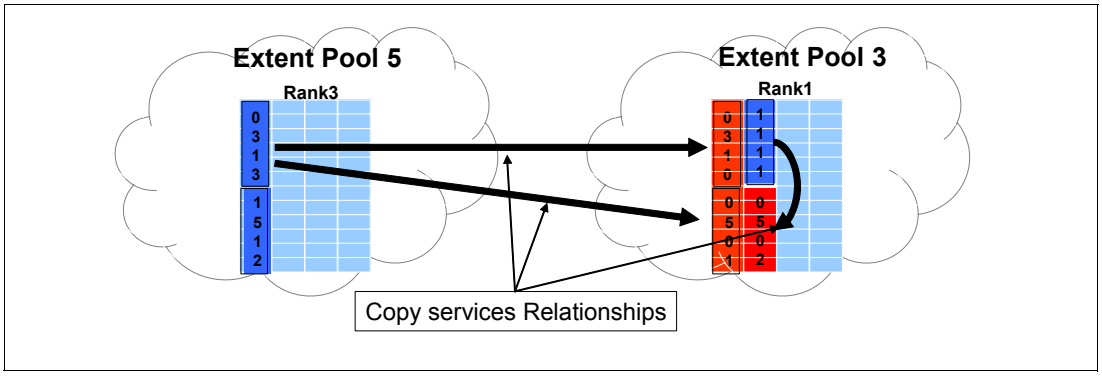



*Figure 9-6 Example of Copy Services relationship between LSSs in the same storage unit*

- 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 2 address groups for the DS6000 at GA; however, more will be available at a later time.
- One address group has 4096 addresses (16 LSS x 256 logical volumes = 4096 logical volumes or addresses). This means that if you had two address groups defined in the DS6000, you would have 8192 (2 X 4096 = 8192) addresses or logical volumes.

**Attention:** The address groups and LSSs are predefined in the DS6000. Initially, they do not have any data format attributes (FB or CKD). Their data format attributes are set when a first logical volume is added to a first LSS in an 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 in which LSS this FB volume will be placed 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.

## 9.1.2 Summary of the DS Storage Manager logical configuration steps

It is our recommendation that the client consider 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 the client approach the configuration from the host and work up to the DDM (raw physical disk) level. This is just the opposite way that the client would configure the raw DDMs into host volumes. Refer to Figure 9-7 to understand the hierarchy in the virtualization layers in the DS6000.

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 pools 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.

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.

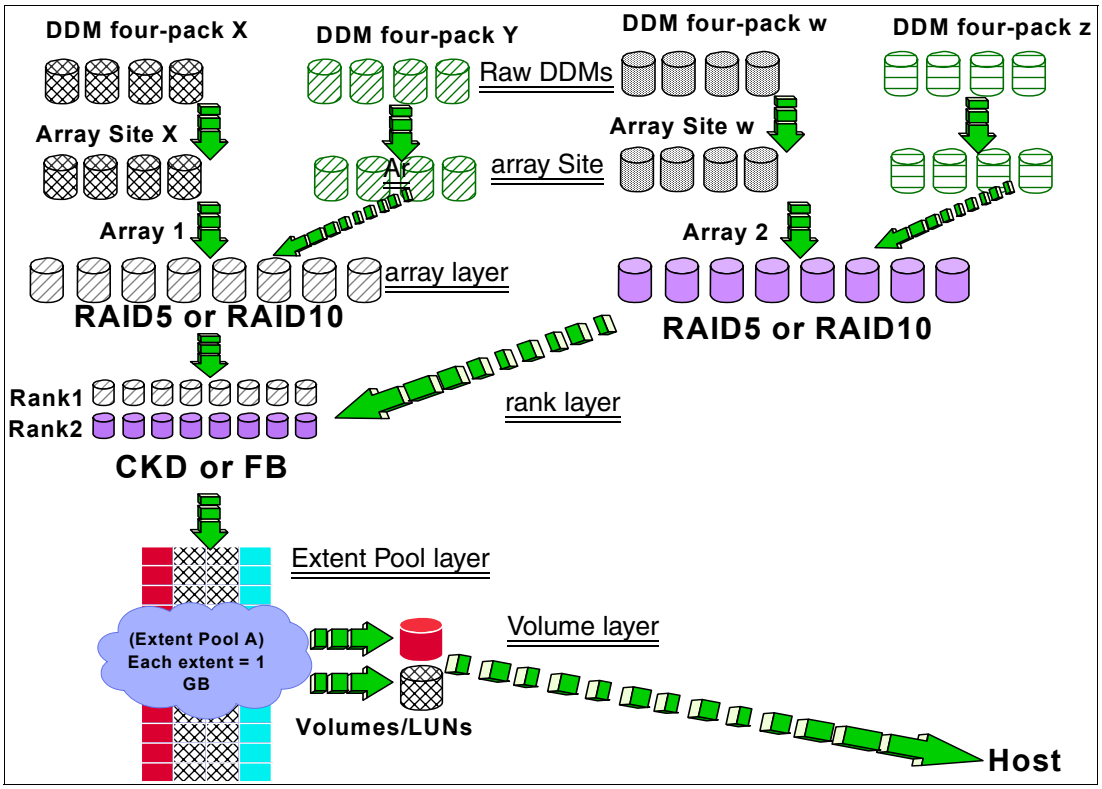



*Figure 9-7 View of the raw DDM to LUN relationship*

In Figure 9-7, you can see that several layers make up the final result, that being the LUN that is presented to the host as usable disk space capacity.

## Raw or physical DDM layer

At the very top, you can see the raw DDMs. There are four DDMs in a group called a four-pack. They are placed into the DS6000 in pairs of two, making eight DDMs. DDM X represents one four-pack and DDM Y represents a pair from another four-pack. Upon placing the four or eight-packs into the DS6000 each four or eight-pack is grouped into array sites, shown as Layer 2. The DS6000 does support 4 DDM arrays; this is not described here.

## Array site layer

At the array site level, predetermined groups of four DDMs of the same speed and capacity are arranged. All DDMs reside in the same Disk Enclosure (Array On Loop configuration).

## Array layer

This level is where the format is placed on the array. Sparing 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 9-8. 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 choose 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 sparing rules are met.

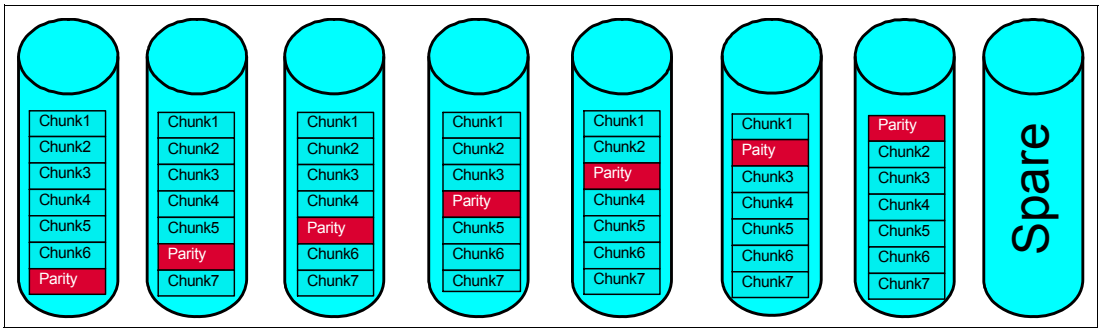



*Figure 9-8 Diagram of how parity is striped across physical disks*

## Rank layer

At this level the ranks are formed. Presently only one array can make up a rank, as shown in Figure 9-7 on page 161.

## Extent pool layer

At this level the extent pools are formed. In Figure 9-7 on page 161 we show that extent pool A is made up of 2 ranks, rank 1 and rank 2. The extents in the pool are 1 GB.

## Logical volume layer

Layer 6 is the final level of the LUN formation. We show in Figure 9-7 on page 161 that two LUNs were created out of extent pool A. The top LUN is made up of 12 extents, making it a 12GB LUN. The bottom LUN is made up of 24 extents, making it a 24GB LUN.

## Logical Configuration flow

Figure 9-9 on page 163 shows the recommended flow for performing logical configuration using the GUI.

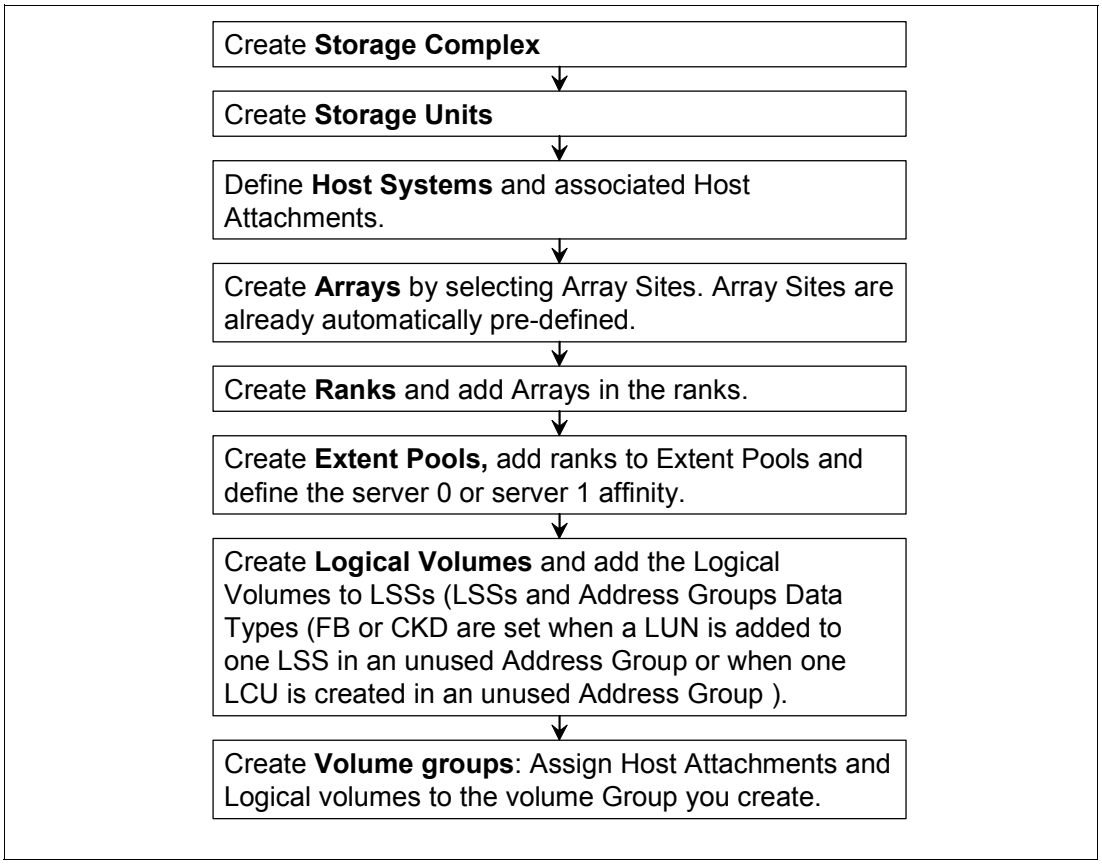



*Figure 9-9 Recommended Logical Configuration steps*

# 9.2 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) using the install shield. 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.

## 9.2.1 Connecting to your DS6000

To connect to the DS6000 through the browser, enter the URL of a PC or the MC you may have purchased. The URL consists of the TCP/IP address as shown in Figure 9-10 on page 164, or a fully qualified name that the DNS server can resolve as shown in Figure 9-11 on page 164.

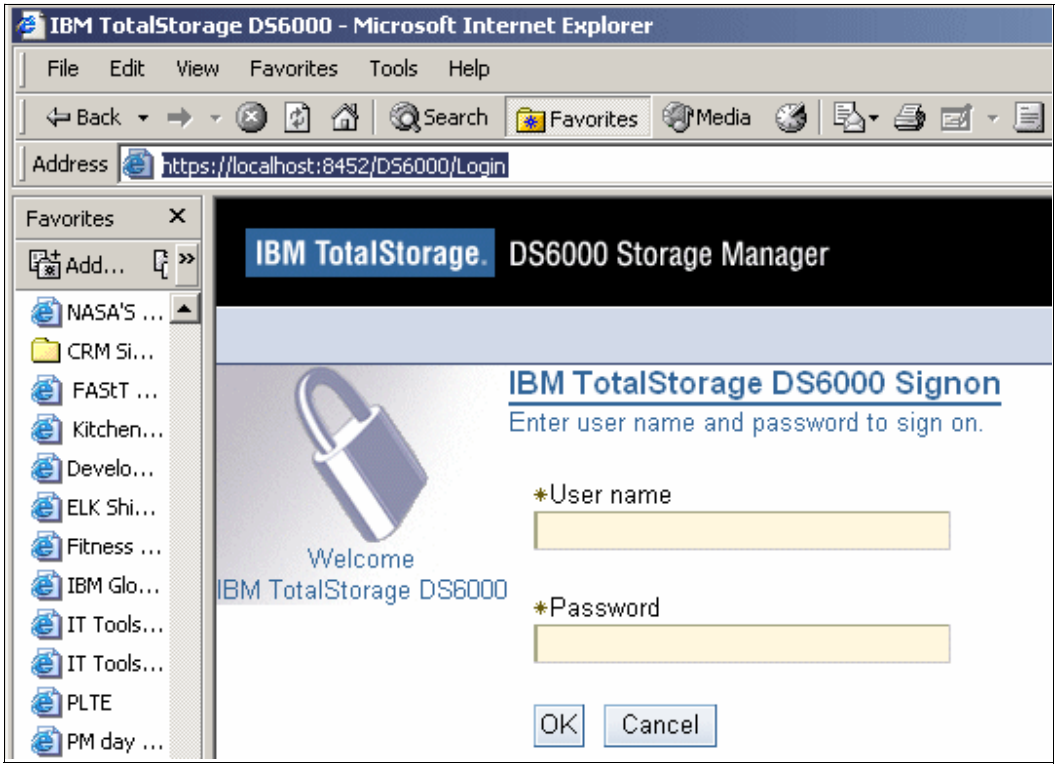



*Figure 9-10 Entering the URL using the TCP/IP address for the SMC*

In Figure 9-10, we show the TCP/IP address and the port number 8452 afterwards.


*Figure 9-11 Entering the URL using the fully qualified name of your MC*

In Figure 9-11, 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, MC_0 for the default MC as shown in Figure 9-11. Then bookmark it for ease of use. When assigning the number, we recommend that you make the last field of the optional MC only one digit higher than the default MC.

## 9.2.2 Introduction and Welcome panel

The IBM TotalStorage DS6000 Storage Manager (DS Storage Manager) is a software application that runs on an MC. It is the interface provided for the user to define and maintain the configuration of the DS6000. The DS Storage Manager can be accessed using a Web browser running on a remote machine connected into the user's network.

Once the GUI is started and the user has successfully logged on, the Welcome panel will display as shown in Figure 9-12.


*Figure 9-12 View of the Welcome panel*

Figure 9-12 shows the Welcome panel's two menu choices. By clicking on the triangles, the menu expands to show the menu options needed to configure the storage.

The DS6000 Storage Manager configurator can be used either in Real-time (online), or Simulated (offline), with an Express Configuration Wizard available in both modes. Either mode is used to manipulate the storage configuration process for a DS6000, defining CKD and fixed block (FB) storage.

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, to be 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 enter the OLE license activation key. You can obtain the OLE license activation key and all 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 9-13 on page 166.


*Figure 9-13 View of the fully expanded Real-time Manager menu choices*

– **Copy Services**

You can use the Copy Services selections of the DS Storage Manager if you choose **Real-time** during the installation of the DS Storage Manager and you purchased these optional features. A further requirement for using the Copy Services features is the activation of license activation keys. You need to obtain the Copy Services license activation keys (including the one for the use of PAVs) from the 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 DSFA Web site.

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. A view of the fully expanded Simulated Manager menu choices is shown in Figure 9-14 on page 167.


*Figure 9-14 View of the fully expanded Simulated Manager menu choices*

- **Express configuration**

  Express configuration provides the simplest and fastest method to configure a storage complex.

  Some configuration methods require extensive time. Because there are many complex functions available to you, you are required to make several decisions during the configuration process. However, with Express configuration, the storage server makes several of those decisions for you. This reduces the time that configuration takes, and simplifies the task for you.

  The Express Configuration Wizard is ideal for the following users:

  – Novice users with little knowledge of storage concepts who want to quickly and easily set up and begin using storage.

  – Expert users who want to quickly configure a storage complex by allowing the storage server to make decisions for the best storage appropriation.

  Using Express configuration, you can perform the following tasks:

  – Configure open systems, iSeries, and zSeries volumes

  – Create a volume group

  – Create a host

  – Map a volume group to a host attachment

The following items should be considered as first steps in the use of either configuration mode.

- **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 user ID and password are as follows:

  userid=admin

  password=admin

- **Create and define the users and passwords**

  Select **User administration** → **Add user** and click **Go**, as shown in Figure 9-15. The screen that is returned is shown in Figure 9-16. On this screen you can add users and set passwords, subject to the following guidelines:

  – 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 can not be part of the password. This entry will appear as asterisks.

  You can grant user access privileges from this screen as well.

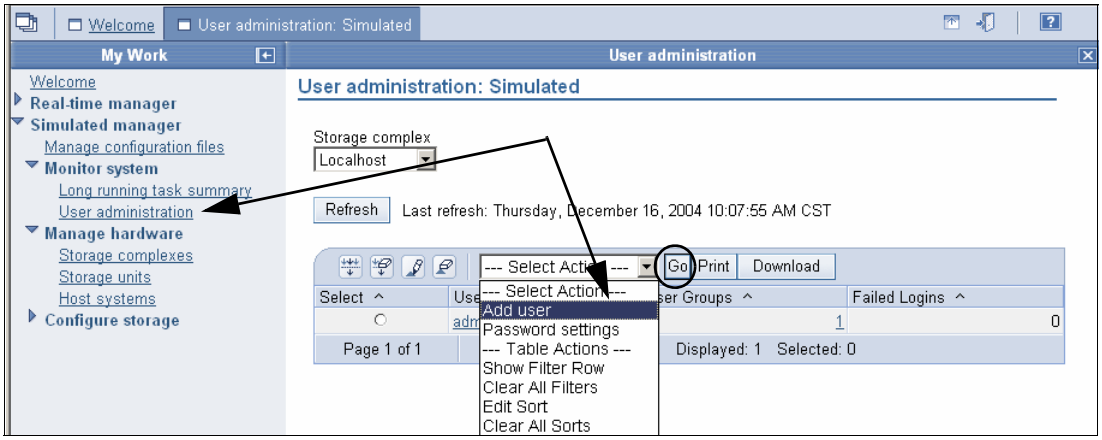



*Figure 9-15 View of the User administration panel*

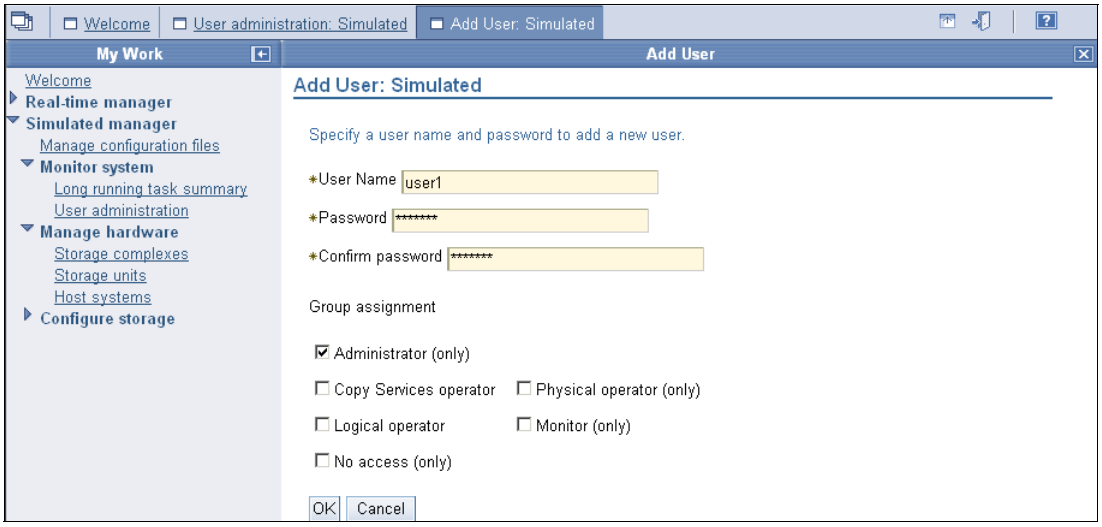



*Figure 9-16 View of the Add User panel*

### *Using the help panels (Software information center)*

The information center displays product or 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 shown in Figure 9-17.

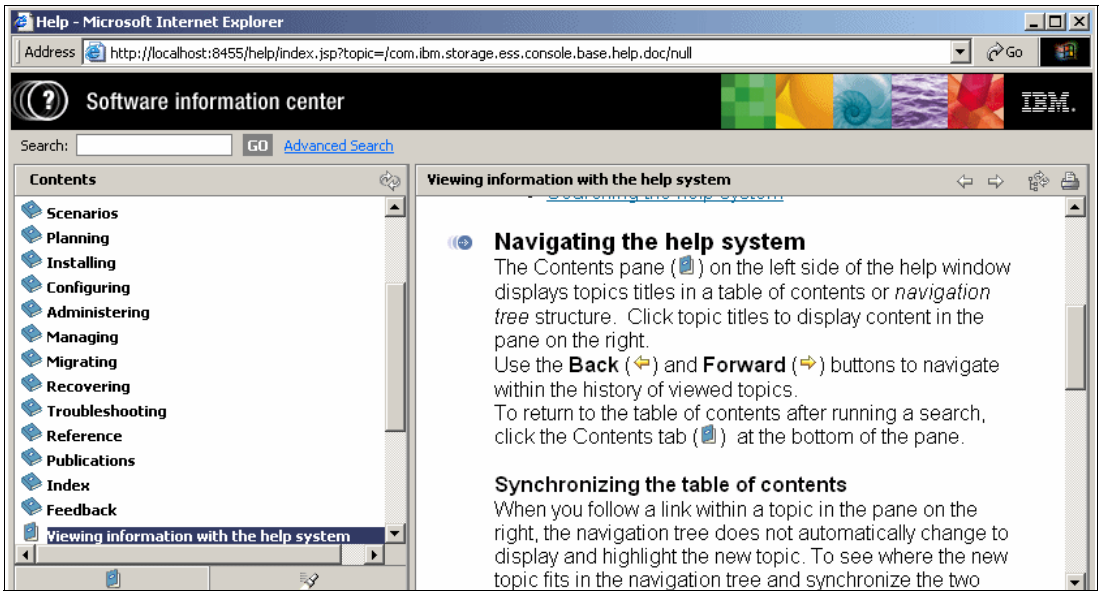



*Figure 9-17 View of the information center*

## 9.2.3 Navigating the GUI

Knowing what icons, radio buttons, and check boxes to click in the GUI, will help you efficiently navigate your way through the configurator and successfully configure your storage.

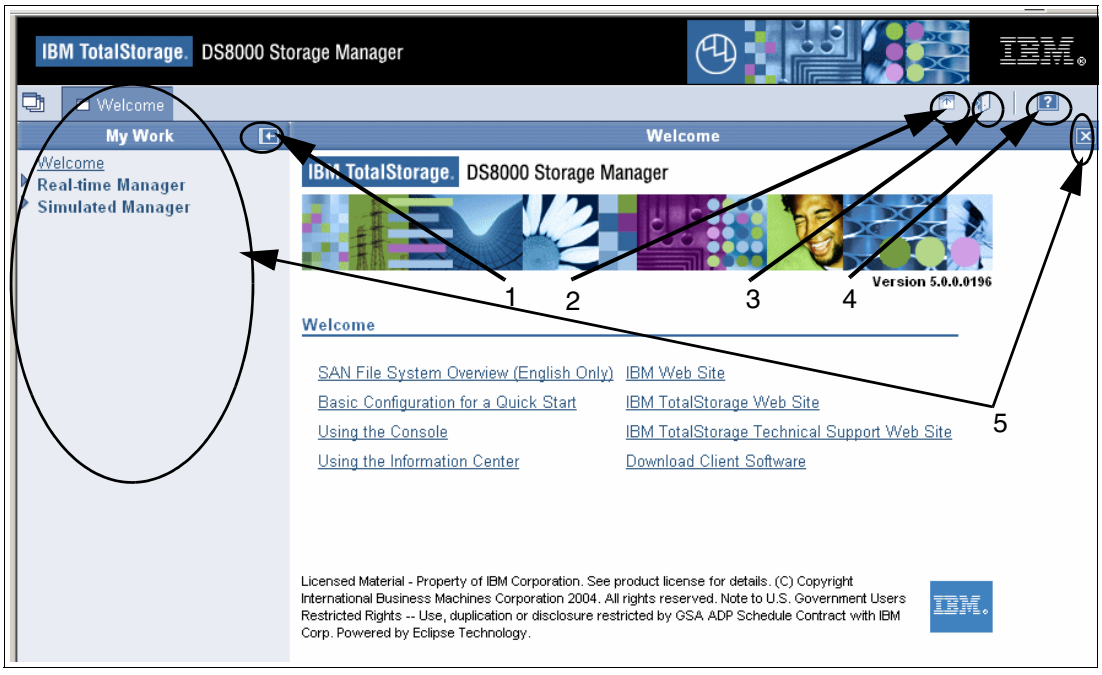



*Figure 9-18 The DS Storage Manager Welcome panel*

On the DS Storage Manager Welcome panel, several buttons appear that need some explanation.

With reference to the numbers shown in Figure 9-18, the icons shown have the following meanings:

1. Click icon 1 to hide the My Work menu area. This increases the space for displaying main work area of the panel.
2. Icon 2 hides the black banner across the top of the screen, again to increase the space to display the panel you're working on.
3. Icon 3 allows you to properly Log out and exit the DS Storage Manager GUI.
4. Use icon 4 to access the Info Center. You get a help menu screen that prompts you for input on help topics.
5. Click icon 5 to close the panels that open when you click the configuration menus in the My Work area.

Figure 9-19 shows that by clicking the expand work area buttons, both the banner and the My Work area can be reduced to allow an expanded view of the work area.

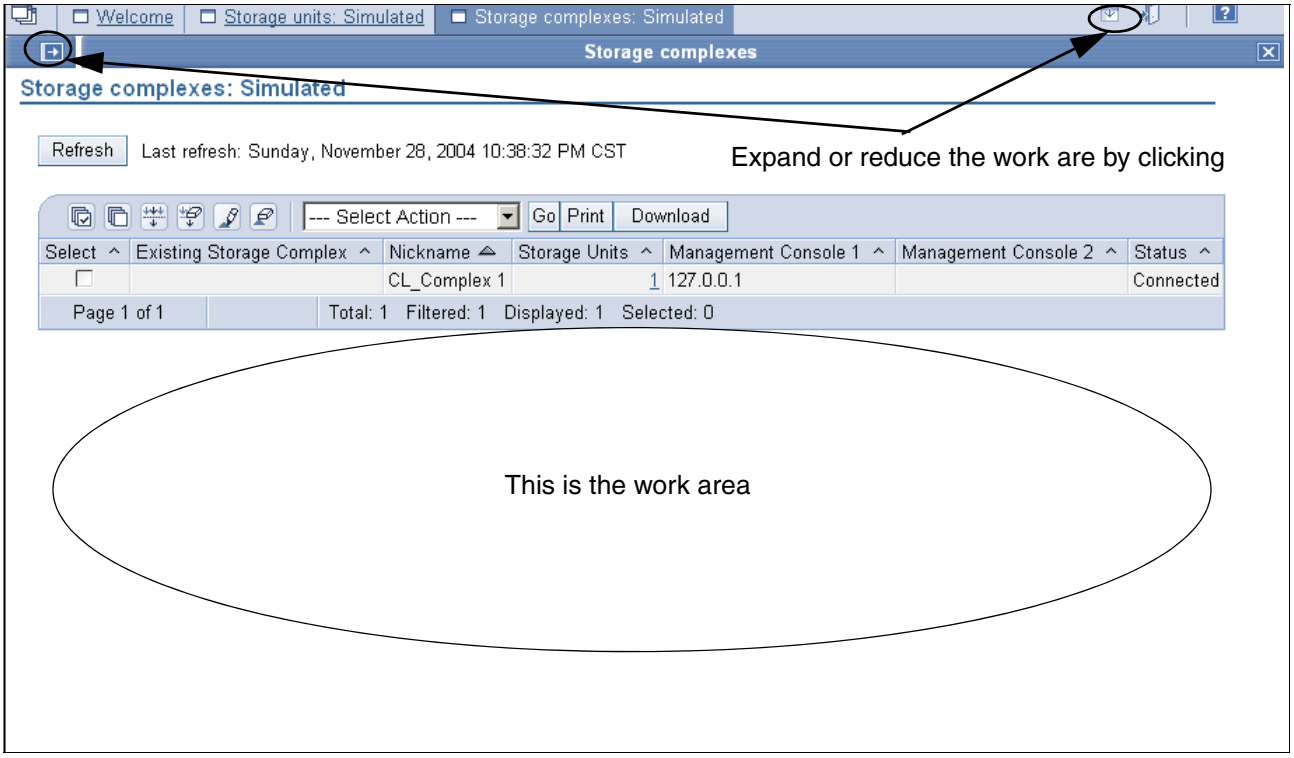



*Figure 9-19 View of the storage complexes in the work area*

If you want to reduce the work area and work from the Real-time or Simulated Manager menu selection again, simply click the button shown on the upper left of Figure 9-19.

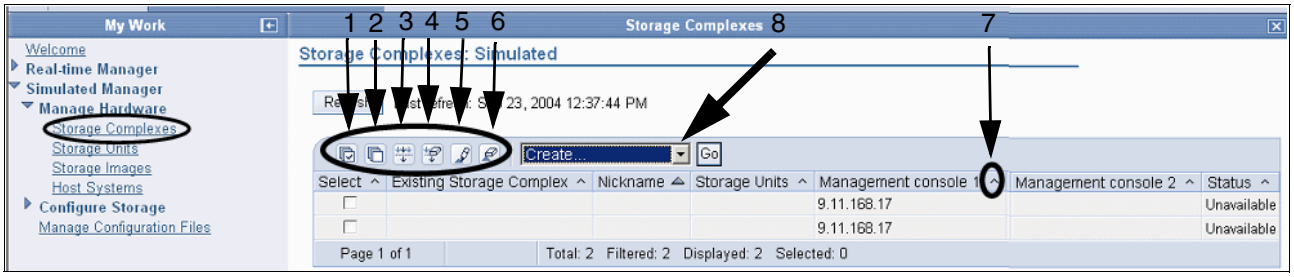



*Figure 9-20 View of the Storage Complexes section*

The buttons displayed on the Storage complexes screen, shown in Figure 9-19 and called out in detail in Figure 9-20, have the following meanings.

- Boxes 1 through 6 are for selecting and filtering:
  - 1 - Select all
  - 2 - Deselect all
  - 3 - Show filter row
  - 4 - Clear all filters
  - 5 - Edit sort
  - 6 - Clear all sorts
- The caret called out as item 7 is a simple ascending/descending sort on a single column.
- Clicking the pull-down menu identified as item 8 results in display of the actions list shown in Figure 9-21, which are the identical actions as represented by boxes 1 through 6 in Figure 9-20.

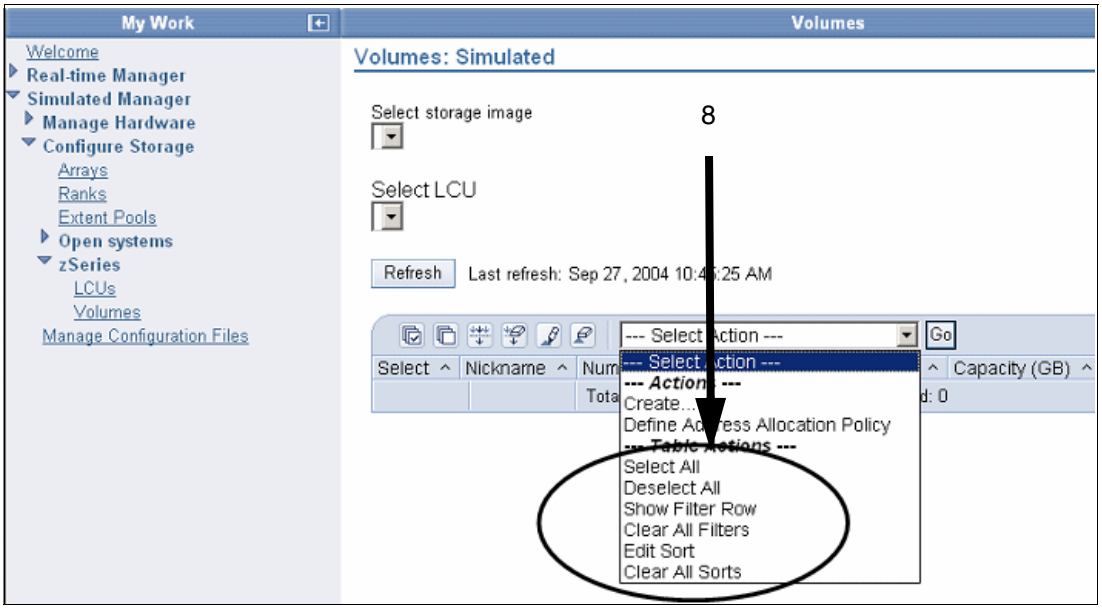



*Figure 9-21 Storage unit view of the drop down box*

## Radio buttons and check boxes

Figure 9-22 shows the difference between radio buttons and check boxes.

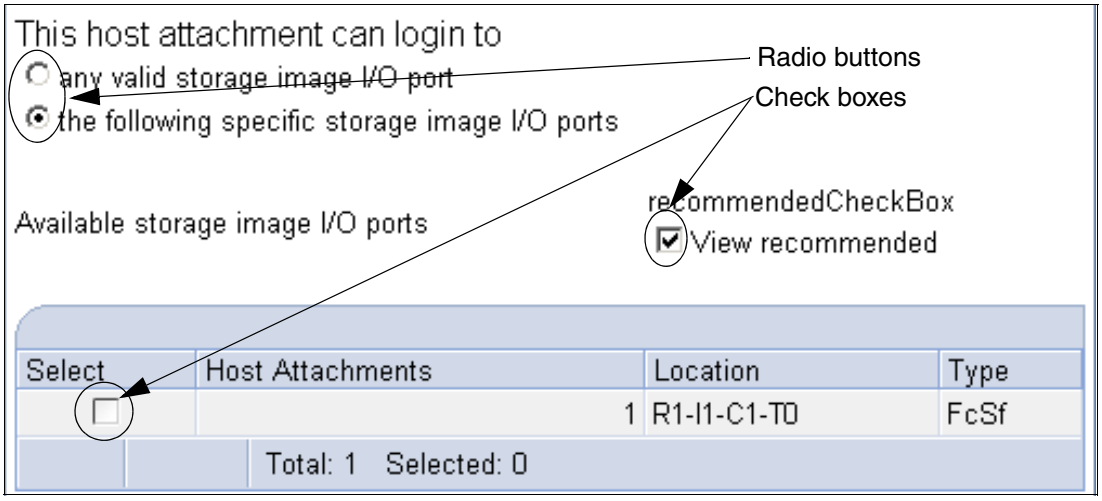



*Figure 9-22 View of radio buttons and check boxes in the host attachment panel*

In the example shown in Figure 9-22, the radio button is checked to allow specific host attachments for selected storage unit I/O ports only. The check box has also been selected to show the recommended location view for the attachment. Only one radio button can be selected in the group, but multiple check boxes can be selected.

# 9.3 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, so long as you define your storage complex, storage unit first.

## 9.3.1 Configuring a storage complex

To create the storage complex in simulated mode, 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 9-23. Follow the panel directions with each advancing window.

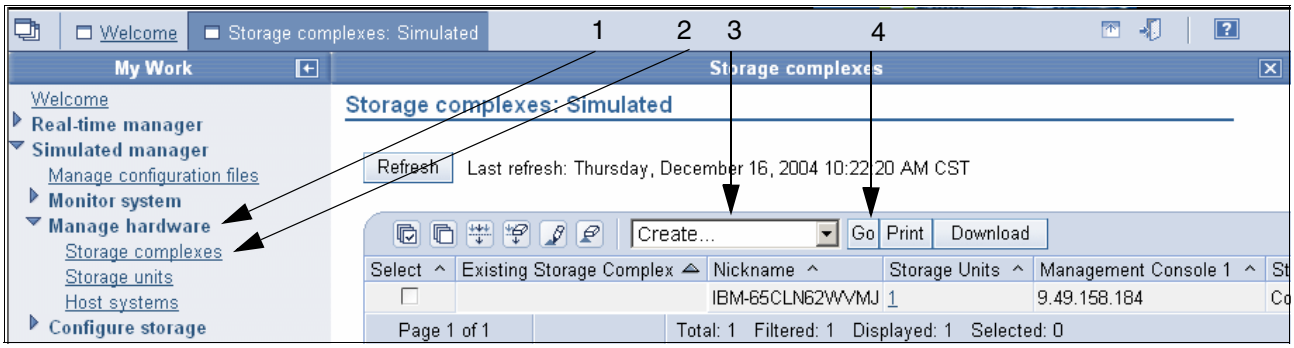



*Figure 9-23 View of the Select Action pull-down menu with Create, selected*

Under the Define Properties panel, assign the Storage complex Nickname.

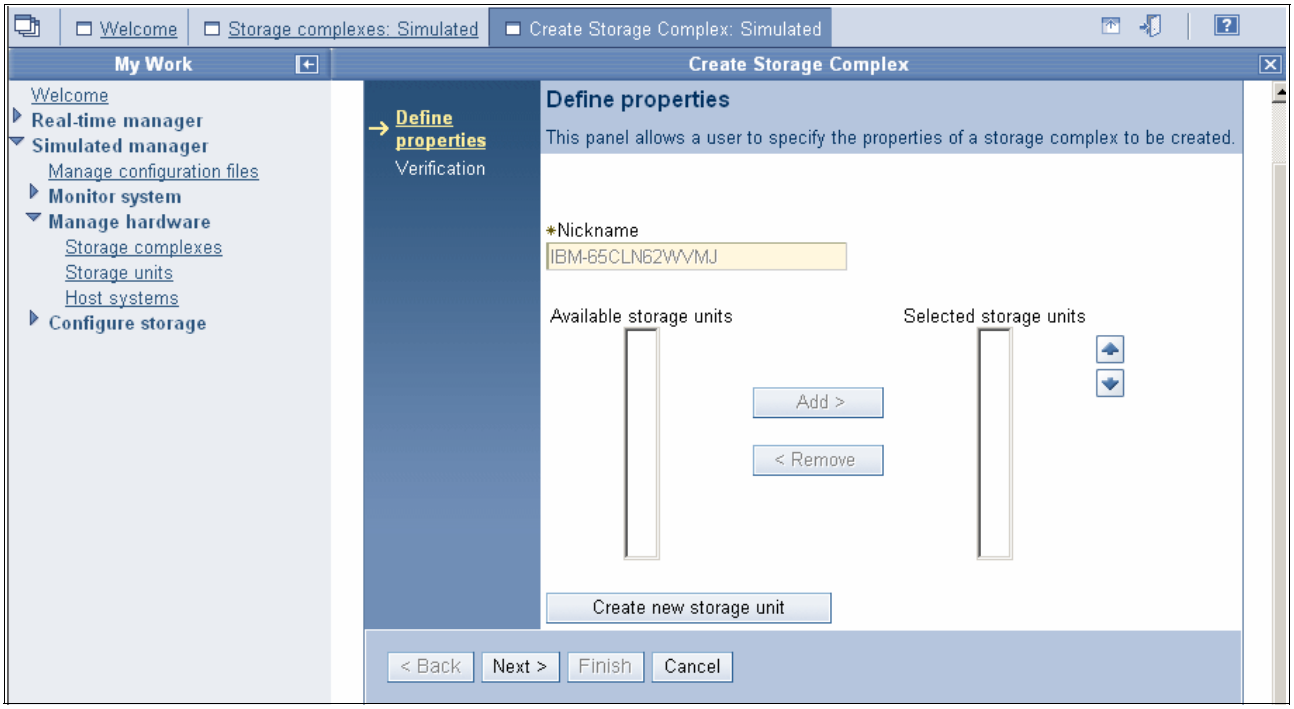



*Figure 9-24 View of the Define properties panel, with the Nickname defined*

Do not click **Create new storage unit** at the bottom of the screen shown in Figure 9-24. Click **Next** and **Finish** in the verification step.

## 9.3.2 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 window.

After clicking **Go**, you will see the General storage information panel as shown in Figure 9-25 on page 174.

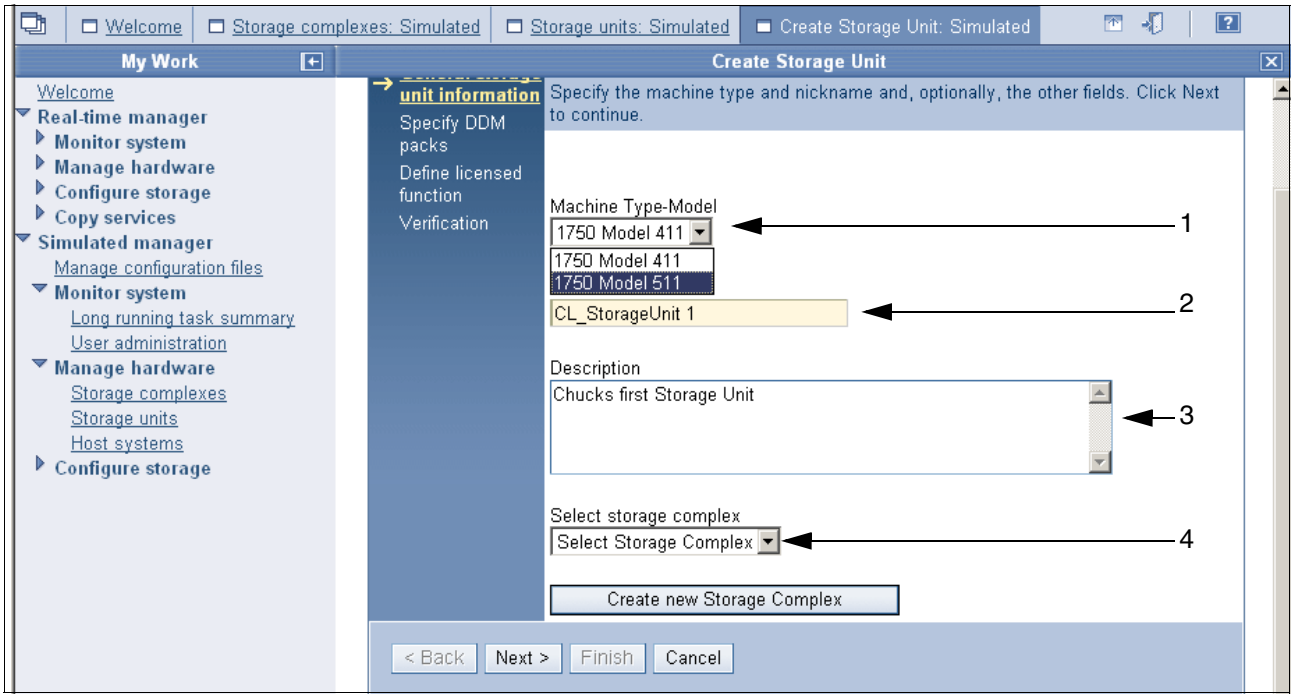



*Figure 9-25 View of the General storage unit information panel*

As illustrated in Figure 9-25, fill in the required fields as follows:

1. Click the **Machine Type-Model** from the pull-down list.
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 unit.

Clicking **Next** will advance you to the Specify DDM packs panel, as shown in Figure 9-26. Fill in the proper information for your specific environment.

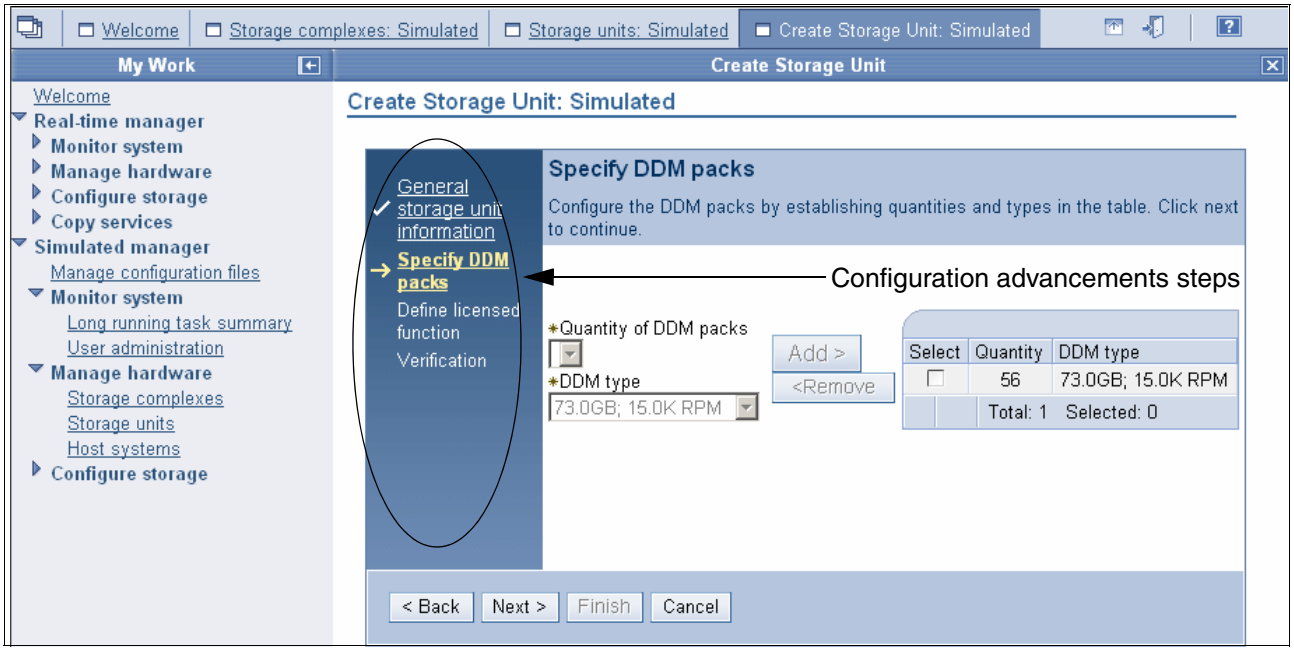



*Figure 9-26 View of Specify DDM packs panel, with the Quantity and DDM type added*

Click **Next** to advance to the Define licensed function panel, under the Create storage unit path, as shown in Figure 9-27.

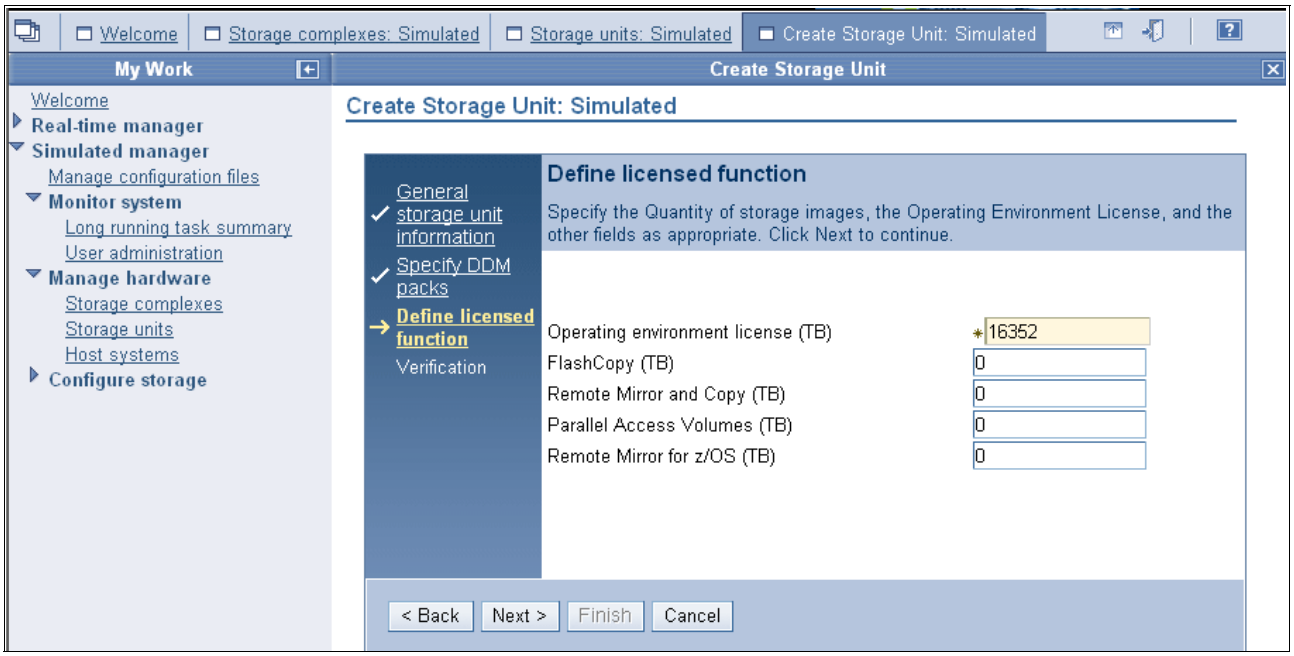



*Figure 9-27 View of the Defined licensed function panel*

Fill in the fields shown in Figure 9-27 as follows:

- The number of licensed TB for the Operation environment.
- The quantity of storage covered by a 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.

Follow the steps specified on the panel shown in Figure 9-27. The next panel, shown in Figure 9-28, requires you to enter the storage type details.

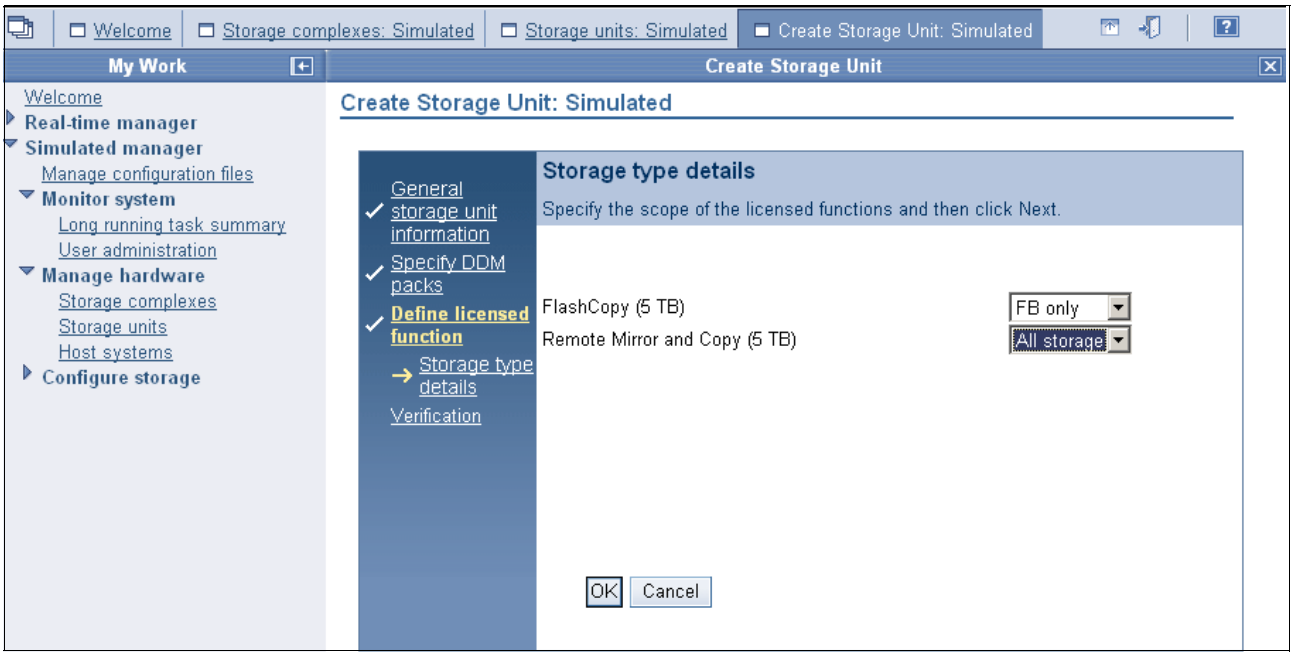



*Figure 9-28 Specify the I/O adapter configuration panel*

Enter the appropriate information and click **OK**.

## 9.3.3 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 9-29. You may want to expand the work area.

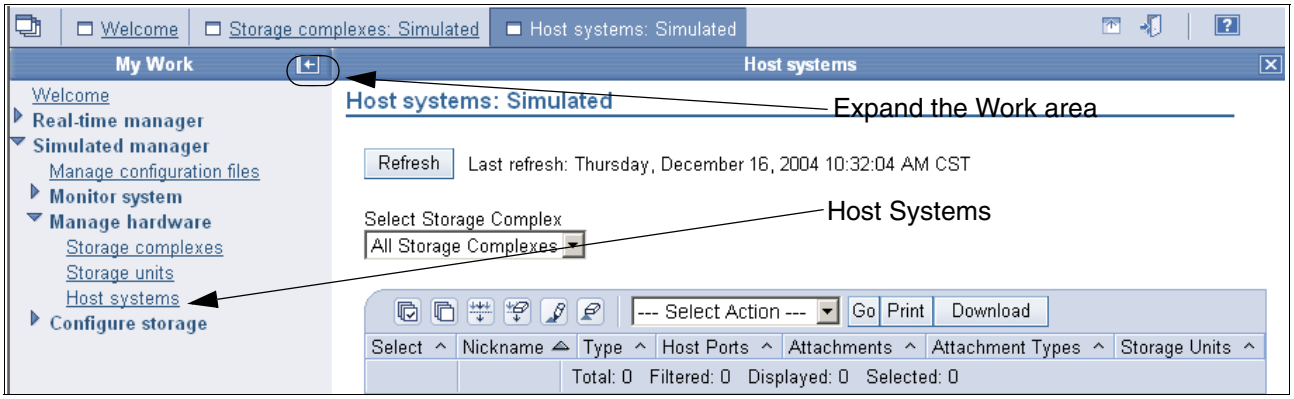



*Figure 9-29 View of create Host systems*

You can expand the view by clicking on the left arrow in the My Work area as shown in Figure 9-29; the expanded view is shown in Figure 9-30.

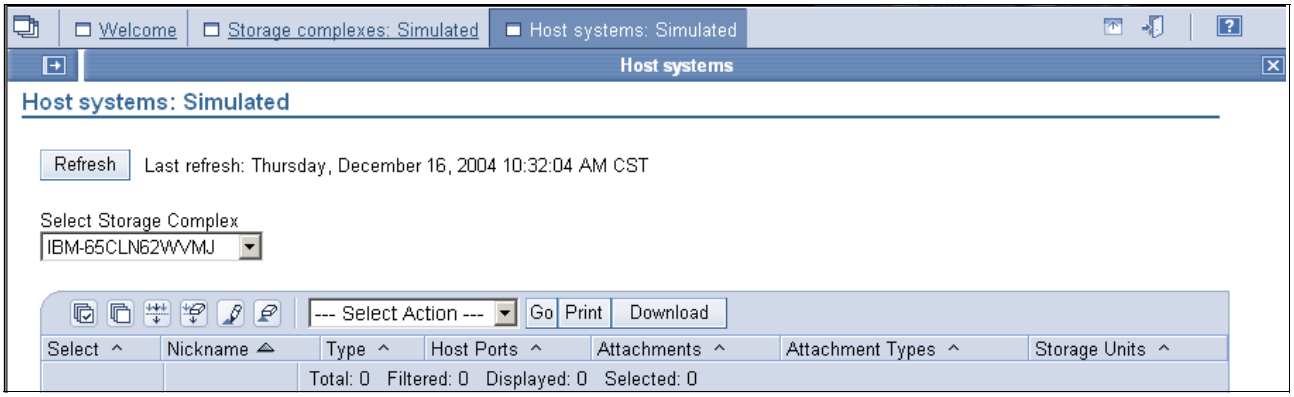



*Figure 9-30 View of Host Systems panel, with the **Go** button selected*

Click the **Select Storage Complex** action pull-down, and highlight the storage complex on which you wish to configure, click **Create** and **Go**. The screen will advance to the General host information panel shown in Figure 9-31.

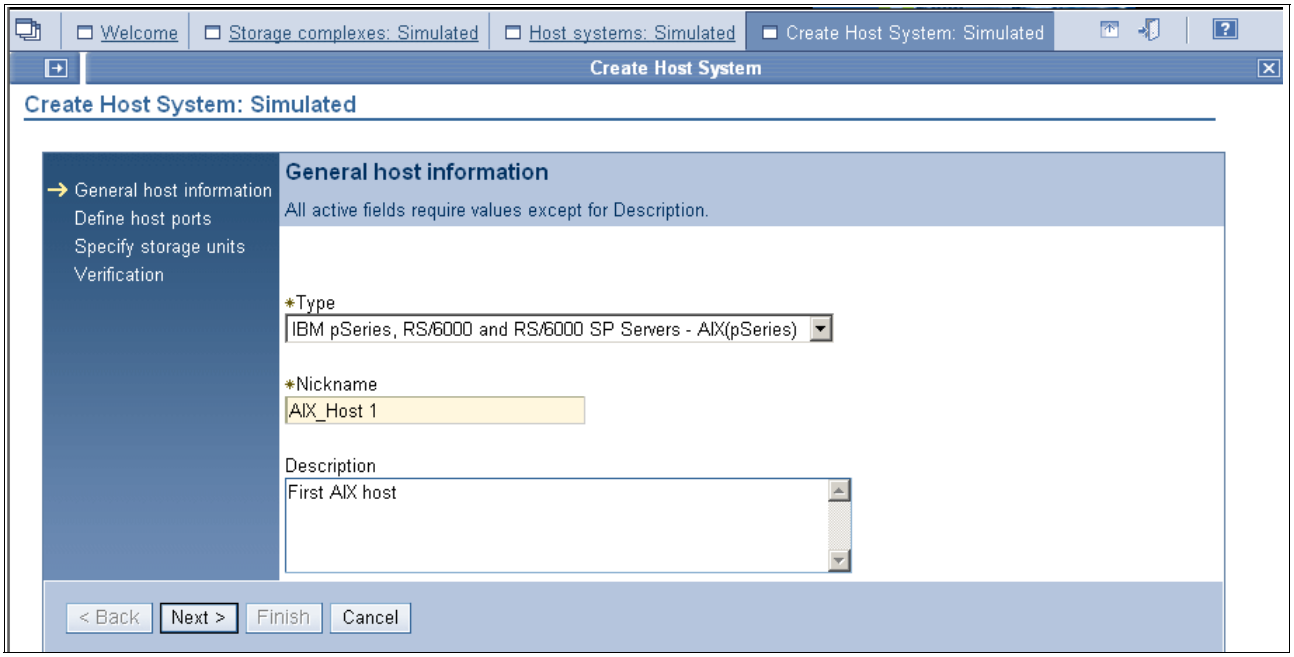



*Figure 9-31 View of the General host information panel*

Click **Next** to advance the screen to the Define Host System panel shown in Figure 9-32 on page 178.

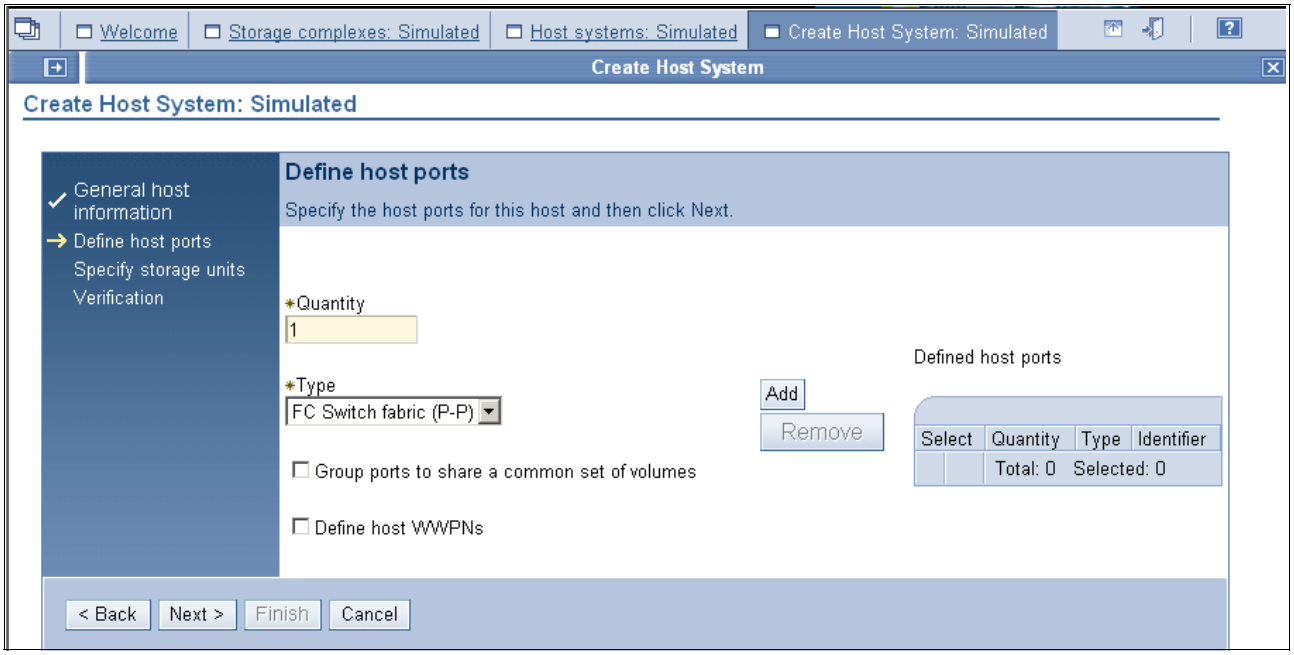



*Figure 9-32 View of Define Host Systems panel*

Enter the appropriate information in the Define host ports panel, as shown in Figure 9-32.

**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 update the new information as shown in Figure 9-33.

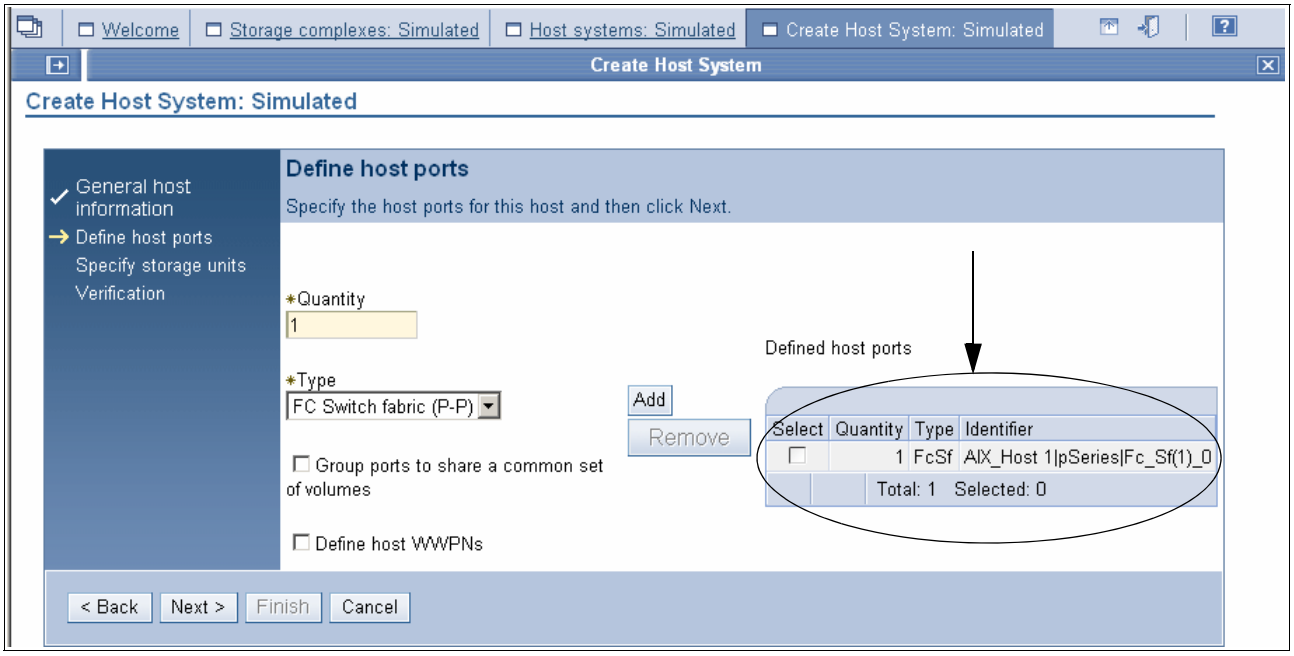



*Figure 9-33 View of Define host ports panel with updated host information*

Click **Next**, and the screen will advance to the Select storage units panel shown in Figure 9-34**.**

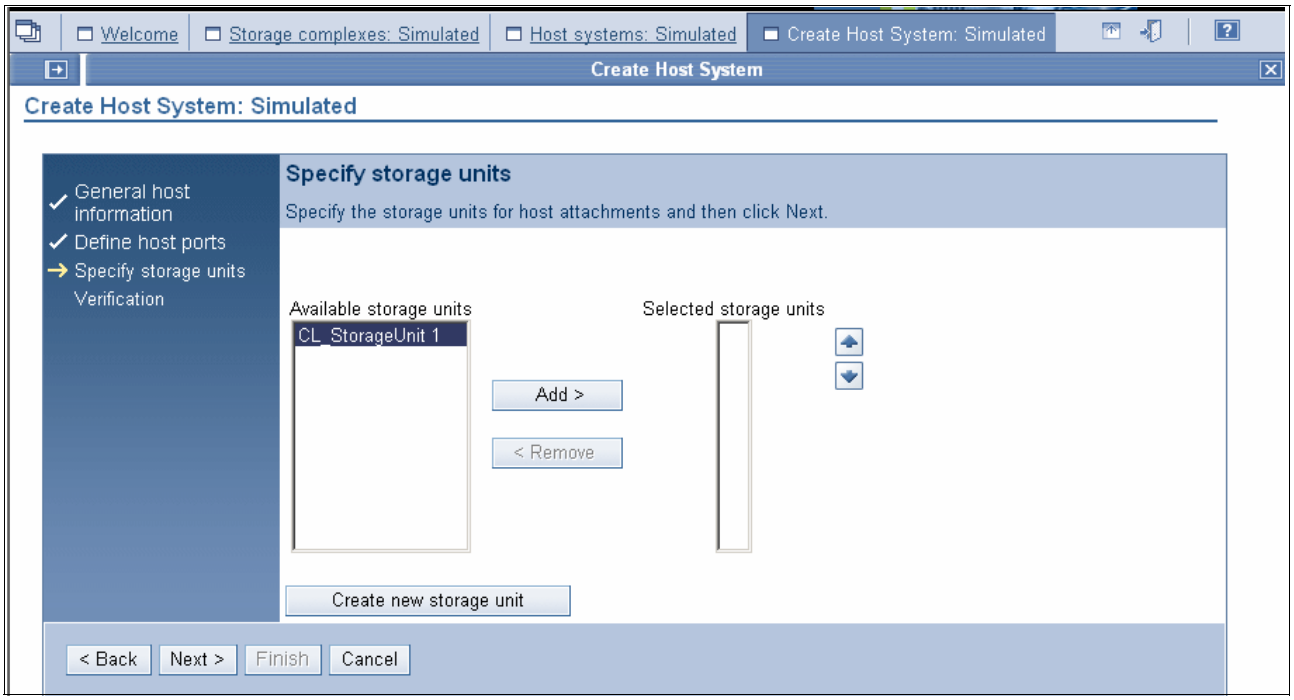



*Figure 9-34 View of the Select storage unit panel*

Highlight the **Available storage units** that you wish, click **Add** and **Next**.

The screen will advance to the Specify storage units parameters shown in Figure 9-35 on page 180.

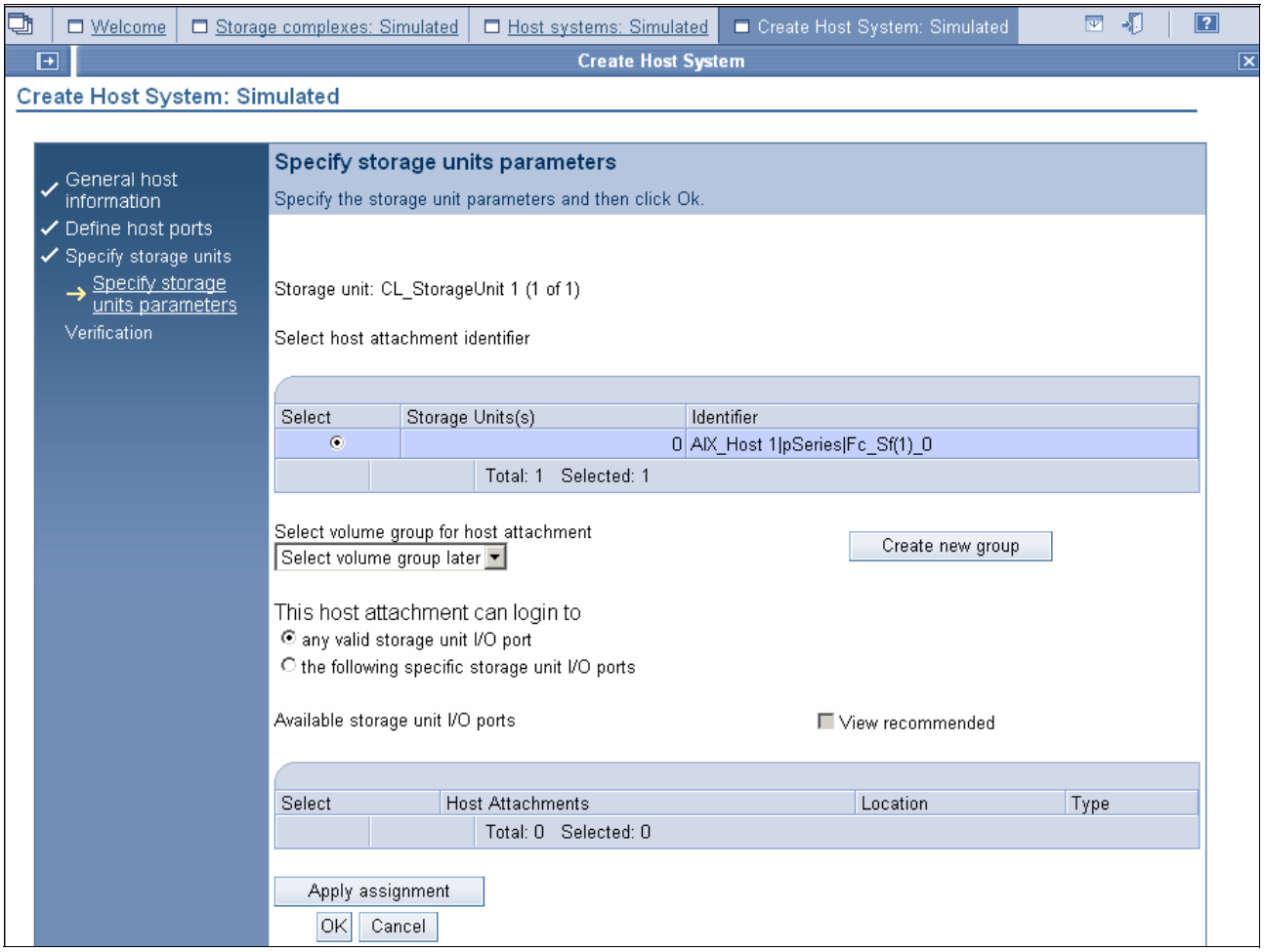



*Figure 9-35 View of the Specify storage units parameters panel*

Under the Specify storage units parameters, do the following:

1. Click the **Select volume group for host attachment** pull-down, and highlight **Select volume group later**.
2. Click **any valid storage unit I/O ports** under the This host attachment can login to field.
3. Click **Apply assignment** and **OK**.
4. Verify and click **Finish**.

## 9.3.4 Creating arrays from array sites

Under **Configure Storage**, click **Arrays**. The screen will advance to the Create Arrays: Simulated panel (not shown here).

Click the **Storage complex** pull-down, and highlight the storage complex on which you wish to configure, click **Create** and **Go**. The screen will advance to the Definition method panel shown in Figure 9-36.

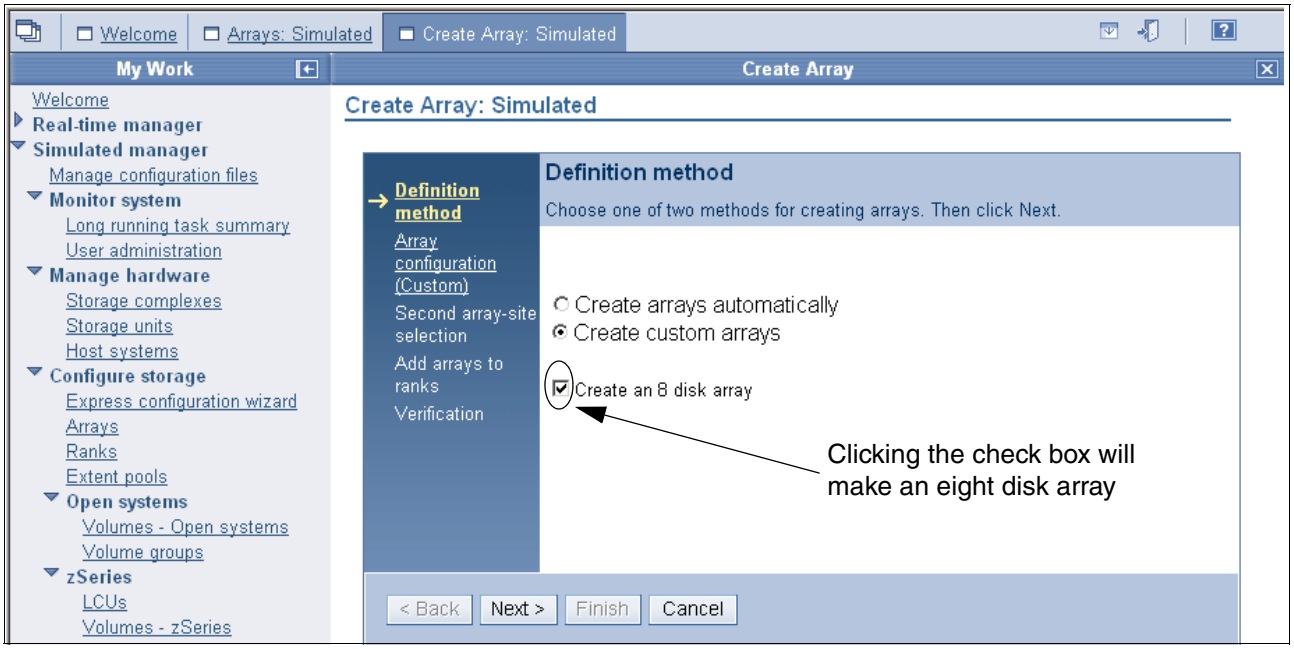



*Figure 9-36 View of the Definition method panel*

From the Definition method panel, if you choose **Create arrays automatically**, the system will automatically take all the space from an array site and place it into an array. You will also notice that by clicking the check box next to **Create an 8 disk array**, that two 4 disk array sites will be placed together to form an eight disk 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, as shown in Figure 9-37.

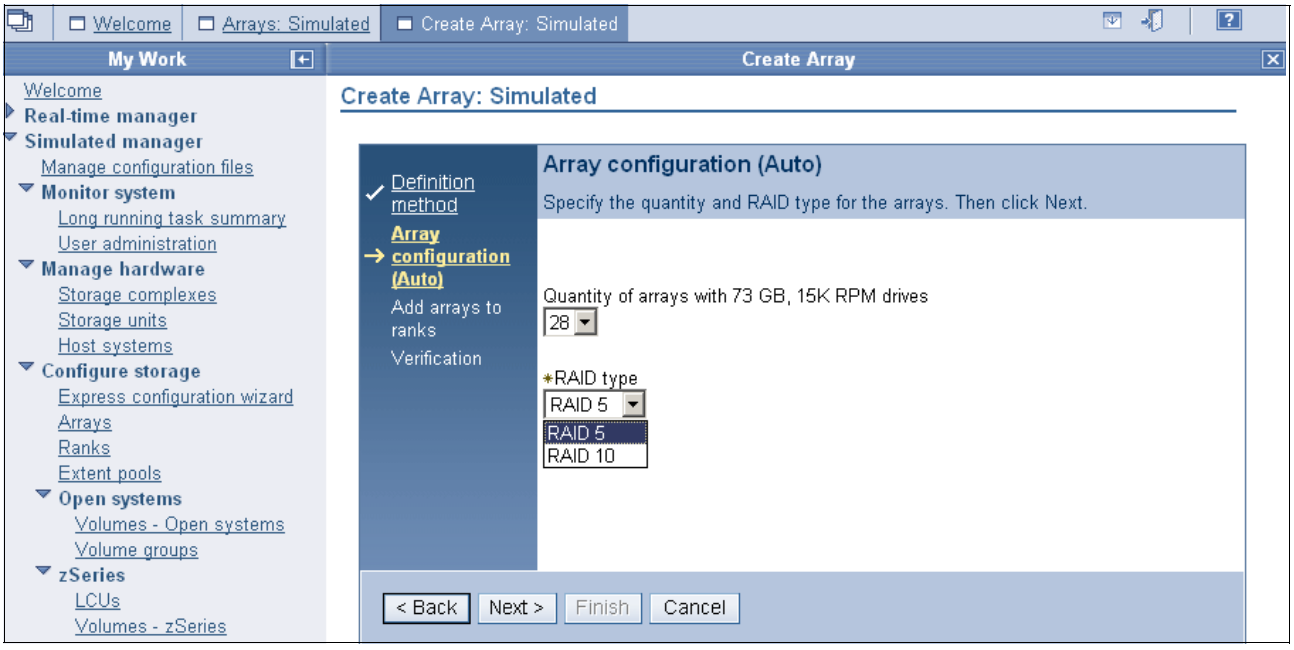



*Figure 9-37* View of the Array configuration (Auto) *panel*

Enter the appropriate information for the quantity of the arrays and the RAID type.

**Note:** If you choose to create eight disk arrays, then you will only have half as many arrays and ranks as if you would have chosen to create four disk arrays.

Click **Next** to advance to the Add arrays to ranks panel shown in Figure 9-38.

If you click the check box next to the **Add these arrays to ranks**, then you will not have to configure the ranks separately at a later time. The ranks take on either storage type **FB** or **CKD**, as shown in Figure 9-38.

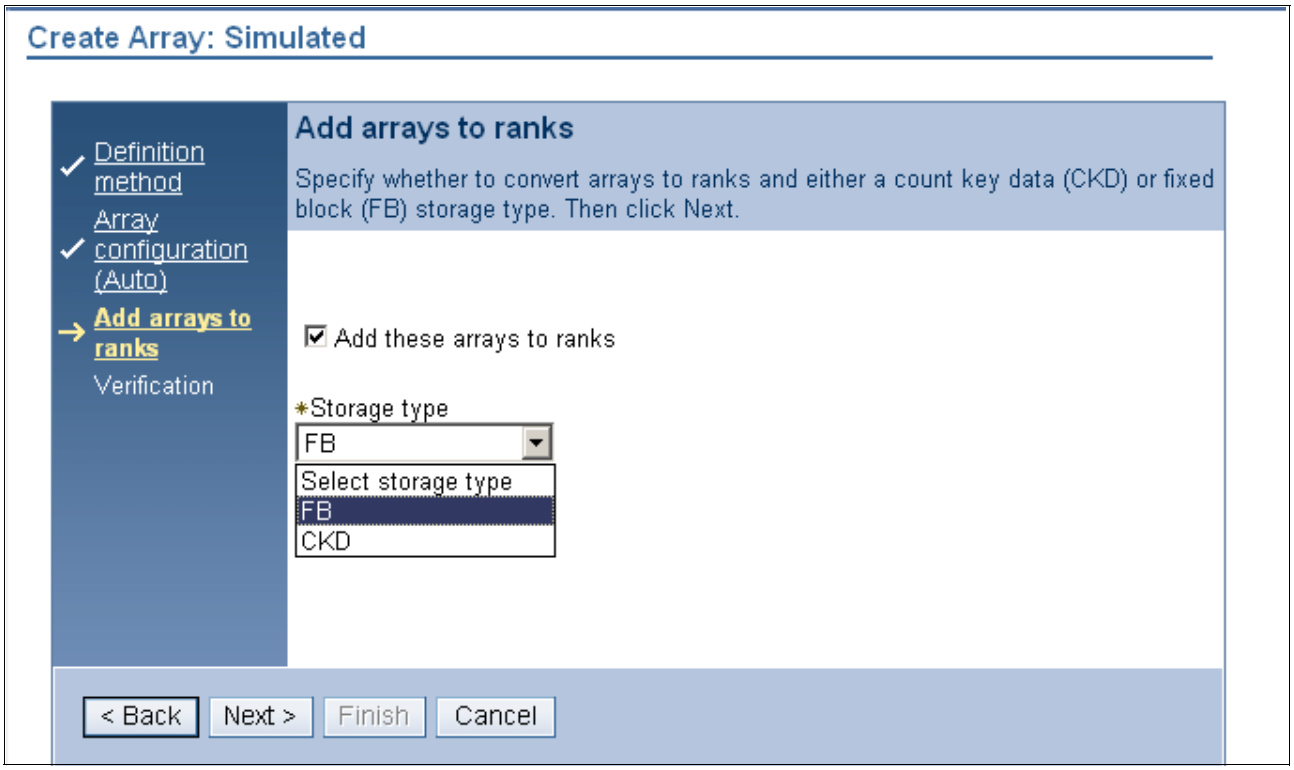



*Figure 9-38 View of the Add arrays to ranks panel, with FB selected*

Click **Next** to bring up the Verification panel, then click **Finish** to configure the arrays and ranks in one step.

If you choose to create custom arrays you will be directed to choose the four disk array sites from which to create your eight disk arrays as shown in Figure 9-39.

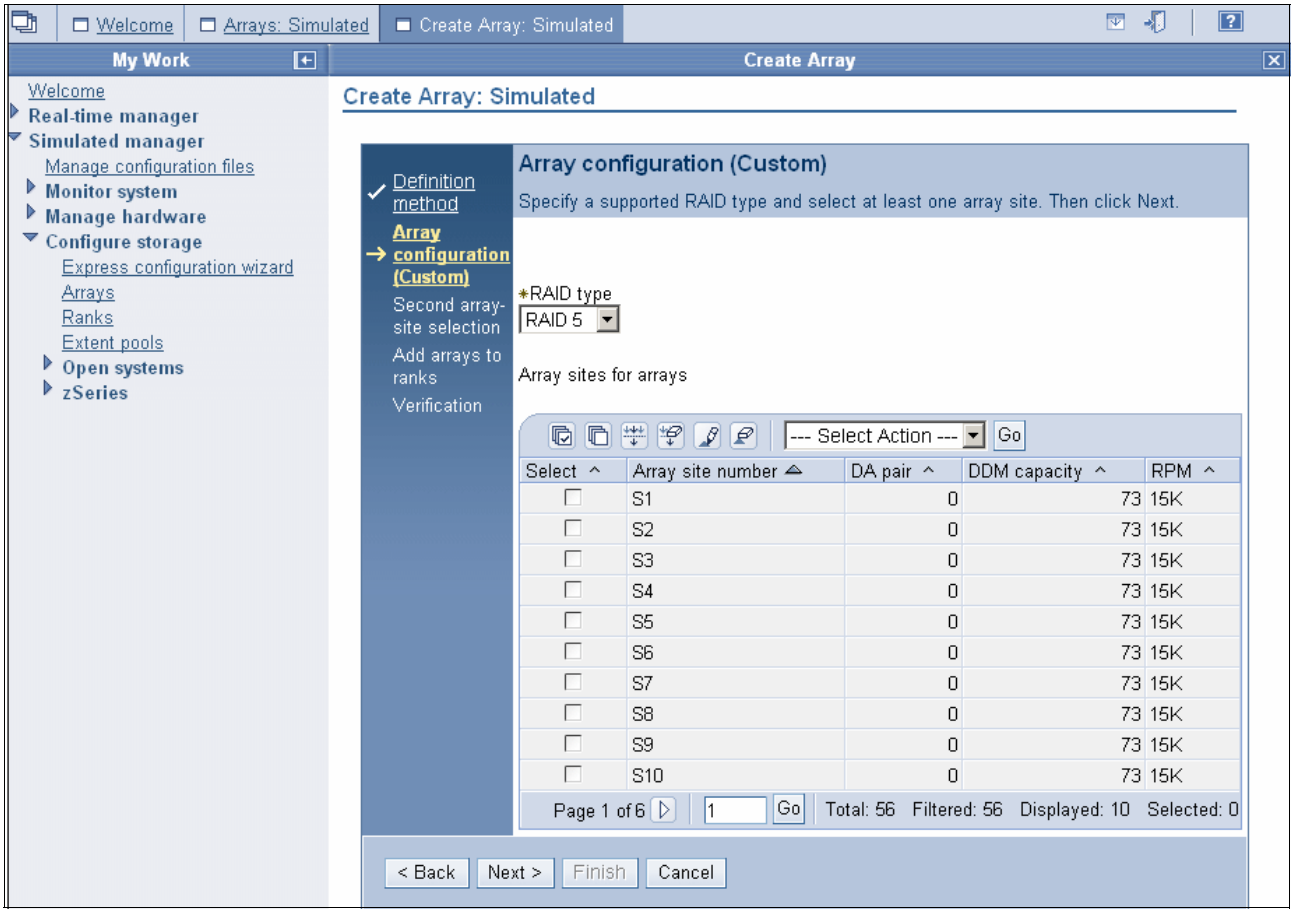



*Figure 9-39 View of creating custom arrays from four disk array sites.*

At this point you can select from the list of four disk array sites to put together to make an eight disk array. If you click **Next** the second array-site selection panel is displayed, as shown in Figure 9-40 on page 184.

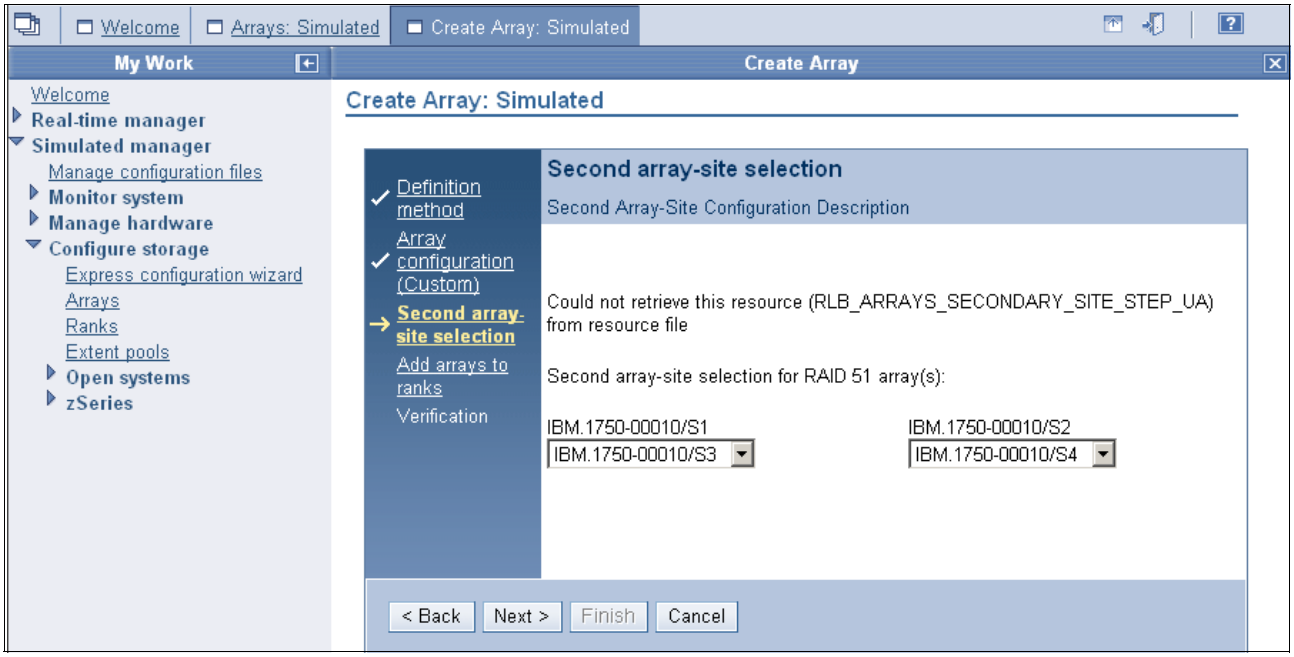



*Figure 9-40 View of the second array-site selection panel*

From this panel you can select the array sites from the pull-down list to make an eight disk array.

## 9.3.5 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 9-41.

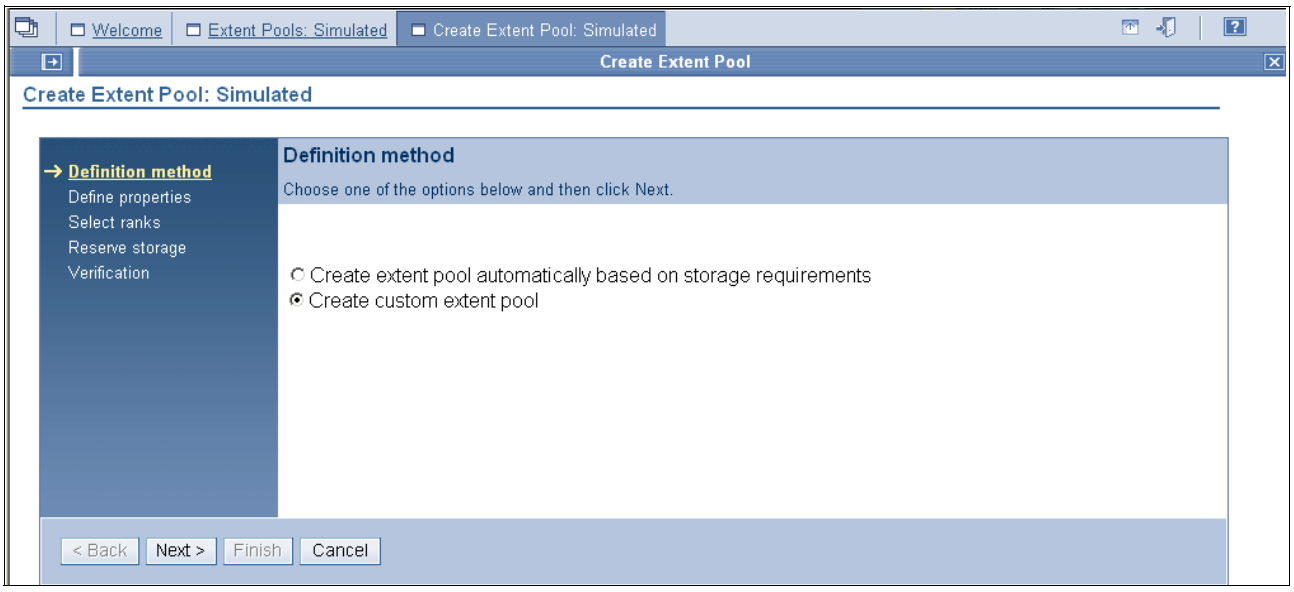



*Figure 9-41 View of the Definition method panel*

The extent pools will take on either a server 0 or server 1 affinity at this point, as shown in Figure 9-42.

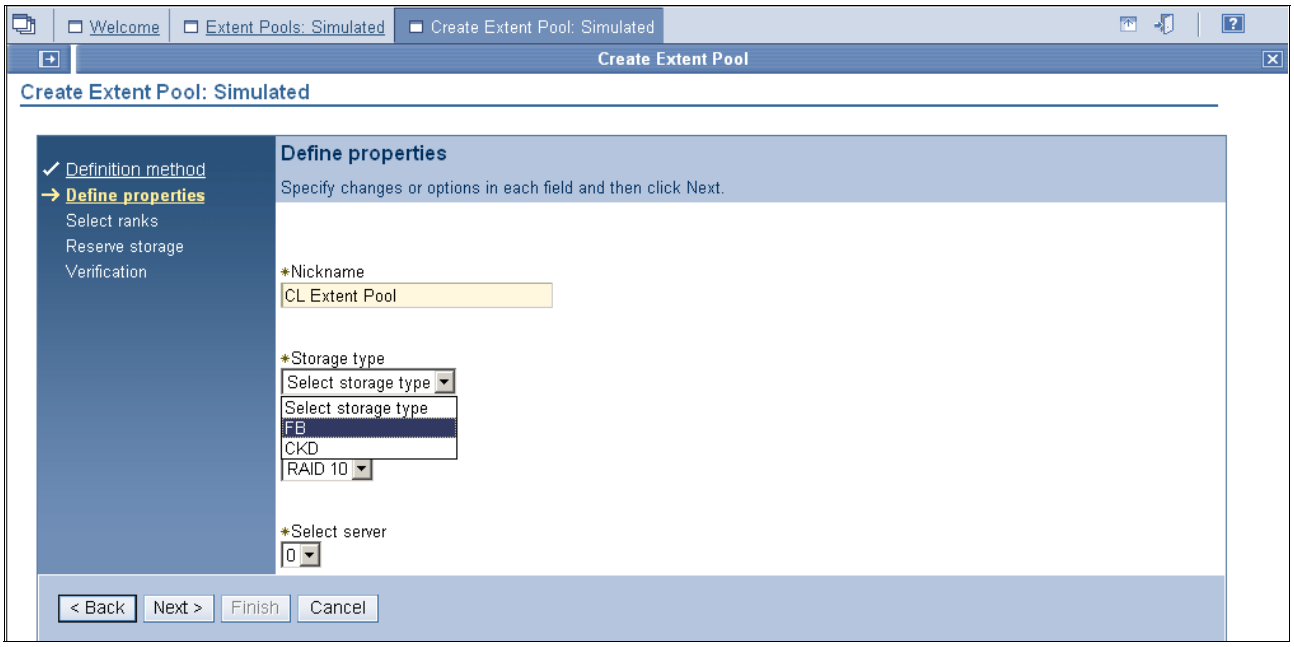



*Figure 9-42 View of Define properties panel*

Click **Next** and **Finish**.

## 9.3.6 Creating FB volumes from extents

Under **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 9-43 on page 186.

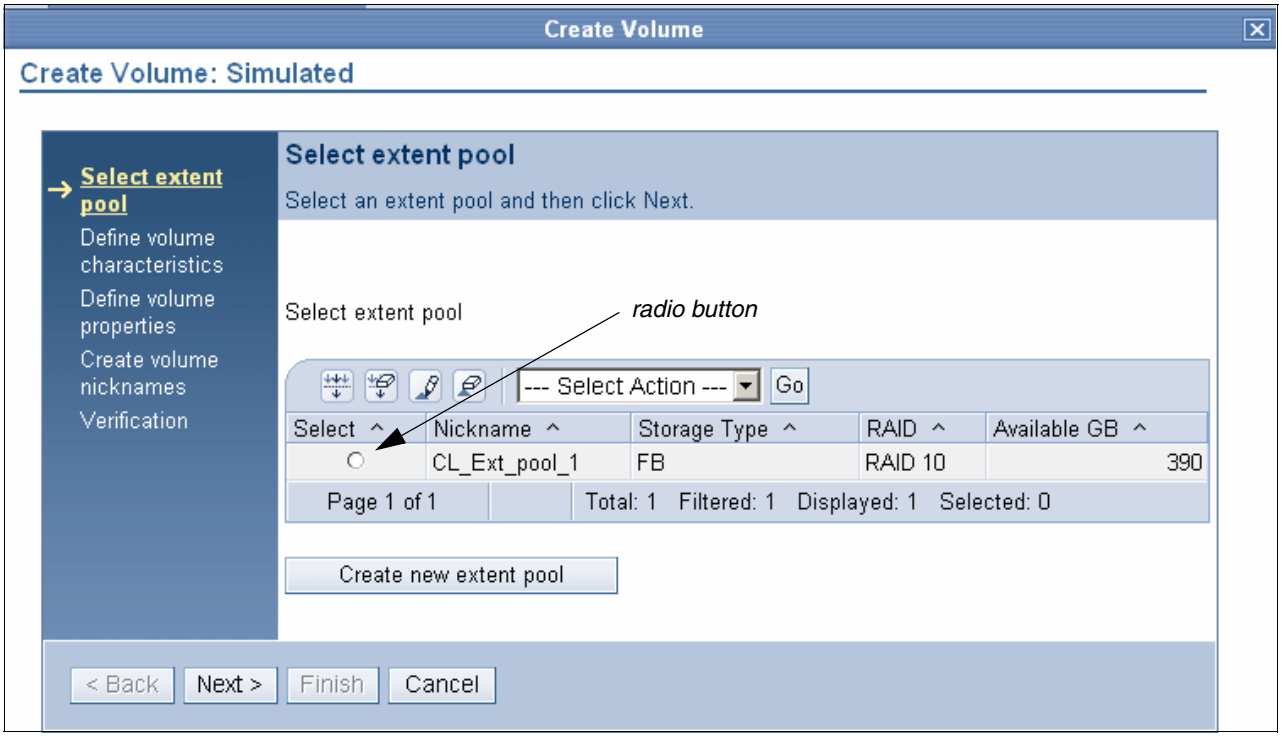



*Figure 9-43 View of the Select extent pool panel*

To determine the quantity and size of the volumes, use the calculators to determine the max size versus quantity as shown in Figure 9-44.

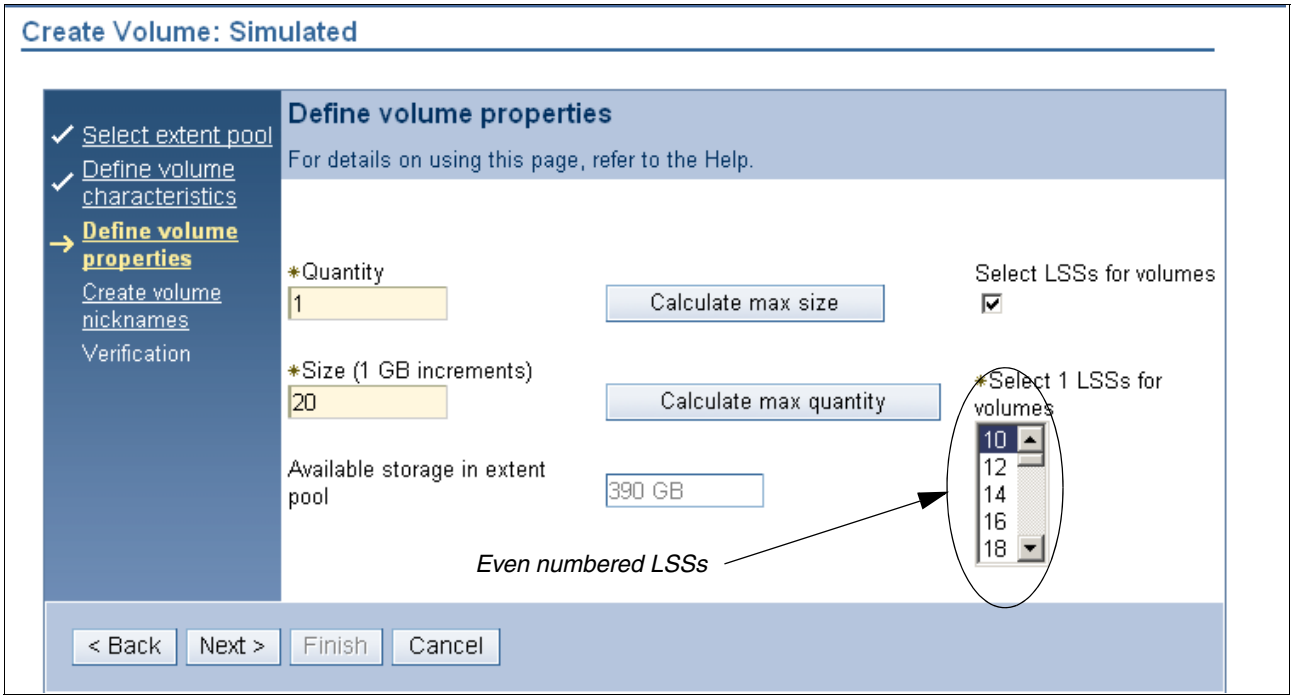



*Figure 9-44 The Define volume properties panel*

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 9-44.

You can give the volume a unique name and number that may help you manage the volumes, as shown in Figure 9-45.

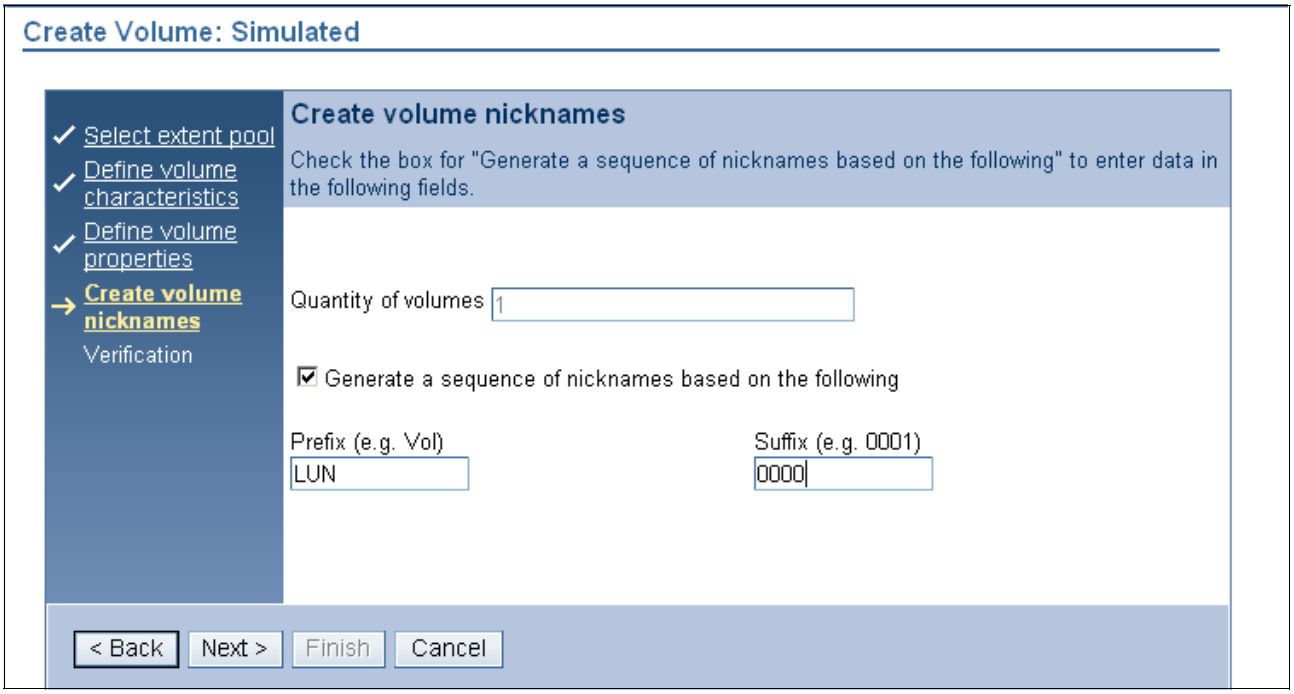



*Figure 9-45 View of the Create volume nicknames panel*

Click **Finish** to end the process of creating the volumes.

## 9.3.7 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 9-46 on page 188.

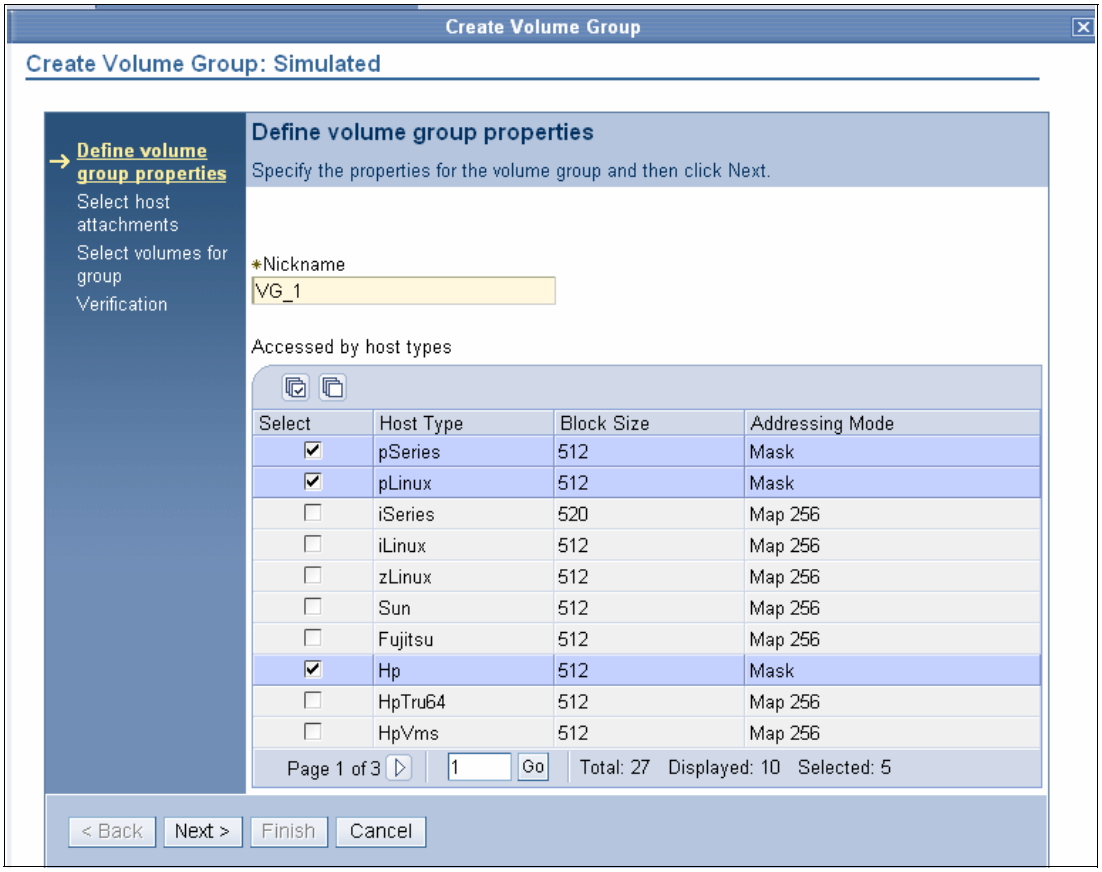



*Figure 9-46 The Define volume group properties filled out*

Select the host attachment you wish to associate the volume group with. See Figure 9-47.

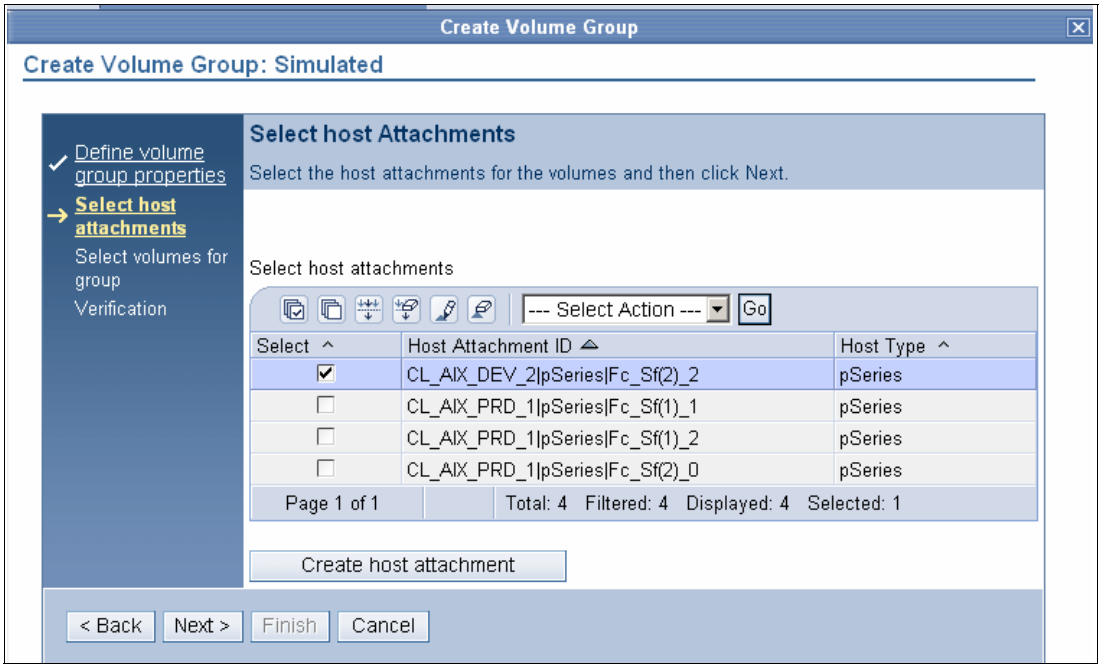



*Figure 9-47 The Select host Attachments panel, with an attachment selected*

Select the volumes for the group panel, as shown in Figure 9-48.

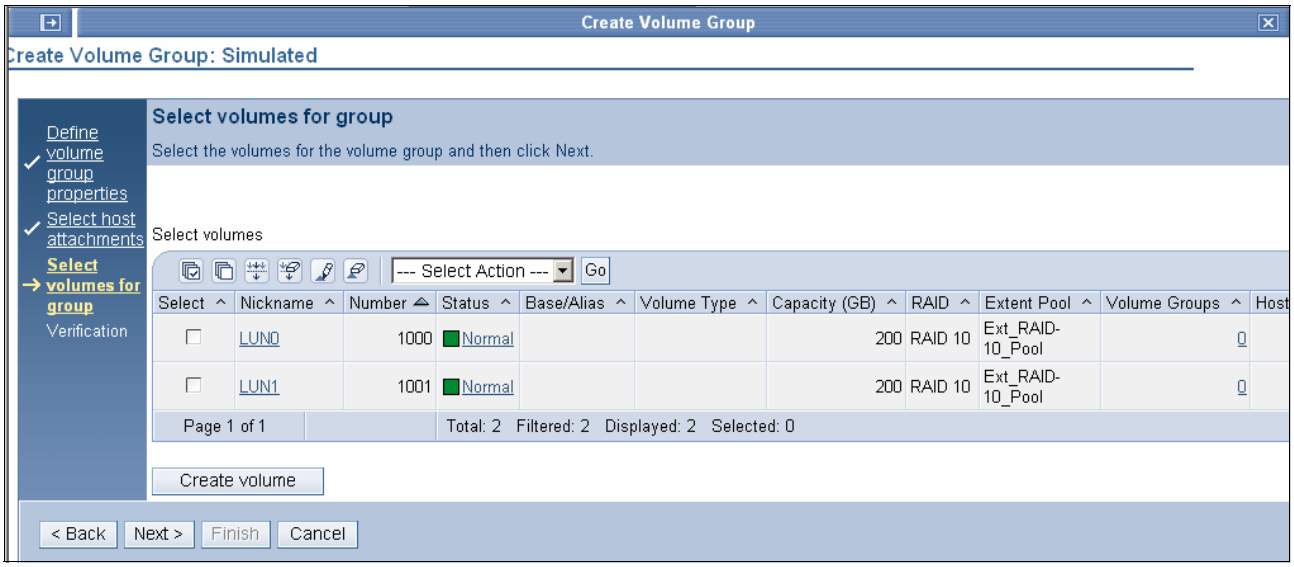



*Figure 9-48 The Select volumes for group panel*

Click **Finish**.

## 9.3.8 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 the **Apply** button.
6. Click **OK**.

You can verify that the volume is now assigned to the desired host's 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.

## 9.3.9 Deleting LUNs and recovering space in the extent pool

Under **Simulated Manager**, perform the following steps to configure the volumes:

1. Click **Volumes**.
2. Click the check box next to the targeted volume you want to delete.
3. Click the Select Action pull-down, and highlight **Delete**.

4. Click **Go**.
5. Click **OK**.

## 9.3.10 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 this panel do the following:
   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**.

## 9.3.11 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, 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. Select the order of Ascending or Descending.
    d. Select the aliases per volumes. For example, 1 alias to every 4 base volumes, or 2 aliases to every 1 base volume.
    e. Click **Next**.
13. Under the Verification panel, click **Finish**.

## 9.3.12 Using the Express Configuration Wizard

After you have created and defined the storage complex, unit, arrays, ranks, and extent pools, you can use the Express Configuration Wizard to create FB or CKD volumes. Under **Configure Storage** do the following:

1. Click **Express Configuration Wizard**.
2. Select the storage unit and volume type and click **Next**.
3. Fill out the appropriate information as shown in Figure 9-49.

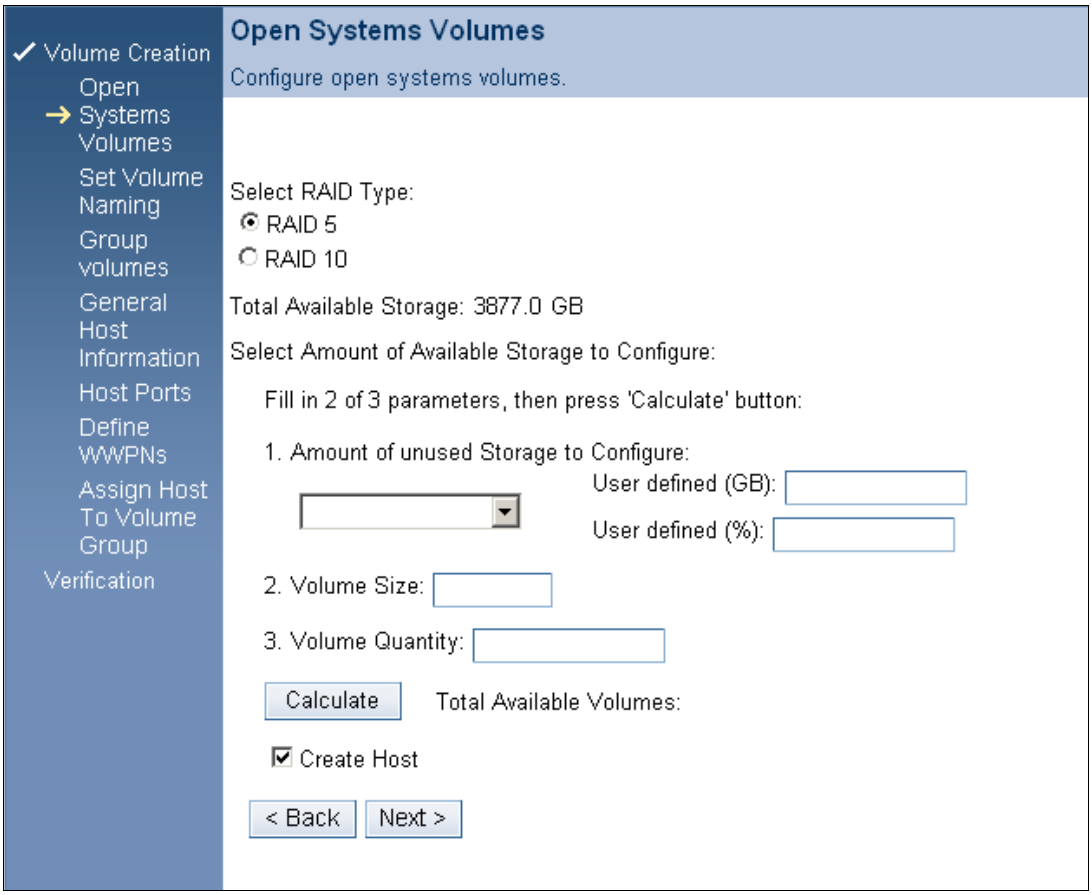



*Figure 9-49 View of the Open System Express volume creation*

4. Continue to follow the menu prompts.

## 9.3.13 Displaying the storage units WWNN in the DS Storage Manager GUI

Under **Simulated manager**, perform the following steps to display the WWNN of the storage unit:

1. Click **Simulated manager** as shown in Figure 9-50.

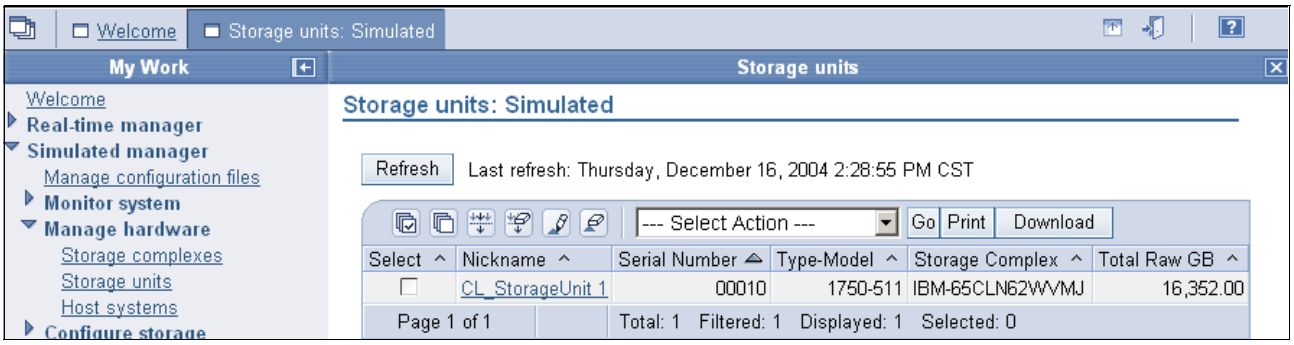



*Figure 9-50 View of the Real-time Manager panel*

2. Click **Storage units**.
3. Select the check box beside the storage unit name, as shown in Figure 9-51.

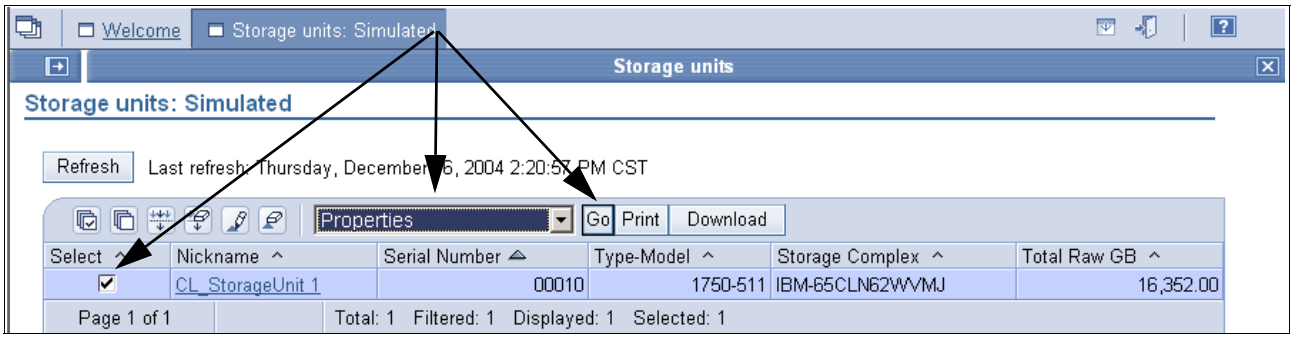



*Figure 9-51 View of the storage unit with the radio button and the Properties selected*

4. Select **Properties** from the pull-down as shown in Figure 9-51.
5. Click **Go**. The panel will advance to the General panel shown in Figure 9-52.

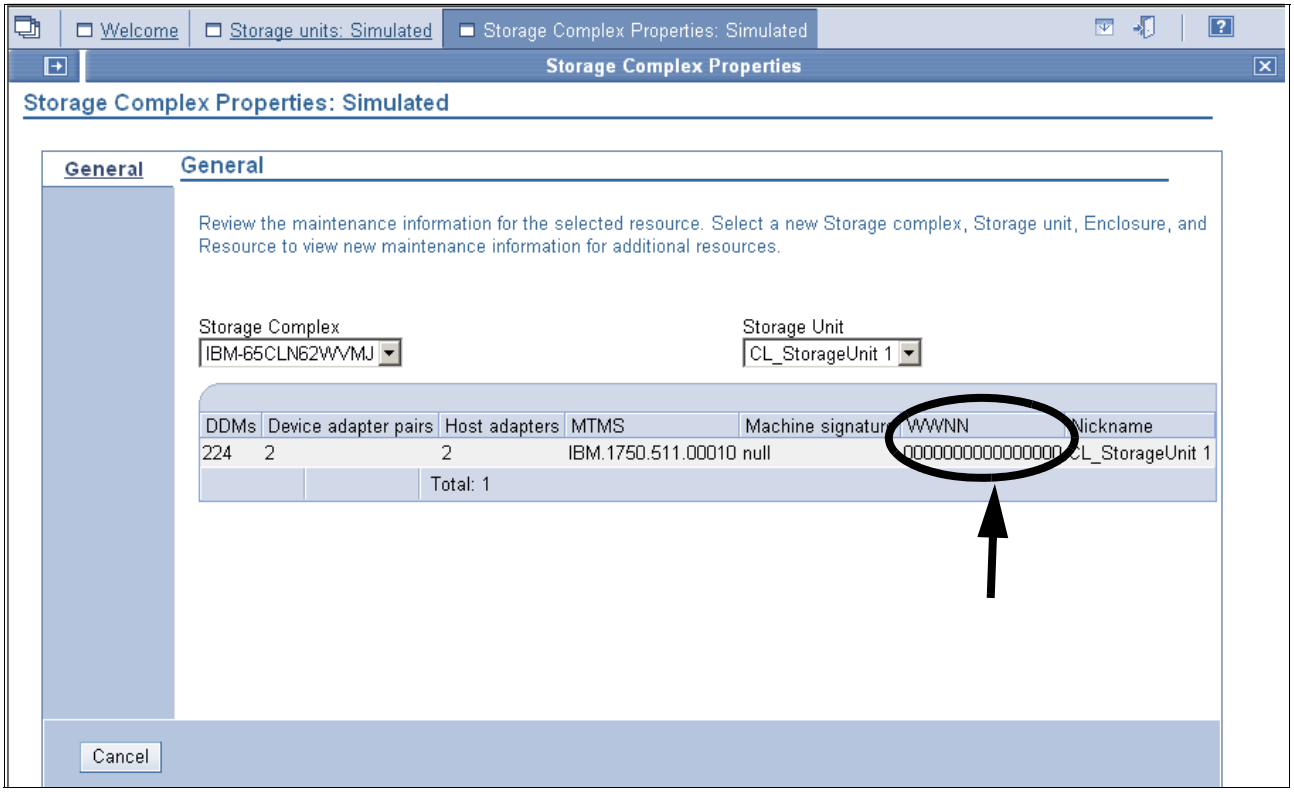



*Figure 9-52 View of the WWNN in the General panel*

The WWNN number is displayed as shown in Figure 9-52.

## 9.4 Summary

In this chapter we have discussed the configuration hierarchy, terminology and concepts. We have recommended an order and methodology for configuring the DS6000 storage server. We have included some logical configuration steps and examples and explained how to navigate the GUI.

10

# 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 it 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

## 10.1 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`

## 10.2 Functionality

The DS CLI can be used to invoke the following storage configuration tasks:

- Create userids that can be used with the GUI and the DS CLI
- 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

## 10.3 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, Windows XP 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

## 10.4 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

## 10.5 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 10-1 on page 199 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.

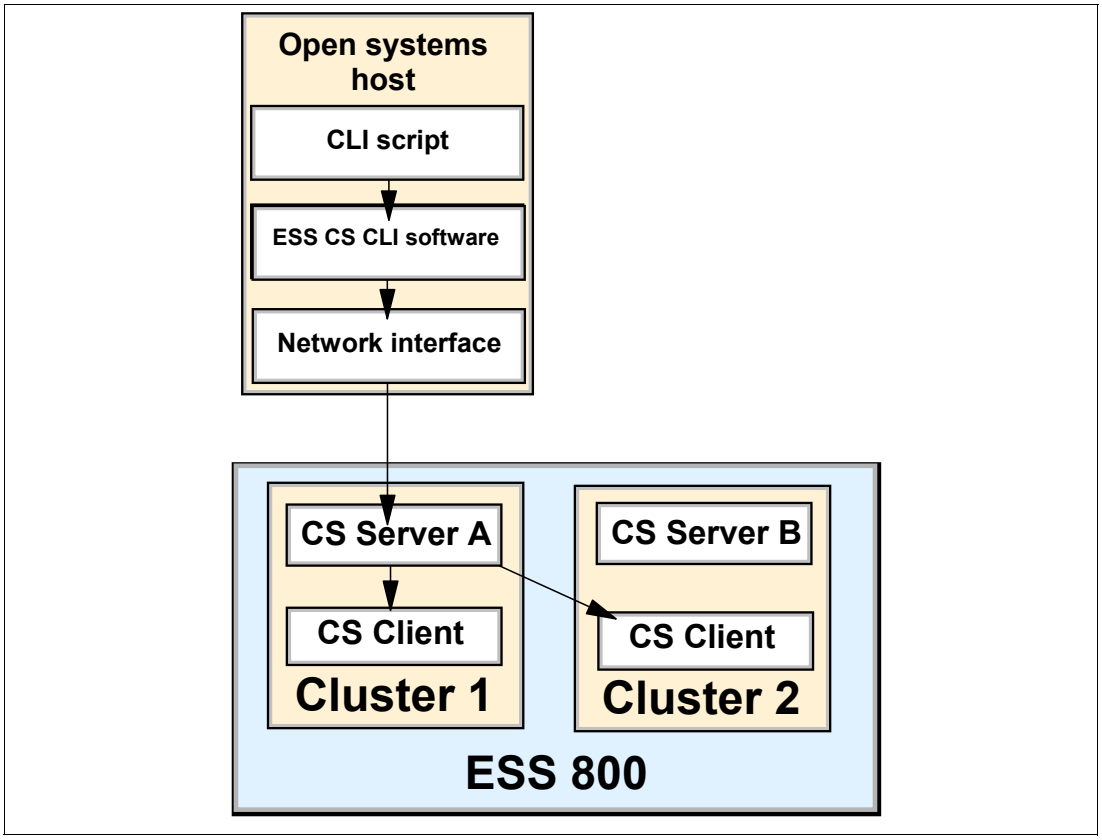



*Figure 10-1 Command flow for ESS 800 Copy Services commands*

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 8, “Configuration planning” on page 125. This is important for planning considerations, where a firewall may exist between the client and the server.

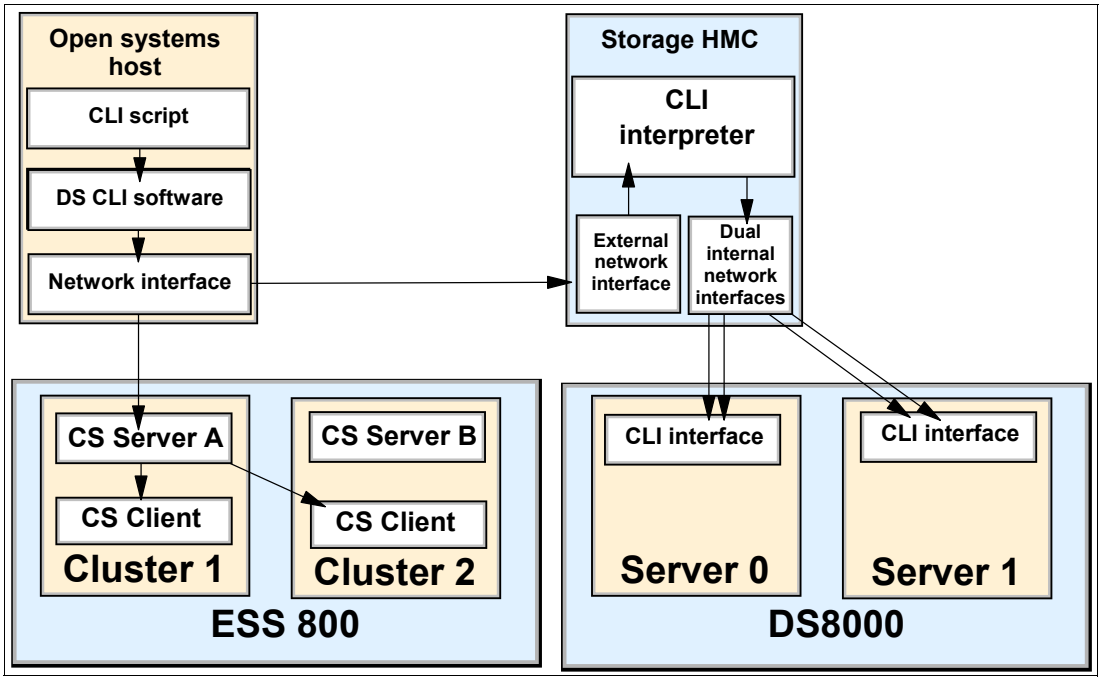



*Figure 10-2 DS CLI Copy Services command flow*

## DS8000 split network

One thing that you may notice about Figure 10-2 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 10-3 on page 201.

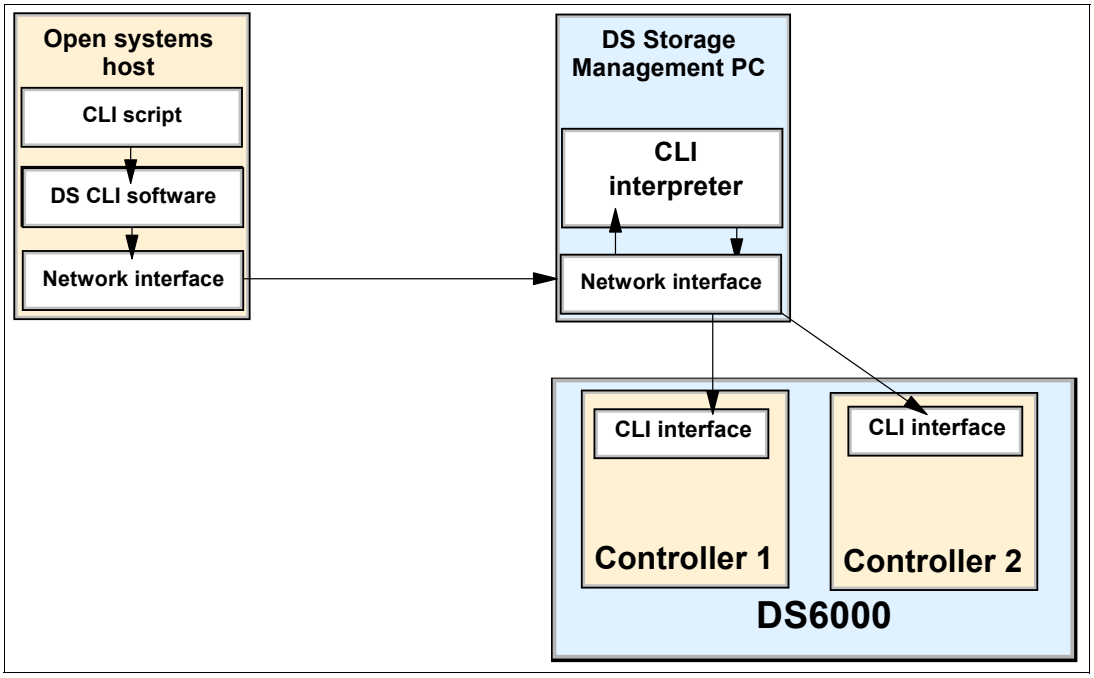



*Figure 10-3 Command flow for the DS6000*

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 10-4 on page 202. 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.

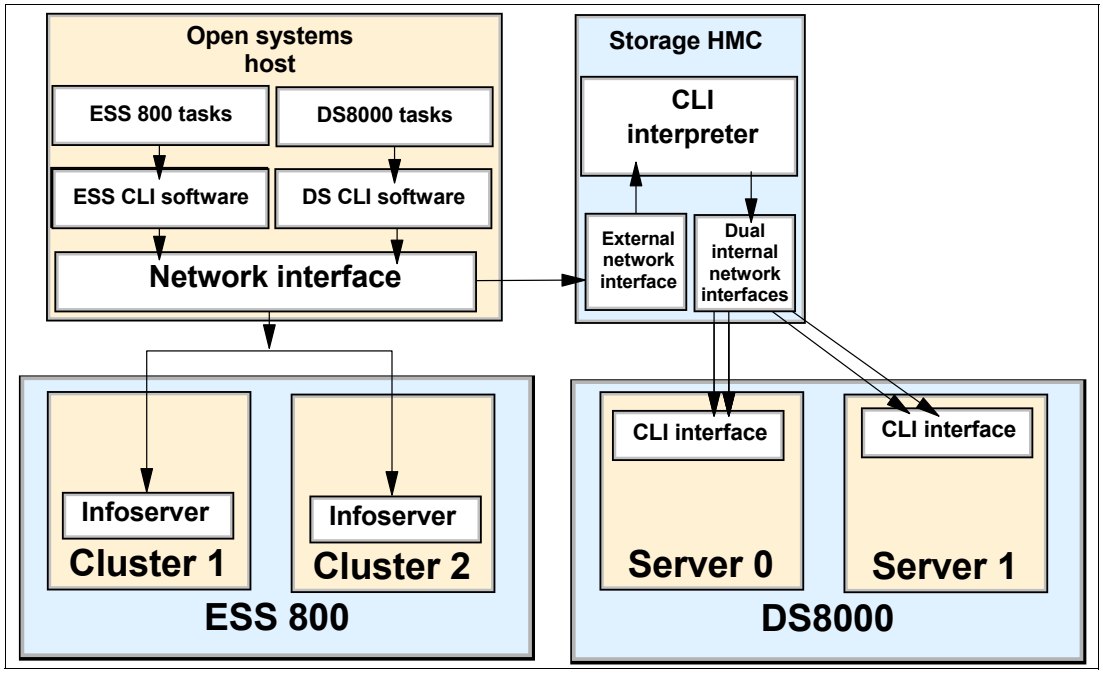



*Figure 10-4 CLI co-existence*

## 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 10-5.

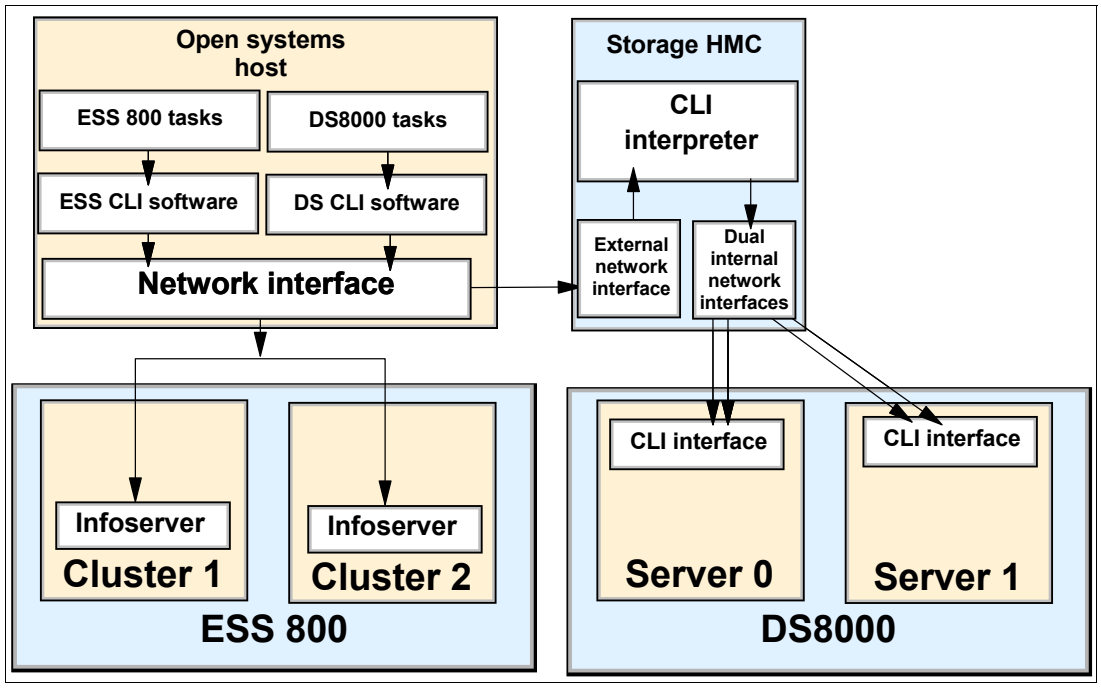



*Figure 10-5 Storage management command flow*

## 10.6 User security

The DS CLI software must authenticate with the DS MC 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 then 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 default password is also *admin*). During setup it is advisable that a new userid be created in the admin group (for use if the password for the *admin* userid is lost). 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 213.

## 10.7 Usage concepts

It is important to understand the various concepts that frame DS CLI usage.

### 10.7.1 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.

#### *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 or a password file would have to be created first. This is shown in Example 10-1.

*Example 10-1 Using DS CLI via a single command*

```
C:\Program Files\IBM\dscli>dscli 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 10-2.

*Example 10-2 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
```

If you are familiar with UNIX, then a simple example of creating a script is shown in Example 10-3.

*Example 10-3 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 or a password file would have to be created first. The use of the interactive mode is shown in Example 10-4.

*Example 10-4 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 10-5.

*Example 10-5 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 10-6.

*Example 10-6 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.

## 10.7.2 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.

## 10.7.3 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:

**help** Lists all available DS CLI commands.

**help -s** Lists all available DS CLI commands with brief descriptions of each.

**help -l** 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 10-7, the **-s** parameter is used to get a short description of the **mkflash** command's purpose.

*Example 10-7 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 10-8, the **-l** parameter is used to get a list of all the parameters that can be used with the **mkflash** command.

*Example 10-8 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**, **-help**, or **-?**, for example:

dscli> **mkflash -help**

or

dscli> **help mkflash**

## 10.7.4 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 10-1.

*Table 10-1 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 10-9 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 10-9 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 10-10 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. 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 10-10 Return code examples*

```
C:\Program Files\IBM\dscli> checkflash.bat
CMUN00018E lsflash: Unable to connect to the management console server
exit status of dscli = 3
A connection error occurred. Try pinging 10.0.0.1
If this fails call network support on 555-1001

C:\Program Files\IBM\dscli>
```

## 10.8 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*, GC26-7681. Or you can use the online help. Example 10-11 gives a sample configuration script showing most of the storage management commands that are used on a DS6000 or DS8000.

*Example 10-11 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 -arsite 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 0200-021F
# 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 0400-041F


#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 -dev IBM.2107-9999999 I0010
setioport -topology ficon -dev IBM.2107-9999999 I0011
```

## 10.9 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 10-2 shows which CLI should be used based on which hardware is installed in a particular environment.

*Table 10-2 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.

## 10.9.1 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.

# 10.10 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.

## 10.10.1 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 10-6 and Example 10-12 show two ways to get a list of saved tasks on the ESS CS Server. In Figure 10-6 the **Tasks** button has been selected in the ESS 800 Web Copy Services GUI.

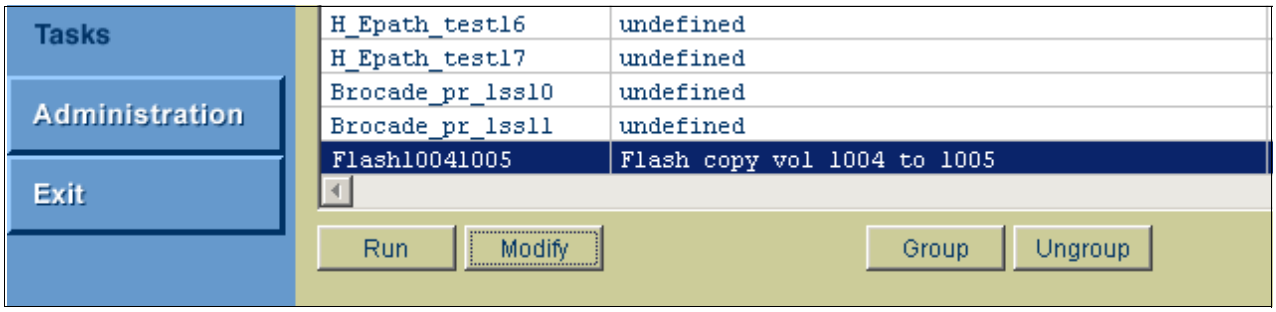



*Figure 10-6 A portion of the tasks listed by using the GUI*

In Example 10-12, the **`list task`** command is used. This is an ESS CLI command.

*Example 10-12 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 passw0rd
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_lss10               PPRCEstablishPair              NotRunning
Brocade_pr_lss11               PPRCEstablishPair              NotRunning
Flash10041005                  FCEstablish                    NotRunning
```

## 10.10.2 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 10-7 on page 211.

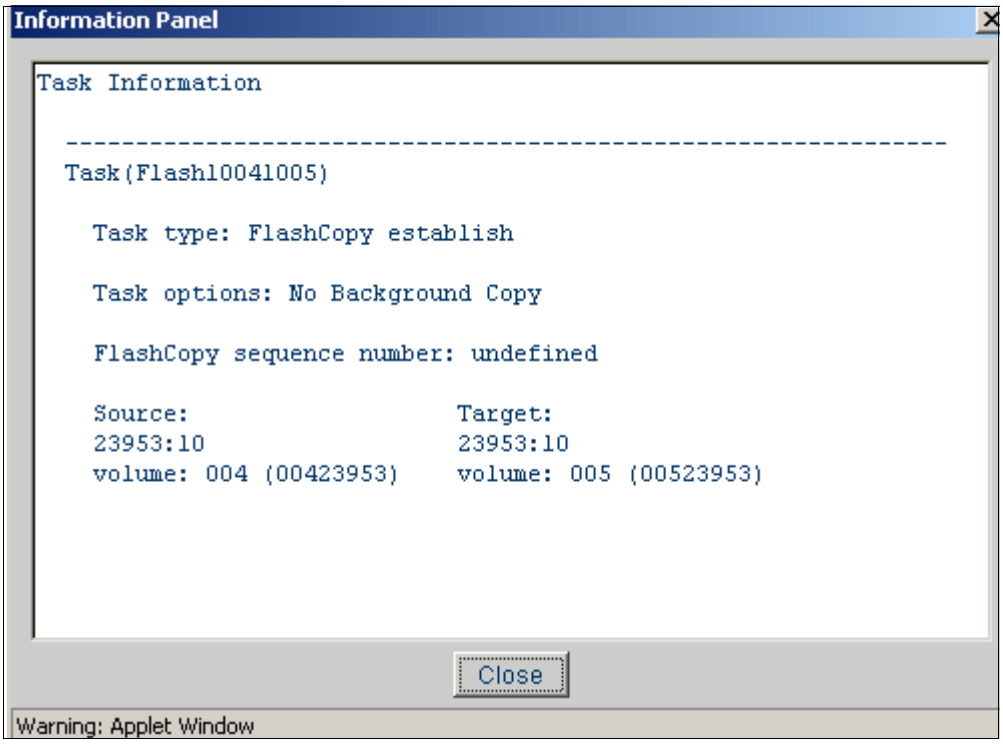



*Figure 10-7 Using the GUI to get the contents of a FlashCopy task*

It makes more sense, however, to use the ESS CLI **show task** command to list the contents of the tasks, as depicted in Example 10-13.

*Example 10-13 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
```

## 10.10.3 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 10-13, each parameter is translated to the correct value for a DS CLI command in Table 10-3.

*Table 10-3 Converting a FlashCopy task to DS CLI*

| ESS CS CLI parameter | Saved task parameter | DS CLI coversion | 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**

**Important:** On the ESS 800, open systems volume IDs are given in an 8 digit format, xxx-sssss where xxx is the LUN ID and sssss is the serial number of the ESS 800. In the example used in this appendix the volumes shown are 004-23953 to 005-23953. These volumes are open systems, or fixed block volumes. When referring to them in the DS CLI, you must add 1000 to the volume ID, so volume 004-23953 is volume ID 1004 and volume 005-23953 is volume ID 1005. This is very important because on the ESS 800, the following address ranges are actually used:

0000 to 0FFF CKD z/Series volumes (4096 possible addresses)
1000 to 1FFF Open systems fixed block LUNs (4096 possible addresses)

If we intend to use FlashCopy to copy ESS LUN 004-23953 onto 005-23953 using DS CLI, we must specify 1004 and 1005. If instead we specify 0004 and 0005, we will actually run the FlashCopy against CKD volumes. This may result in an unplanned outage on the zSeries system that was using CKD volume 0005.

The ESS CLI command, **show task**, will show the correct value for the volume ID

In Example 10-14 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 10-14 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 10-8.

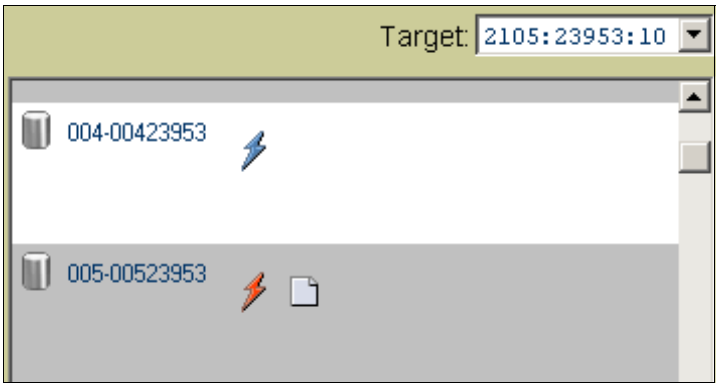



*Figure 10-8 FlashCopy status via the ESS 800 Web Copy Services GUI*

## 10.10.4 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 tempw0rd 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 *tempw0rd* 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. When the userid is first used, the password has to be changed. In this example, the user changes the password to *passw0rd*:

```
chuser -pw passw0rd csasdmin
```

Now if a password file is needed, it will need to be created with the **managepwfile** command. An example is:

```
managepwfile -action add -name csadmin -pw passw0rd
```

Having added the userid called *csadmin*, the password has been saved in an encrypted file called *security.dat*. By default, the file is placed in:

- c:\Documents and settings\<Windows user name>\DSCLI\security.dat

or

- $HOME/dscli/security.dat

You can however use the **-pwfile** parameter to place this file anywhere.

## 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. By 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 10-15.

*Example 10-15 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

# Username must be specified if a password file is used
username: csadmin
# Password filename is the name of an encrypted password file
# This file is located at C:\Program Files\IBM\dscli
pwfile: security.dat

# 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 real DS MC, then you can specify the 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 10-16.

*Example 10-16 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: passw0rd
```

```
# 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 10-17. In this example the **lsuser** command is issued directly to a DS MC. 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 10-17 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 passw0rd 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 10-18.

*Example 10-18 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 10-19.

*Example 10-19 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 10-20.

*Example 10-20 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 >
```

## 10.11 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.

# Part 4

# Implementation and management in the z/OS environment

In this part we discuss considerations for the DS6000 series when used in the z/OS environment. The topics include:

- z/OS software
- Data migration in the z/OS environment

11

# Performance considerations

This chapter discusses early performance considerations regarding the DS6000 series.

Disk Magic modelling for DS6000 is going to be available in early 2005. 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 DS6000 addresses this challenge
- Specific considerations for open systems and z/OS

## 11.1 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, introduced its Seascape® architecture, and brought the ESS into the marketplace in 1999 based on this architecture.

### 11.1.1 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.

### 11.1.2 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 solved and are the driving force to develop storage servers such as the new DS6000 series.

With the advent of the DS6000 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 11-1 on page 221 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.

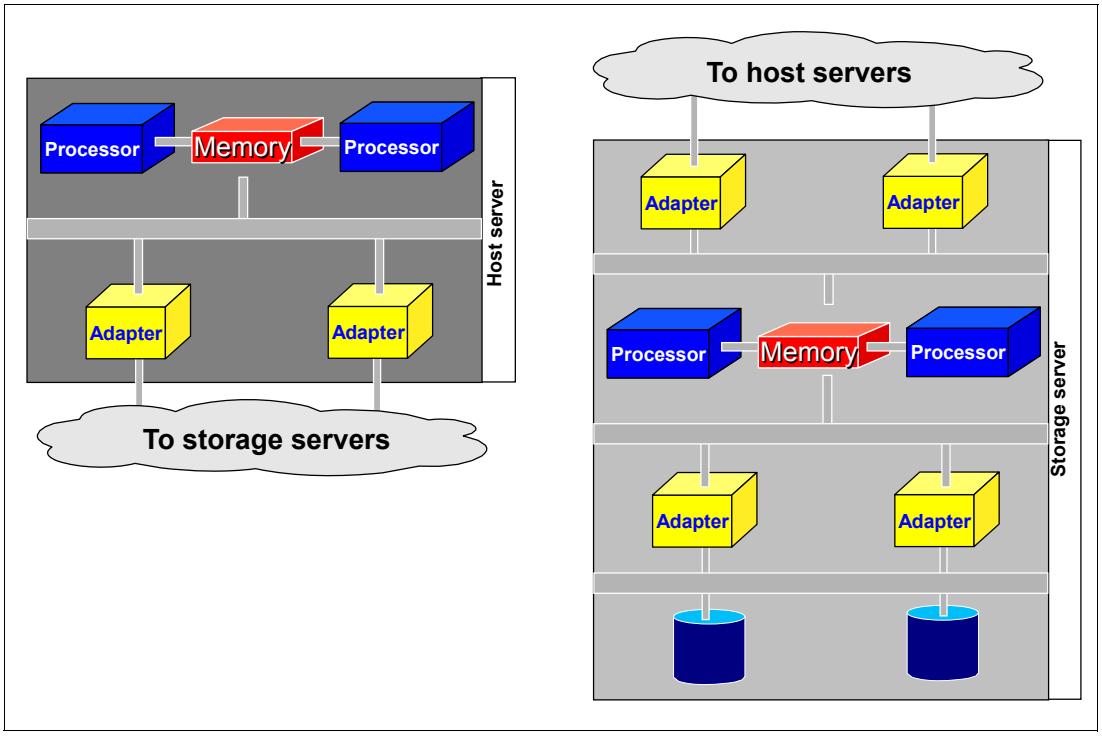



*Figure 11-1 Host server and storage server comparison: Balanced throughput challenge*

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 respect to 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.

## 11.2 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.

### 11.2.1 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/sec 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.

### 11.2.2 Arrays across loops

Advancing from the F20 to the 800 model, the internal structures and buses increased two fold and so did the backend 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 backend 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.

### 11.2.3 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.

### 11.2.4 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.

### 11.2.5 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.

## 11.3 How does the DS6000 address the challenge?

The DS6000 overcomes the architectural limits and bottlenecks which developed over time in the ESS due to the increasing number of I/Os and MB/sec from application servers.

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.

### 11.3.1 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 DS6000 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 DS6000 series overcomes FC-AL shortcomings

The DS6000 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 11-2 on page 224. 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.

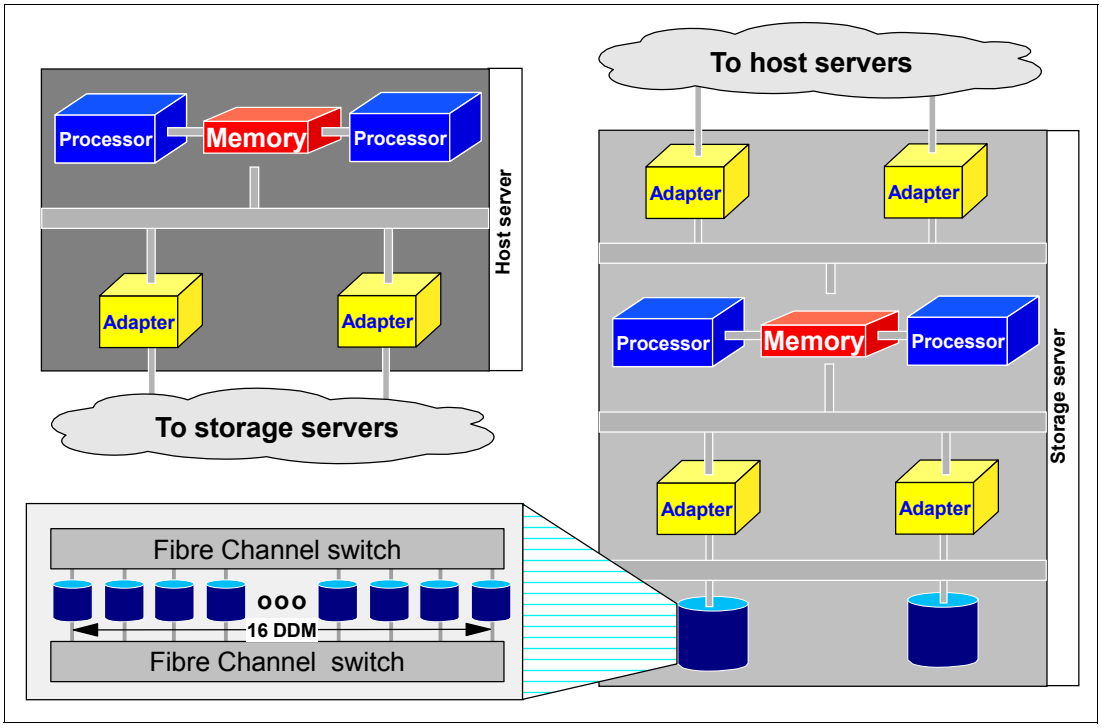



*Figure 11-2 Switched FC-AL disk subsystem*

Performance is enhanced as both DAs connect to the switched Fibre Channel disk subsystem backend as displayed in Figure 11-3 on page 225. Note that each DA port can concurrently send and receive data.

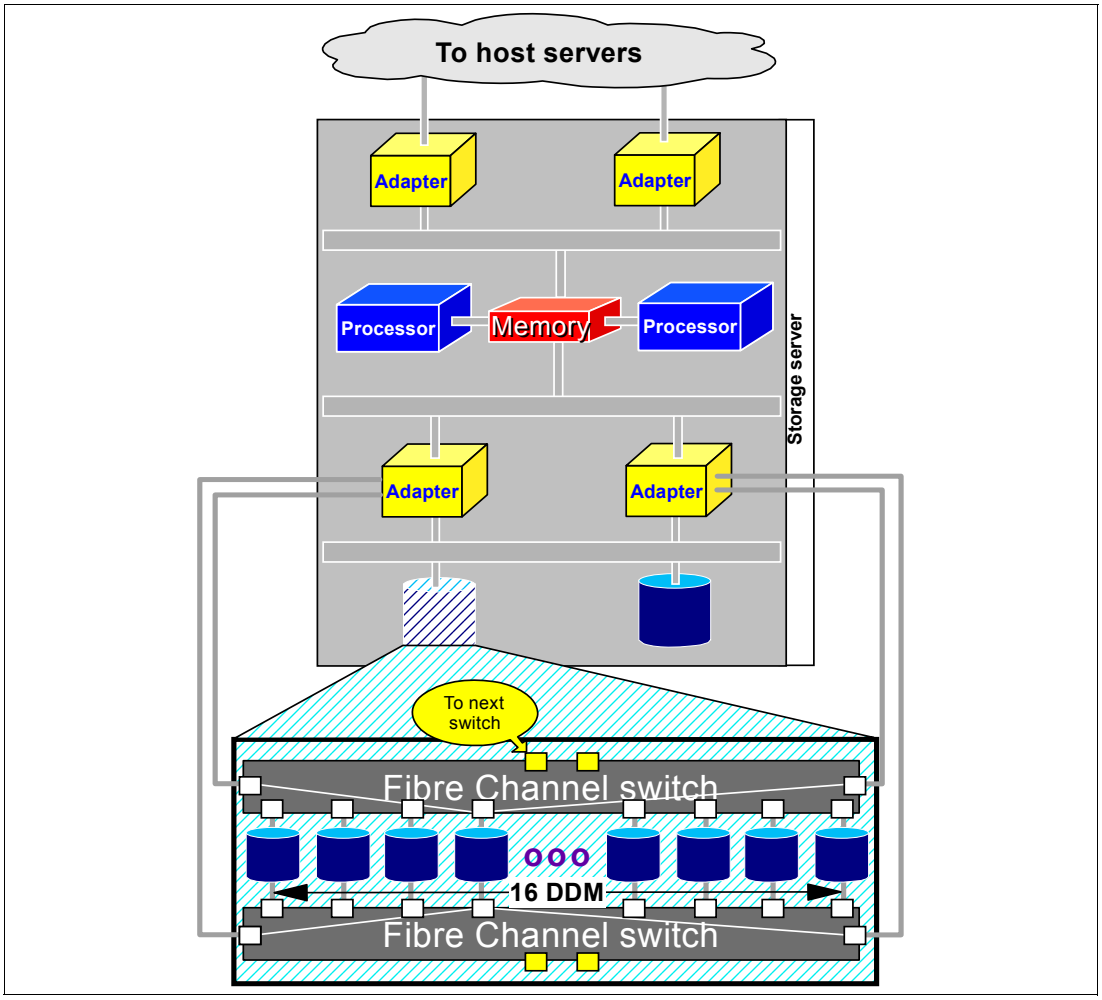



*Figure 11-3 High availability and increased bandwidth connecting both DA to two logical loops*

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 4, “Virtualization concepts” on page 65.

## 11.3.2 Fibre Channel device adapter

The DS6000 still relies on eight DDMs to form a RAID-5 or a RAID-10 array. 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 4, “Virtualization concepts” on page 65.

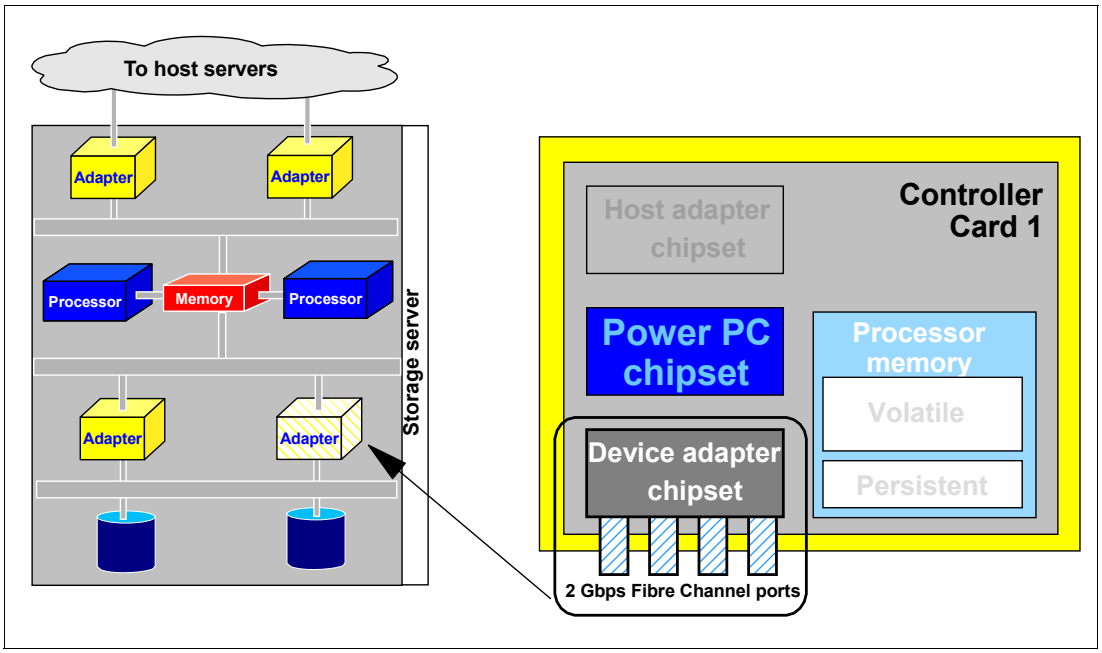



*Figure 11-4 Fibre Channel device adapter with 2 Gbps ports*

The new RAID device adapter chipset connects to 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.

Note that each DA chipset 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 used hardware, but also by the firmware efficiency.

## 11.3.3 New four-port host adapters

Before looking into the server complex we briefly review the new host adapters and their enhancements to address performance. Figure 11-5 on page 227 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.

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 DS6000 series can hold up to 8,192 devices. So all devices within a DS6000 series can be reached through a FICON port. Whether this is desirable is a different question and is discussed further in “Configuration recommendations for z/OS” on page 237.

With two sets of HA chip sets the DS6000 series can configure up to eight FICON or FCP ports.

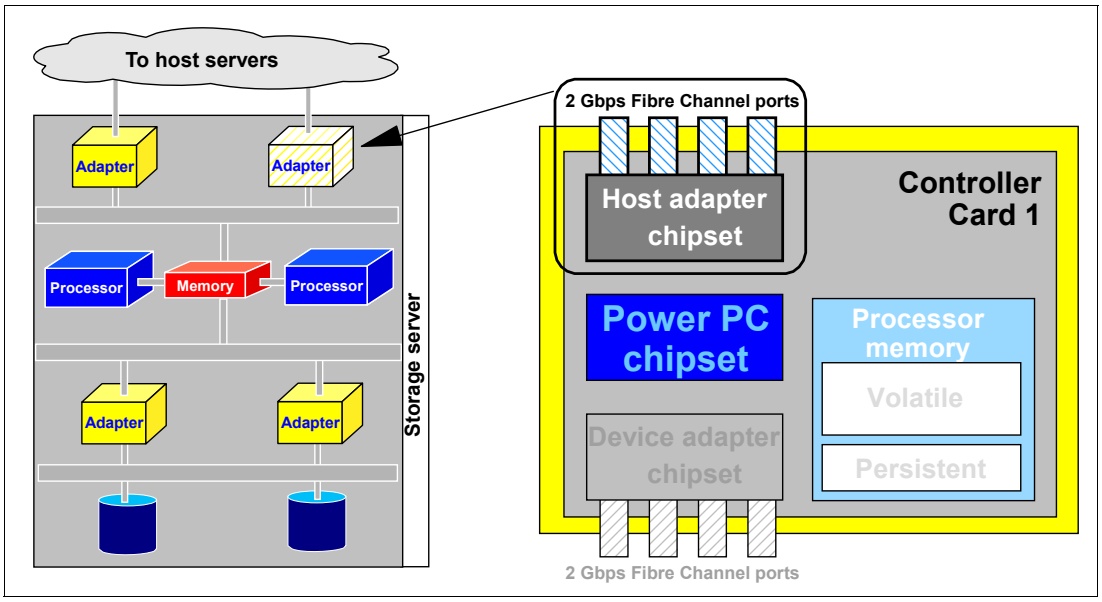



*Figure 11-5 Host adapter with 4 Fibre Channel ports*

With eight 2 Gbps ports the DS6000 series provides a theoretical aggregated host I/O bandwidth of 8 times 2 Gbps. Note that besides the adapter used and port technology, throughput depends also on the firmware efficiency and how the channel protocol is implemented.

## 11.3.4 Enterprise-class dual cluster design for the DS6800

The DS6000 series provides a dual cluster or rather a dual server design, which is also found in the ESS and DS8000 series. This offers an enterprise-class level of availability and functionality in a space efficient, modular design at a low price.

The DS6000 series incorporates the latest PowerPC processor technology. A simplified view is in Figure 11-6 on page 228. The dual-processor complex approach allows for concurrent microcode loads, transparent I/O failover and failback support, and redundant, hot-swapable components.

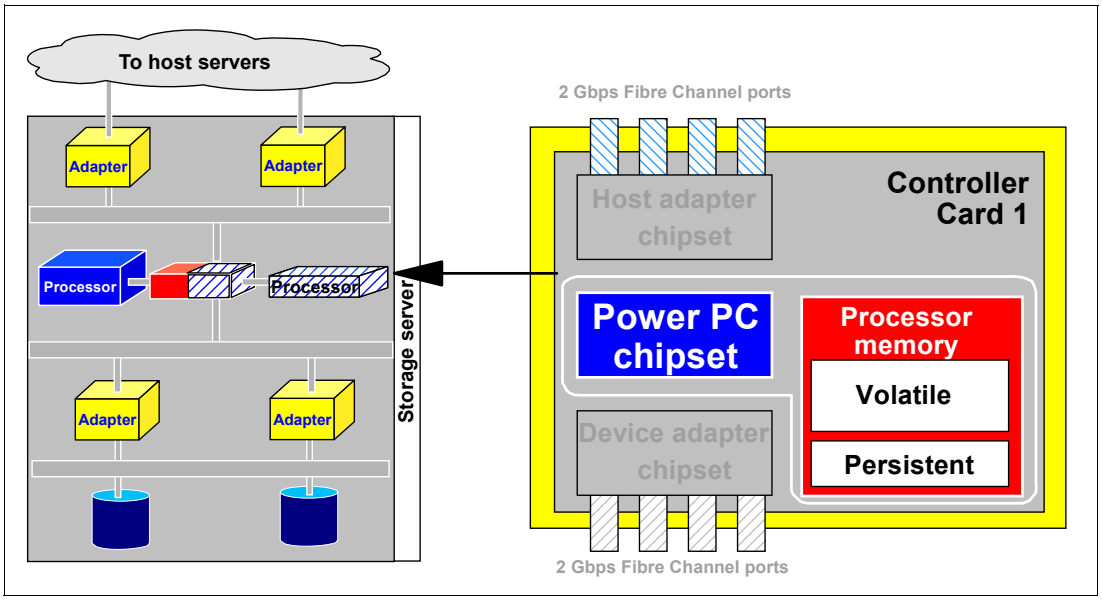



*Figure 11-6 Standard PowerPC processor complexes for DS6800-511*

The next figure, Figure 11-7, provides a less abstract view. It outlines some details on the dual processor complex of the DS6800 enclosure and its gates to host servers through HAs, and its connections to the disk storage backend through the DAs.

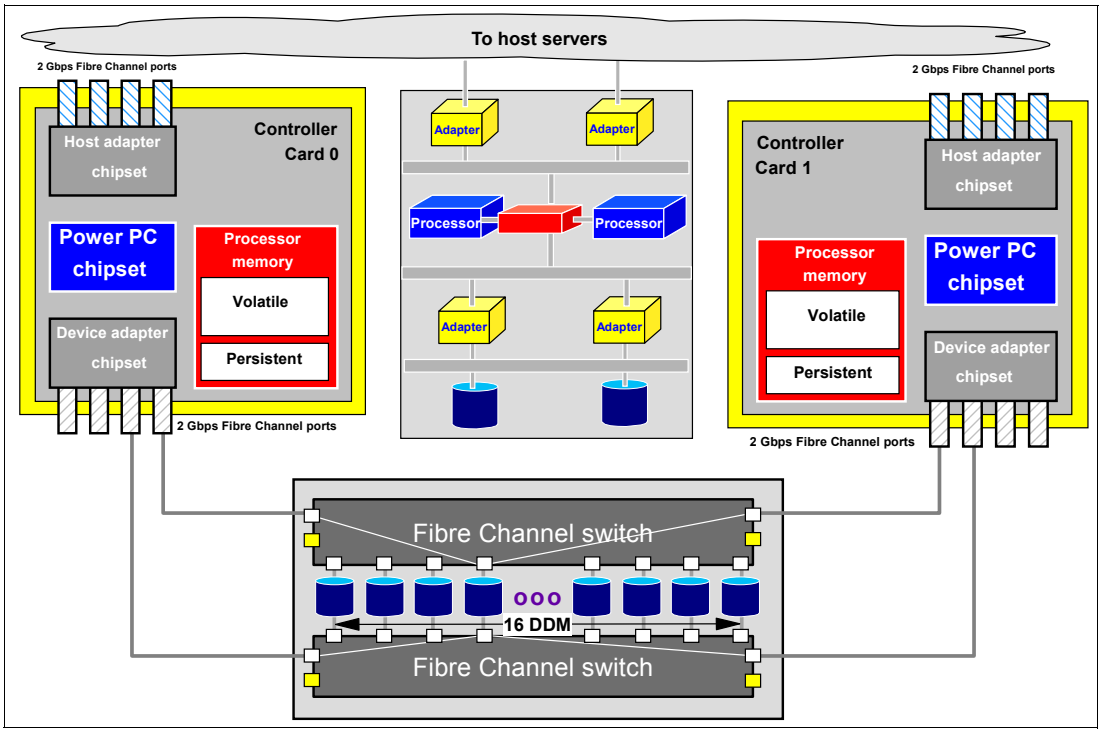



*Figure 11-7 DS6800 server enclosure with its Fibre Channel switched disk subsystem*

The DS6800 controls, through its two processor complexes, not only one I/O enclosure as Figure 11-7 displays, but can connect to up to 7 expansion enclosures. Figure 11-8 on page 229 shows a DS6800 with one DS6000 expansion enclosure.

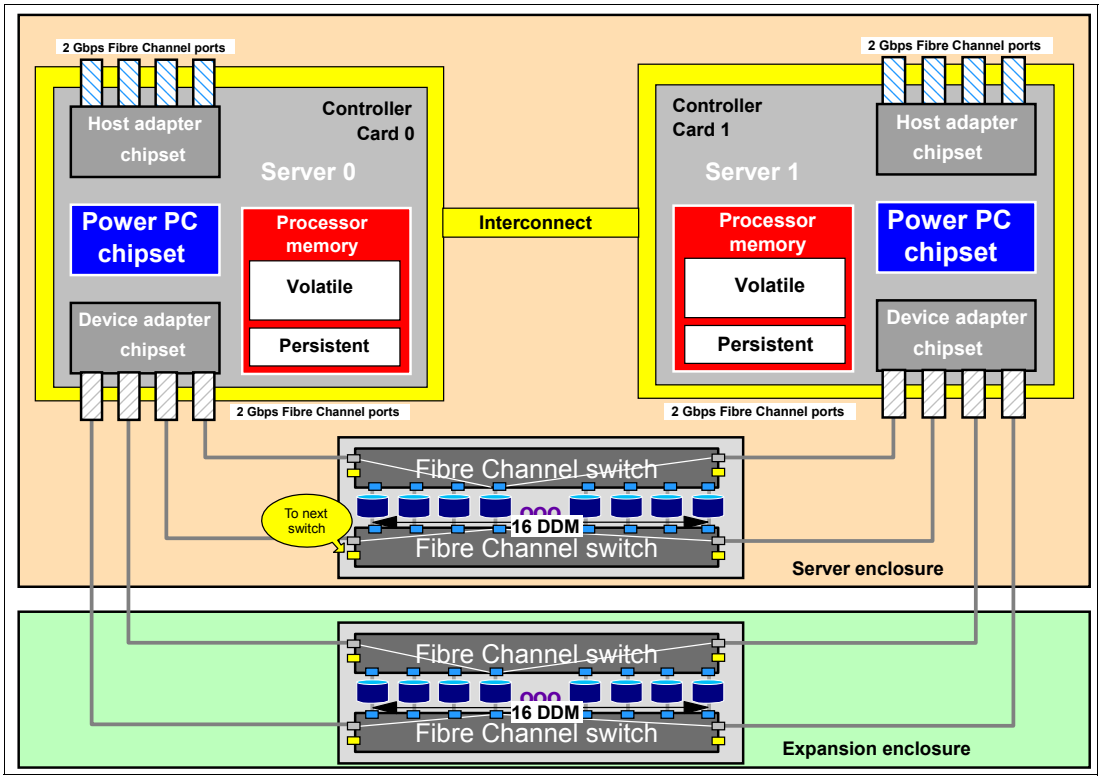



*Figure 11-8 DS6800 with one DS6000 expansion enclosure*

Note that each Fibre Channel switch in the disk subsystems from here on connects to the next pair of Fibre Channel switches through its two remaining ports. This is similar to inter-switch links between Fibre Channel switches.

Through the affinity of extent pools to servers, the DA in a server is used to drive the I/O to the disk drives in the host extent pools owned by its server.

When creating volumes in extent pools, these volumes get an affinity to a certain server through the extent pool affinity to a server (see Chapter 4, “Virtualization concepts” on page 65). This suggests even distribution of volumes across all ranks in the disk subsystems and all loops to balance the workload. Although each HA port can reach any volume in the disk subsystem, Figure 11-8 indicates also a server affinity to its local HA and its Fibre Channel ports. This introduces the concept of a preferred path. When a volume has an affinity, for example, to server 0, and is accessed through a port in the HA of server 0, then the I/O is locally processed. When this volume is accessed through the HA of the other server, in this example from server 1, then the I/O is routed to the server which owns the extent pool, which here is server 0.

## 11.3.5 Vertical growth and scalability

Figure 11-9 on page 230 shows a simplified view of the basic DS6000 structure and how it accounts for scalability.

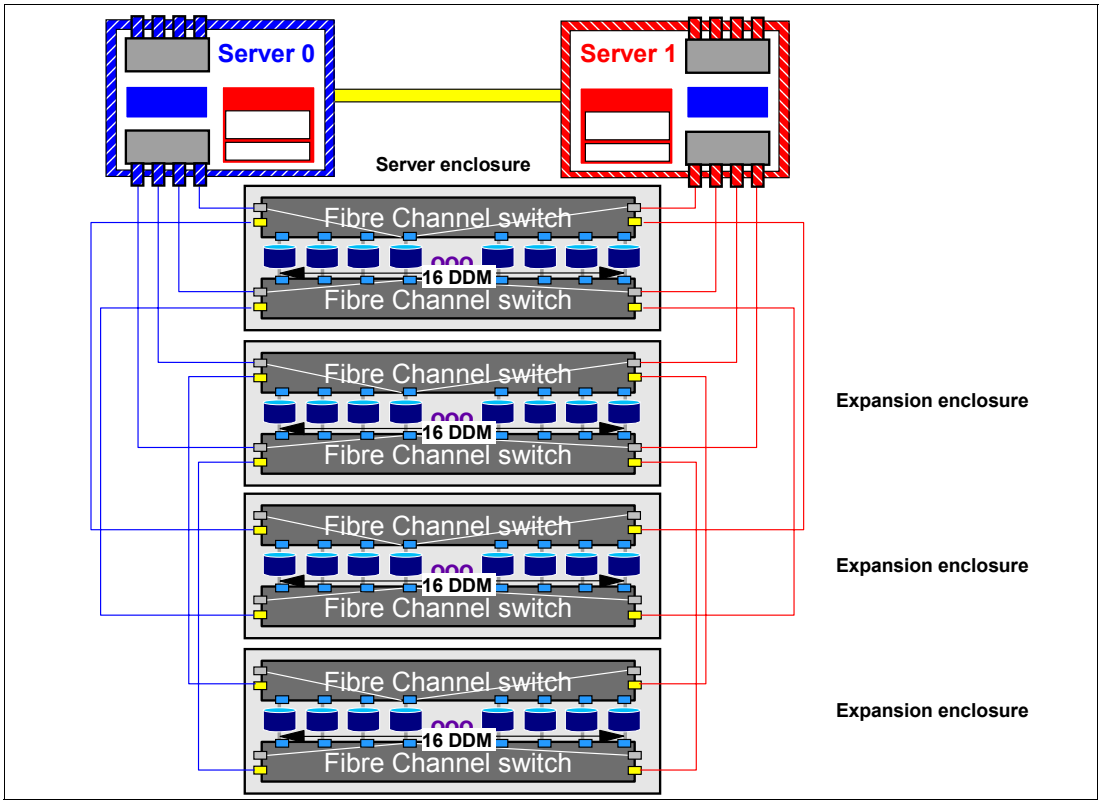



*Figure 11-9 DS6000 interconnects to expansion enclosures and scales very well*

Figure 11-9 outlines how expansion enclosures connect through inter-switch links to the server enclosure. Note the two Fibre Channel loops which are evenly populated as the number of expansion enclosures grow.

## 11.4 Performance and sizing considerations for open systems

To determine the most optimal DS6000 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.

### 11.4.1 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

## 11.4.2 Data placement in the DS6000

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 DS6000 resources utilization, you should:

- Equally spread the LUNs across the DS6000 servers.

  Spreading the LUNs equally on rank group 0 and 1 will balance the load across the DS6000 servers.
- Use as many disks as possible.
- Distribute across DA pairs and 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.

## 11.4.3 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.

DS6000 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 11-10 on page 232.

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 DS6000 as shown in Figure 11-10.

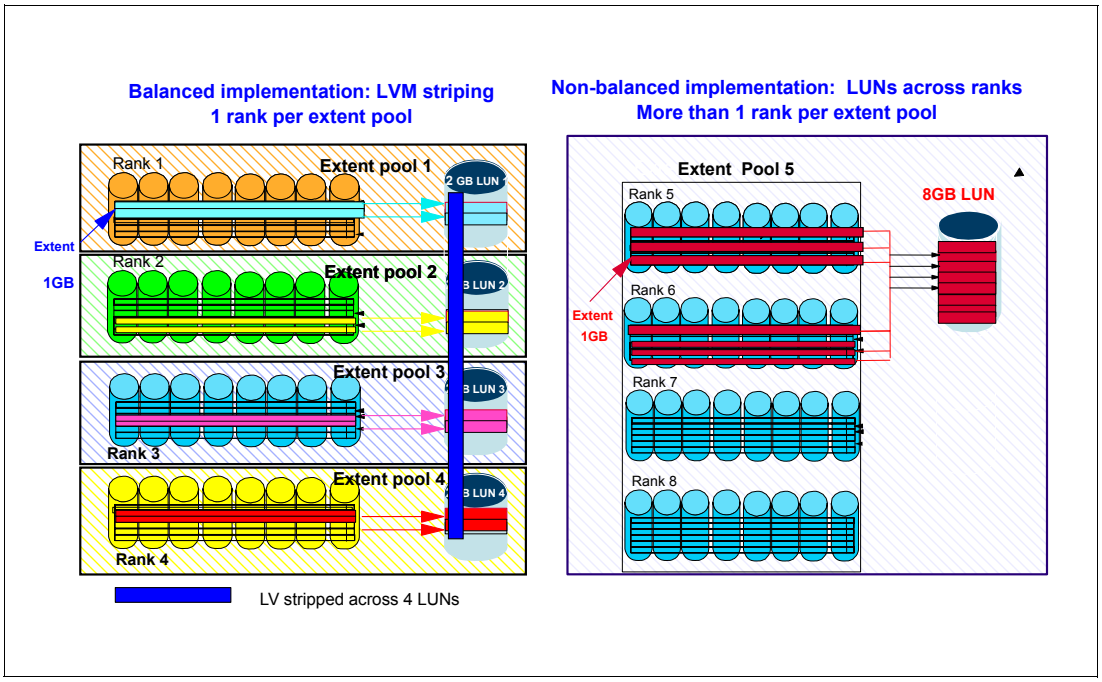



*Figure 11-10 Spreading data across ranks*

**Note:** The recommendation is to use host striping wherever possible to distribute the read and write I/O access patterns across the physical resources of the DS6000.

### 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 DS6000 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 4MB to 64MB.

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.

## 11.4.4 Determining the number of connections between the host and DS6000

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 DS6000 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 DS6000.

### 11.4.5 Determining the number of paths to a LUN

When configuring the IBM DS6000 for an open systems host, a decision must be made regarding the number of paths to a particular LUN, because the multipath 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 multipath 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 DS6000. It resides in a host system with the native disk device driver as described in Chapter 14, “Open systems support and software” on page 275.

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.

### 11.4.6 Determining where to attach the host

When determining where to attach multiple paths from a single host system to I/O ports on the DS6000, the following considerations apply:

- Enure that the host has at least two connections to  the DS6000, using one host I/O port on DS6000 controller 0 and one host I/O port on controller 1.
- If you need more than two paths from a host to the DS6000, spread the attached I/O ports evenly between the two DS6000 controllers.

The DS6000 host adapters, device adapters and ranks all have affinity to one DS6000 controller card or the other.

## 11.5 Performance and sizing considerations for z/OS

Here we discuss some z/OS-specific topics regarding the performance potential of the DS6000. We also address what to consider when you configure and size a DS6000 to replace older storage hardware in z/OS environments.

## 11.5.1 Connect to zSeries hosts

Figure 11-11 displays a configuration fragment on how to connect a DS6800 to a FICON host.

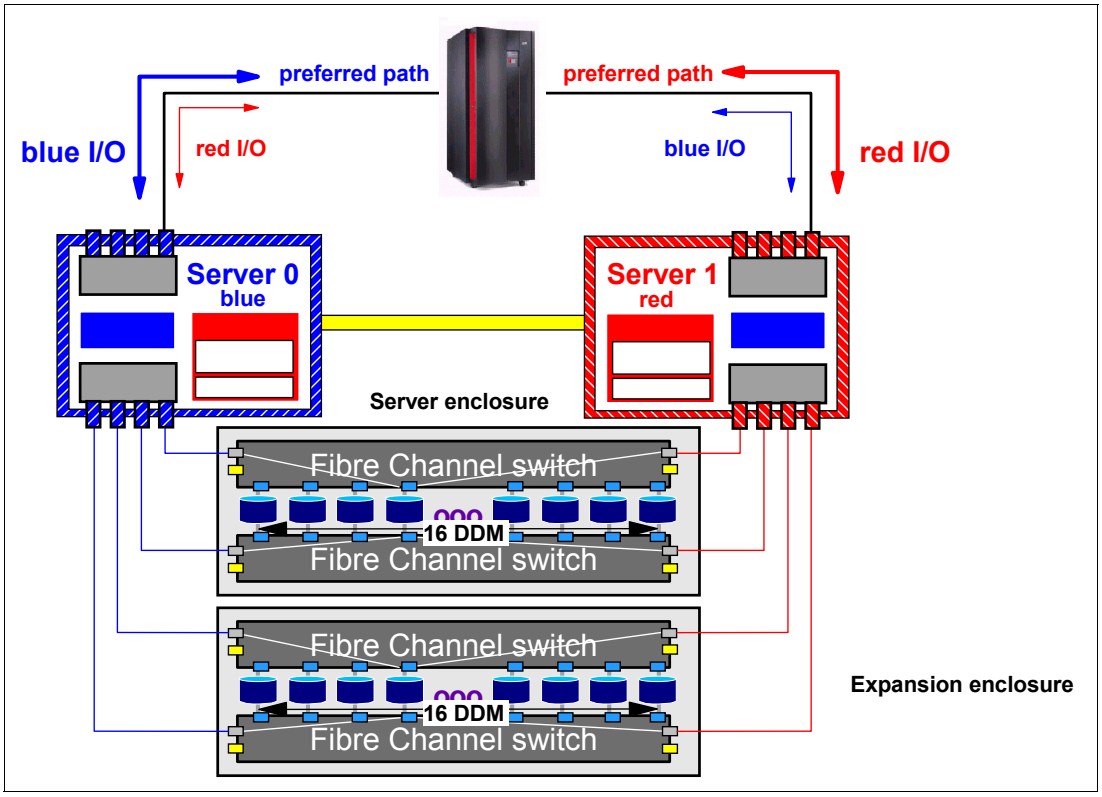



*Figure 11-11 DS6800 front end connectivity example - partial view*

The physical connectivity does not reveal what is important to consider when creating logical volumes and assigning these volumes to LCUs. Before considering different configuration approaches, we provide some general discussion on the potential of the DS6000 series in the following sections.

## 11.5.2 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 DS6000 series provides only 2 Gbps FCP ports, which can be configured either as FICON to connect to zSeries servers or as FCP ports to connect to Fibre Channel-attached open systems hosts. The example in Figure 11-11 shows only two FICON Express channels. But just two FICON Express channels have the potential to provide roughly a bandwidth of 2 x 175 MB/sec, which equals 350 MB/sec. This is a very conservative number. Some measurements show up to 206 MB/sec per 2 Gbps FICON Express channel and 406 MB/sec aggregated for this particular example with just two 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 11-11 with only two FICON Express channels has the potential of over

13,600 I/Os per second with the conservative numbers. These numbers vary depending on the server type used.

The ESS 800 has an aggregated bandwidth of about 500 MB/sec for highly sequential reads and about 350 MB/sec for sequential writes. The DS6800 can achieve higher data rates than an ESS 800.

In a z/OS environment a typical transaction workload might perform on an ESS 800 Turbo II with a large cache configuration slightly better than with a DS6800. This is the only example where the ESS 800 outperforms the DS6800. In all open systems environments, the DS6800 performs better than the ESS 800. This is also true for sequential throughput in z/OS environments.

## 11.5.3 An appropriate DS6000 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 in mirrored cache

Without discrete DS6000 benchmark figures, a sizing approach to follow could be to propose how many ESS 800s might be consolidated into a DS6000 model. From that you can derive the number of ESS 750s, ESS F20s, and ESS E20s which can collapse into a DS6000. The older ESS models have a known relationship to the ESS 800.

Further considerations are, for example, the connection technology used, like FICON or FICON Express channels, and the number of channels.

Generally speaking, a properly configured DS6000 has the potential to provide the same or better numbers than an ESS 800, except for transaction workloads with a large cache in the ESS. Since the ESS 800 has the performance capabilities of two ESS F20s, a properly configured DS6000 can replace two ESS F20s.

### Processor memory size considerations for z/OS environments

Processor memory or cache in the DS6000 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 and smart, adaptive caching algorithms are just as important to guarantee outstanding performance. This is implemented in the DS6000 series, except for the cache segment size, which is currently 68 KB.

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 staged to disk.

The IBM Tucson performance evaluation lab suggests a certain ratio between cache size to backstore capacity. In general, the recommendation is:

- 0.5% cache to backstore ratio for z/OS high performance

- 0.2% cache to backstore ratio for high performance open systems
- 0.2% for z/OS for standard performance
- A ratio of 0.1% between cache size and backstore capacity for open system environments for standard performance

### S/390 or zSeries channel consolidation

The number of channels plays a role as well when sizing DS6000 configurations and when we know from where we are coming. The total number of channels that were used where you are coming from has to be considered in the following way:

- ESCON channels are not supported for the DS6000. When coming from an ESCON environment and switching to FICON channels, a four to one ratio is very conservative. Consider, for example, replacing 16 ESCON channels with four FICON Express channels. Plan for four FICON channels as a minimum.
- 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 an ESS 800 with eight FICON channels each connected to IBM 9672 G5 or G6 servers, might end up in a single DS6000 also with eight FICON channels.
- When migrating not only to the DS6000 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 DS6000. (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 then keep a minimum of four FICON ports. You might consider using 25% fewer FICON ports in the DS6000 than the aggregated number of FICON 2 Gbps ports from the source environment. For example, when you consolidate an ESS 800 with 10 FICON 2 Gbps ports to a DS6000, plan for all eight possible FICON ports on the DS6000.

### 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 and 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 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. Note the DS6000 can also be configured with a RAID array comprised of four DDMs.

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 226 we stated that the disk storage subsystem DA port in a DS6000 has about five times more sequential throughput capability than an ESS 800 DA port provides. Based on the 2 Gbps Fibre Channel connectivity to a DS6000 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 the following numbers:

- 146 GB DDM with 10K RPM delivers 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. Note that this discussion takes a theoretical approach, but it is sufficient to get a first impression.

At GA the IBM internal tool, 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 DS6000 accordingly.

Use Capacity Magic to find out about usable disk capacity. This tool is also available at an IBM internal intranet sales site.

## 11.5.4 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 and the ranks might be on different loops.

A horizontal pooling approach assumes that volumes within a logical volume pool, like all DB2 volumes, are evenly spread across all ranks and loops. 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 11-11 on page 234 displays some aspects regarding the disk subsystem within a DS6000:

- Chapter 4, "Virtualization concepts" on page 65, 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 DS6000. 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. For the DS6000 this would have the advantage that you can plan for proper volume placement with respect to preferred paths.

- In the example in Figure 11-12 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.

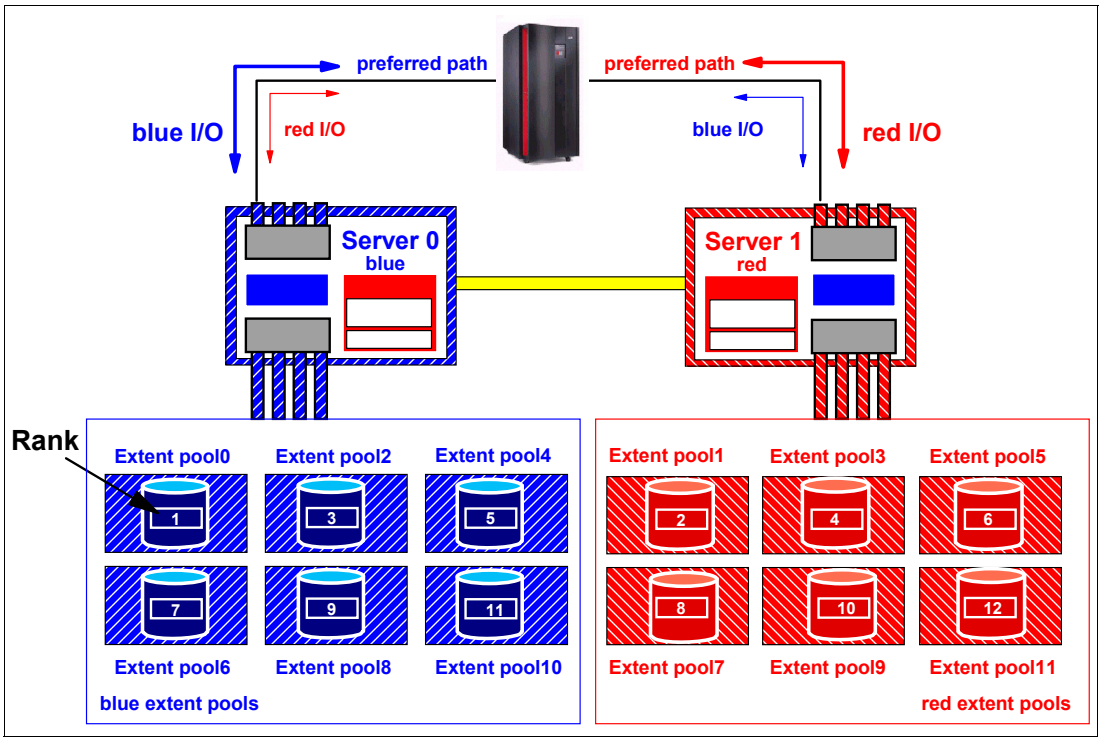



*Figure 11-12 Extent pool affinity to processor complex with one extent pool for each rank*

Figure 11-12 indicates that there is an affinity between FICON ports and certain extent pools and, therefore, an affinity between FICON ports and certain volumes within these extent pools.

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. But the HA and DA affinity to a server prefers one path over the other. Now z/OS is able to notice the preferred path and then schedule an I/O over the preferred path as long as the path is not saturated.

## Minimize the number of extent pools

The other extreme is to create just two extent pools when the DS6000 is configured as CKD storage only. You would then subdivide the disk subsystem evenly between both processor complexes or servers as Figure 11-13 shows.

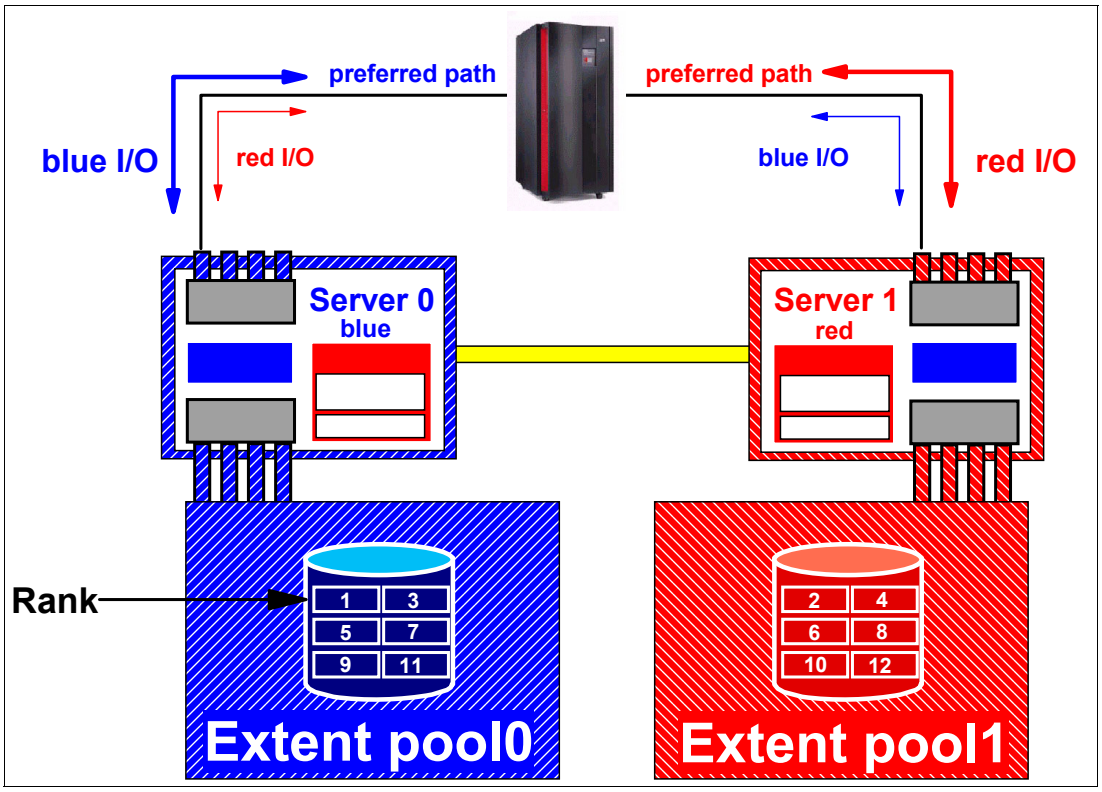



*Figure 11-13 Extent pool affinity to processor complex with pooled ranks in two extent pools*

Again what is obvious here is the affinity between all volumes residing in extent pool 0 to the left processor complex, server 0, including its HA, and the same 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. You may have the desire to control the placement of the most critical performance volumes and also configure for preferred paths. This might lead to a compromise between both approaches, as Figure 11-14 on page 240 suggests.

## Plan for a reasonable number of extent pools

Figure 11-14 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.

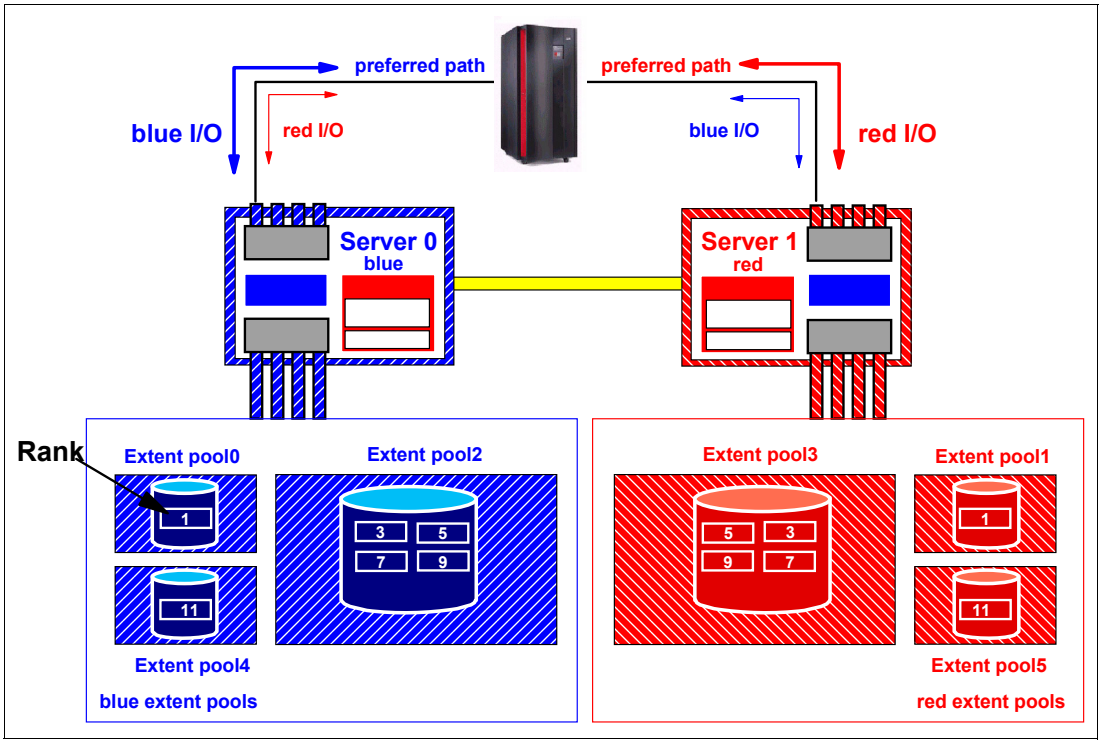



*Figure 11-14 Mix of extent pools*

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 DS6000. Then you might consider two or four smaller extent pools with dedicated ranks for high performance workloads and their 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 DS6000 you might consider configuring each DS6000 in a uniform fashion. We recommend grouping all volumes from all the large extent pools into one large SMS storage group. Cover the smaller, high performance extent pools through discrete SMS storage groups for each DS6000. For example, in a dual logging database environment allow assignment of extent pool0 to the first logging volume and extent pool1 for the second logging volume. Consider a respective channel path configuration which takes preferred paths into account.

## 11.6 Summary

The DS6000 high performance processor complex configuration is the base for a maximum of host I/O operations per second. The DS6000 can handle I/O rates of about what an ESS 800 can deliver at maximum speed. With the introduction of the smart, switch-based Fibre Channel disk back end, which overcomes the FC-AL arbitration overhead and operates at 2 Gbps speed, the DS6000 provides a better sequential throughput than an ESS 800.

The DS6000 series is designed to deliver enterprise-class storage capabilities in a space efficient, modular design at a low price. It provides a wide capacity range from 16 DDMs up to

128DDMs. Depending on the DDM size this reaches a total of up to 67.2 TB. Just the base enclosure provides up to 4.8 TB of physical storage capacity with 16 DDMs and 300 GB per DDM.

The small and fast DS6000 with its rich functionality and compatibility with the ESS 750, ESS 800, and DS8000, in all functional respects makes this a very attractive choice.

12

# zSeries software enhancements

This chapter discusses z/OS, z/VM, z/VSE™, and Transaction Processing Facility (TPF) software enhancements that support the DS6000 series. The enhancements include:

- Scalability support
- Large volume support
- Hardware configuration definition (HCD) to recognize the DS6000 series
- Performance statistics
- Resource Measurement Facility (RMF)
- Preferred pathing

## 12.1 Software enhancements for the DS6000

A number of enhancements have been introduced into the z/OS, z/VM, z/VSE,VSE/ESA and TPF operating systems to support the DS6000. The enhancements are not just to support the DS6000, but also to provide additional benefits that are not specific to the DS6000.

## 12.2 z/OS enhancements

The DS6000 series simplifies system deployment by supporting major server platforms. The DS6000 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 DS6000 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 DS6000 as a 2105 device.

**Important:** Always review the latest Preventative Service Planning (PSP) 1750DEVICE bucket for software updates.

The PSP information can be found at:

http://www-1.ibm.com/servers/resourcelink/svc03100.nsf?OpenDatabase

Basic device support has been enhanced in the following areas:

- Scalability support
- Large Volume Support (LVS)
- Read availability mask support
- Initial Program Load (IPL) enhancements
- DS6000 definition to host software
- Read Control Unit and Device Recognition for DS6000
- New performance statistics
- Resource Measurement Facility (RMF)
- Preferred pathing
- Migration considerations
- Coexistence considerations

### 12.2.1 Scalability support

The Input/Output subsystem (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 DS6000 that has the capability to scale up to 8192 devices. With the current support, we may have CPU or spin lock contention, or exhaust storage below the 16M line at device failover.

Now with z/OS 1.4 and higher with the DS6000 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 as the DS6000 has a much higher number of devices compared to the IBM 2105. In the IBM 2105, we have 4096 devices and in the DS6000 we have up to 8192 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 is done 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.

## 12.2.2 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.

## 12.2.3 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 command to retrieve the information from the control unit. By doing this, there is no need to maintain a table for this information.

### 12.2.4 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 (or example, 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.

### 12.2.5 DS6000 definition to host software

The DASD Unit Information Module (UIM) is changed to define the new control unit type of 1750. 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 1750 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 1750 control unit type will allow this storage facility to be uniquely reflected in the Hardware Configuration Manager (HCM). The definition of the 1750 control unit type in the HCD is not required to define an IBM 1750 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 32 for the DS6000. The number of devices per LSS is still limited to 256. The number of CKD logical volumes is increased from 4096 to 8192 devices per DS6000.

### 12.2.6 Read Control Unit and Device Recognition for DS6000

The host system will inform the attached DS6000 of its capabilities, such that it emulates a DS8000. This does not limit any DS6000 functions. The DS6000 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 messages and command output display DS6000 information:

- EREP messages
- DEVSERV QDASD and PATHS command responses

The output from the IDCAMS LISTDATA COUNTS, DSTATUS, STATUS and IDCAMS will display emulated DS8000s. 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)

- ICKDSF
- DFSORT
- EREP

## 12.2.7 New performance statistics

There are two new sets of performance statistics that will be reported by the DS6000. 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 DS6000. These reports will consist of back-end counters that capture the activity between the cache and the ranks in the DS6000 for each individual logical volume. The new rank and extent pool statistics will be disk system wide instead of volume wide only.

**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 DS6000. Figure 12-1 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 12-1 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.

## 12.2.8 Resource Measurement Facility (RMF)

RMF support for the DS6000 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).

## 12.2.9 Preferred pathing

In the DS6000, host ports have a fixed assignment to a server (or controller card). The DS6000 will notify the host operating system, in this case DFSMS (device support), if a path is preferred or not. Device support will then identify preferred paths to the IOS. I/Os will be directed to preferred paths to avoid crossing the PCI-X connection. The only time this will not be honored is when there are no preferred paths available. The software will then switch over to use non-preferred paths. There will be a slight performance penalty if the I/O is not executed over the preferred path. The I/O request and the data would have to be transferred across the bridge interface that connects both servers. These transfers add some latency to the response time. Furthermore, the bridge interface is also used to mirror the persistent memory and for other inter-server communication. It could become a bottleneck if too many normal I/O requests ran across it, although it is a high bandwidth, low latency, PCI-X connection. If the IOS support for preferred pathing is not implemented, sequential reads may drop by up to 50%. The response time in low stress environments may also increase by up to 10% to 20%.

New messages will inform the user when all preferred or the last preferred path is varied offline. Figure 12-2 shows the output from the DEVSERV PATHS command. Now, this output displays preferred paths information.

```
DS P,9E02
IEE459I 09.14.03 DEVSERV PATHS 943
UNIT DTYPE  M CNT VOLSER  CHPID=PATH STATUS
     RTYPE   SSID CFW TC   DFW   PIN  DC-STATE CCA  DDC   ALT  CU-TYPE
9E02,33909 ,O,000,339R53,38=+ 42=+
     PATH ATTRIBUTES     PF   NP
     1750    9900  Y  YY.  YY.    N   SIMPLEX   02   02        2107
*********************** SYMBOL DEFINITIONS ************************
O = ONLINE                        + = PATH AVAILABLE
PF = PREFERRED                    NP = NON-PREFERRED
```

*Figure 12-2 DEVSERV PATHS command showing preferred pathing*

Figure 12-3, shows the output from the DISPLAY DEV command. The output has been enhanced to display the path attributes as preferred path (PF) or non-preferred path (NP).

```
D M=DEV(9E02)
IEE174I 09.14.23 DISPLAY M 945
DEVICE 9E02   STATUS=ONLINE
CHP                   38   42
DEST LINK ADDRESS     2F   2F
ENTRY LINK ADDRESS    27   27
PATH ONLINE           Y    Y
CHP PHYSICALLY ONLINE Y    Y
PATH OPERATIONAL      Y    Y
PATH ATTRIBUTES       PF   NP
MANAGED               N    N
MAXIMUM MANAGED CHPID(S) ALLOWED:  0
DESTINATION CU LOGICAL ADDRESS = 00
CU ND      = 001750.000.IBM.13.000000048156
DEVICE NED = 001750.000.IBM.13.000000048156
PAV BASE AND ALIASES  7
```

*Figure 12-3 D M=DEV command output*

### 12.2.10 Migration considerations

A DS6000 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.

### 12.2.11 Coexistence considerations

Support for the DS6000 running in 2105 mode on systems without this SPE installed will be provided. It will consist of the recognition of the DS6000 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.

## 12.3 z/VM enhancements

z/VM is an IBM operating system that supplies a virtual machine to each logged-on user. The DS6000 will be supported on z/VM 4.4 and higher.

**Important:** Always review the latest Preventative Service Planning (PSP) 1750DEVICE bucket for software updates.

The PSP information can be found at:

http://www-1.ibm.com/servers/resourcelink/svc03100.nsf?OpenDatabase

## 12.4 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 DS6000 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) 1750DEVICE 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 DS6000.

## 12.5 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 DS6000 will be supported on TPF 4.1 and higher.

**Important:** Always review the latest Preventative Service Planning (PSP) 1750DEVICE bucket for software updates.

The PSP information can be found at:

http://www-1.ibm.com/servers/resourcelink/svc03100.nsf?OpenDatabase

13

# Data Migration in zSeries environments

This chapter describes several methods for migrating data from existing disk storage servers onto the DS6000 disk storage server. This includes migrating data from the ESS 2105 as well as from other disk storage servers to the new DS6000 disk storage server. 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
- Large volume support
- 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.

# 13.1 Define migration objectives in z/OS environments

Data migration is an important activity that needs to be planned well to ensure the success of DS6000 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 to guarantee data integrity at any time.

## 13.1.1 Consolidate storage subsystems

In the course of a data migration you might consider consolidating the volume environment from which you are coming.

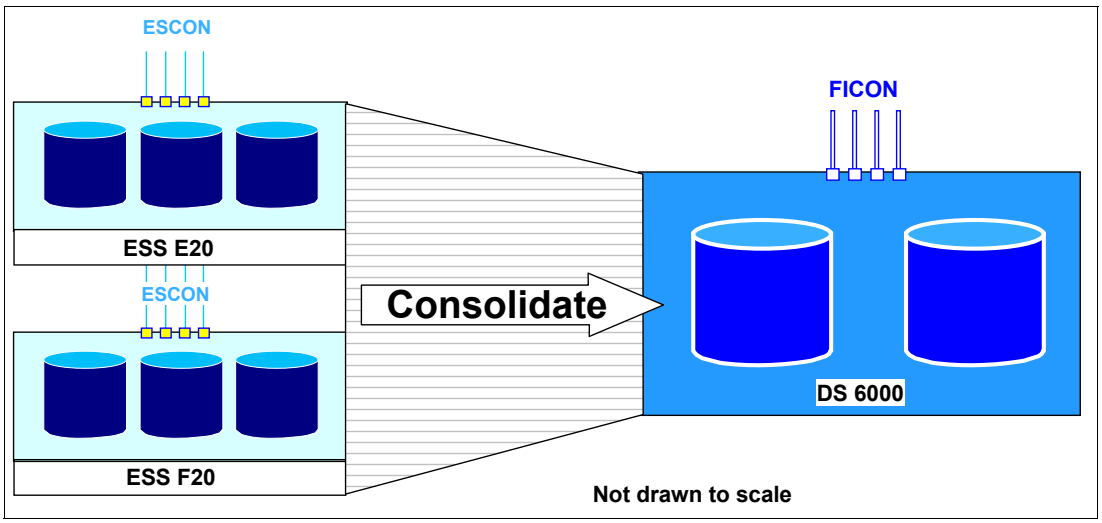



*Figure 13-1 Consolidation opportunities when migrating to DS6000*

A DS6800 can contain up to 32 logical control units (LCUs) at GA. This allows the DS6800 to simulate up to 32 times an IBM 3390-6 control unit image with 256 devices within each single control unit image. The number of supported volumes within a DS6800 is 32 x 256 devices which equals 8,192 logical volumes. This is twice as much as what an ESS 2105 can configure for a z/OS-based configuration. This suggests consolidation of multiple control units into a single DS6800 disk storage server. This requires careful planning, and depending on the configuration complexity, it might take weeks or months to complete the planning and perform the actual migration.

## 13.1.2 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 to now most customers have relied on a standard sized volume, that being an IBM 3390-3 with 3,339 cylinders. 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 DS6800 model due to its flexibility to configure its backend storage independently from DDM sizes. There is no affinity any longer between an LSS and physical

8-packs. The allocation of a volume happens in extents or increments of the size of an IBM 3390-1 volume, or 1,113 cylinders. So a 3390-3 consists of exactly three extents from an extent pool. A 3390-9 with 10,017 cylinders comprises 9 extents out of an extent pool. There is no affinity any longer to an 8-pack 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 DS6800, you might consider planning such a move in the course of a data migration effort to reduce the number of volumes.

Although with the newly announced DS6800 storage servers we now support logical 3390 volumes of up to 65,520 cylinders, you might still plan for a standard sized logical volume, which is smaller than volumes as big as the new maximum volume size allows. Consider that full volume operations will take longer to copy or dump when increasing the volume size. 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 30,051 cylinders might be fine for most of the data. This is the space of nine 3390-3 volumes or 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 1,113 cylinders or the equivalent of a 3390-1 model.

Utilizing big volumes will require the use of dynamic parallel access volumes (PAV) to allow many concurrent accesses to the very same volume. On such a big volume we see about 9 times as many concurrent I/Os than what we might see on a single 3390-3. Or to put it differently, we 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 installations still performing full volume operations to a significant extent you might plan for smaller volumes. For example, to fully dump nine 3390-3 volumes 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 classical 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.

### 13.1.3 Keep source and target volume at the current size

When the number of volumes does not reach the current z/Series 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. 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 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.

### 13.1.4 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. It usually needs a few but brief service interruptions, when moving the remaining data sets which are usually open and active 24 hours every day.

## 13.2 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 device geometry is defined by the track capacity and the number of tracks per cylinder. The same device geometry means that the source and target device 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 a 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 some more detail.

### 13.2.1 Physical migration with DFSMSdss and other storage software

Full volume copy through the DFSMSdss 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 except when the products are owned by the installation. When the number of volumes is in the range of up to a few hundreds, then standard-based software like DFSMSdss is an option, although 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 and so does not require extra costs for software.

To summarize: 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 way over to the target device, like TDMF and FDRPAS do. When brief service interruptions can be tolerated the standard software is still a popular solution.

## 13.2.2 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, and connects to the customer configuration during the migration.

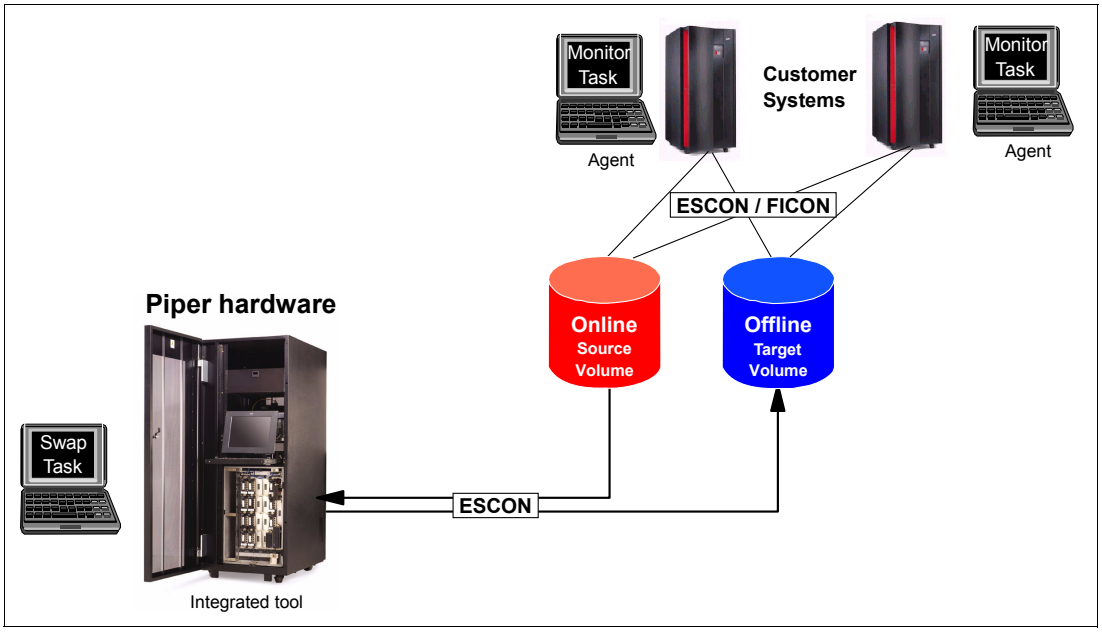



*Figure 13-2 Piper for z/OS environment configuration*

Currently this server is a Mulitprise 3000, which can connect through ESCON channels only. This will exclude this approach to migrate data to the DS6800, which provides pure FICON and Fibre Channel connectivity only. 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
- Preloaded with OS/390 2.10
- 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 and tasks 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.

To summarize: 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. Piper will soon provide a FICON-only configuration since the DS6800 only supports FCP FICON and not ESCON.

## 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 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 D/R solutions based on XRC.

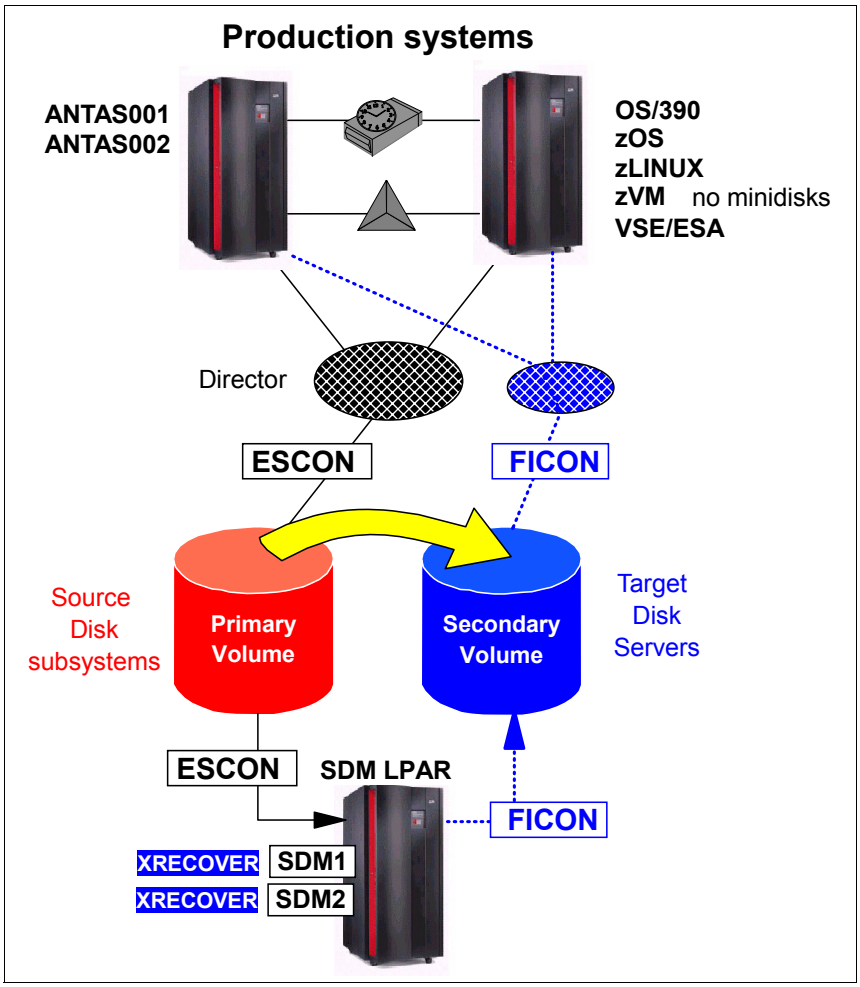



*Figure 13-3 Data Migration with z/OS Global Mirror (XRC)*

When coming from ESCON connectivity, it might be a good time to switch to FICON during the data migration, even though this makes the migration a bit more complex. XRC is the right vehicle to perform the migration transparently, and for an online data migration approach it 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 for XRC is extreme scalability. Figure 13-3 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 which 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. In a medium sized configuration, one or two SDMs is most likely sufficient.

Assume the migration consolidates two medium sized ESS F20s with a total of about 5 TB to a DS6800 disk server. This would suggest connecting the SDM LPAR to each ESS F20 with two dedicated ESCON channels. Configure for each F20 an SDM within the SDM LPAR. The DS6800 FICON channels might be shared between the SDMs because there is no potential bottleneck when coming from ESCON. This would take about one day to replicate all data from the two F20s to the DS6800, provided there is not too much application write I/O during the initial full copy. Otherwise it takes just a few more hours, depending on the amount of application write I/O during the first full volume copy, and the entire migration can be completed over a weekend.

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.

## 13.2.3 Hardware- and 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.

### 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 server supports only PPRC over Fibre Channel links and does not support PPRC ESCON links. The ESS Model 800 actually supports both PPRC link technologies.

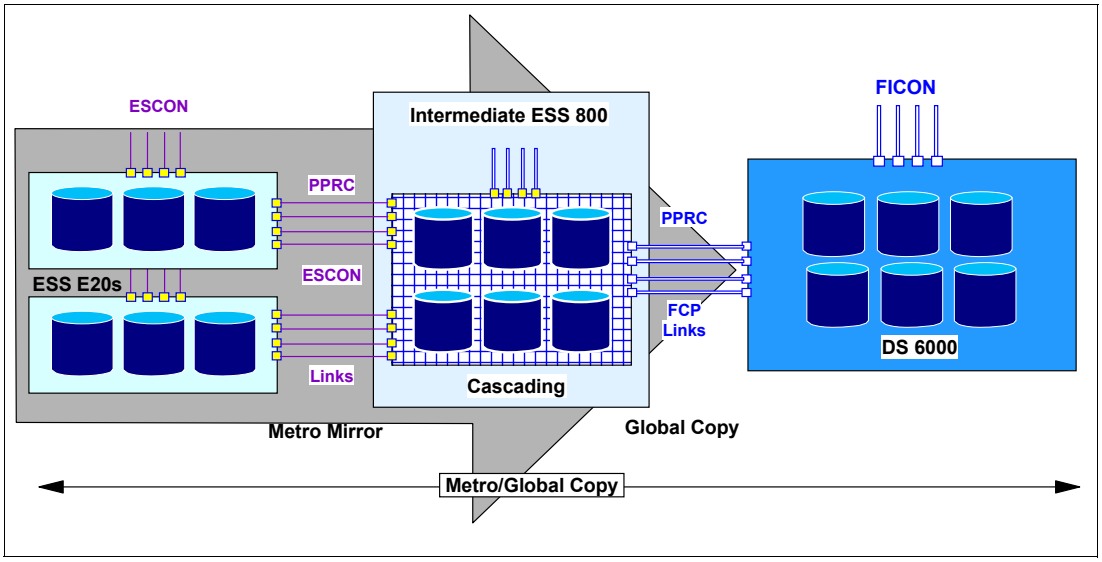



*Figure 13-4 Intermediate ESS 800 used to migrate data with PPRC over ESCON*

To utilize the advantage of PPRC with concurrent data migration, on a physical volume level though, 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 Metro Mirror between the intermediate ESS 800 and the target DS6800 disk server.

Dynamic Address Switch (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 per 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 accepted, 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 ESS 750 or ESS 800 to the DS6800.

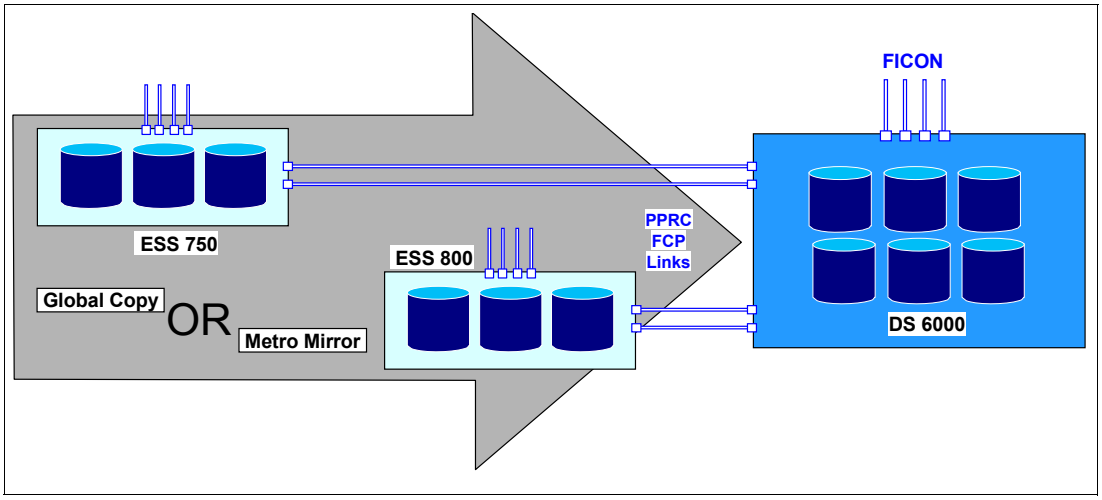



*Figure 13-5 Metro Mirror or Global Copy from ESS 750 or ESS 800 to DS6800*

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 to 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 Global 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 Global 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 Servers application on an ESS 800 this would look like Figure 13-6 on page 261.

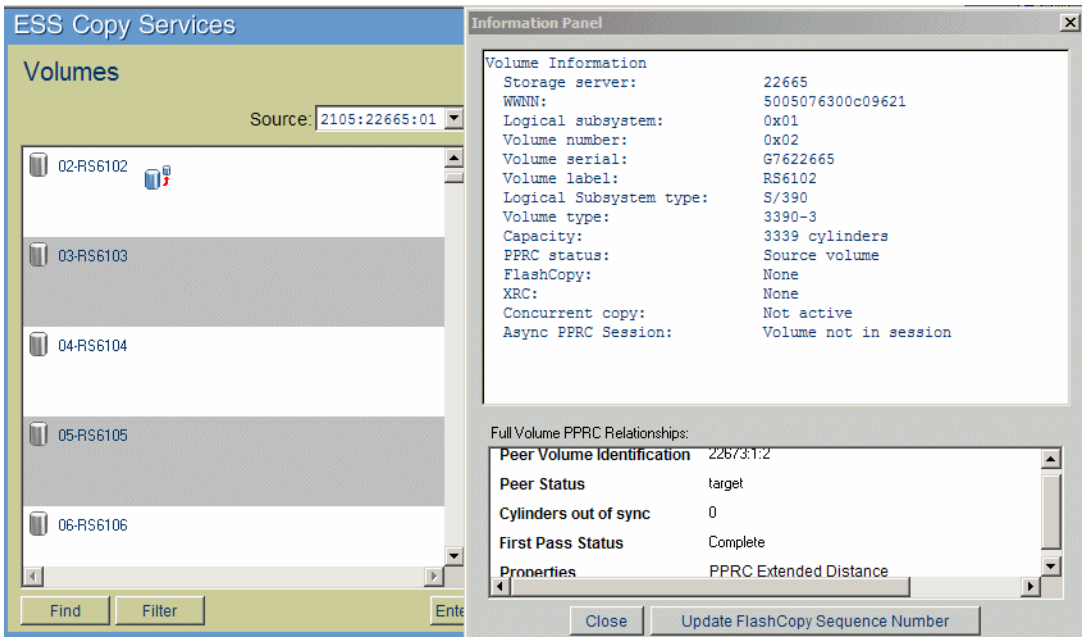



*Figure 13-6 Check with Global Copy whether all data was replicated to the new volume*

This approach is not really practical though. ICKDSF also 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 13-1.

*Example 13-1 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

                                                (PRIMARY)    (SECONDARY)
                                                SSID    CCA  SSID    CCA
DEVICE  LEVEL      STATE           PATH STATUS  SER #   LSS  SER #   LSS
------  ---------  --------------  -----------  ----------  ----------
6102    PRIMARY    PENDING    XD   ACTIVE       6100    02   8100    02
                                                22665   01   22673   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 13-2 on page 262.

*Example 13-2 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

                                                (PRIMARY)    (SECONDARY)
                                                SSID    CCA  SSID    CCA
DEVICE  LEVEL      STATE           PATH STATUS  SER #   LSS  SER #   LSS
------  ---------  --------------  -----------  ----------  ----------
6102    PRIMARY    PENDING    XD   ACTIVE       6100    02   8100    02
                                                22665   01   22673   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 catches the concerned numbers.

*Example 13-3 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         WWNN                   LIC LEVEL             *
* -----------  ----------------           -----------            *
```

```
* PRIMARY....  50050763300C09621            2.4.01.0062             *
* SECONDARY.1  50050763300C09629                                    *
**********************************************************************
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 13-4 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         WWNN                  LIC LEVEL               *
* -----------  ----------------           -----------             *
* PRIMARY....  50050763300C09621           2.4.01.0062             *
* SECONDARY.1  50050763300C09629                                   *
**********************************************************************
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 DS6000. 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.

## 13.3 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
- DFSMShsm™
- 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.

### 13.3.1 Data Set Services Utility (DFSMSdss)

A common utility is DFSMSdss. In this case it is not used for full 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 13-5 DFSMSdss for logical data migration from certain input volumes*

```
//MIGRATE  EXEC PGM=ADRDSSU
//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 CATALOG                                 -
       ALLDATA(*) ALLX WAIT(0,0) ADMIN OPT(3) CANCELERROR
/*
//
```

Example 13-5 depicts how to migrate all data sets from certain volumes. The keyword is LOGINDDNAME 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 13-5 assumes a system-managed environment. Here the system would automatically place the output data sets according to what the Automatic Class Selection Routine (ACS) suggests. The next paragraph is more specific on how this approach works.

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) and then restore the aggregate. This is nothing but the group of data sets which have been selected through filtering, put back onto the new DS8000 storage servers. For more information, see the redbook *DFSMShsm ABARS and Mainstar Solutions,* SG24-5089.

System utilities don't play an import role any longer since DFSMSdss incorporated the capability to manage all data set types for copy and move operations in a very efficient way. In a VSE/ESA environment, VSE/VSAM functions are still used like 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.

## 13.3.2 Data migration within the system-managed storage environment

System-managed storage or SMS manages volumes and storage groups (SG) and allocates data sets based on a policy or rules to 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 which decides whether this volume or that SG is eligible to receive new data set allocations.

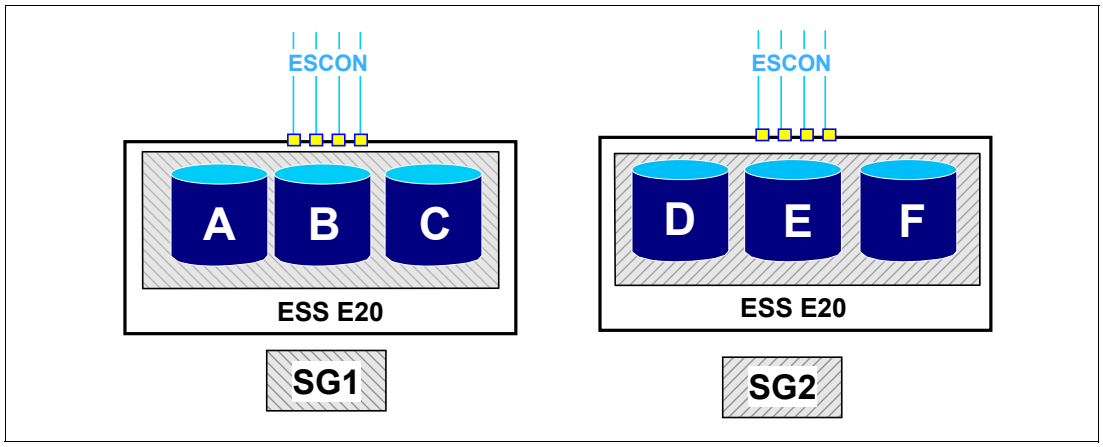



*Figure 13-7 SMS Storage Groups - migration source environment*

To explain this approach, Figure 13-7 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 DS6800 storage server, and that the number of volumes will be consolidated from 6 down to two, with the respective capacity as displayed in Figure 13-8.

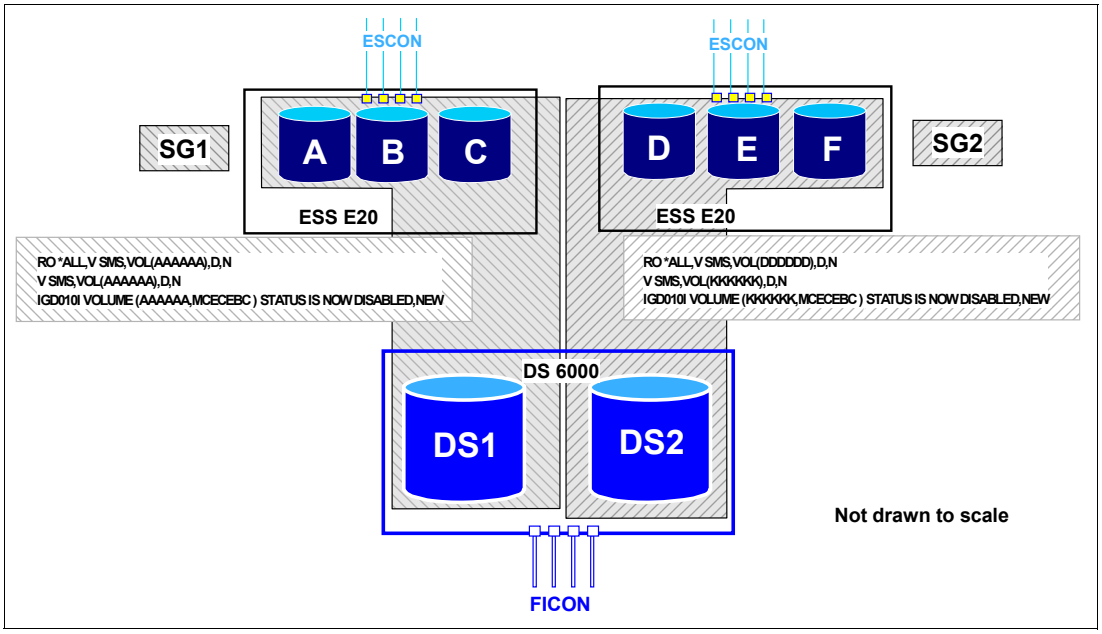



*Figure 13-8 Utilize SMS SG and Volume status to direct all new allocation to new volumes*

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 F 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 through F but must go to volumes DS1 or DS2. This allows you to gradually migrate the data from the old devices A to F onto the new and larger devices DS1 and DS2. This SMS system command 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 SCDS and reload the ACDS accordingly. To make this change permanent you have to change the SCDS through the ISMF panels.

We go briefly through the ISMF panel sequence to make the SMS volume status complete and permanent.

Example 13-6 is the first ISMF Storage Group Application panel where you choose the actual Source Control Data Set (SCDS), the storage group, XC, and select option 4 to alter the SMS volume status.

*Example 13-6 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 13-7. Option 3 allows you to alter the SMS volume status. Here you select the volumes which ought to receive the change.

*Example 13-7 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 13-7 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 13-8, 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 13-8 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      SMS Vol     System/Sys      SMS Vol    ( Possible SMS Vol
  Group Name      Status      Group Name      Status       Status for each:
  ----------      -------     ----------      -------      NOTCON, ENABLE,
```

```
 MCECEBC    ===> ENABLE      MZBCVS2    ===> ENABLE       DISALL, DISNEW,
            ===>                        ===>              QUIALL, QUINEW )
            ===>                        ===>
            ===>                        ===>            * SYS GROUP = sysplex
            ===>                        ===>              minus systems in the
            ===>                        ===>              sysplex explicitly
            ===>                        ===>              defined in the SCDS
            ===>                        ===>
```

In this panel we overtype the SMS volume status with the desired status change. This shows in the following panel, shown in Example 13-9.

*Example 13-9 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:
 ----------      -------     ----------      -------      NOTCON, ENABLE,
 MCECEBC    ===> disnew      MZBCVS2    ===> disnew       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 displayed in Example 13-10.

*Example 13-10 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 13-11 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 13-11 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 over 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 13-5 on page 264. Then allocation runs through the standard SMS allocation, 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. Example 13-12 illustrates this. You might specify for STORGRP a list of storage group names to address multiple storage groups at the same time.

*Example 13-12 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=ADRDSSU
//SYSPRINT DD SYSOUT=*
```

```
//SYSIN    DD *
    COPY  STORGRP(SG1)                                      -
          DS(INC(**)                                        -
             EXCLUDE(SYS1.VTOCIX.*,SYS1.VVDS.*))            -
          DELETE CATALOG SELECTMULTI(ANY) SPHERE-
          ALLDATA(*) ALLX 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 13-12 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 through jobs like in Example 13-12, in which you can run more than one instance concurrently.

### 13.3.3 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.

## 13.4 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,

EXTVTOC, 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 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.

## 13.5 z/VM and VSE/ESA data migration

DFSMS/VM® provides a set of software utility and command functions which are suited 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 3390 to 3390.
- DIRMAINT CMDISK moves data between minidisks in VM from any device type to any device type which is supported by VM.
- CMS COPYFILE offers a logical copy on a file level from any minidisk device to any minidisk device which VM supports and is, therefore, 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 DS6800. Because PPRC is a host server independent approach, similar considerations apply as outlined under “Hardware- and microcode-based migration” on page 258.

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 DS6800 disk storage server through PPRC as outlined under “Hardware- and microcode-based migration” on page 258.
- Logical copy operations under VSE/ESA which allow data movement from source storage servers to new DS6800 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.

## 13.6 Summary of data migration

The route which an installation takes to migrate data to one or more DS6800 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 or no impact when the actual switch utilizes 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 DS6800 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 DS6800 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 1,113 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 which will serve you with promising performance and excellent scalability combined with rich functionality and high availability.

# Part 5

# Implementation and management in the open systems environment

In this part we discuss considerations for the DS6000 series when used in an open systems environment. The topics include:

- Open systems support and software
- Data migration in open systems

14

# Open systems support and software

In this chapter we describe how the DS6000 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 DS6000
- IBM solutions and services that integrate DS6000 functionality

# 14.1 Open systems support

The scope of open systems support of the new DS6000 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 14.1.2, “Where to look for updated and detailed information” on page 277.

## 14.1.1 Supported operating systems and servers

Table 14-1 provides an overview of the open system platforms, operating systems, and high availability clustering applications that are generally supported for attachment to the DS6000. 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 14.1.2, “Where to look for updated and detailed information” on page 277.

*Table 14-1 Platforms, operating systems and applications supported with DS6000*

| 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 | |

The DS6000 and DS8000 have the same open systems support matrix. There are only a few exceptions, with respect to the timing.

### 14.1.2 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 DS6000 Interoperability Matrix

The *DS6000 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/ds6000/pdf/ds6000-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 DS6000 Host Systems Attachment Guide

The *DS6000 Host Systems Attachment Guide,* SC26-7680, guides you in detail through all the steps that are required to attach an open system host to your DS6000 storage system. It is available at:

http://www.ibm.com/servers/storage/disk/ds6000

#### 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

### *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 DS6000 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 DS6000. 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.

## 14.1.3 Differences to ESS 2105

For DS6000 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 DS6000. Therefore all parallel SCSI support had to be dropped. No host system can be connected to the DS6000 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 MDS9000 (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)

### 14.1.4 Boot support

For most of the supported platforms and operating systems you can use the DS6000 as a boot device. The *DS6000 Interoperability Matrix* provides detailed information about boot support. Refer to “The DS6000 Interoperability Matrix” on page 277.

The *DS6000 Host Systems Attachment Guide,* SC26-7680, helps you with the procedures necessary to set up your host in order to boot from the DS6000. See “The DS6000 Host Systems Attachment Guide” on page 277.

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 14.2, “Subsystem Device Driver” on page 280.

### 14.1.5 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.

### 14.1.6 Differences in interoperability between DS6000 and DS8000

The DS6000 and DS8000 have the same open systems support matrix. There are only a few exceptions, with respect to the timing.

## 14.2 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.

Most 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 DS6000 attachment, IBM makes the IBM Subsystem Device Driver (SDD) available to provide the following functionality:

- Enhanced data availability through automatic path failover and failback
- 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

In the DS6000, host ports have a fixed assignment to a server (or controller card). Therefore there is a slight performance penalty if data from a logical volume managed by one server is accessed from a port that is located on the other server. The request for the logical volume and the data would have to be transferred across the bridge interface that connects both servers. These transfers add some latency to the response time. Furthermore, this interface is also used to mirror the persistent memory and for other inter-server communication. It could become a bottleneck if too many normal I/O requests ran across it, although it is a high bandwidth, low latency, PCI-X connection.

**Important:** Although this implementation may look similar to that of the DS4000 (formerly known as the FAStT) series, it is fundamentally different. You can access any volume through any path at any time without the volume being reassigned to the alternate server. As long as a server is available, it will keep its volumes, to avoid switchover times, which would have far more impact on performance.

Therefore it is advisable to access a volume primarily through the four host ports that are located on the server it is associated with. DS6000 supports a range of SCSI-3 commands that allow you to identify the optimum and backup paths to each logical volume. In SCSI standard terms this is called *Asymmetrical Logical Unit Access*.

SDD uses these commands to discover the preferred paths to a volume. It will exclusively use these, as long as at least one of them is available. Only if none of the preferred paths can be used will it switch to the alternate paths. Automatic failback to the preferred path is performed as soon as at least one becomes available after a failure.

Of course, SDD performs dynamic load balancing across all available preferred paths to ensure full utilization of the SAN and HBA resources.

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, "Operating systems specifics" on page 299.

**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 303 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 the DS6000 with iSeries" on page 329.

## 14.3 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 303)
- IBM OS/400 V5R3 multipath support (see Appendix B, "Using the DS6000 with iSeries" on page 329)

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 DS6000 attachment, although the scope may vary. There may be limitations for certain host bus adapters or operating system versions. Always consult the DS6000 Interoperability Matrix for the latest information.

## 14.4 DS CLI

The DS Command-Line Interface (DS CLI) provides a command set to perform almost all monitoring and configuration tasks on the DS6000. 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 DS6000 Licensed Internal Code and is delivered with the Customer Software Packet. It is closely tied to the installed version of code and is therefore not available for download. The CLI code must be obtained from the same software bundle as the current microcode update. When DS6000 code updates occur, the DS CLI must also be updated.

The DS CLI is available for most of the supported operating systems. The DS6000 Interoperability Matrix contains a complete list.

There are some pre-requisites for the host system that the DS CLI is running on, specifically:

- 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 *DS6000 Command-Line Interface User's Guide*, SC26-7681 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 DS MC 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 DS-MC for immediate execution.

There also is a section in this book describing the usage of the DS CLI, including some examples (Chapter 10, "DS CLI" on page 195).

The *DS6000 Command-Line Interface User's Guide*, SC26-7681 contains a complete command reference.

## 14.5 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.

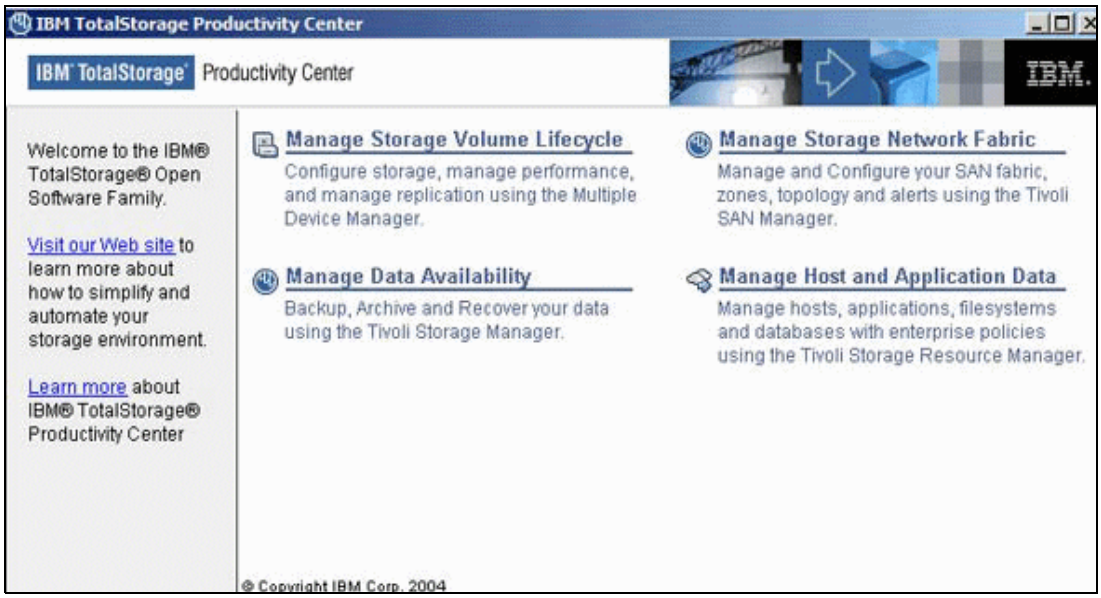



*Figure 14-1 IBM TotalStorage Productivity Center*

TPC is the integration point for storage and fabric management, and replication, as depicted in Figure 14-1. 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 lifecycle, 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 filesystem 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 14-2 on page 284 shows the MDM main panel.

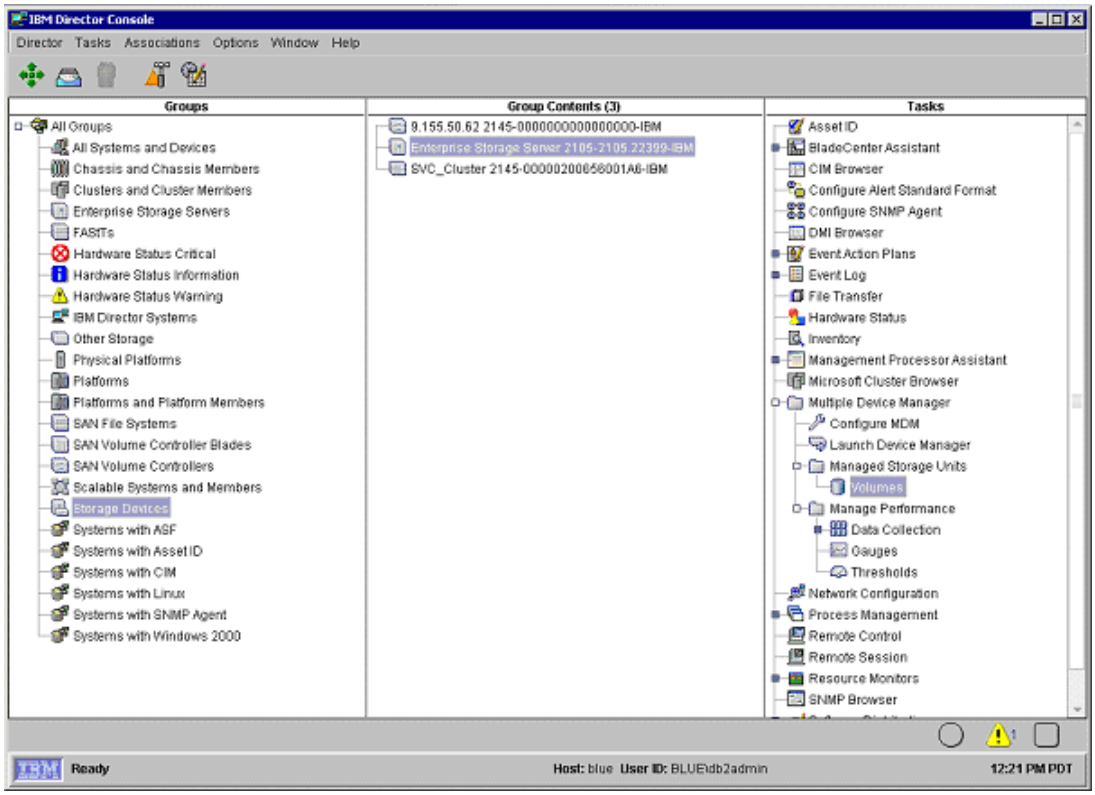



*Figure 14-2 MDM main panel*

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.

## 14.5.1 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 14-3 shows an example of a DM view.

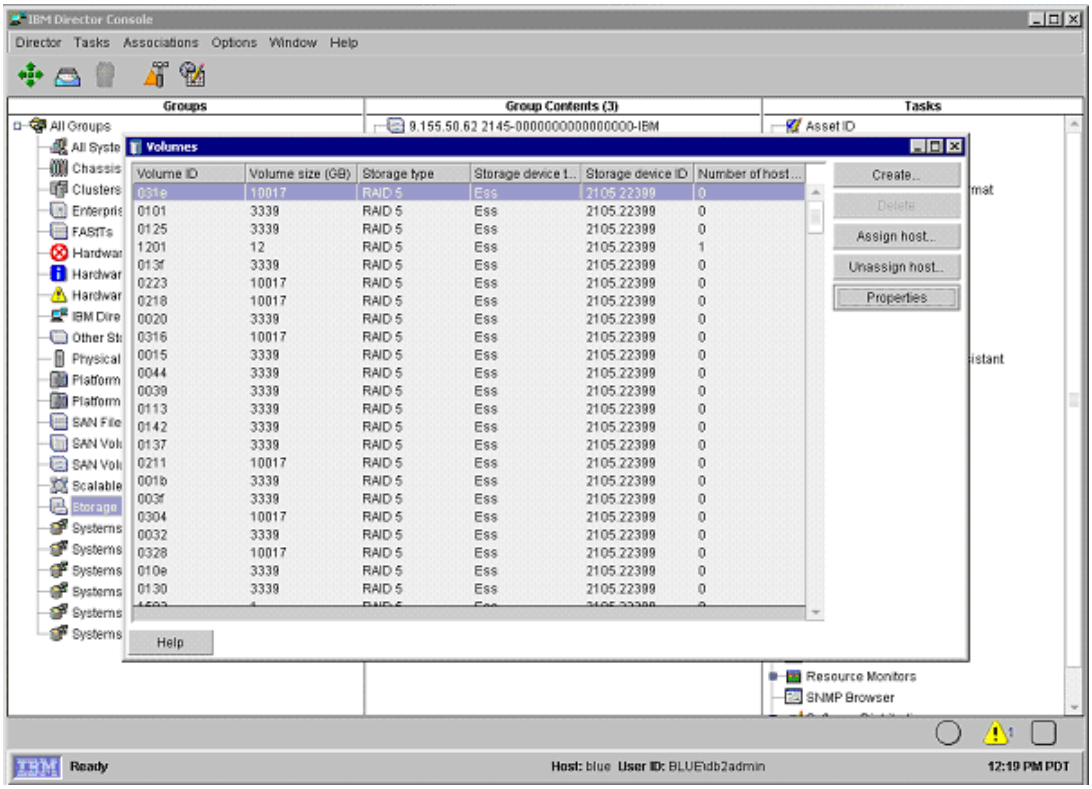



*Figure 14-3 Sample Device Manager view*

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

## 14.5.2 TCP 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.
- 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 14-4.

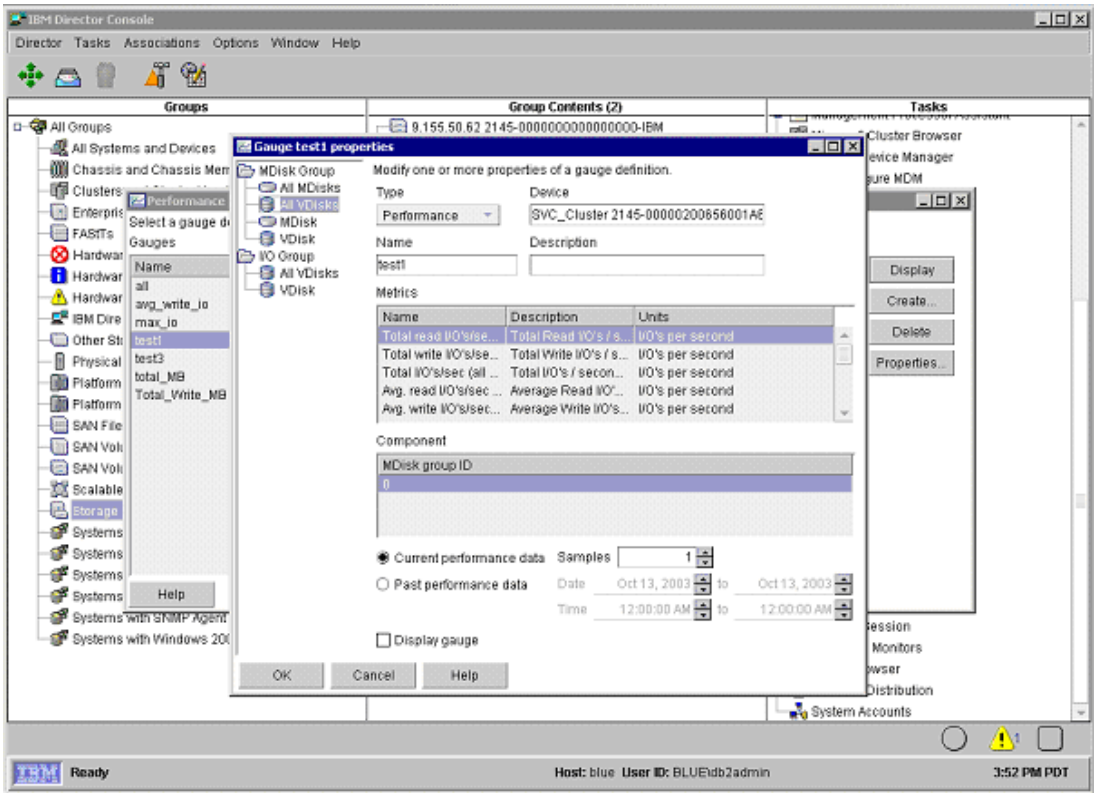



*Figure 14-4 Example screenshot of TPC for Disk*

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 use of the Performance Manager refer to the redbook *IBM TotalStorage Multiple Device Manager Usage Guide,* SG24-7097.

### 14.5.3 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 volume 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.

## 14.6 Global Mirror Utility

The *DS6000 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 specific service arrangement.

The GMU is distributed on a separate CD with the DS6000 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 DS6000 Storage 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 DS6000 volume status and the Global Mirror session status, you can either use the DS6000 Storage Manager or CLI.

## 14.7 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 DS6000 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 DS6000 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

## 14.8 Summary

The new DS6000 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 DS6000 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.

15

# Data migration in the open systems environment

In this chapter we discuss important concepts for the migration of data to the new DS6000:

- Data migration considerations
- Data migration and consolidation
- Comparison of the different methods

## 15.1 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 DS6000. 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 15-1:

- 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

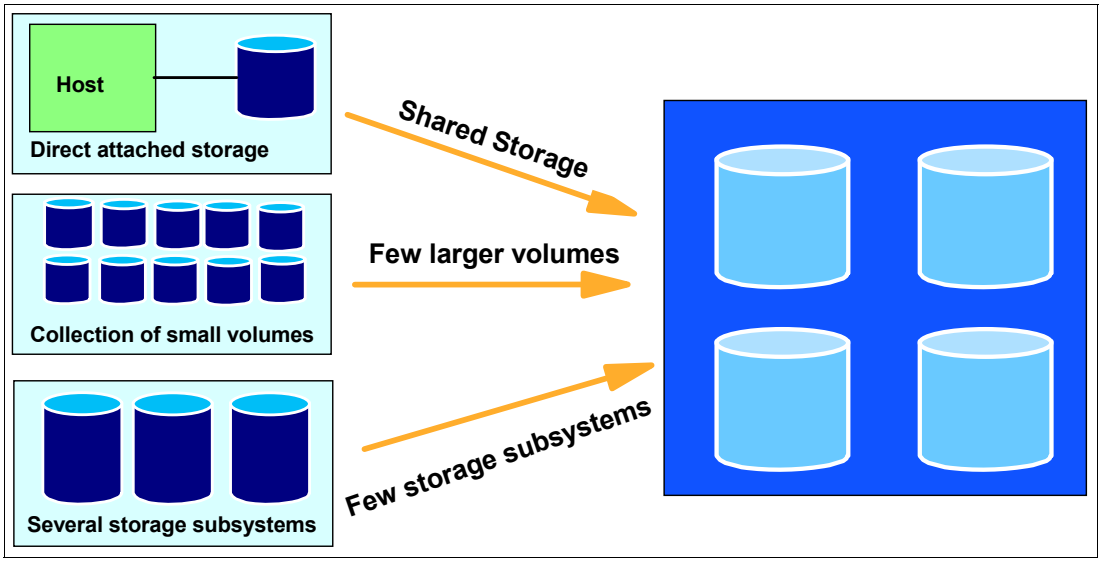



*Figure 15-1 Different ways of consolidation*

Very often, the goal of a consolidation effort is a combination of more than one of these types.

There are many different methods for data migration. To decide on 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 DS6000 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 DS6000 can be found in the *DS6000 Introduction and Planning Guide*, GC26-7679 at:

`http://www-1.ibm.com/servers/storage/disk/ds6000/pdf/DS6000_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.

## 15.2 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 the DS6000 with iSeries” on page 329.

### 15.2.1 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 is 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 DS6000 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 15-2 outlines the process.

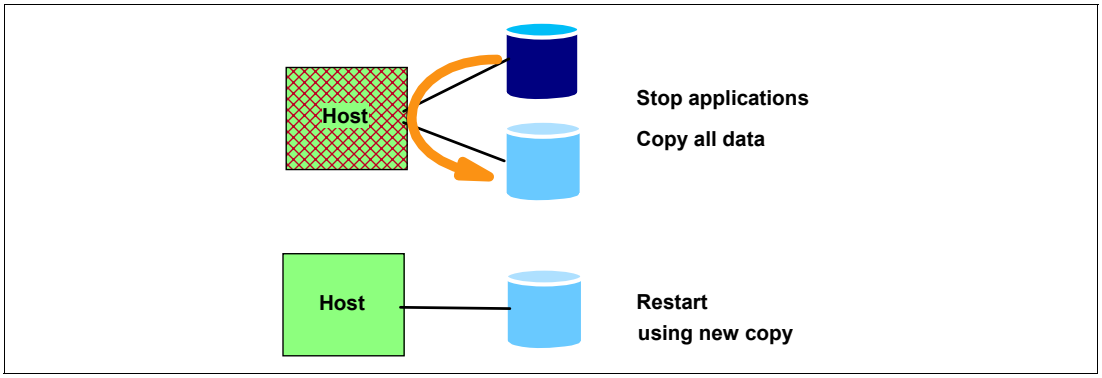



*Figure 15-2 Migration with copy commands*

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 15-3 shows the steps required.

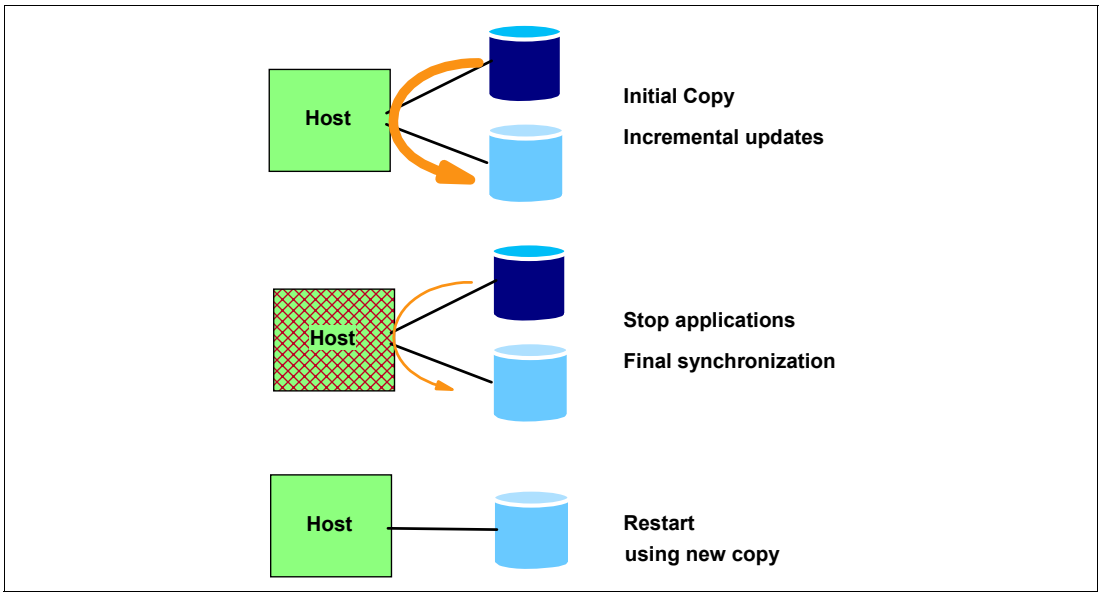



*Figure 15-3 Data migration with rsync*

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 **rsnyc** 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 DS6000 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 15-4 shows the process.

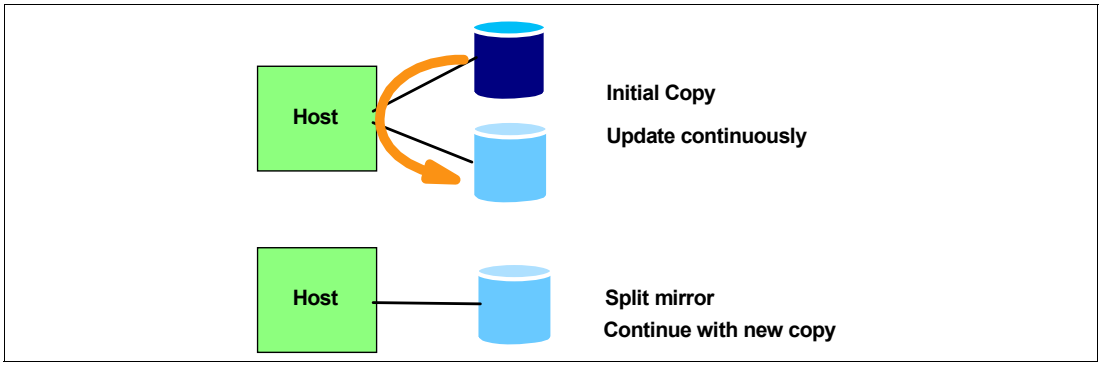



*Figure 15-4 Migration using LVM mirroring*

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 15-5 on page 295 illustrates this method.

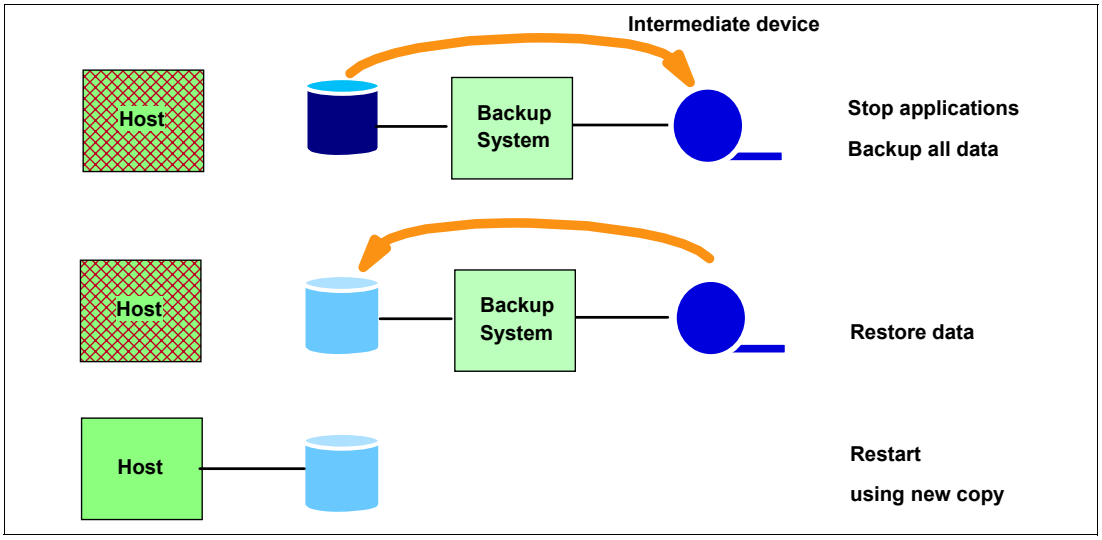



*Figure 15-5 Migration using backup and restore*

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.

## 15.2.2 Subsystem-based data migration

The DS6000 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

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 15-6 outlines the migration steps.

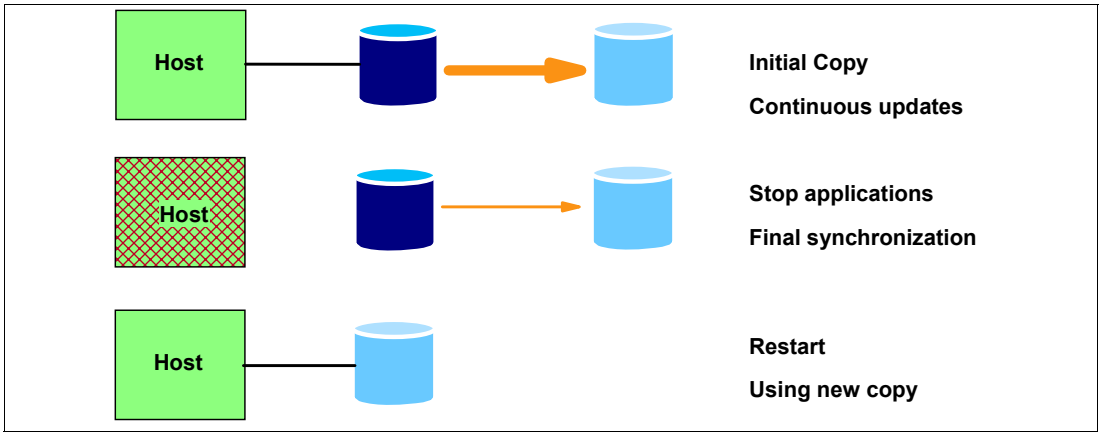



*Figure 15-6 Migration with Metro Mirror or Global Copy*

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 support for remote copy over Fibre Channel and can therefore not be mirrored directly to a DS6000.

## 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 15-7 shows the setup and steps to take for this method.

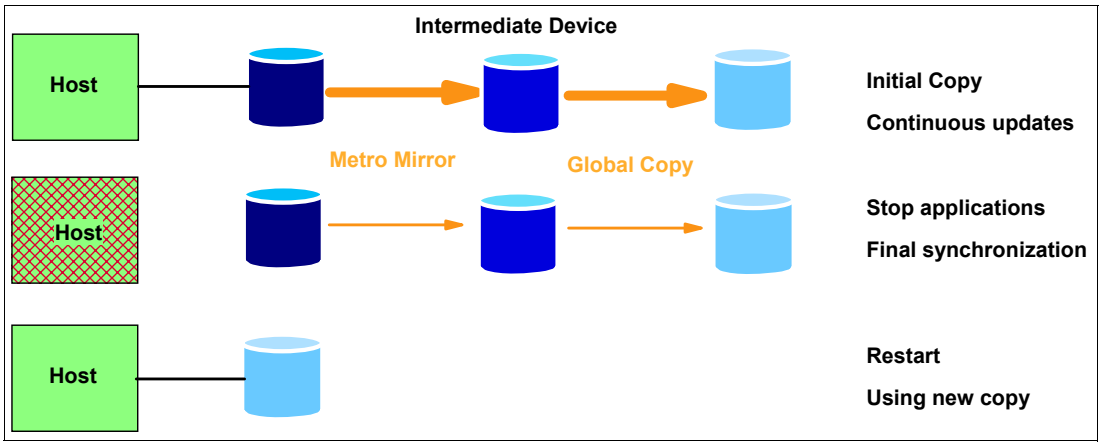



*Figure 15-7 Migration with Metro Mirror, Global Copy and an intermediate device*

### 15.2.3 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 13, “Data Migration in zSeries environments” on page 251.

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 15-8 illustrates how Piper works.

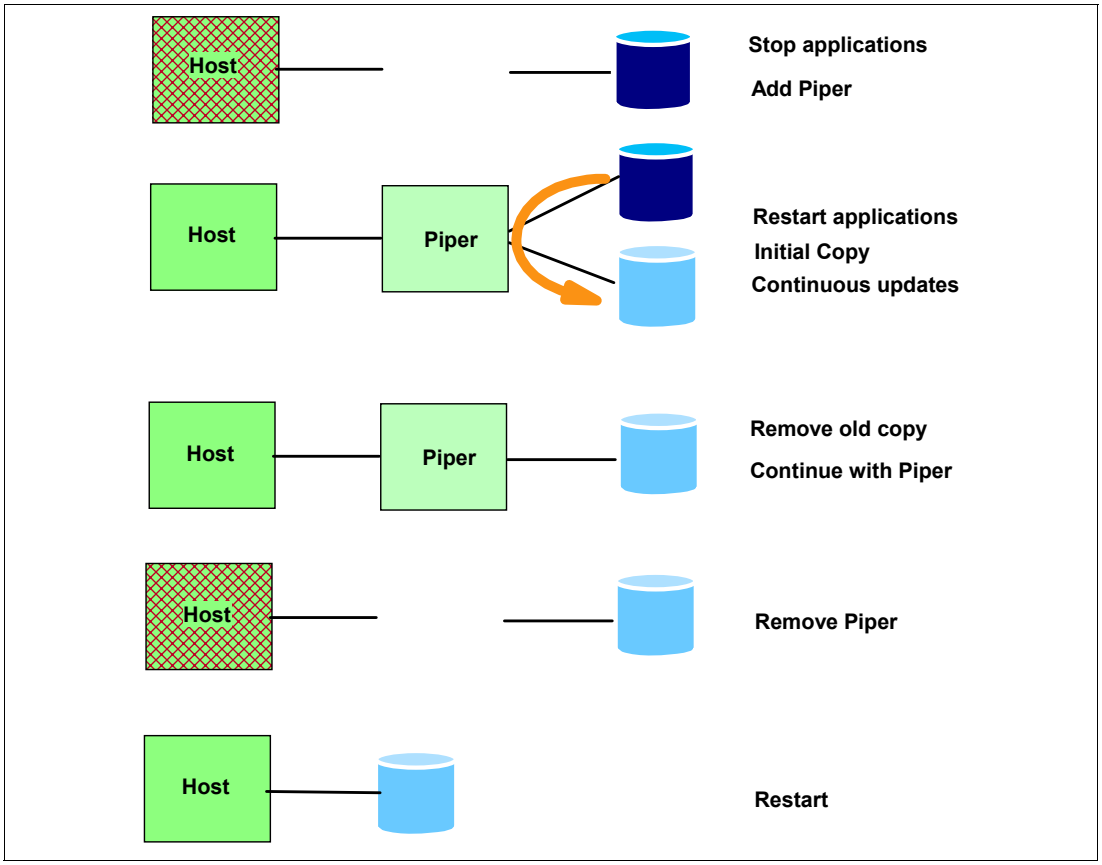



*Figure 15-8 Piper migration*

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 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

### 15.2.4 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.

## 15.3 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 DS6000 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 DS6000 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.

## 15.4 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

# Operating systems specifics

In this appendix, we describe the particular issues of some operating systems with respect to the attachment to a DS6000. 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 DS6000 Host Systems Attachment Guide

The *DS6000 Host Systems Attachment Guide,* SC26-7680, provides instructions to prepare a host system for DS6000 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 DS6000 (when supported)

It varies in detail for the different platforms. Download it from:

`http://www.ibm.com/servers/storage/disk/ds6000`

Many more publications are available from IBM and other vendors. Refer to Chapter 14, “Open systems support and software” on page 275 and to the operating system-specific sections in this chapter.

## Planning

Thorough planning is necessary to ensure that your new DS6000 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 11, “Performance considerations” on page 219.

### Capacity planning considerations

For proper sizing of your DS6000 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?
- 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?

### Data placement considerations

A DS6000 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 DS6000 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 301 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 request 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** man page 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 303. 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** man page 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** man page 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 *DS6000 Host Systems Attachment Guide*, GC26-7680, 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 DS6000 disks. They prepare the *Object Data Manager* (ODM) with information about the subsystem. That way the DS6000 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 DS6000 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 DS6000 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 DS6000 disks to a pSeries server, the World Wide Port Name (WWPN) of each of the pSeries Fibre Channel adapters has to be registered in the DS6000. 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 -vl <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 DS6000 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 DS6000 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 DS6000. SDD is available for AIX 5.1, 5.2, and 5.3.

Refer to 14.2, “Subsystem Device Driver” on page 280 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: 1750 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: 1750 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 <volumegroup 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 DS6000 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 14.2, “Subsystem Device Driver” on page 280).

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 DS6000 Host Attachment Script and install the MPIO version of it. This script identifies the DS6000 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.

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

The management of MPIO devices is described in the “Managing MPIO-Capable Devices” section of the *System Management Guide: Operating System and Devices for AIX 5L*:

http://publib16.boulder.ibm.com/pseries/en_US/aixbman/baseadmn/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-03.ibm.com/servers/storage/support/software/sanvc/installing.html

### *Determine the installed SDDPCM level*

You use the same command as for SDD, **`lslpp -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 man pages 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 DS6000 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 DS6000, LVM striping can be used to distribute data across more than one array (rank).

Refer to Chapter 11, “Performance considerations” on page 219 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 15, “Data migration in the open systems environment” on page 289.

## 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 **pstat -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 DS6000 is supported as a boot device on RS/6000 and pSeries that support Fibre Channel boot capability. This support is also available for the IBM eServer BladeCenter. Refer to *DS6000 Host Systems Attachment Guide*, SC26-7680, 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 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 DS6000, 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 DS6000, see Appendix B, “Using the DS6000 with iSeries” on page 329.

For more information on running AIX in an i5 partition, refer to the i5 Information Center at:

http://publib.boulder.ibm.com/infocenter/iseries/v1r2s/en_US/index.htm?info/iphat/iphatlparkickoff.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 301 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 -O al; 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/u02lv                /u02
  0.01      0  16896 1004.3  /dev/u04lv                /u04
  0.00      0   3968  235.9  /dev/u03lv                /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 1750
  0.87    496    800   77.0  /dev/hdisk79              IBM MPIO FC 1750
  0.87    672    776   86.1  /dev/hdisk81              IBM MPIO FC 1750
  0.86    392    960   80.4  /dev/hdisk63              IBM MPIO FC 1750
  0.86    328    776   65.6  /dev/hdisk83              IBM MPIO FC 1750
  0.86    528    624   68.5  /dev/hdisk69              IBM MPIO FC 1750
  0.86    480    656   67.5  /dev/hdisk55              IBM MPIO FC 1750
  0.86    408    536   56.1  /dev/hdisk73              IBM MPIO FC 1750
  0.86    456    720   69.9  /dev/hdisk77              IBM MPIO FC 1750
  0.86    440    720   68.9  /dev/hdisk59              IBM MPIO FC 1750

...

------------------------------------------------------------------------
Detailed Physical Volume Stats   (512 byte blocks)
------------------------------------------------------------------------

VOLUME: /dev/hdisk65  description: IBM MPIO FC 1750
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 DS6000 Interoperability Matrix:

http://www.ibm.com/servers/storage/disk/ds6000/pdf/ds6000-matrix.pdf

See also “The DS6000 Interoperability Matrix” on page 277.

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 DS6000 storage subsystem.

### The DS6000 Host Systems Attachment Guide

The *DS6000 Host Systems Attachment Guide,* GC26-7680 provides instructions to prepare an Intel IA-32-based machine for DS6000 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 DS6000

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 DS6000. The details provided for the attachment to the IBM Enterprise Storage Server (ESS 2105) are mostly valid for DS6000, 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`

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

Most of the information provided in these publications is valid for DS6000 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/sd*x* 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 | /dev/sda | /dev/sdp |
| 65 | /dev/sdq | /dev/sdaf |
| 66 | /dev/sdag | /dev/sdav |
| 71 | /dev/sddi | /dev/sddx |
| 128 | /dev/sddy | /dev/sden |
| 129 | /dev/sdeo | /dev/sdfd |
| 135 | /dev/sdig | /dev/sdiv |

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 man page of the **mknod** command for more details. Example A-10 shows the creation of special device files for the 17th SCSI disk 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 DS6000 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 DS6000 attachment to Linux on Intel IA-32 and IA-64 architectures, IBM pSeries and iSeries is the IBM Subsystem Device Driver (SDD) (see 14.2, “Subsystem Device Driver” on page 280). 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 DS6000 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 DS6000 provides the volumes to the host with a fixed Fibre Channel address and varying LUN. Therefore RH-EL 3 will see only one DS6000 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 shows how the /etc/modules.conf file should look 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 DS6000 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_lpfc_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/rel/201g/force_lpfc_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 rescan 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 DS6000 volume, seen by HBA 0
/dev/sdc - 2nd DS6000 volume, seen by HBA 0
/dev/sdd - 1st DS6000 volume, seen by HBA 1
/dev/sde - 2nd DS6000 volume, seen by HBA 1
```

Example A-14 shows the SCSI disk assignment after one more DS6000 volume is added.

*Example: A-14 SCSi disks after dynamic addition of another DS6000 volume*

```
/dev/sda - internal SCSI disk
/dev/sdb - 1st DS6000 volume, seen by HBA 0
```

```
/dev/sdc - 2nd DS6000 volume, seen by HBA 0
/dev/sdd - 1st DS6000 volume, seen by HBA 1
/dev/sde - 2nd DS6000 volume, seen by HBA 1
/dev/sdf - new DS6000 volume, seen by HBA 0
/dev/sdg - new DS6000 volume, seen by HBA 1
```

The mapping of special device files is now different than it would have been if all three DS6000 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 316.

*Example: A-15 SCSi disks after dynamic addition of another DS6000 volume and reboot*

```
/dev/sda - internal SCSI disk
/dev/sdb - 1st DS6000 volume, seen by HBA 0
/dev/sdc - 2nd DS6000 volume, seen by HBA 0
/dev/sdd - new DS6000 volume, seen by HBA 0
/dev/sde - 1st DS6000 volume, seen by HBA 1
/dev/sdf - 2nd DS6000 volume, seen by HBA 1
/dev/sdg - new DS6000 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 DS6000 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 DS6000, 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 DS6000, see Appendix B, “Using the DS6000 with iSeries” on page 329.

More information on running Linux in an iSeries partition can be found in the iSeries Information Center at:

http://publib.boulder.ibm.com/iseries/v5r2/ic2924/index.htm

For running Linux in an i5 partition check, the i5 Information Center at:

http://publib.boulder.ibm.com/infocenter/iseries/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: 1750511          Rev: .545
  Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi2 Channel: 00 Id: 00 Lun: 01
  Vendor: IBM      Model: 1750511          Rev: .545
  Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi2 Channel: 00 Id: 00 Lun: 02
  Vendor: IBM      Model: 1750511          Rev: .545
  Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 00
  Vendor: IBM      Model: 1750511          Rev: .545
  Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 01
  Vendor: IBM      Model: 1750511          Rev: .545
  Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Id: 00 Lun: 02
  Vendor: IBM      Model: 1750511          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 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 man page 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/sg*x*** 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 DS6000 support is provided on RPQ only), we do not discuss Windows NT here.

DS6000 supports FC attachment to Microsoft Windows 2000/2003 servers. For details regarding operating system versions and HBA types see the *DS6000 Interoperability Matrix*, available at:

http://www.ibm.com/servers/storage/disk/ds6000/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 *DS6000 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.mspx

    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.

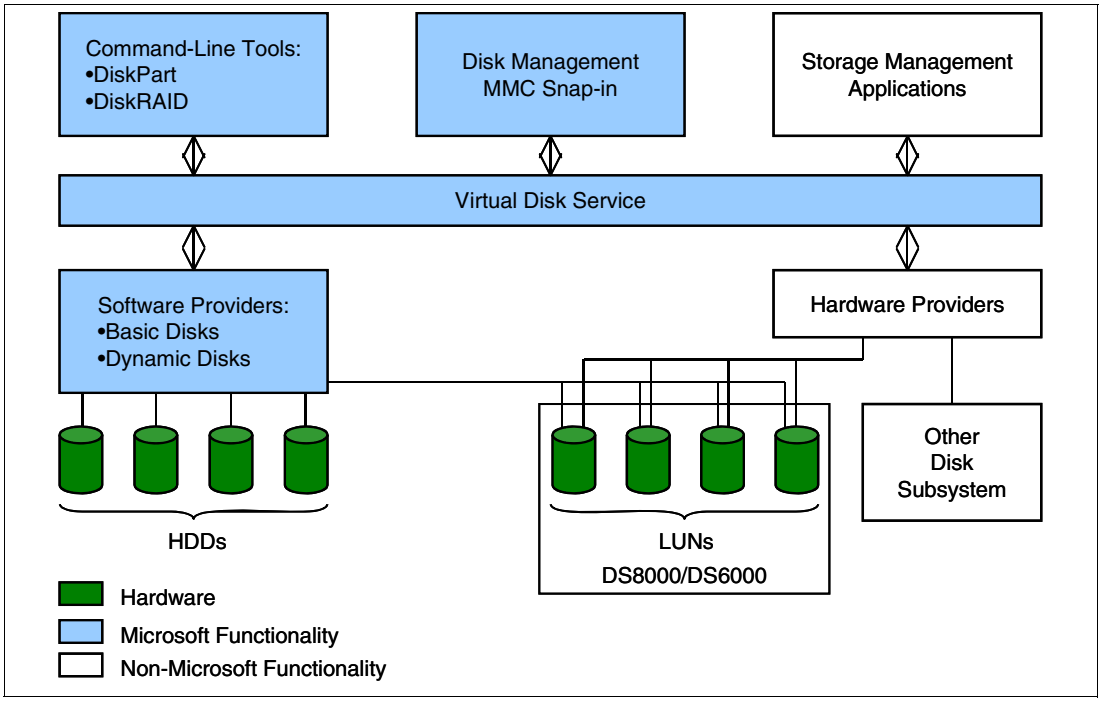



*Figure A-1 Microsoft VDS Architecture*

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`

The DS6000 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 DS6000 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 DS6000. 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

### *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

## HP OpenVMS

DS6000 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 *DS6000 Interoperability Matrix*, available at:

http://www.ibm.com/servers/storage/disk/ds6000/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 DS6000 host adapter microcode. As a result, DS6000 FC ports cannot be shared between OpenVMS and non-OpenVMS hosts.

**Important:** The DS6000 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 the DS6000 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 DS6000 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 DS6000, 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 DS6000 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 DS6000, 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 DS6000 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 DS6000 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 revectored 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 (WRITEL), 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/WRITEL, 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/wrl/DECarchives/DTJ/DTJ301/DTJ301SC.TXT`

The DS6000 provides volumes as SCSI-3 devices and thus does not implement a forced error indicator. It also does not support the READL and WRITEL command set for data integrity reasons.

Usually the OpenVMS SCSI Port Driver recognizes if a device supports READL/WRITEL, 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 the DS6000 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 DS6000. This section describes the hardware and software pre-requisites for attaching the DS6000.

## Hardware

The DS6000 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 DS6000.

There are two Fibre Channel adapters for iSeries. Both support the DS6000:

- 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

## 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 DS6000 to iSeries, you should check for the latest PTFs, which may have superseded those shown here.

# Logical volume sizes

OS/400 is supported on DS6000 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 | | | | | |
| 1750-A81 | 1750-A01 | 8.5 | 16,777,216 | 8 | 0.00 | 100.00 |
| 1750-A82 | 1750-A02 | 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 | | | | | |
| 1750-A85 | 1750-A05 | 35.1 | 68,681,728 | 33 | 0.25 | 99.24 |
| 1750-A84 | 1750-A04 | 70.5 | 137,822,208 | 66 | 0.28 | 99.57 |
| 1750-A86 | 1750-A06 | 141.1 | 275,644,416 | 132 | 0.56 | 99.57 |
| 1750-A87 | 1750-A07 | 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 330 in conjunction with Figure 8-9 on page 141 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 353.

# 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 DS6000 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 DS6000. In this case, only one LUN should be defined as unprotected; otherwise, when mirroring is started to mirror the LSU to the DS6000 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 DS6000, 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 than removing an internal disk from an OS/400 configuration. Indeed, deleting a logical volume on the DS6000 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 DS6000. 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 DS6000 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 333.

```
                          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*

4. 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.

```
                       Work with Disk Configuration

Select one of the following:

     1. Display disk configuration
     2. Add units to ASPs
     3. Work with ASP threshold
     4. Include unit in device parity protection
     5. Enable remote load source mirroring
     6. Disable remote load source mirroring
     7. Start compression on non-configured units
     8. Add units to ASPs and balance data
     9. Start device parity protection


Selection
     8

F3=Exit         F12=Cancel
```

*Figure B-3 Work with Disk Configuration menu*

5. Figure B-4 on page 334 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.

```
                        Specify ASPs to Add Units to

Specify the ASP to add each unit to.

Specify  Serial                                Resource
  ASP    Number      Type  Model  Capacity    Name
         21-662C5    4326   050      35165    DD124
         21-54782    4326   050      35165    DD136
  1      75-1118707  1750   A85      35165    DD006


F3=Exit     F5=Refresh     F11=Display disk configuration capacity
F12=Cancel
```

*Figure B-4 Specify ASPs to Add Units to*

6. The Confirm Add Units panel will appear for review as shown in Figure B-5. If everything is correct, press Enter to continue.

```
                              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.

            Serial                  Resource
ASP  Unit   Number      Type Model  Name        Protection
  1                                             Unprotected
       1    02-89058    6717  074   DD004       Device Parity
       2    68-0CA4E32  6717  074   DD003       Device Parity
       3    68-0C9F8CA  6717  074   DD002       Device Parity
       4    68-0CA5D96  6717  074   DD001       Device Parity
       5    75-1118707  1750  A85   DD006       Unprotected


F9=Resulting Capacity           F12=Cancel
```

*Figure B-5 Confirm Add Units*

7. 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.

1. Start iSeries Navigator. Figure B-6 on page 335 shows the initial panel.

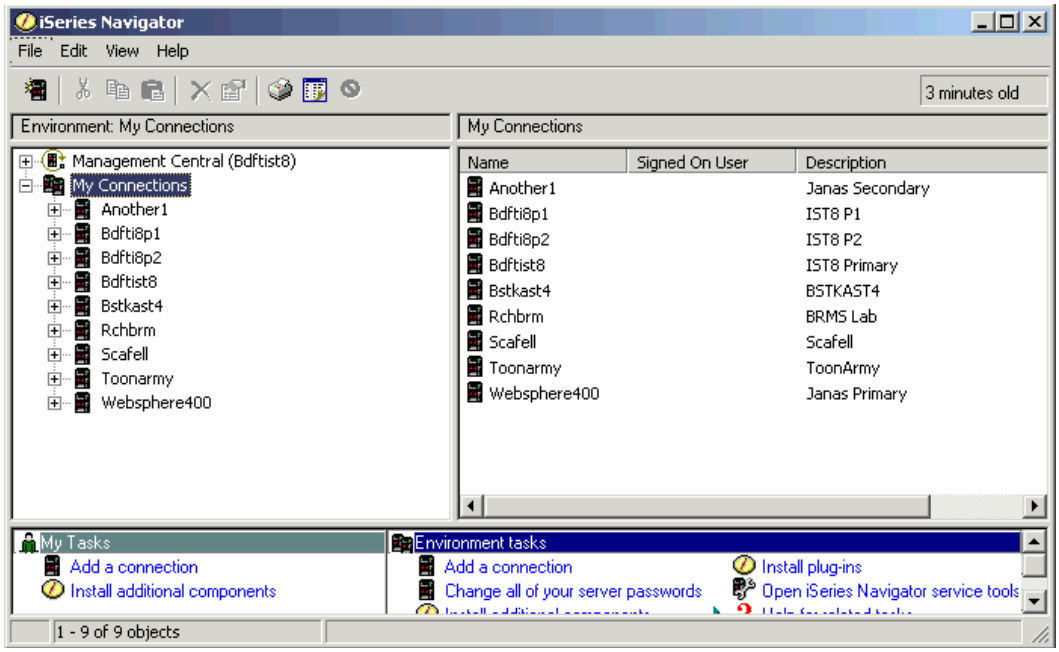



*Figure B-6 iSeries Navigator initial panel*

2. Expand the iSeries to which you wish to add the logical volume and sign on to that server as shown in Figure B-7.

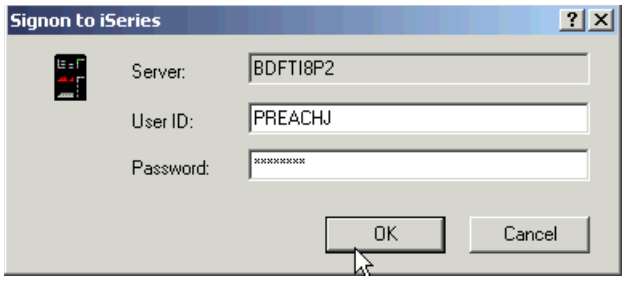



*Figure B-7 iSeries Navigator Signon to iSeries panel*

3. Expand **Configuration and Service**, **Hardware**, and **Disk Units** as shown in Figure B-8 on page 336.

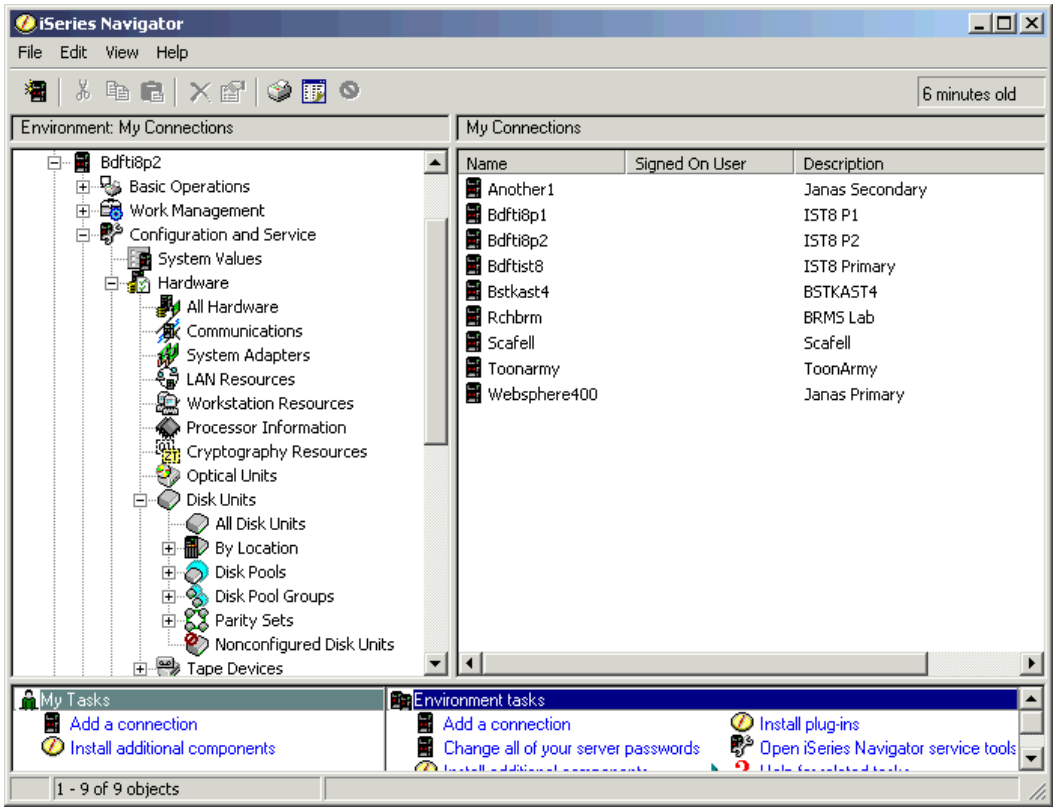



*Figure B-8 iSeries Navigator Disk Units*

4. 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**.

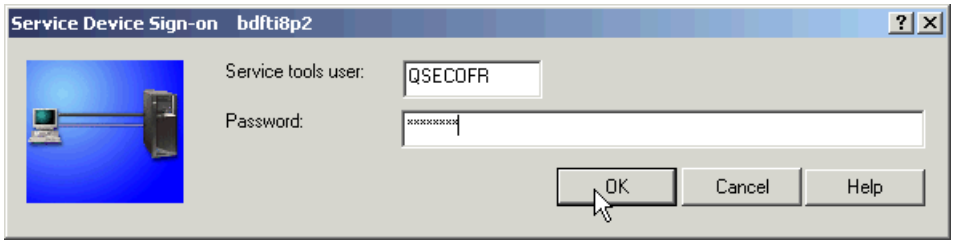



*Figure B-9 SST Signon*

5. Right-click **Disk Pools** and select **New Disk Pool** as shown in Figure B-10 on page 337.

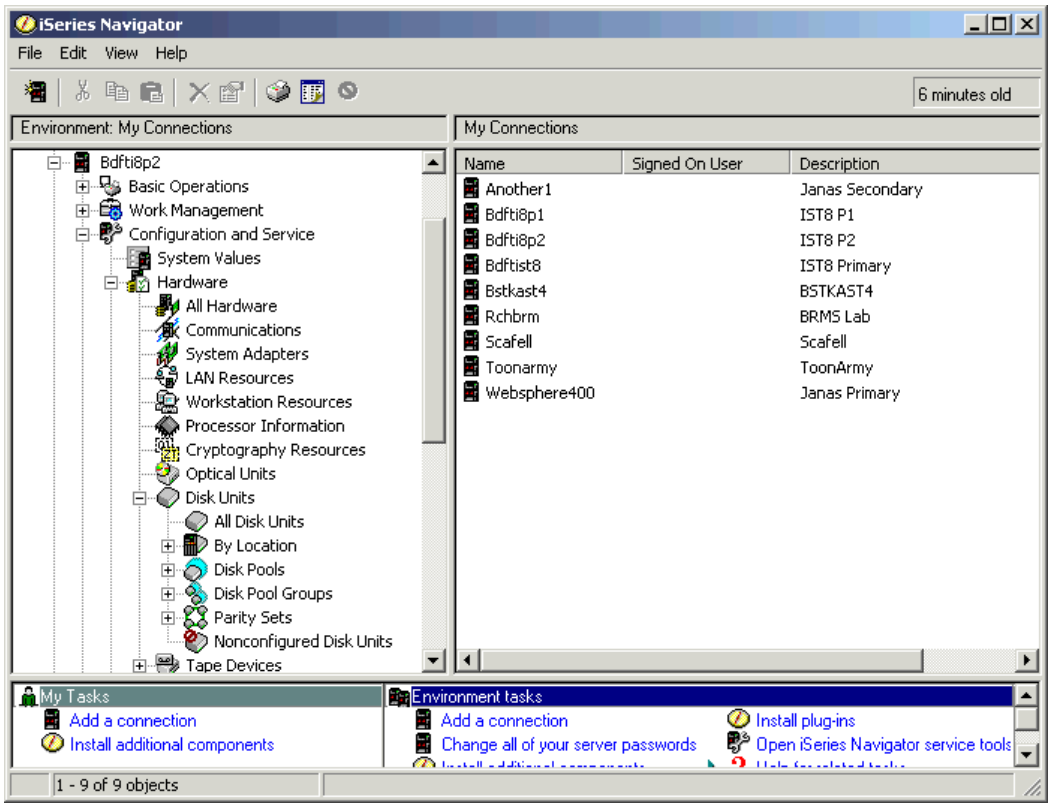



*Figure B-10 Create a new disk pool*

6. The New Disk Pool wizard appears as shown in Figure B-11. Click **Next**.

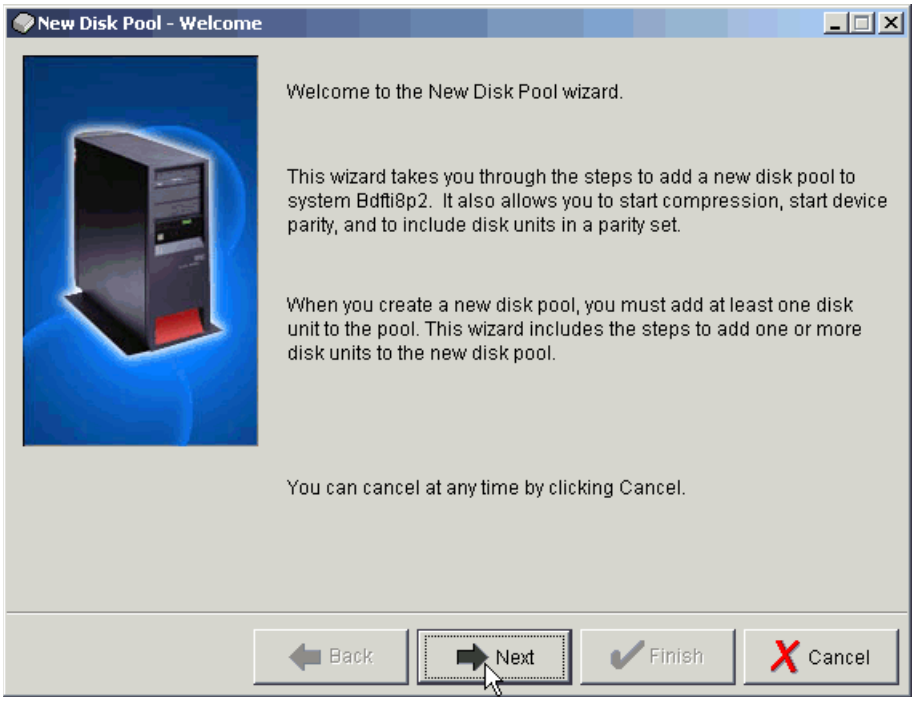



*Figure B-11 New disk pool - welcome*

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.

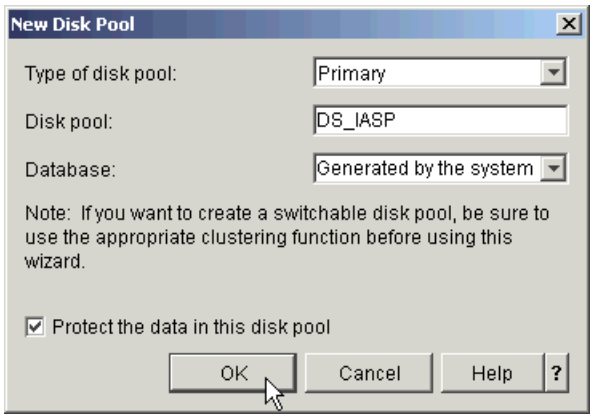



*Figure B-12 Defining a new disk pool*

8. A confirmation panel like that shown in Figure B-13 will appear to summarize the disk pool configuration. Select **Next** to continue.

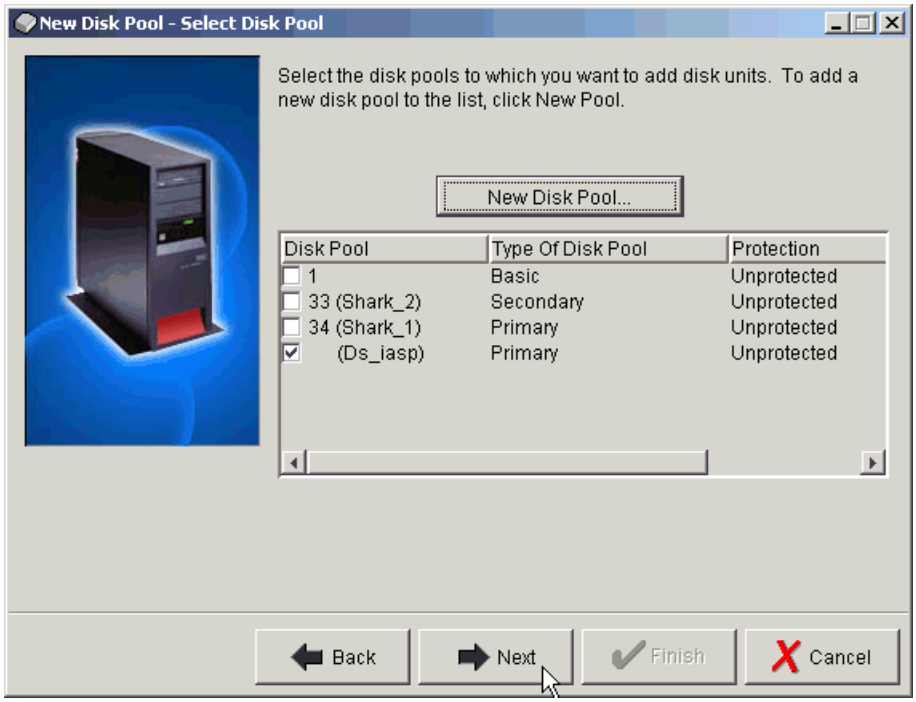



*Figure B-13 Confirm disk pool configuration*

9. 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 339.

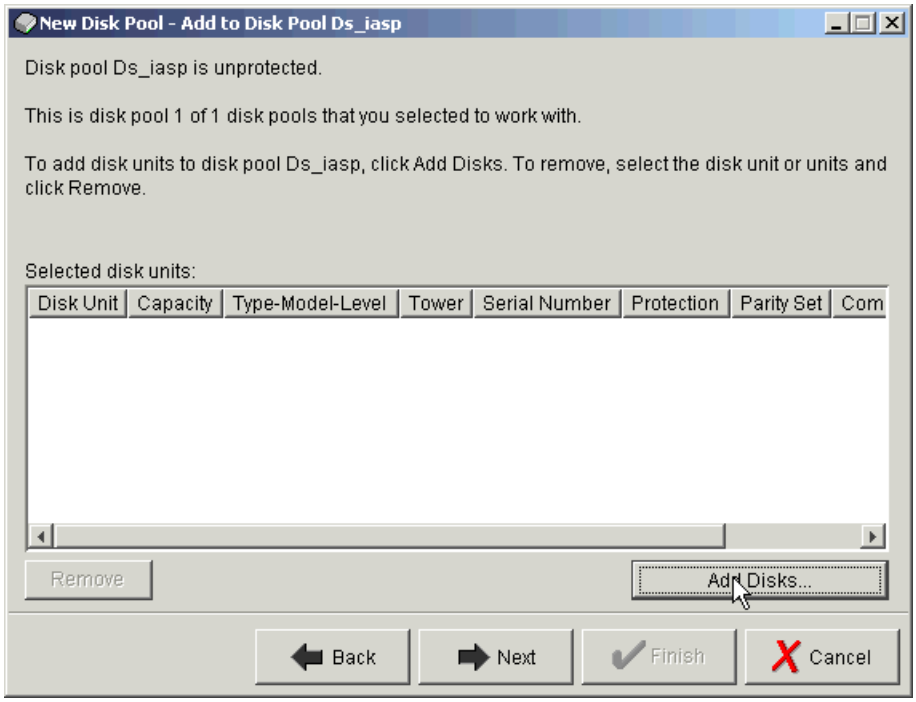



*Figure B-14 Add disks to Disk Pool*

10. 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**.

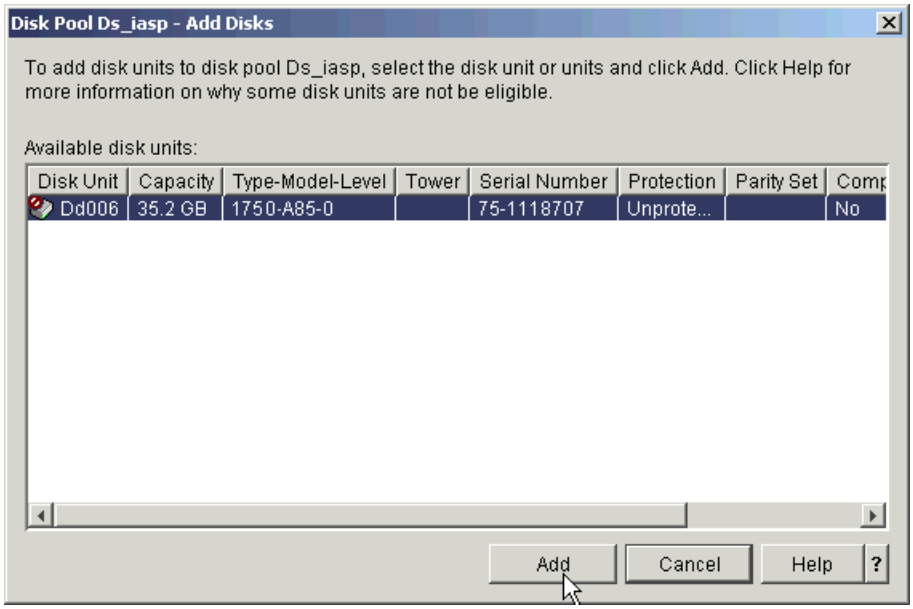



*Figure B-15 Choose the disks to add to the Disk Pool*

11. A confirmation screen appears as shown in Figure B-16 on page 340. Click **Next** to continue.

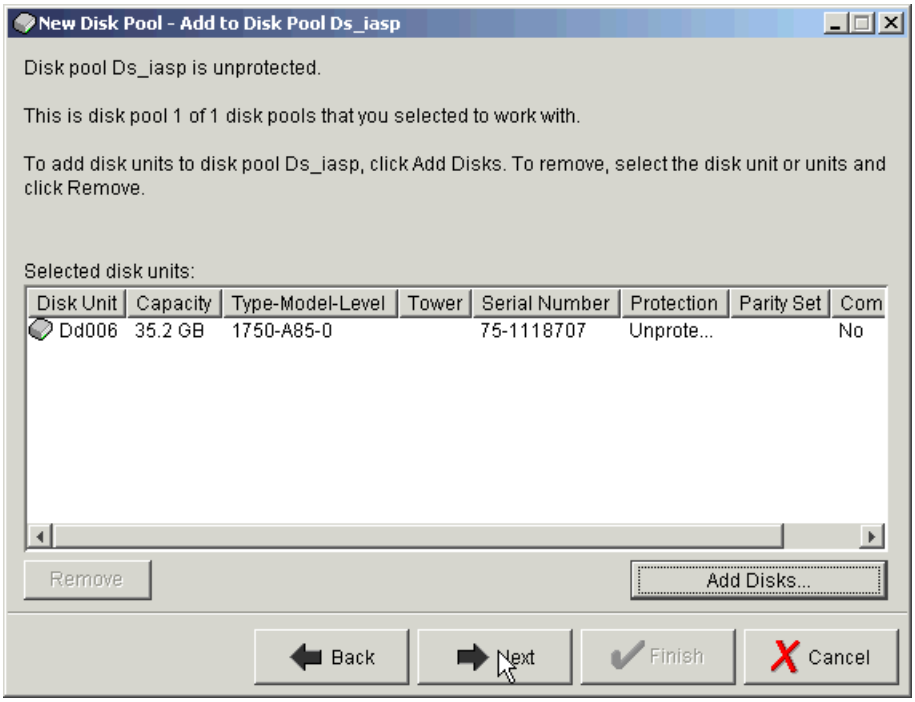



*Figure B-16 Confirm disks to be added to Disk Pool*

12. A summary of the Disk Pool configuration similar to Figure B-17 appears. Click **Finish** to add the disks to the Disk Pool.

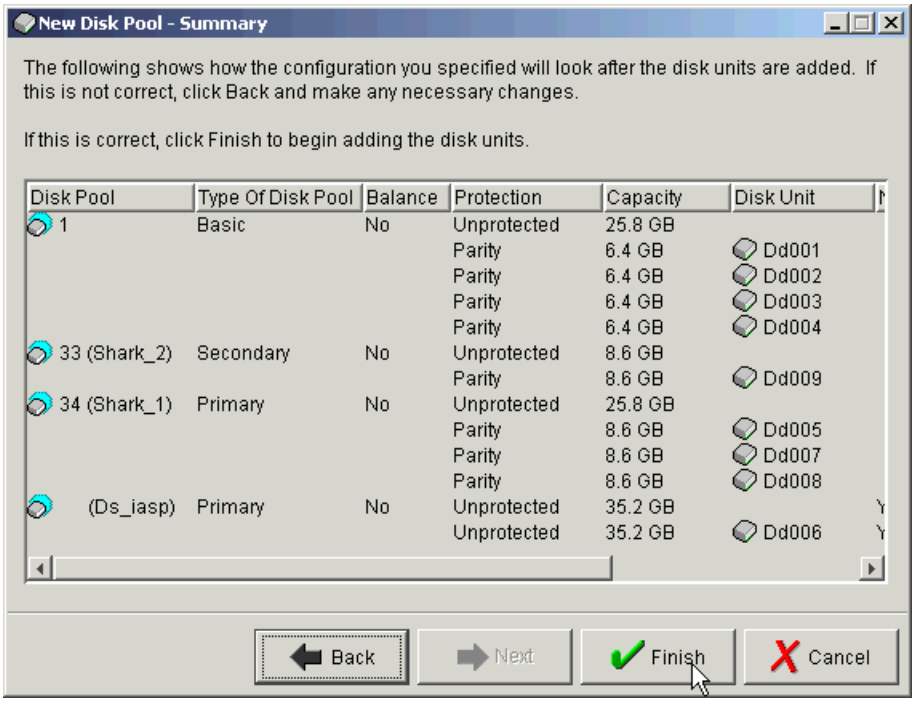



*Figure B-17 New Disk Pool Summary*

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 341 will appear showing progress. This step may take some time, depending on the number and size of the logical units being added.

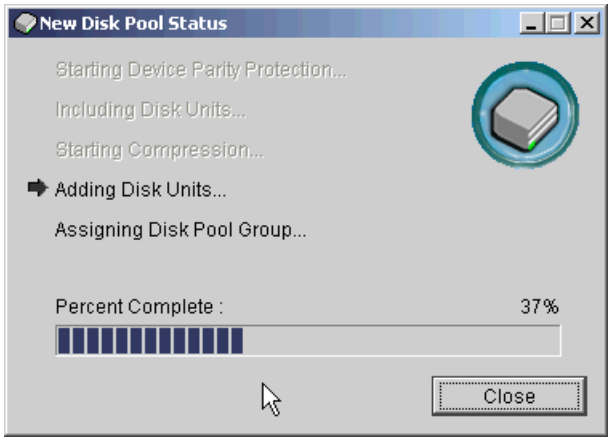



*Figure B-18 New Disk Pool Status*

14. When complete, click **OK** on the information panel shown in Figure B-19.

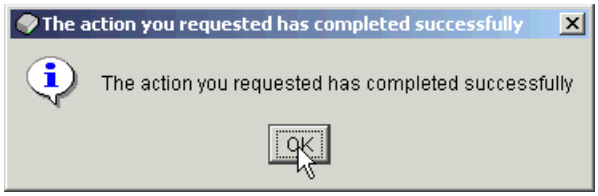



*Figure B-19 Disks added successfully to Disk Pool*

15. The new Disk Pool can be seen on iSeries Navigator **Disk Pools** in Figure B-20.

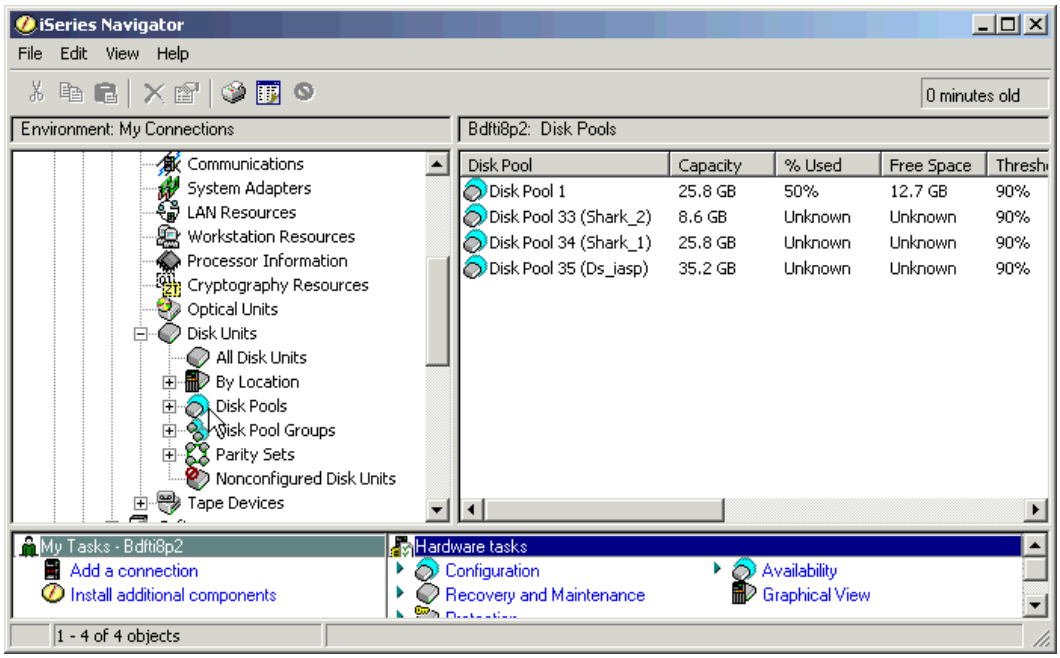



*Figure B-20 New Disk Pool shown on iSeries Navigator*

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.

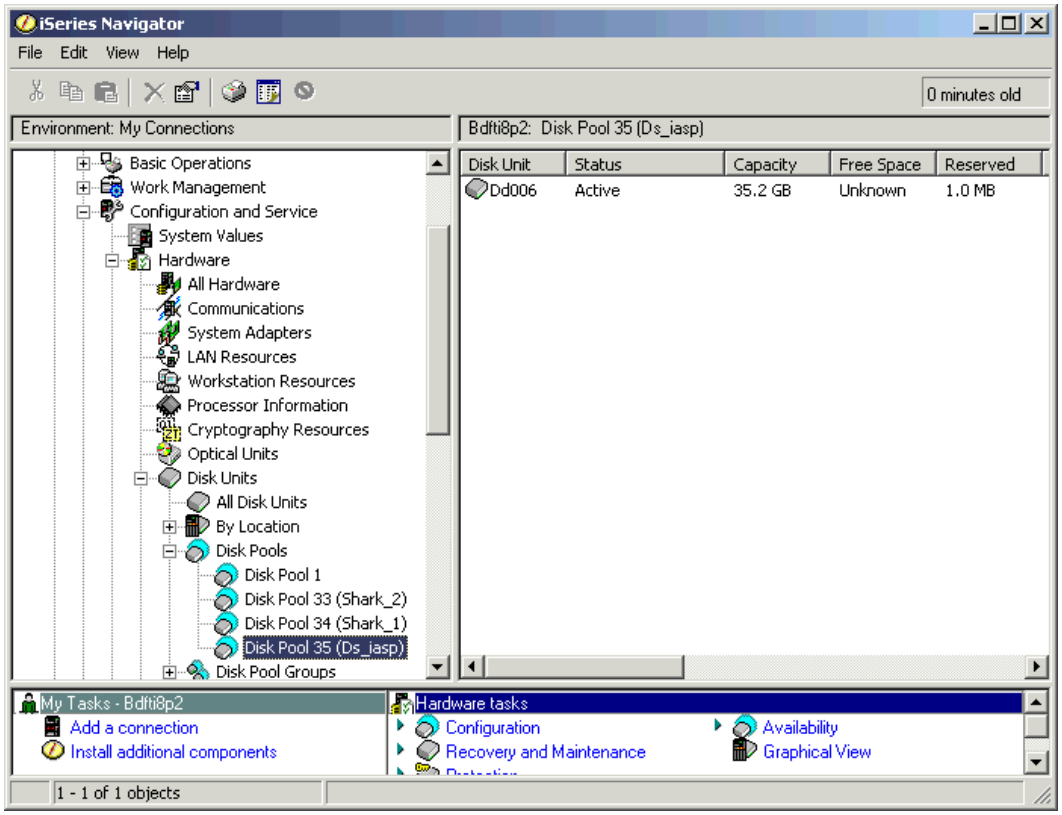



*Figure B-21 New logical volume shown on iSeries Navigator*

# 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 DS6000. 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 DS6000 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 DS6000, 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 DS6000 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).

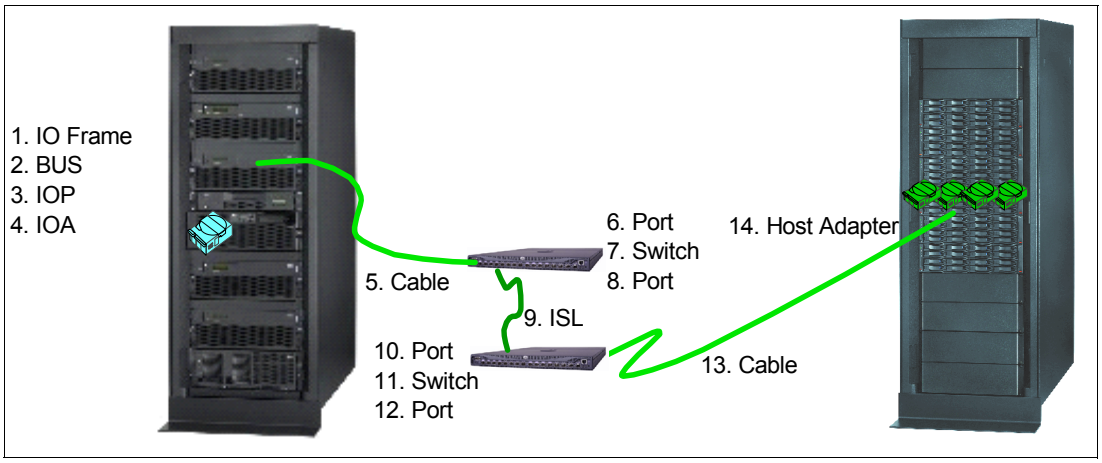



*Figure B-22 Single points of failure*

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 DS6000. Figure B-23 on page 344 shows this.

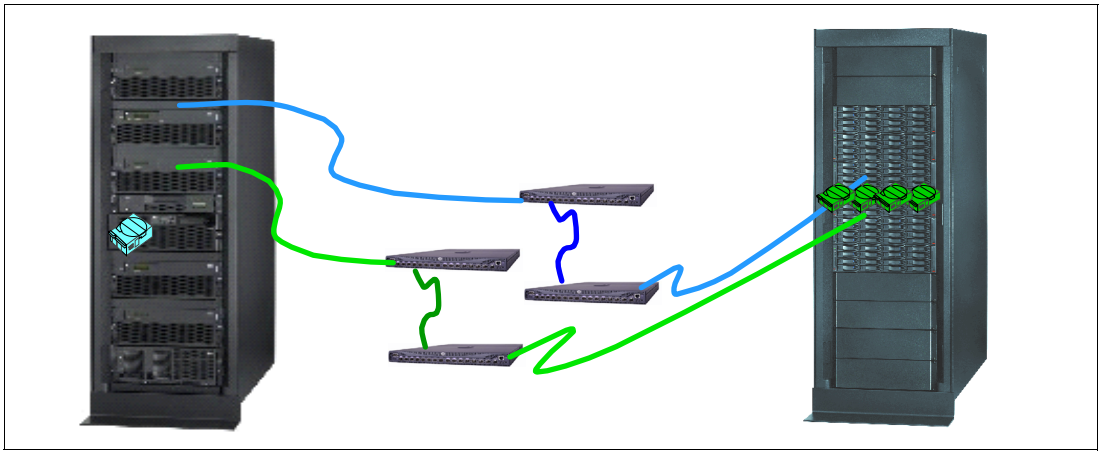


*Figure B-23 Multipath removes single points of failure*

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 DS6000:

- 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 353 for more information on sizing and performance guidelines.

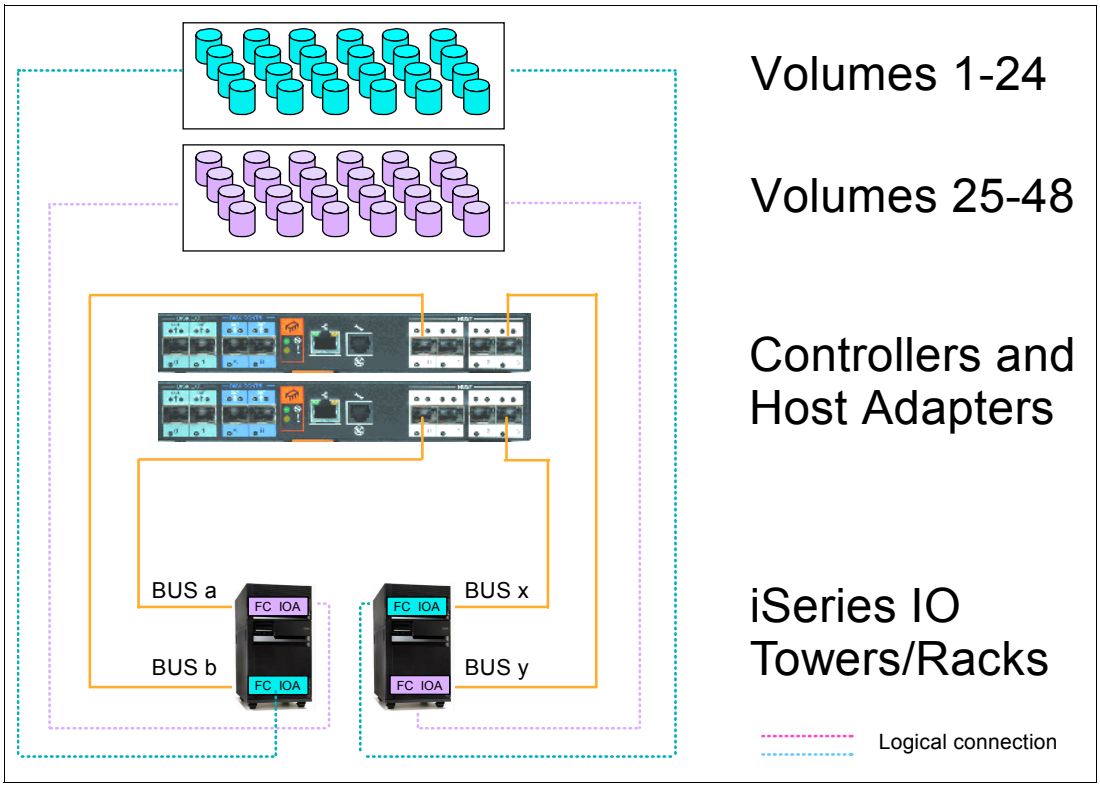



*Figure B-24 Example of multipath with iSeries*

Figure B-24 shows an example where 48 logical volumes are configured in the DS6000. The first 24 of these are assigned via a host adapter in the top controller card in the DS6000 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 controller card in the DS6000 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 controller card in the DS6000. 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 controller card.

## 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 332.

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 346. 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 1750-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  Serial                                Resource
  ASP    Number      Type  Model  Capacity    Name
         21-662C5    4326   050      35165    DD124
         21-54782    4326   050      35165    DD136
  1      75-1118707  1750   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 349.

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.

            Serial                  Resource
ASP  Unit   Number      Type Model  Name        Protection
  1                                             Unprotected
       1    02-89058    6717  074   DD004       Device Parity
       2    68-0CA4E32  6717  074   DD003       Device Parity
       3    68-0C9F8CA  6717  074   DD002       Device Parity
       4    68-0CA5D96  6717  074   DD001       Device Parity
       5    75-1118707  1750  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 334.

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**.

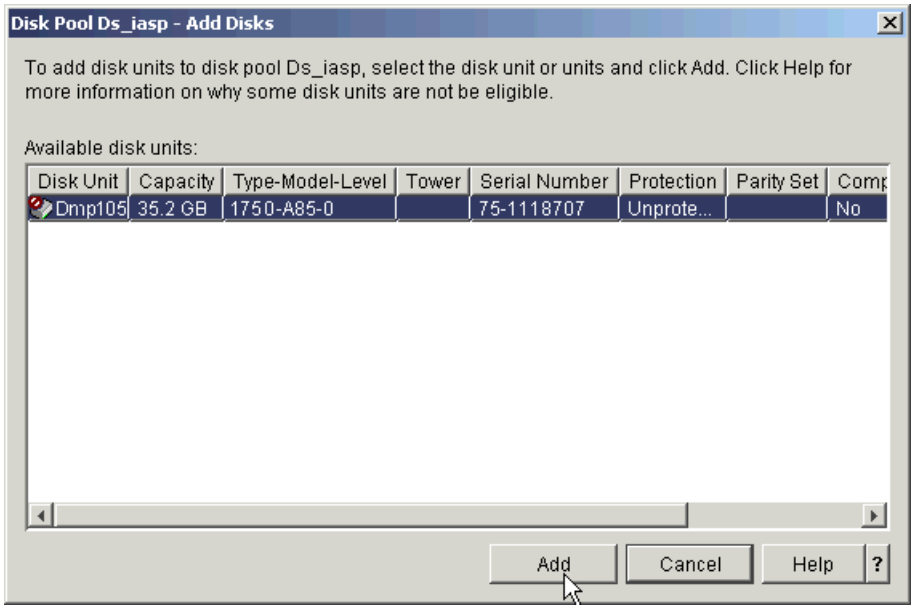



*Figure B-27 Adding a multipath volume*

**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 349.

The remaining steps are identical to those in “Adding volumes to an Independent Auxiliary Storage Pool” on page 334.

When you have completed these steps, the new Disk Pool can be seen on iSeries Navigator Disk Pools in Figure B-28.

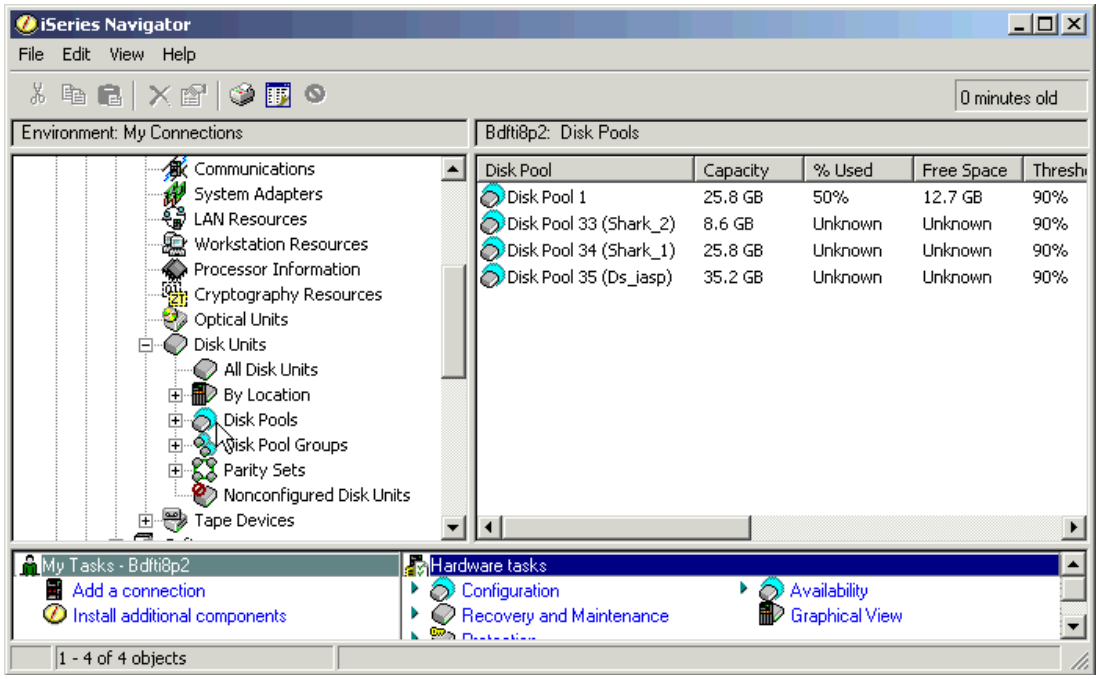



*Figure B-28 New Disk Pool shown on iSeries Navigator*

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.

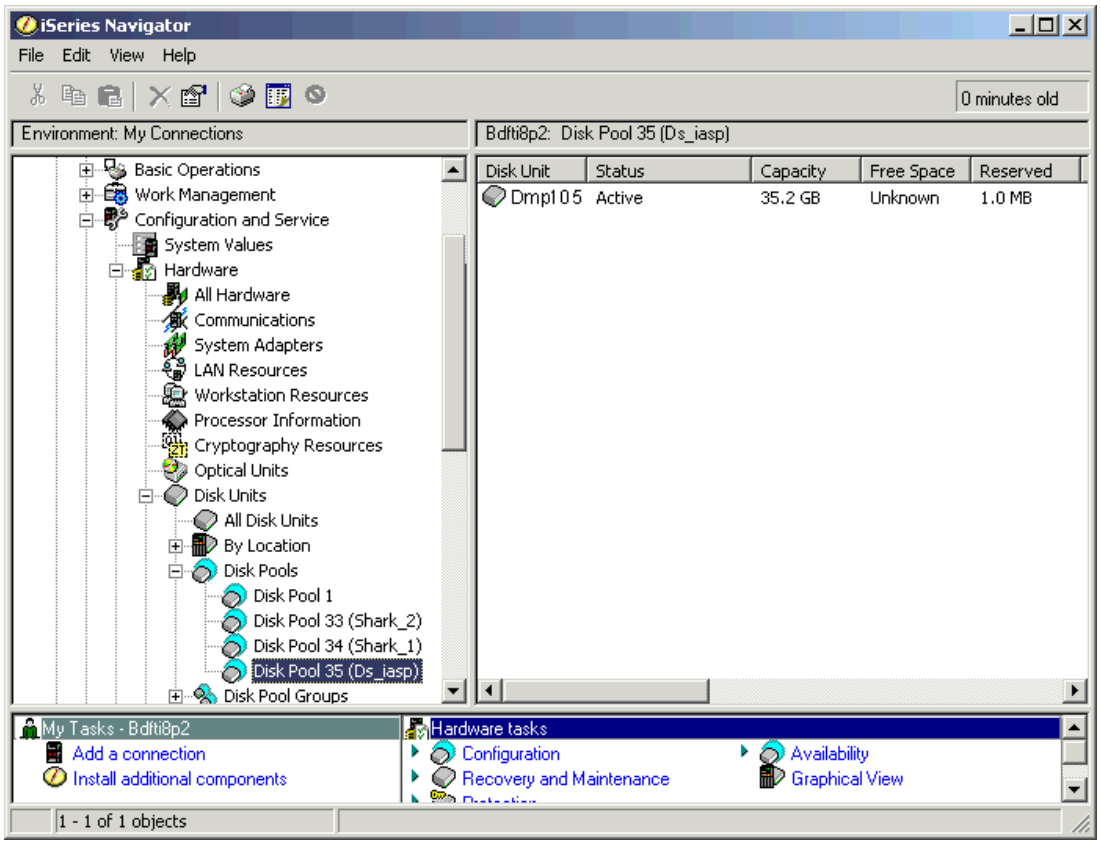



*Figure B-29 New logical volume shown on iSeries Navigator*

## 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.

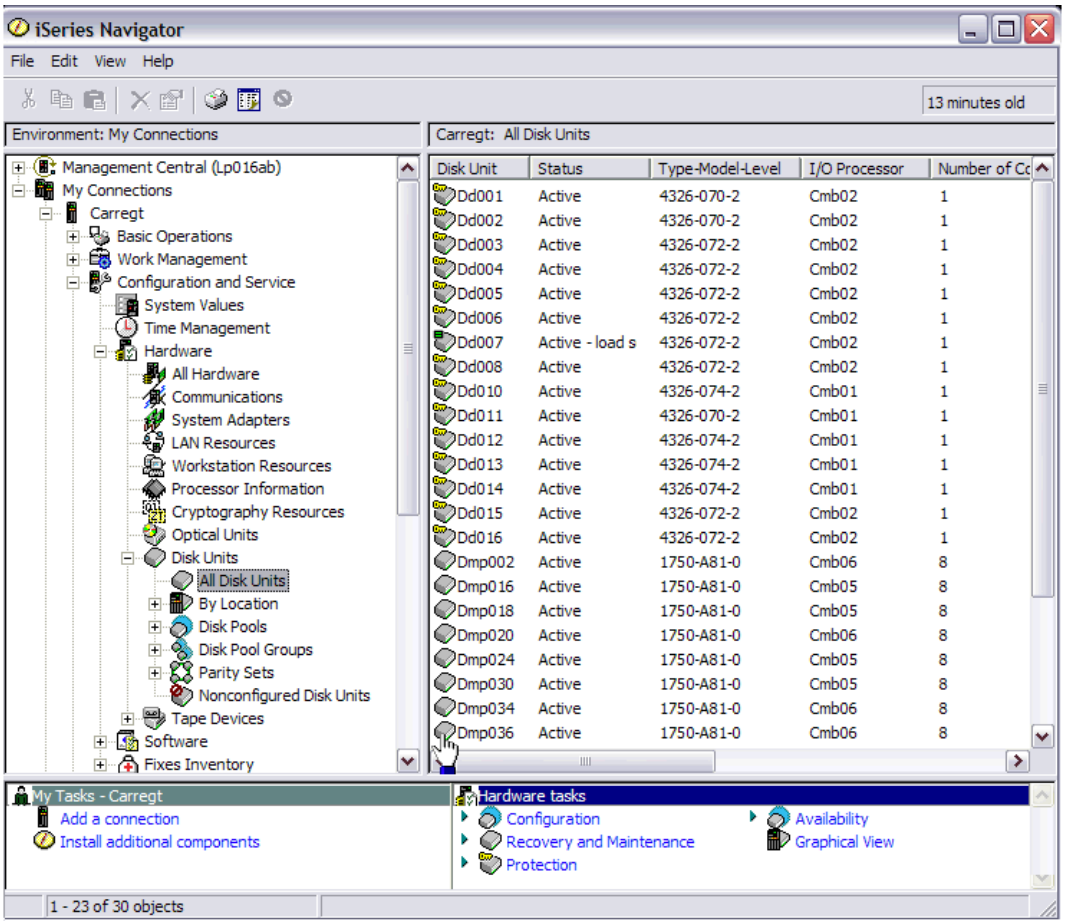



*Figure B-30 Example of multipath logical units*

To see the other connections to a logical unit, right click the unit and select **Properties**, as shown in Figure B-31 on page 350.

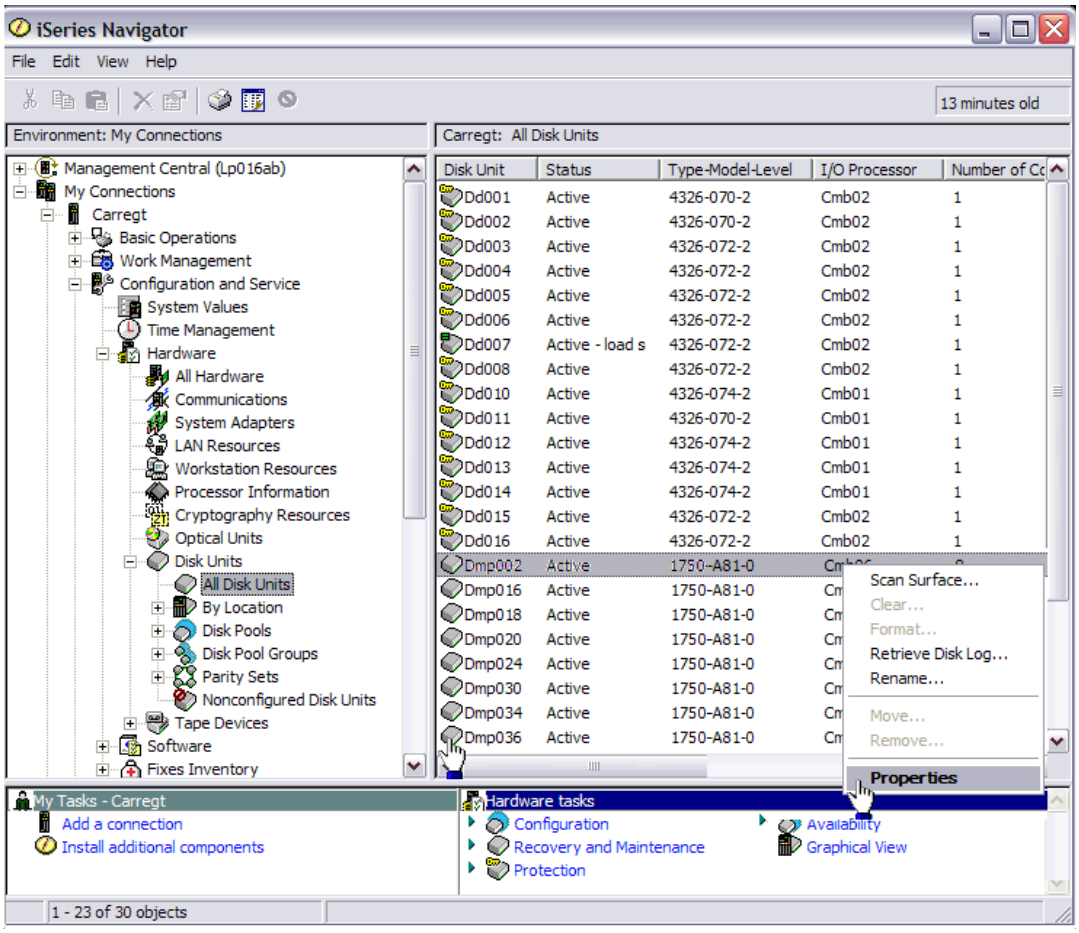



*Figure B-31 Selecting properties for a multipath logical unit*

You will then see the General Properties tab for the selected unit, as in Figure B-32. 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, where you can see the other seven connections for this logical unit.

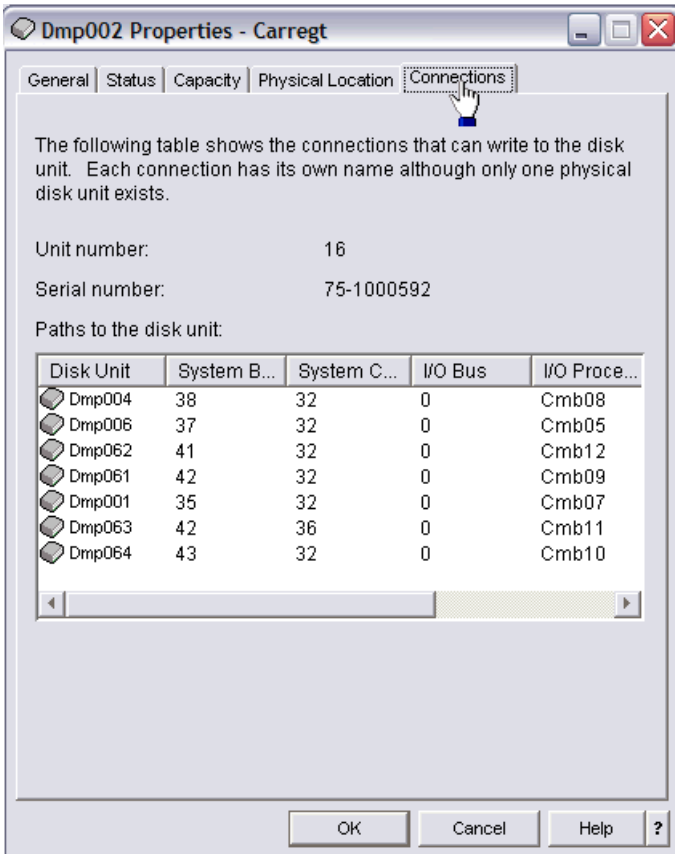



*Figure B-33 Multipath connections*

## 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 DS6000, 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 DS6000 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.

### Preferred path for DS6000

As discussed previously in “Avoiding single points of failure” on page 343, iSeries multipath can be implemented to allow a logical volume to be accessed via multiple connections. However, the *Preferred Path* facility for DS6000 discussed in Chapter 3, “RAS” on page 45, is currently not implemented in OS/400 and this may lead to some small performance degradation for large workloads. This is because a longer path will be used when accessing logical volumes across the midplane in the DS6000. Preferred path would access these volumes over the shortest path. However, as OS/400 does not support this, it uses a round-robin algorithm to spread I/O to a logical volume over all available paths.

## Sizing guidelines

In Figure B-34 on page 354, 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.

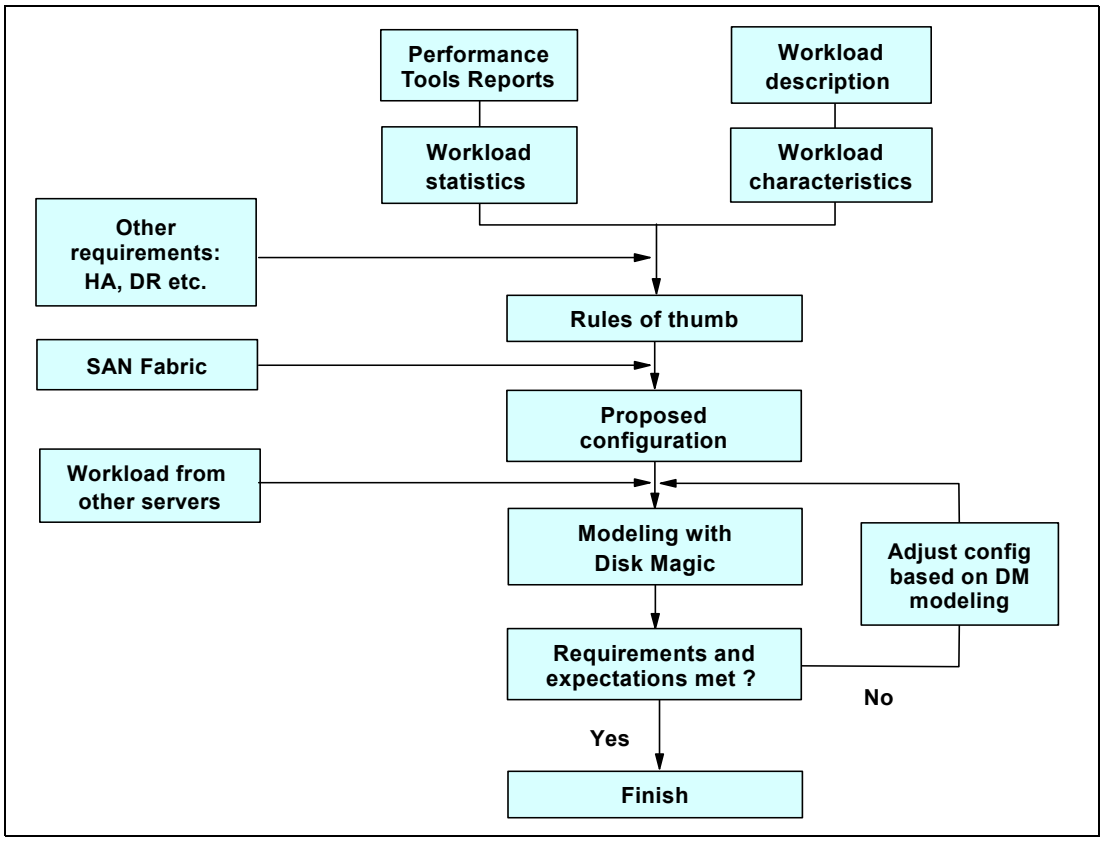



*Figure B-34 Process for sizing external storage*

## 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 per 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 356. However, the majority of iSeries workloads do not require this.

## Cache

In general, iSeries workloads do not benefit from large cache. Since there is no choice for cache size on DS6000, you should ensure that your workload is not excessively cache-friendly. This can be seen in OS/400 Performance Tools System, Component and Resource Interval reports. 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 DS6000. (The same guidelines are valid for ESS 800).

**Note:** Access Density is the capacity of occupied disk space divided by the average I/O per 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 330, 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 35.1GB 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 2TB 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 500GB 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 15.6GB LUNs per I/O adapter. However, since OS/400 only supports fixed volume sizes, we could support 28 17.5GB volumes to give us approximately 492GB per adapter.

## Recommended number of ranks

As a general guideline, you may consider 1500 disk operations/second 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 DS6000 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 DS6000 systems to iSeries via switches, you should plan that I/O traffic from multiple iSeries adapters can go through one port on a DS6000 and zone the switches accordingly. DS6000 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 DS6000.

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

# Migration

For many iSeries customers, migrating to the DS6000 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 DS6000 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 the DS6000 logical volumes.

Once an external logical volume has been created, it will always keep its model type and be either protected or unprotected. 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. See Table B-1 on page 330 for DS6000 logical volume models and types.

## 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 DS6000 (or indeed, any combination of external storage units which support Metro Mirror and Global Copy). For further discussion on Metro Mirror and Global Copy, see 15.2.2, “Subsystem-based data migration” on page 295. Consider the example shown in Figure B-35 on page 358.

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 DS6000.

When ready to migrate from the ESS to the DS6000, you should do a complete shutdown of the iSeries, unassign the ESS LUNs and assign the DS6000 LUNs to the iSeries. After IPLing the iSeries, the new DS6000 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 because this is an invalid configuration. Careful planning and implementation is required to ensure this does not happen, otherwise unpredictable results may occur.

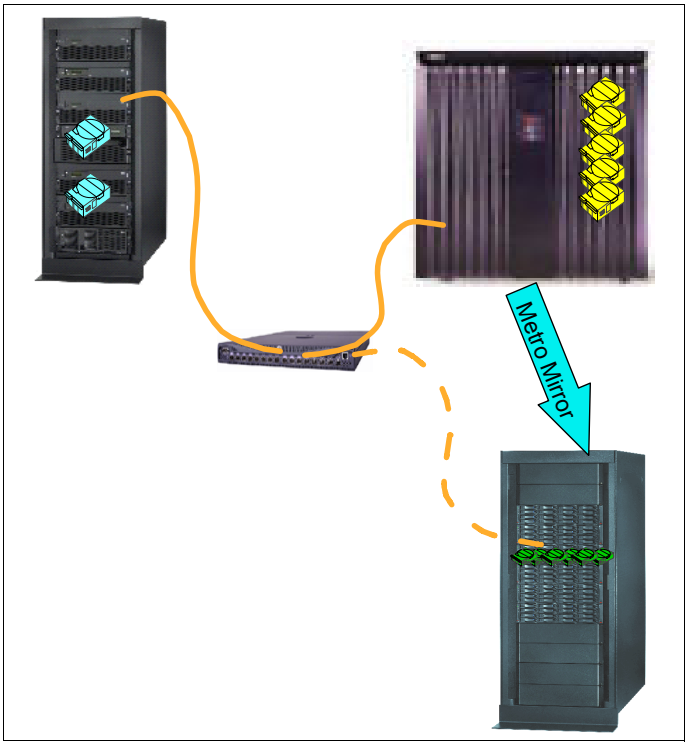



*Figure B-35 Using Metro Mirror to migrate from ESS to the DS6000*

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 DS6000 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 DS6000. 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 DS6000, whether the existing disks are internal or external. When you assign the new DS6000 logical volumes to the iSeries, initially they are non-configured (see “Adding volumes to iSeries configuration” on page 332 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. 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 DS6000.

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 `PWRDWNSYS` 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/iseries/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:rchclst@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 DS6000, 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/iseries/v1r2s/en_US/index.htm?info/iphat/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 DS6000 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 DS6000, 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/iseries/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/iseries/v1r2s/en_US/info/iphbi/iphbi.pdf

C

# 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 Services 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 255. Additional information about this offering is on the following Web site:

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 15.2.3, “IBM Piper migration” on page 297. For the latest information about this service, visit the following Web site:

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-1.ibm.com/services/us/index.wss/offering/its/a1000189

## 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

### 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 361.

## 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 15-9 on page 367 you can see an example for the Support Product List (SPL) of the IBM Support Line with the DS6800 and the DS8100 support highlighted. You get the complete SPL on the following Web site:

http://www.ibm.com/services/sl/products/java3.html

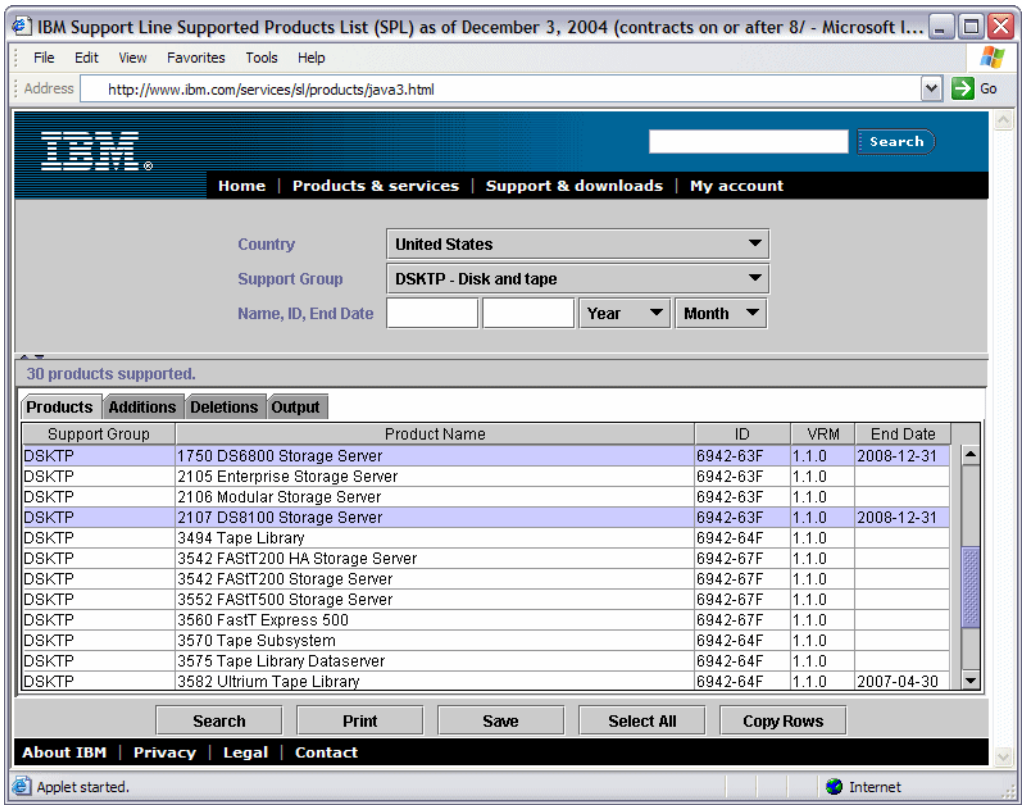



*Figure 15-9 Example of the Supported Product List (SPL) from the IBM Support Line*

# 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 371. Note that some of the documents referenced here may be available in softcopy only.

- *The IBM TotalStorage DS6000 Series: Implementation,* SG24-6781
- *The IBM TotalStorage DS6000 Series: Copy Services with IBM @server zSeries,* SG24-6782
- *The IBM TotalStorage DS6000 Series: Copy Services in Open Environments,* SG24-6783
- *The IBM TotalStorage DS6000 Series: Performance Monitoring and Tuning*, SG24-7145
- *IBM TotalStorage: Integration of the SAN Volume Controller, SAN Integration Server, and SAN File System,* SG24-6097
- *IBM TotalStorage: Introducing the SAN File System,* SG24-7057
- *The IBM TotalStorage Solutions Handbook*, SG24-5250
- *IBM TotalStorage Business Continuity Solutions Guide*, SG24-6547
- *iSeries and IBM TotalStorage: A Guide to Implementing External Disk on eServer i5*, SG24-7120
- *IBM TotalStorage Productivity Center V2.3: Getting Started*, SG24-6490
- *Managing Disk Subsystems using IBM TotalStorage Productivity Center for Disk*, SG24-7097
- *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

## 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 DS6000 Command-Line Interface User’s Guide,* SC26-7625
- *IBM TotalStorage DS6000: Host Systems Attachment Guide,* SC26-7628
- *IBM TotalStorage DS6000: Introduction and Planning Guide,* GC35-0495
- *IBM TotalStorage Multipath Subsystem Device Driver User’s Guide*, SC30-4096
- *IBM TotalStorage DS6000: User’s Guide*, SC26-7623
- *IBM TotalStorage DS Open Application Programming Interface Reference*, GC35-0493
- *IBM TotalStorage DS6000 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:

- Documentation for DS6800:
  http://www.ibm.com/servers/storage/support/disk/ds6800/
- SDD and Host Attachment scripts
  http://www.ibm.com/support/
- 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 DS6000:
  http://www.ibm.com/servers/storage/support/disk/1750.html
- Supported servers for the DS6000:
  http://www.storage.ibm.com/hardsoft/products/DS6000/supserver.htm
- The interoperability matrix:
  http://www.ibm.com/servers/storage/disk/DS6000/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
- 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

# Help from IBM

IBM Support and downloads

**ibm.com**/support

IBM Global Services

**ibm.com**/services

# Index

## E



## F



## G



## H



## I

## L



## M

## N



## O



## P



## R



## S

## T



## U



## V



## W



## X



## Z

# DS6000 Series: Concepts and Architecture

(0.5" spine)
0.475"<->0.873"
250 <-> 459 pages

# The IBM TotalStorage DS6000 Series: Concepts and Architecture

**Enterprise-class storage functions in a compact and modular design**

**On demand scalability and multi-platform connectivity**

**Enhanced configuration flexibility with virtualization**

This IBM Redbook describes the IBM TotalStorage DS6000 storage server series, its architecture, its logical design, hardware design and components, advanced functions, performance features, and specific characteristics. The information contained in this book is useful for those who need a general understanding of this powerful new disk subsystem, as well as for those looking for a more detailed understanding of how the DS6000 series is designed and operates.

The DS6000 series is a follow-on product of the IBM TotalStorage Enterprise Storage Server with new functions related to storage virtualization and flexibility.

The DS6000 series is a storage product targeted for the midrange market, but it has all the functions and availability features that normally can be found only in high end storage systems. In a very small enclosure, which fits in a standard 19-inch rack, the DS6000 series offers capacity, reliability functions, and performance similar to those of an ESS 800 or comparable high-end storage systems.

**INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION**

**BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE**

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:**
**ibm.com**/redbooks

SG24-6471-00 ISBN