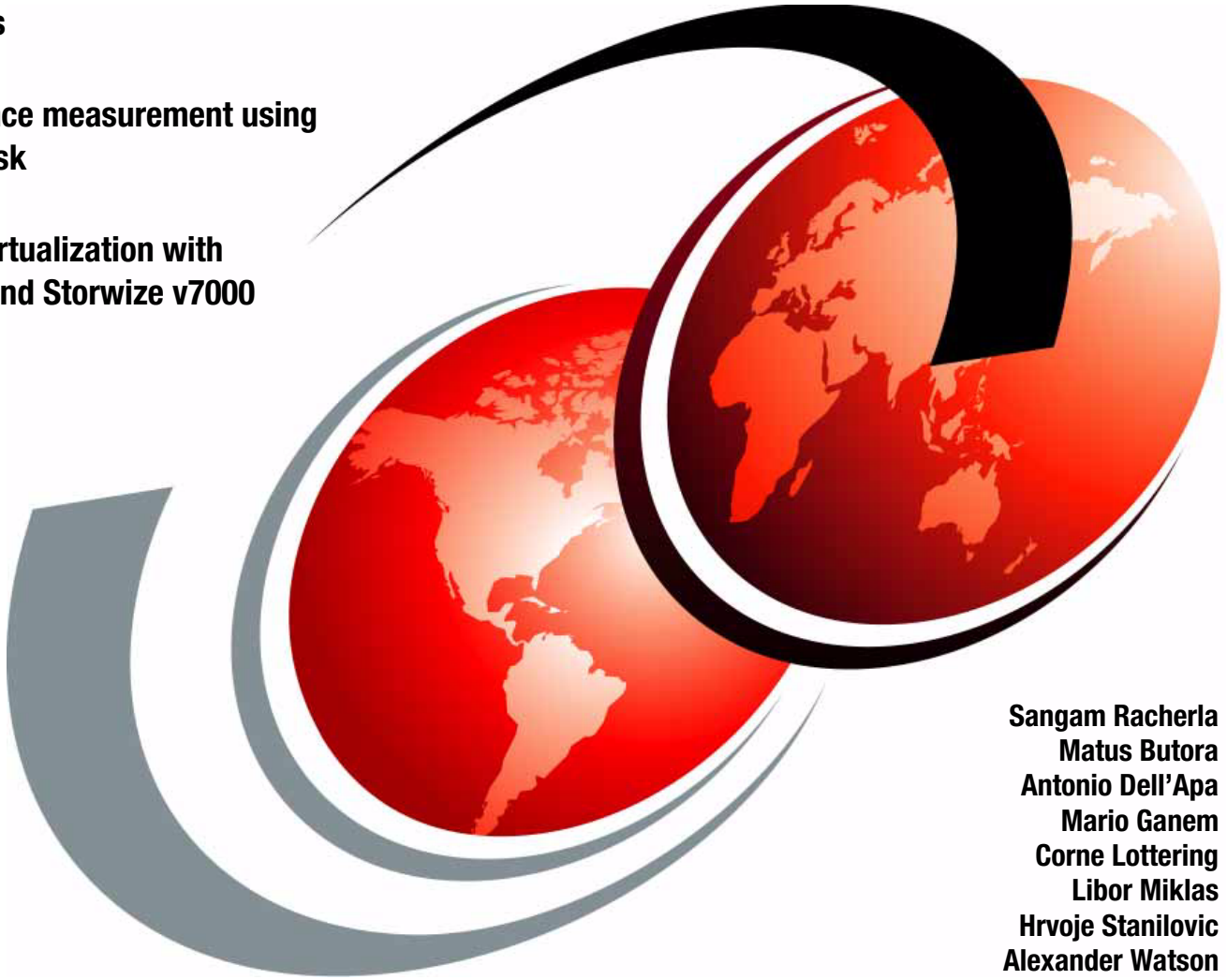
IBM

IBM System Storage DS5000 Series Implementation and Best Practices Guide

- Host configuration and performance tuning tips
- Performance measurement using TPC for Disk
- Storage virtualization with IBM SVC and Storwize v7000



Sangam Racherla
Matus Butora
Antonio Dell'Apa
Mario Ganem
Corne Lottering
Libor Miklas
Hrvoje Stanilovic
Alexander Watson

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM System Storage DS5000 Series Implementation and Best Practices Guide

December 2012

SG24-8024-00

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

First Edition (December 2012)

This edition applies to:
- IBM System Storage® DS5000 series running Firmware V7.77.
- IBM System Storage DS Storage Manager V10.77.

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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

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

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

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

COPYRIGHT LICENSE:

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

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

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

Active Memory™
AIX 5L™
AIX®
BladeCenter®
DB2®
DS4000®
DS6000™
DS8000®
Easy Tier®
Enterprise Storage Server®
ESCON®
eServer™
Express Storage™
FICON®
FlashCopy®
GPFS™
HACMP™
i5/OS™
IBM®
Informix®
iSeries®
Micro-Partitioning®
Netfinity®
Power Systems™
POWER6®
POWER7®
PowerHA®
PowerVM®
POWER®
pSeries®
Real-time Compression™
Redbooks®
Redpapers™
Redbooks (logo) ®
solidDB®
Storwize®
System i®
System p®
System Storage DS®
System Storage®
System x®
System z®
Tivoli®
WebSphere®
XIV®
z/OS®
zSeries®

The following terms are trademarks of other companies:

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

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

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

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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

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

Preface

This IBM® Redbooks® publication represents a compilation of best practices for deploying and configuring the IBM System Storage® DS5000 Series family of products. This book is intended for IBM technical professionals, Business Partners, and customers responsible for the planning, deployment, and maintenance of the IBM System Storage DS5000 Series family of products. We realize that setting up DS5000 Storage Servers can be a complex task. There is no single configuration that will be satisfactory for every application or situation.

First, we provide a conceptual framework for understanding the hardware in a Storage Area Network. Then we offer our guidelines, hints, and tips for the physical installation, cabling, and zoning, using the Storage Manager setup tasks. Next we provide a quick guide to help you install and configure the DS5000 using best practices.

After that, we turn our attention to the performance and tuning of various components and features, including numerous guidelines. We look at performance implications for various application products such as IBM DB2®,Oracle, IBM Tivoli® Storage Manager, Microsoft SQL server, and in particular, Microsoft Exchange server.

Then we review the various tools available to simulate workloads and to measure, collect, and analyze performance data. We also consider the IBM AIX® environment, including IBM High Availability Cluster Multiprocessing (IBM HACMP™) and IBM General Parallel File System (IBM GPFS™). This edition of the book also includes guidelines for managing and using the DS5000 with the IBM System Storage SAN Volume Controller (SVC) and IBM Storwize® V7000.

This book is designed specifically to help you with the implementation and best practice scenarios. It can be used in conjunction with these other IBM Redbooks publications:

- *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023
- *IBM System Storage DS Storage Manager Copy Services Guide*, SG24-7822

The team who wrote this book

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

Sangam Racherla is an IT Specialist and Project Leader working at the ITSO in San Jose, CA. He has 12 years of experience in the IT field and has been with the ITSO for the past eight years. Sangam has extensive experience in installing and supporting the ITSO lab equipment for various IBM Redbooks publications projects. He has expertise in working with Microsoft Windows, Linux, IBM AIX®, IBM System x®, and IBM System p® servers, and various SAN and storage products. Sangam holds a degree in Electronics and Communication Engineering.

Matus Butora is an IT Specialist and leader of Storage Support in the IBM ITD Delivery Center in the Czech Republic. He works with enterprise storage environments providing solutions and support for global strategic customers across the globe. Matus has eight years of experience with open storage hardware and software including IBM DS8000®, IBM Midrange Storage DS3000/DS4000/DS5000, SVC, NetApp, Tivoli Storage Management, and Tivoli Storage Productivity Center. Matus is a certified IBM Professional and NetApp NCDA certified Administrator.

Antonio Dell'Apa is an IBM Senior Accredited Product Service Specialist and a Team Leader in the MTS Technical Support Team in Rome. He joined IBM in 1989 and spent the first 10 years as a Customer Engineer for IBM iSeries®, IBM pSeries®, and IBM zSeries® product families. In 2002, he joined the Technical Support group for open systems storage, SAN, and virtualization products. Since 2006 he has been a member of the Virtual EMEA Team (VET) providing Level 2 support for DS3000, DS4000®, and DS5000 products to the EMEA region. During the past years, Antonio has also been in charge of deploying education and training to CE and FE specialists to maintain, service, and implement IBM storage products, such as the DS3000, DS5000, and IBM N series.

Mario Ganem is an IT professional, specialized in cloud computing and storage solutions. He has 15 years of experience in the IT industry. Mario resides in Buenos Aires, Argentina, where he works as an Infrastructure IT Architect in the Delivery Center in Argentina. Prior to starting his career in IBM in 2006, Mario worked in many companies such as Hewlett Packard, Compaq, and Unisys. He developed the internal virtualization products curriculum training, which he is currently teaching to DCA professionals. He holds, among others, many industry certifications from Microsoft, Red Hat, VMWare, Novell, Cisco, CompTIA, Hewlett Packard, and Compaq.

Corne Lottering is a Technical Storage Sales Specialist at Saudi Business Machines, the IBM General Marketing and Sales Representative in Saudi Arabia. His job includes customer assessment, planning, design, and delivery of IBM Storage Solutions involving IBM System Storage platforms including IBM San Volume Controller, IBM DS5000 Midrange storage, IBM XIV®, IBM DS8000 Enterprise Storage, and IBM N series. His previous experience includes working as a Systems Storage Sales Specialist in the IBM Sub Saharan Africa Growth Market Region for the Systems and Technology Group. His primary focus was Sales in the Central African countries, but he also provided pre-sales support to the Business Partner community. Corne has more than ten years of experience with IBM working with a wide variety of storage technologies including the DS4000, DS5000, DS8000, IBM XIV. IBM SAN switches, IBM Tape Systems, and storage software.

Libor Miklas is a Team Leader and an experienced IT Specialist working at the IBM Global Services Delivery Center in Czech Republic. He demonstrates ten years of practical experience in the IT industry. During the last six years, his main focus has been on backup and recovery and on storage management. He has already written other IBM Redbooks publications related to the IBM storage products and SAN. Libor and his team support midrange and enterprise storage environments for various global and local clients, worldwide. He is an IBM Certified Deployment Professional of the Tivoli Storage Manager family of products and holds a Masters degree in Electrical Engineering and Telecommunications.

Hrvoje Stanilovic is an IBM Certified Specialist - Midrange Storage Technical Support and Remote Support Engineer working for IBM Croatia. He is a member of the CEEMEA VFE Midrange Storage Support team and EMEA PFE Support team, providing Level 2 support for DS3000, DS4000, and DS5000 products in Europe, the Middle East, and Africa. His primary focus is post-sales Midrange Storage, SAN, and Storage Virtualization support, but he is also very active in supporting local projects, mentoring, and knowledge sharing. Hrvoje has been with IBM for four years, during which he has transitioned through various roles, including IBM System p hardware support and Cisco networking support, before working with Midrange Storage systems.

Alexander Watson is a Senior IT Specialist for Storage ATS Americas in the United States. He is a Subject Matter Expert on SAN switches and the DS4000/DS500 products. He has over ten years of experience in planning, managing, designing, implementing, problem analysis, and tuning of SAN environments. Alexander has worked at IBM for ten years. His areas of expertise include SAN fabric networking, open system storage I/O, and the IBM Midrange Storage Subsystems family of products.

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

Jon Tate
Bertrand Dufrasne
Ann Lund
Mary Lovelace
Alex Osuna
Karen Orlando
Larry Coyne

International Technical Support Organization, San Jose Center

Fred Scholten
Joseph F Bacco
Paul Goetz
Harold Pike
Danh Le
Noah J Seller
Reginald Phillips
John Sanner
Joyce Mercado
Barry Haddon
John Fasano
John Murtagh
Roger Bullard
Pete Urbisci
Gene Cullum
James Elliott
Bill Willson
Brenda Robinson
Rebecca C Swingler
Sharyn D Wolfe

IBM

John Bish
Doug Merrill
David Worley

NetApp

Brian Steffler
Steven Tong
Yong Choi

Brocade Communications Systems, Inc.

Thanks to the authors of the current and previous edition of the following Redbooks publications:

- *IBM Midrange System Storage Hardware Guide*, SG24-7676
- *IBM Midrange System Storage Implementation and Best Practices Guide*, SG24-6363
- *IBM System Storage DS4000 and Storage Manager V10.30*, SG24-7010
- *VMware Implementation with IBM System Storage DS4000/DS5000*, REDP-4609

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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 books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

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

  ibm.com/redbooks
- Send your comments in an email to:

  redbooks@us.ibm.com
- Mail your comments to:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:

  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction to IBM Midrange System Storage and SAN

In this chapter, we introduce IBM Midrange System Storage products with a brief description of the various models, their features, and their position on the market and in the concrete small to medium business solutions. In addition, we summarize the functions and benefits of the DS Storage Manager software.

But first, let us look at basic concepts and topologies of Storage Area Network (SAN) and System Storage Networking.

Readers already familiar with the IBM System Storage DS5000 Series product line and SAN concepts can skip this chapter.

1.1 Introduction to SAN

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

A SAN is a high-speed network that enables the establishment of switched, routed, or direct connections between storage devices and hosts (servers) within the specific distance supported by the designed environment. At the basic level, the SAN is a Fibre Channel (FC) network; however, new technology now enables this network to be routed or tunneled over many other networks as well. Typical example is lossless TCP/IP network protocol, that can accommodate FC traffic using iSCSI or Fibre Channel over Ethernet (FCoE).

The SAN can be viewed as an extension of the storage bus concept, which enables storage devices to be interconnected using concepts similar to that of local area networks (LANs) and wide area networks (WANs). A SAN connection to specific storage device can be shared across multiple host servers or dedicated to one single server. We are talking about SAN zoning. It can be local or extended over geographical distances.

Figure 1-1 shows a brief overview of a consolidated SAN network using multiple protocols and interconnecting multiple servers to various storage systems (disks, tapes). The example given here represents the enterprise datacenter using core Fibre Channel connections and remote branch office with midrange SAN components. Different communication protocols are used to consolidate main DC with remote offices into single, easily manageable and storage efficient SAN network.

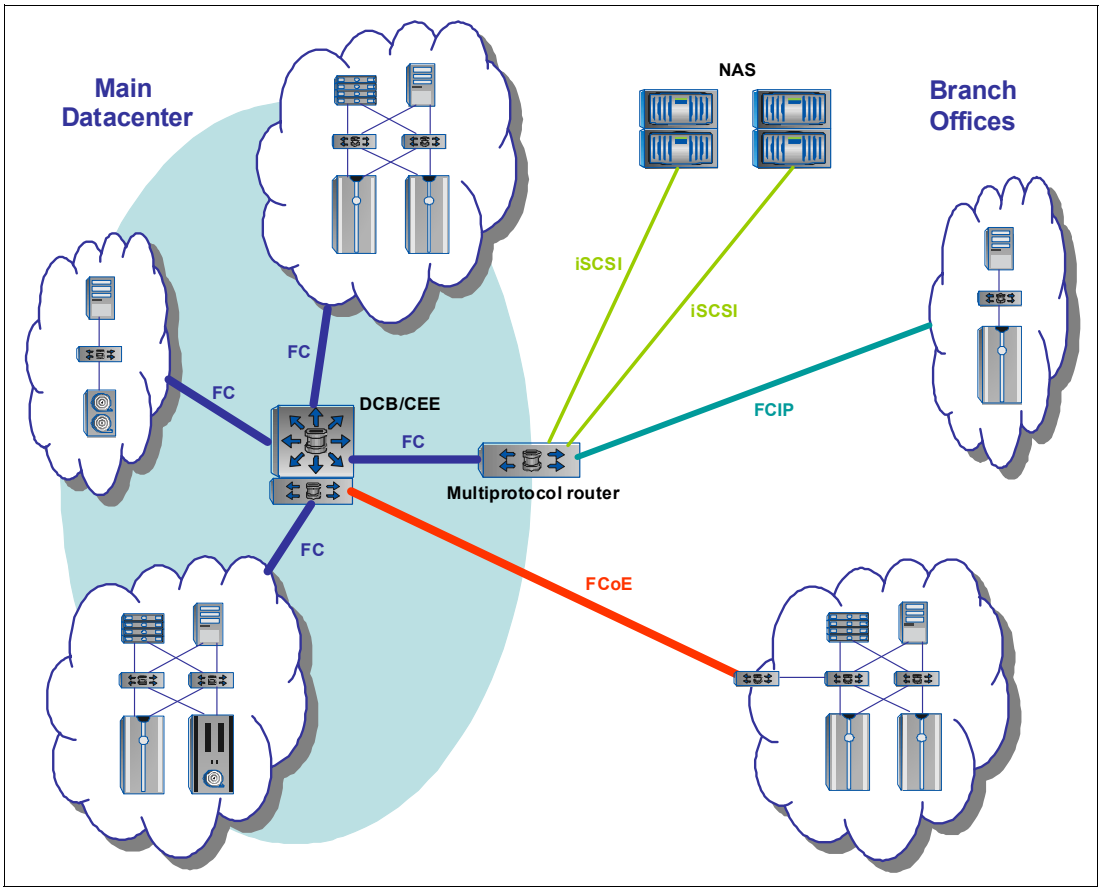


Figure 1-1 Example of consolidated SAN environment

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

1.1.1 Storage networking protocols

In this section, we explain communication protocols and technologies used for storage networking.

Fibre Channel

Today, Fibre Channel (FC) is the architecture on which most SAN implementations are built. Fibre Channel is a technology standard that enables data to be transferred from one network node to another at very high speeds. Current implementations transfer data at 1 Gbps, 2 Gbps, 4 Gbps, 8 Gbps, and 16 Gbps. At the moment there are limited FC Host Bus Adapters (HBA) for host connections at 16 Gbps, thus this speed is eligible for Inter-switch Links (ISL) between SAN directors, switches, and routers.

Fibre Channel architecture is sometimes referred to as the fibre version of SCSI. Fibre Channel is an architecture that can carry Intelligent Peripheral Interface (IPI) traffic, IP traffic, IBM FICON® traffic, FCP (SCSI) traffic, and possibly traffic using other protocols, all on the standard FC transport.

iSCSI

Internet SCSI (iSCSI) is a transport protocol that carries SCSI commands from an initiator to a target. It is a data storage networking protocol that transports standard Small Computer System Interface (SCSI) requests over the standard Transmission Control Protocol/Internet Protocol (TCP/IP) networking technology.

iSCSI enables the implementation of IP-based storage area networks (SANs), enabling customers to use the same networking technologies — for both storage and data networks. As it uses TCP/IP, iSCSI is also well suited to run over almost any physical network. By eliminating the need for a second network technology just for storage, iSCSI has the potential to lower the costs of deploying storage networks.

FCP

The Fibre Channel Protocol (FCP) is the interface protocol of SCSI on Fibre Channel. It is a gigabit speed network technology primarily used for Storage Networking. Fibre Channel is standardized in the T11 Technical Committee of the InterNational Committee for Information Technology Standards (INCITS), an American National Standard Institute (ANSI) accredited standards committee. It started for use primarily in the supercomputer field, but has become the standard connection type for storage area networks in enterprise storage. Despite its name, Fibre Channel signaling can run on both twisted-pair copper wire and fiber optic cables.

FCIP

Fibre Channel over IP (FCIP) is also known as Fibre Channel tunneling or storage tunneling. It is a method to allow the transmission of Fibre Channel information to be tunnelled through the IP network. Because most organizations already have an existing IP infrastructure, the attraction of being able to link geographically dispersed SANs, at a relatively low cost, is enormous.

FCIP encapsulates Fibre Channel block data and subsequently transports it over a TCP socket. TCP/IP services are utilized to establish connectivity between remote SANs. Any congestion control and management, as well as data error and data loss recovery, is handled by TCP/IP services, and does not affect FC fabric services.

The major point with FCIP is that is does not replace FC with IP, it simply allows deployments of FC fabrics using IP tunnelling. The assumption that this might lead to is that the “industry” has decided that FC-based SANs are more than appropriate, and that the only need for the IP connection is to facilitate any distance requirement that is beyond the current scope of an FCP SAN.

FICON

FICON architecture is an enhancement of, rather than a replacement for, the now relatively old IBM ESCON® architecture. As a SAN is Fibre Channel based, FICON is a prerequisite for IBM z/OS® systems to fully participate in a heterogeneous SAN, where the SAN switch devices allow the mixture of open systems and mainframe traffic.

FICON is a protocol that uses Fibre Channel as its physical medium. FICON channels are capable of data rates up to 200 MBps full duplex, they extend the channel distance (up to 100 km), increase the number of control unit images per link, increase the number of device addresses per control unit link, and retain the topology and switch management characteristics of ESCON.

Fibre Channel over Ethernet

Fibre Channel over Ethernet (FCoE) is an encapsulation of Fibre Channel frames into the lossless 10 Gbps Ethernet networks. FCoE maps Fibre Channel directly over Ethernet while being independent of the Ethernet forwarding scheme. The greatest benefit of using FCoE is significant reduction of networking ports in the datacenter and associated cabling, reduced power and cooling requirements, and finally simplified operation and maintenance of one common network with minimal impact to the existing procedures and processes.

1.1.2 SAN components

In this section, we provide a basic description of the SAN storage components and building blocks as shown in Figure 1-2, and as we further explain in the following text.

Host servers

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

SAN storage subsystems

The storage infrastructure is the foundation on which information relies, and therefore, must support a company’s business objectives and business model. In this environment, simply deploying more and faster storage devices is not enough. A SAN infrastructure provides enhanced availability, performance, scaleability, data accessibility, and system manageability.

It is important to remember that a good SAN begins with a good design. The SAN liberates the storage device, so it is not on a particular server bus, and attaches it directly to the network. In other words, storage is externalized and can be functionally distributed across the organization. The SAN also enables the centralization of storage devices and the clustering of servers, which has the potential to allow easier and less expensive centralized administration that lowers the total cost of ownership (TCO).

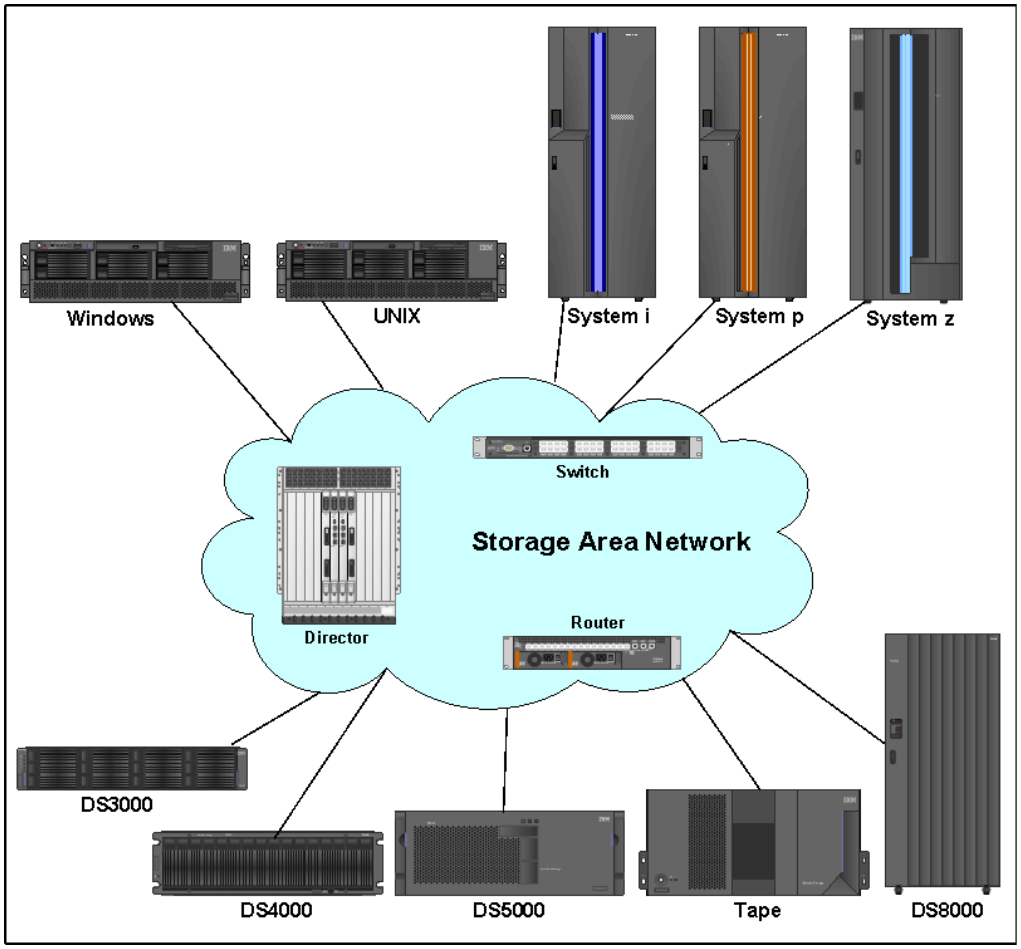


Figure 1-2 SAN components

SAN topologies

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

- Point-to-point: The point-to-point topology consists of a single connection between two nodes. All the bandwidth is dedicated to these two nodes.
- Loop: In the loop topology, the bandwidth is shared between all the nodes connected to the loop. The loop can be wired node-to-node, and is how the DS5000 subsystems perform their direct connection to single host without switch attachment; however, if a node fails or is not powered on, the loop is out of operation, which can be overcome by using a hub. A hub opens the loop when a new node is connected, and closes it when a node disconnects.
- Switched or fabric: A switch enables multiple concurrent connections between nodes. There are two types of switches: circuit switches and frame switches. Circuit switches establish a dedicated connection between two nodes, whereas frame switches route frames between nodes and establish the connection only when needed, which is also known as switched fabric.

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

SAN interconnections

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

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

Inter-Switch Links (ISLs)

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

Trunking

Inter-Switch Links can be combined into logical groups to form trunks. In IBM TotalStorage switches, trunks can be groups of up to four ports on a switch connected to four ports on a second switch. At the outset, a trunk master is defined, and subsequent trunk slaves can be added, which has the effect of aggregating the throughput across all links. Therefore, in the case of switches with 8 Gbps ports, if we trunk up to four ports, we allow for a 32 Gbps Inter-Switch Link.

1.1.3 SAN zoning

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

Typically, you can use zones to do the following tasks:

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

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

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

Zone types

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

Port level zone A zone containing members specified by switch ports (domain ID, port number) only. Port level zoning is enforced by hardware in the switch.

WWPN zone A zone containing members specified by device World Wide Port Name (WWPN) only. WWPN zones are hardware enforced in the switch.

Mixed zone A zone containing various members specified by WWPN and certain members specified by switch port. Mixed zones are software enforced through the fabric name server.

Zones can be hardware enforced or software enforced:

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

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

For more complete information regarding Storage Area Networks, see the following Redbooks publications:

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

Zoning configuration

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

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

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

Keep in mind that you will need to update the zoning information, if you ever need to replace a Fibre Channel adapter in one of your servers. Most storage systems such as the DS5000, Enterprise Storage Subsystem, and IBM Tape Libraries have a WWN tied to the Vital Product Data of the system unit, so individual parts can usually be replaced with no effect on zoning.

For more details on configuring zoning with your particular switch, see *Implementing an IBM/Brocade SAN with 8 Gbps Directors and Switches*, SG24-6116 or other vendor specific documentation.

Multiple fabrics

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

- Each server can have two or more HBAs. In a two-HBA configuration, each HBA can connect to a separate fabric.
- Each DS5000 can use separate host ports or mini hubs to connect to multiple fabrics, thus giving a presence in each fabric.
- Zoning in each fabric means that the server can have many paths to its resources, which also means that the zoning needs to be done in each fabric separately.
- The complete loss of a fabric means that the host can still access the resources through the other fabric. The multiple fabric increases the complexity, resiliency, and redundancy of the SAN infrastructure. This, however, comes at a larger cost due to the duplication of switches, HBAs, zoning administration, and fiber connections. This trade-off needs to be carefully examined to see whether your SAN infrastructure requirements require multiple fabrics.

Virtual Fabrics

The Virtual Fabrics (VSANs) feature uses the proven security and fault isolation features of single physical fabric, enabling clients to create logical groups of separately managed devices, ports, and switches within a physical SAN infrastructure. This deployment is an appreciated benefit in large datacenter solutions with multiple clients sharing the same physical infrastructure. Change management processes (storage provisioning, software upgrades, and so on) are dedicated to each of the clients, to each single VSAN. It means that changes made in one VSAN do not impact different VSANs and different clients, respectively.

1.2 Position of the DS5000 family

IBM has brought together into one family, known as the DS family, a broad range of disk systems to help small, medium, and enterprise environments choose the right solution for their rapidly growing storage needs. The IBM DS family combines the high performance IBM System Storage DS8000 series of enterprise storage device subsystems that inherit from the industry proven IBM Enterprise Storage Server® (ESS), with the DS5000 series of midrange systems, and entry-level DS3000 family of products.

Apart of the DS family members, the IBM offers wider portfolio of the storage systems for medium to enterprise business solutions, that enable datacenter virtualization and cloud computing, two components of IBM Smarter Datacenters with efficient storage provisioning. These products include IBM Storwize V7000 for small to medium businesses and IBM XIV System Storage for enterprise solutions. The specific category of storage products is Network Attached Storage (NAS), that provides a wide range of network attachment capabilities to a broad range of host and client systems. It includes IBM System Storage N series, IBM Scale Out Network Attached Storage (SONAS), and IBM Real-time Compression™ Appliance (RtCA).

Throughout this publication, we describe in detail, best practices and typical implementation scenarios of IBM Midrange storage systems, particularly the DS5000 family of products. They address the continuously growing needs for storage capacities within small to medium business solutions. Figure 1-3 explains the position of the DS5000 products on the market.

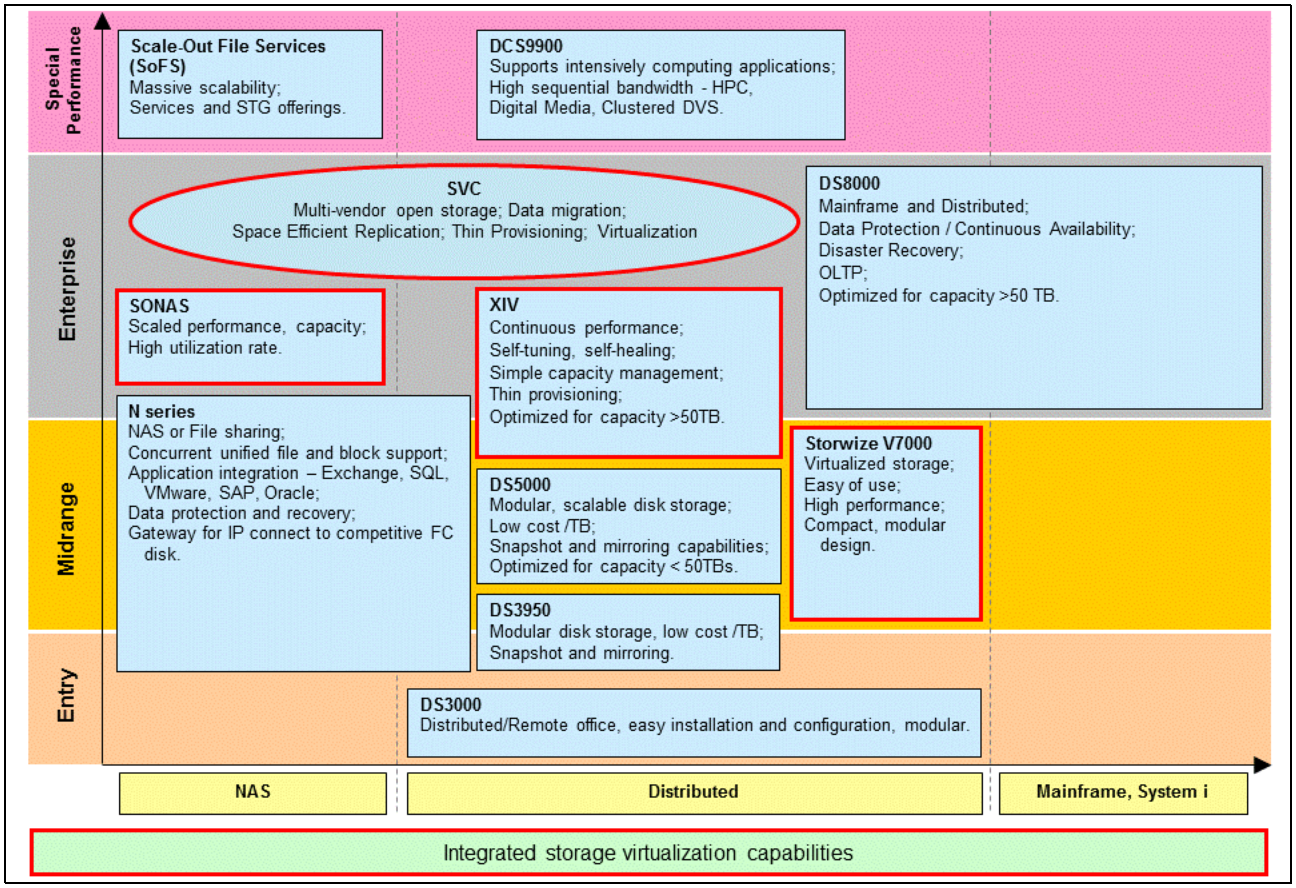


Figure 1-3 Position of IBM storage systems on the market

The IBM DS5000 series of subsystems support various types of disk drives. Beginning with the cost-effective Serial Advanced Technology Attachment (SATA) drives, through the 6 Gbps Serial Attached SCSI (SAS), 4 Gbps Fiber Channel (FC) disks, FC-SAS, up to Solid State Drives (SSD disks). The IBM Midrange disk storage systems offer the enhanced data security with Self Encrypting Disks (SED) - 4 Gbps Fiber Channel Full Disk Encryption (FDE) drives.

The DS5000s use Redundant Array of Independent Disks (RAID). RAID technology offers various levels of performance and protection for the user data from disk drive failures.

The IBM DS5000 storage subsystem offers 8 Gbps Fibre Channel (FC) interfaces to connect the host systems and external disk drive enclosures. Additionally, there is also a 10 Gbps iSCSI interface available for host attachment.

1.3 DS5000 features and family members

The DS5000 series provide high system availability using the hot-swappable and redundant components. It is crucial functionality, when the storage subsystem is placed in business critical client environments or connected to the servers consolidated in Storage Area Networks (SAN) using one of the available virtualization technics. The current models also offer a number of advanced features and functions that can be implemented dynamically without stopping normal operations:

- Dynamic Capacity Expansion: Allows for adding additional drives to an array group. Automatically re-stripes the LUNs to make use of the additional drive resources immediately.
- Dynamic Volume Expansion: Allows for increasing the size of a specific LUN which is already defined and in use. The additional space will be used when the host allows it to be recognized.
- Dynamic Segment Size: Allows for better handling of the host IO when alignment and IO block size issues are encountered.
- Dynamic Cache Block Size: Allows for dynamic change to be made to the selected cache block size to better handle host IO block size with minimal management.
- Dynamic RAID type: Allows for RAID type to be changed dynamically from one RAID type to another to improve performance and availability as needed.

Many of these features can be enabled together to resolve configuration based issues discovered after implementing the storage into production; or in cases where growth has exceeded the overall expectations.

1.3.1 Available DS5000 models

The current DS5000 series consists of the following models:

- IBM System Storage DS5020 Storage Subsystem:

  The DS5020 (machine type 1814-20A) is the newest member of the DS5000 series. It is designed to help address the storage requirements of small to medium business environments. The DS5020 is a 3U rack-mountable enclosure that delivers high performance, advanced functions, and high availability, as well as modular and scalable storage capacity. It supports RAID levels 0, 1, 3, 5, and 6 up to over 67 TB when using 600 GB FC, or 100 TB when using 900 GB FC-SAS hard drives, 44.8 TB when using 400 GB SSD drives, and up to 224 TB when using 2 TB SATA Enhanced Disk Drive Modules (E-DDM). The maximum number of storage partitions is 128.

The DS5020 houses redundant, dual-active RAID controllers with either two Fibre Channel ports, four Fibre Channel ports, or two Fibre Channel and two iSCSI ports per controller. The DS5020 can be configured for the attachment of host servers and up to 6 EXP520 and EXP810 storage expansion enclosures. It can be ordered in configuration with 1 or 2 GB memory in each controller. Figure 1-4 shows the front view of the DS5020 base enclosure.

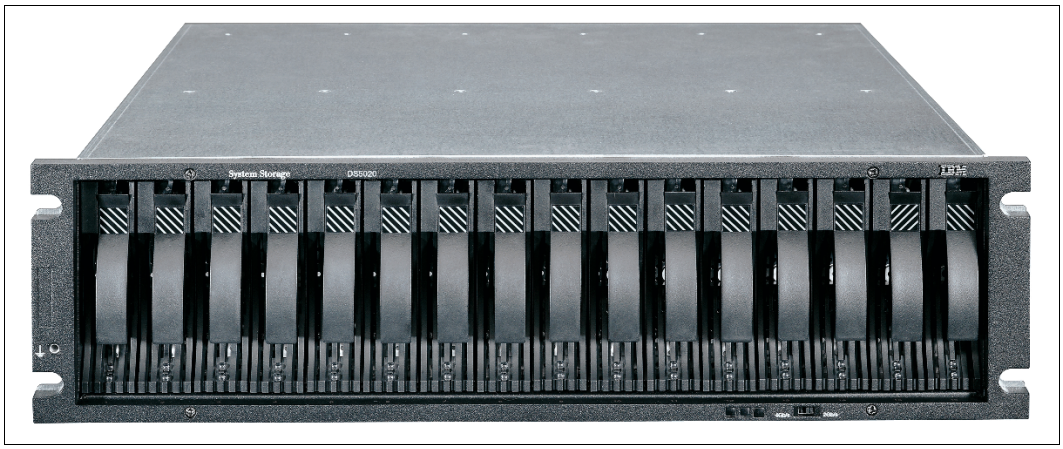


Figure 1-4 Front view of IBM System Storage DS5020 Storage Subsystem

- IBM DS5100 Storage Subsystem:

  The DS5100 (machine type 1818-51A) supports the large and growing data storage requirements of business critical applications. The storage subsystem offers you data access and protection to meet your existing enterprise storage requirements and prepare for the future.

  This storage subsystem is a 4U rack-mountable enclosure, has sixteen 4 Gbps FC drive interfaces, and can hold a maximum of twenty eight EXP5000 expansion units, or up to twenty eight expansion units composed of a mixture of EXP5000 and EXP810 (for migration purposes) for a total of up to 448 disk drives. With 8 high density EXP5060 expansion units and SATA drives it scales up to 480 disks showing the overall capacity of 1440 TB. DS5100 supports the configurations with intermixing drive types: FC, FC-SAS, SED, SATA and SSD. The system incorporates 4 GB memory cache per controller, upgradable up to 32 GB per controller unit.

  The DS5100 storage systems are equally adept at supporting transactional applications such as databases and On Line Transaction Processing (OLTP), throughput-intensive applications such as high performance computing (HPC) and rich media, and concurrent workloads for consolidation and virtualization. Figure 1-5 shows the front view of DS5100.

Figure 1-5 Front view of IBM System Storage DS5100 Storage Subsystem

- IBM DS5300 Storage Subsystem:

  The DS5300 (machine type 1818-53A) server is primarily the same machine as DS5100. It consists of the same components, but has greater memory cache in base configuration with the size of 8 GB per controller, similarly upgradable up to 32 GB per controller. It also demonstrates greater performance of internal processor bus. As same as DS5100 it supports up to 512 partitions of storage array.

- IBM DS3950 Express Storage™ Subsystem:

  At first sight, one can argue that the DS3950 does not fall into the category of DS5000 family members, and looking at the naming convention of the machine itself, that might be obviously right. On the other hand, the DS3950 (machine type 1814 models 94A and 98A) is conceptually similar to the DS5020 Storage Subsystem, it is positioned as a midrange storage product, and that is being proven even by machine type number. Because of that, we decided not to include any further details about this product in this publication, but rather to focus on DS5000 devices. Readers interested in the DS3950 Express model can follow the information in *IBM System Storage DS3950 Introduction*, REDP-4702.

At a glance, the storage products of the DS5000 family introduce the modular design that avoids over-configuration for an affordable price, while offering seamless “pay-as-you-grow” scalability as required by growing needs. Its efficient utilization lowers the raw capacity requirements, and support for intermixing high performance and high capacity drives enables tiered storage in your environment. With the Enhanced Remote Mirroring and IBM FlashCopy® functions, it uses the Business Continuity and Disaster Recovery protection in small, medium, and enterprise business deployments. The comprehensive information about IBM Midrange storage products is available at the following website:

`http://www.ibm.com/systems/storage/disk/midrange/`

1.3.2 Host connectivity options

Each model of DS5000 supports the host connectivity through the following interfaces:

- IBM System Storage DS5020 Storage Subsystem:
  - Up to four standard 8 or 4 Gbps FC host ports
  - Optional two 1 Gbps iSCSI host ports

- IBM System Storage DS5100 Storage Subsystem:
  - Up to eight standard 8 or 4 Gbps FC host ports
  - Up to four 10 or 1 Gbps iSCSI ports
  - The optional combination of FC and iSCSI ports
- IBM System Storage DS5300 Storage Subsystem:
  - Up to eight standard 8 or 4 Gbps FC host ports
  - Up to four 10 or 1 Gbps iSCSI ports
  - The optional combination of FC and iSCSI ports

The DS5000 devices have two slots per controller for Host Interface Cards (HIC). When combining multiple HICs in one system, one needs to consider the following conditions:

- The 1 Gbps and the 10 Gbps HIC should not be installed on the same storage subsystem.
- Each controller must have the same type HICs in identical slot positions. For example, if controller A has 4 Gbps FC and 10 Gbps iSCSI HICs in HIC slots 1 and 2, respectively, controller B should also have 4 Gbps FC and 10 Gbps iSCSI HICs in HIC slots 1 and 2, respectively.

1.4 Expansion enclosures

At the time of writing, the DS5000 series expansion enclosure offers a 4 Gbps FC interface. Four models are available:

- EXP810 Expansion Enclosure:

  This expansion unit (machine type 1812-81A) is packaged in a 3U rack-mountable enclosure, and supports up to 16 FC disk drives or E-DMM SATA drives. It contains 16 drive bays, dual-switched 4 Gbps ESMs, and dual power supplies and cooling components. Fully populated with 600 GB FC disk drive modules, this enclosure offers up to 9.6 TB of raw storage capacity or up to 32 TB when populated with 2-TB E-DDM SATA drives. Through the proper firmware level, this expansion unit is able to host both FC and SATA drives. Intermix of FC and SATA drives is supported within this expansion enclosure.

  Hint: The EXP810 expansion unit is the only one that can be connected to every storage subsystem of the withdrawn DS4000 family and currently available DS5000 products. Therefore it is a good candidate for data migration purposes to new midrange storage systems.

- EXP5000 Expansion Enclosure:

  This 3U rack-mountable expansion enclosure (machine type 1818-D1A) supports up to 16 FC disk drives, E-DMM SATA drives, Full Disk Encryption (FDE) drives, and up to 20 SSDs per subsystem. It contains 16 drive bays, dual-switched 4 Gbps ESMs, and dual power supplies and cooling components. Fully populated with 900 GB FC-SAS disk drive modules, this enclosure offers up to 14.4 TB of raw storage capacity or up to 32 TB when populated with the 2 TB E-DDM SATA drives.

  The EXP5000 expansion unit can be connected to the DS5100 or DS5300 storage subsystems only. Through the proper firmware level, this expansion unit is able to host both FDE, FC, SATA, FC-SAS drives, and SSD as well. Intermix of these drives is supported within a single expansion enclosure with limitations as described further in 2.2.3, “Disk intermix capability” on page 24. Or, for even more detail, see *IBM System Storage DS4000/DS5000 Fibre Channel and Serial ATA Intermix Premium Feature Installation Overview*, GC53-1137.

- EXP520 Expansion Enclosure:

  The EXP520 Expansion Unit (machine type 1814-52A) is a high-capacity 16-drive bay unit packaged in a 3U rack-mountable enclosure. Fully populated with 900 GB FC-SAS disk drive modules, this enclosure offers up to 14.4 TB of raw storage capacity or up to 32 TB when populated with the 2 TB E-DDM SATA drives. The EXP520 expansion unit connects to the DS5020 storage server. With the proper firmware level, this expansion unit is able to host both FDE, FC, FC-SAS, SDD, and SATA drives. Intermix of FC, SATA, FC-SAS, SDD, FDE drives is supported.

- EXP5060 Expansion Enclosure:

  The EXP5060 (machine type 1818-G1A) High Density Disk Enclosure is packaged in a 4U rack-mount enclosure containing 5 drive trays, where each tray holds up to 12 SATA disk drives, for a total disk drive capacity of up to 60 SATA drives per enclosure. With the attachment to DS5100 and DS5300 storage subsystems, the EXP5060 provides continuous, reliable service, using hot-swap technology for easy replacement without interruption, and supports redundant, dual-loop configurations. The storage capacity of the EXP5060 is 180 TB per enclosure using 3 TB SATA drives. Up to eight EXP5060 enclosures can be attached to a DS5000 controller (when neither the EXP5000 expansion or EXP810 expansion enclosures are part of the configuration) to provide a configuration that scales up to 480 SATA disk drives with a physical storage capacity of up to 1440 TB.

  For more information, see the *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679.

1.4.1 Supported disk drives

At the time of writing, the IBM Midrange Storage Subsystem family supports various disk drives:

- *Fibre Channel* (FC): 18, 36, 73, 146, 300, 450, 600 at a speed of 15k rpm (up to 4 Gbps), and 9, 18, 36, 73, 146, and 300 GB, at a speed of 10k rpm (only 2 Gbps).
- *Serial Advanced Technology Attachment* (SATA): 500 GB, 750 GB, 1 TB, 2 TB, and 3 TB disk drives, all at a speed of 7.2k rpm.
- *Serial Attached SCSI* (SAS): 600 and 900 GB disks at a speed of 10k rpm (6 Gbps) connected using 4 Gbps FC interposer.
- *Self Encrypting Disks* (SED): 146, 300, 450, and 600 GB disks at a speed of 15k rpm. They use Full Disk Encryption methodology to avoid accessing the data from stolen or misused disks.
- *Solid State Drives* (SSD): 73 and 300 GB 3.5” disk drives with 4 Gbps FC interface. Or 200 and 400 GB SAS 2.5” SSD disk drives with FC-SAS interposer. Up to 20 SSD drives can be used per system.

Tip: The term FC-SAS refers to a 6 Gbps SAS drives with 4 Gbps FC-SAS interposers.

The following firmware levels are required to support the newest disk drives:

- DS5020: 7.60.13.05 or higher
- DS5100: 7.50.13.00 or higher
- DS5300: 7.50.13.00 or higher
- EXP810: 9898 to 98B5, 98C1 and 98C5 or higher (depends on connected controller)
- EXP520: 98C5 or higher
- EXP5000: 98C1, 98C3, 98C5 or higher

1.4.2 Summary of the DS5000 family

Figure 1-6 shows the position of DS5000 series on the storage market. Therefore, you can plan your budget accordingly, based on current and future storage requirements.

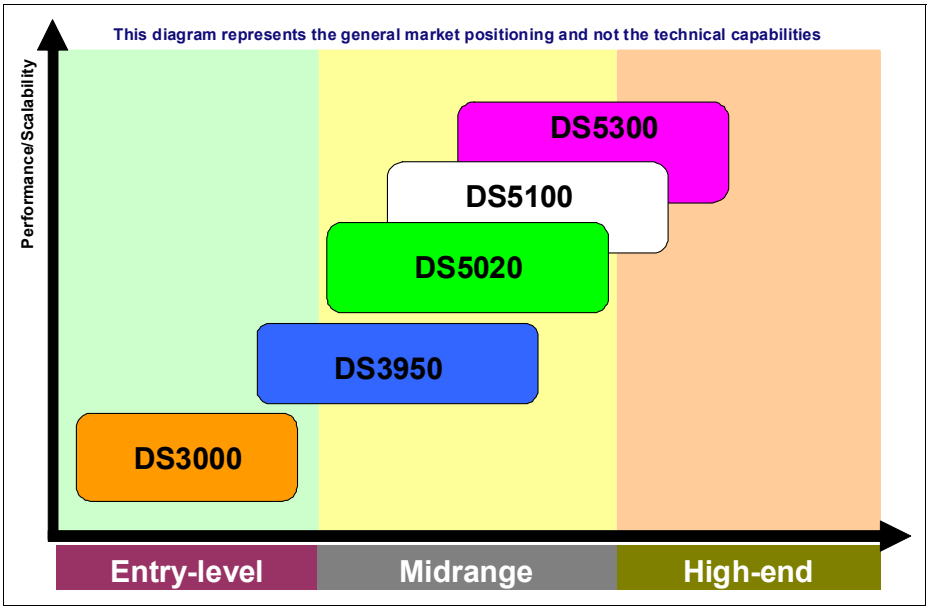


Figure 1-6 DS5000 series positioning

Finally, in Table 1-1, we briefly summarize technical parameters of all three available models of DS5000 family. We mention numbers in the maximum configuration of storage controllers with respective expansion units.

Table 1-1 DS5000 family at a glance

Features	DS5020	DS5100	DS5300
Processors	One 667 Mhz Xscale	Intel Xeon 2.8 GHz	Intel Xeon 2.8 GHz
Memory cache, maximum	2/4 GB	8/64 GB	8/64 GB
Max. host FC ports	Eight 8 Gbps	Sixteen 8 Gbps	Sixteen 8 Gbps
Max. host iSCSI ports	Four 1 Gbps	Eight 10 Gbps	Eight 10 Gbps
Disk FC ports	Four 4 Gbps FC	Sixteen 4 Gbps FC	Sixteen 4 Gbps FC
Max. disk drives	FC[a] - 112, SATA - 112	FC[a] - 448, SATA - 480	FC[a] - 448, SATA - 480
Max. capacity	FC - 67.2 TB SAS[a] - 100.8 TB SATA - 336 TB	FC - 268.8 TB SAS[a] - 403.2 TB SATA - 1440 TB SSD - 8 TB	FC - 268.8 TB SAS[a] - 403.2 TB SATA - 1440 TB SSD - 8 TB
Max. hosts	256	256	256
Max partitions/LUNs	128/2048	512/4096	512/4096
Premium features	FlashCopy, VolumeCopy, ERM, Intermix	FlashCopy, VolumeCopy, ERM, Intermix	FlashCopy, VolumeCopy, ERM, Intermix

Features	DS5020	DS5100	DS5300
Performance[b]:			
► Cached Read IOPS	120k	650k	700k
► Disk Read IOPS	44k	65k	172k
► Write IOPS	9k	20k	45k
► Cached Read MB/s	1500	3200	6400
► Disk Read MB/s	990	3200	6400
► Disk Write MB/s	850	2500	5300

a. SAS drives connected through FC interposers
b. Achieved in lab environment, and results might vary with various applications

1.5 DS Storage Manager

The DS Storage Manager software is the primary tool for managing, configuring, monitoring, and updating firmware, support data collection for the DS3000, DS4000, and DS5000 series of storage subsystems, and repair procedures. This tool provides two interfaces with a user-friendly graphical user interface (GUI), and a command line interpreter (**smcli**) interface for use with the scripts and macros to make the repetitive work easy. Various types of work that can be performed include configuration of RAID arrays and logical drives, assigning logical drives to a host, expanding the size of the arrays and logical drives, and converting disk arrays from one RAID level to another.

The tool can be used for troubleshooting and management tasks, such as checking the status of the storage subsystem components, updating the firmware of the controllers, replacement procedures for failed components, including rebuilding drives for use, and managing the storage subsystem. Finally, it offers implementation and management capabilities for advanced premium feature functions such as FlashCopy, Volume Copy, and Enhanced Remote Mirroring. The Storage Manager software package also includes the required host software components for the specific host environments that are planned to be supported.

The Storage Manager software level is closely tied to the features of the level of the firmware code that is being ran on the subsystem. Newer Storage Manager versions are designed to be backward compatible with the current firmware levels for previous generations of products as well as earlier versions of firmware for the current product line. Newer firmware levels might require a newer version of the Storage Manager to be installed.

Attention: Always consult the System Storage Interoperation Center (SSIC) for the latest supported host types and operating systems at this website:

http://www.ibm.com/systems/support/storage/ssic/interoperability.wss

The Storage Manager software is now packaged as follows:

- *Host-based* software:
  - Storage Manager 10.7*x* Client (SMclient):

    The SMclient component provides the GUI and the “smcli” interfaces for managing storage subsystems through the Ethernet network or from the host computer.
  - Storage Manager 10.7*x* Runtime (SMruntime):

    The SMruntime is a Java runtime environment that is required for the SMclient to function. It is not available on every platform as a separate package, but in those cases, it has been bundled into the SMclient package.
  - Storage Manager 10.7*x* Agent (SMagent):

    The SMagent package is an optional component that allows in-band management of the DS5000 storage subsystems.
  - Storage Manager 10.7*x* Utilities (SMutil):

    The Storage Manager Utilities package contains command line tools for making logical drives available to the operating system for specific host environments.
  - Multipath drivers:

    The Storage Manager offers a choice of multipath drivers, RDAC, or MPIO. This choice might be limited depending on host operating systems. Consult the Storage Manager `readme` file for the specific release being used.

    During the installation you are prompted to choose between RDAC or MPIO. Both are Fibre Channel I/O path failover drivers that are installed on host computers. They are only required if the host computer has a host bus adapter (HBA) installed.
- *Controller-based* software:
  - DS5000 storage subsystem controller firmware and NVSRAM:

    The controller firmware and NVSRAM are always installed as a pair and provide the “brains” of the DS5000 storage subsystem.
  - DS5000 storage subsystem Environmental Service Modules (ESM) firmware:

    The ESM firmware controls the interface between the controller and the drives.
  - DS5000 storage subsystem drive firmware:

    The drive firmware is the software that tells the specific drive types how to perform and behave on the back-end FC loops.

The installation and operation of DS Storage Manager is described in Chapter 4, “Host configuration guide” on page 151 and in the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023. Additional information can be found at the following website:

`http://www.ibm.com/systems/storage/disk/midrange/`

2

IBM System Storage DS5000 storage subsystem planning tasks

Careful planning is essential to any new storage installation. Having a well thought-out design and plan prior to the purchase and implementation can help you get the most out of your investment for the present and protect it for the future.

The DS5000 storage subsystem can support a number of varying capabilities which can be used to meet many different environments. An in-depth look at how these values can be used is provided and is described with considerations for host, storage, and applications throughout this manual.

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

Important: This chapter is aimed at both pre-sale and post-sale decision making points.

In this chapter, we provide you with guidelines to help with the planning process.

2.1 Planning overview

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

- What are my storage needs now and in the near future?
- What are my host connection requirements?
- What are my Storage Area Network (SAN) requirements?
- What additional hardware will I need?
- What reliability and availability do I require?
- What redundancy do I need? (For example, do I need off-site mirroring?)
- What will be the physical layout of the installation? Only local site, or remote sites as well?
- What compatibility issues do I need to address?
- Will I use any storage virtualization product such as IBM SAN Volume controller?
- Will I use any unified storage product such as the IBM System Storage N series?
- What operating systems am I going to use (existing or new installation)?
- Will I use any volume management software (Veritas or LVM for example)?
- What applications will access the storage subsystem?
- What are the hardware and software requirements of these applications?
- What level of performance do I need?
- How much does it cost?

With the DS5000 storage subsystem, you have a number of choices to make concerning how you want it to support your environment. With many of these choices, you will want to make decisions based on your planned workload. Understanding how the DS5000 storage subsystem options and features impact various workloads can help you know what hardware, settings, and choices you want to make before starting your implementation and configuration process. Here are some solution decisions to consider:

- Number and types of hosts and applications to be supported
- Which premium features are needed for your environment's solution
- How the data will be backed up and recovered.
- Internal layout setting considerations:
  - Number of drives in the arrays
  - Size of the logical drives
  - RAID level to be used
  - Cache usage and settings
  - Segment size

To plan the DS5000 storage subsystem layout, you need to know the attached hosts, their operating system, and the applications which will be using the DS5000 storage subsystem. You also need to consider the performance aspects and requirements. Finally, you need to know which special premium features will be running in the background on the DS5000 storage subsystem.

On the other hand, you also need to define the layout of the attached hosts with their host bus adapters and the mappings of the logical drives to specific host group or host partitions.

This list of questions is not exhaustive, and as you can see, certain questions go beyond simply planning and configuring the DS5000 storage subsystem.

2.2 Planning your DS5000 storage layout

With the DS5000 storage subsystem, you can define your storage to be presented to your host in many different manners. Though you can select by default to build all your storage into a common configuration that is used for all workloads, you might want to consider that it might not be the best method for your workloads. All workloads can perform at their best with the right configurations. Also, making good use of your capacity is also an important factor in making wise decisions about how to define the array and logical drive layouts.

Planning for your specific needs is an important part of the successful layout of your storage. In Chapter 7, “IBM Midrange Storage Subsystem tuning with typical applications” on page 309, you can see why you might want to have a variety of different RAID types and different size arrays to best handle your applications. For a description of the different RAID types, see 2.2.5, “DS5000 arrays and RAID levels” on page 26.

2.2.1 Disk expansion enclosures

The DS5000 storage subsystems offer the EXP5060 high capacity expansion or the EXP5000 expansion enclosures, and is compatible with the EXP810 for migration purposes. When planning for which enclosure to use with your DS5000, you must look at the applications and data that you will be using. The EXP520 is a new expansion enclosure that will be used solely with the DS5020 storage subsystem.

The EXP5000 and EXP520 enclosures can accommodate either Fibre Channel, SSD, or SATA II drives. Intermixing of the drives is permitted. The EXP5000 and EXP520 is an 8 or 4 Gbps capable enclosure, offering high performance and value. At the time of this writing the EXP5060 can only support SATA II drives, and there is no intermixing in the enclosure supported. For detailed information on implementation and best practices on the planning and use of the EXP5060 see: *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679.

Enclosure IDs

It is very important to correctly set the tray (enclosure) IDs. They are used to differentiate multiple EXP enclosures that are connected to the same DS5000 storage subsystem. Each EXP enclosure must use a unique value. The DS5000 Storage Manager (SM) uses the tray IDs to identify each EXP enclosure.

For the EXP5060 and EXP5000, the enclosure ID (shown in Figure 2-1) is indicated by a dual seven-segment LED located on the back of each ESM next to the other ESM indicator lights. The storage server firmware automatically sets the enclosure ID number. If needed, you can change the enclosure ID setting through the DS5000 storage management software only. There are no switches on the EXP5060, EXP5000, or EXP810 chassis to manually set the enclosure ID.

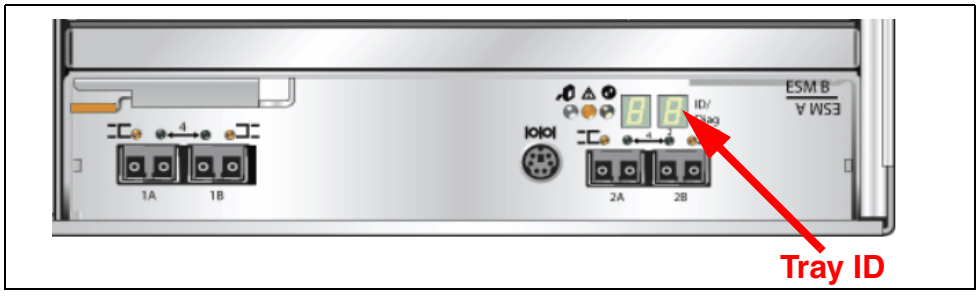


Figure 2-1 Enclosure ID LEDs for EXP5000

Enclosure guidelines

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

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

The enclosure ID on EXP5000 is automatically assigned.

Because the DS5000 storage subsystem follows a specific address assignment scheme for the drives, you need to observe a few guidelines when assigning enclosure IDs to expansion units if you were to assign these manually. Failure to adhere to these guidelines can cause issues with I/O error recovery, and make the troubleshooting of certain drive communication issues more difficult:

- Whenever possible, maintain a number of expansion units on the DS5000 storage subsystem that allows for the configuration of enclosure loss protection. It can be as few as two with a RAID 1 or RAID 10 array group, or however many drives you are striping your RAID 5, or n - 2 for your RAID 6 array configuration. When adding EXP expansion units, consider your plan for the usage of the new drives to determine the best growth planning. With the DS5100 and DS5300, growth is generally suggested in groups of four EXPs at a time.
- When adding drives to a DS5000 storage subsystem, insert the new drives into an expansion enclosure at no more than two at a time.

2.2.2 Drive types

The speed and the type of the drives used will impact the performance. Typically, the faster the drive, the higher the performance. This increase in performance comes at a cost; the faster drives typically cost more than the lower performance drives. The highest performing storage for random small IO is the Solid state disks (SSD) followed by FC and FC-SAS drives which outperform the SATA drives.

There are five different disk drive types available for the DS5000 storage subsystems:

- Fibre Channel (FC) disks (Encryption capable FDE or not)
- Serial Attached SCSI (SAS) disks (with FC to SAS interposer)
- Solid State Drives (SSD,
- SAS SSD (with FC to SAS interposer)
- Serial ATA disks (SATA)

These drives, listed by their interface, speed, and size, are available at the time of writing:

- 4 Gbps FC, 300 GB / 15K Enhanced Disk Drive Module
- 4 Gbps FC, 450 GB / 15K Enhanced Disk Drive Module
- 4 Gbps FC, 600 GB / 15K Enhanced Disk Drive Module

The following SAS drives are currently available:

- SAS 300 GB / 10K (FC-SAS interface)
- SAS 600 GB / 10K(FC-SAS interface)
- SAS 900 GB / 10K(FC-SAS interface)

The following SATA drives are currently available:

- 4 Gbps SATA: 7.2K rpm 1000 GB (1 TB) E-DDM
- 4 Gbps SATA: 7.2K rpm 2000 GB (2 TB) E-DDM

The following Solid State Drives (SSD) are currently available:

- SSD 200 GB (FC-SAS interface)
- SSD 400 GB (FC-SAS interface)

In addition, DS5000 systems support Full Disk Encryption (FDE) drives are drives with built-in disk encryption hardware that prevents unauthorized access to the data on a drive that is physically removed from the storage subsystem:

- Encryption Capable 4 Gbps, 15K rpm FC 300 GB
- Encryption Capable 4 Gbps, 15K rpm FC 450 GB
- Encryption Capable 4 Gbps, 15K rpm FC 600 GB
- Encryption Capable 4 Gbps, 10K rpm SAS 600 GB (FC-SAS interface)

RAID arrays can only be created by using the same disk types. Depending the usage planned for a specific array, you might want to choose a specific disk type to optimize the overall performance.

If your application demands high levels of throughput, then SATA, SAS, and FC disk types can provide similar performance values.

SSD drives, on the other hand, perform much better for I/O intensive applications, but not as good as the above three types for throughput. If your application is critical for I/O operations, your first disk selection is SSD, then FC or SAS; however, avoid using SATA drives.

Tip: Select the best drive types for your array, depending on your application needs:

- For I/O demanding applications, use as first choice SSD, then FC or SAS, and then SATA.
- For throughput demanding applications, use FC, SAS, or SATA as your first options.

Table 2-1 compares the Fibre Channel and SAS 10K, 15K, and SATA 7200 drives (single drive).

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

Table 2-1 Comparison between Fibre Channel and SAS verses SAS-Nearline and SATA

Drive feature	Fibre Channel	SATA	SATA-2	SATA difference
Spin speed	10K and 15K	7.2 K		
Command queuing	Yes 16 Max	No 1 Max	Yes 16 Max	
Single disk I/O Rate (# of 512 bytes IOPS)[a]	280 & 340	88	88	.31 and .25
Read bandwidth (MBps)	69 & 76	60	60	.86 and .78
Write bandwidth (MBps)	68 & 71	30	30	.44

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

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

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

Command queuing (or queue depth) allows for multiple commands to be outstanding to the disk drive at the same time. The drives have a queue where outstanding commands can be dynamically rescheduled or re-ordered, along with the necessary tracking mechanisms for outstanding and completed portions of workload. With the DS5300 and DS5020, the disks have a command queue depth of 16. With the DS5100 and base level subsystems, the queue depth of the disks is limited to 4.

2.2.3 Disk intermix capability

With the DS5000 storage subsystems, all of the above disk types can be intermixed for their use to concurrently meet specific workload needs for your solution with a single DS5000 controller configuration. You can create and manage distinct arrays or logical drives that are built from SSDs, FC, SAS, or SATA disks in a DS5000 storage subsystem, and allocate the drives to the appropriate applications in the attached host servers.

When planning your type of drives, consider these three classes of storage needs:

- *Online (or primary) storage* is used as storage for applications that require immediate access to data, such as databases and frequently accessed user data. Primary storage holds business-critical information and data with the highest value and importance. This storage requires high performance and high availability technologies such as Fibre Channel technology.
- *Near-line (or secondary) storage* is used for applications that do not require immediate access but still require the performance, availability, and flexibility of disk storage. It can also be used to cache online storage to reduce the time required for data backups. Secondary storage represents a large percentage of a company's data and is an ideal fit for both the larger size FC, and FC-SAS disks, or the SATA disks.
- *Offline (archival) storage* is used for backup or long-term storage. For this type of storage, tape remains the most economical solution.

Now that we have identified that SATA technology is best suited to near-line storage usage, we have the following considerations regarding its use in the same storage subsystem as online storage:

- *Performance:* Instead of having a storage subsystem for online storage and another for near-line storage, both types of storage can be combined with one storage subsystem, and use higher performing controllers.
- *Scalability:* The total SATA disk capacity is far greater than the Fibre Channel and SAS offerings. The new EV-DDM and E-DDM drives are 2000 GB, as compared to the largest FC drive being 600 GB and the FC-SAS being 900 GB.
- *Cost:* Consider an existing Fibre Channel environment where you want to implement SATA technology. This consideration is further enhanced with Fibre Channel and SATA drive intermixing within the same enclosures. The following cost considerations apply when intermixing versus implementing a separate SATA storage subsystem:
  - Implementation costs of a new SATA storage subsystem versus intermixing with the existing Fibre Channel enclosures.
  - SAN infrastructure costs for more ports or hardware.
  - Administration costs (effort) depend on the number of controllers managed. Intermixing eliminates controller pairs by leveraging existing controllers.

After these factors have been considered and the criteria for the environment has been identified, the choice to intermix or not can be determined.

2.2.4 Drive Security

Drive Security is a new premium feature where Full Disk Encryption (FDE) protects the data on the disks only when the drives are removed from storage subsystem enclosures. Drive Security requires security capable drives (FDEs) and provides access to data only through a controller that has the correct security key when Drive Security is enabled.

Security requirements

Businesses must comply with a growing number of corporate standards and government regulations, Drive security is one tool that can enhance security, thus complying with these new standards and regulations.

Full Disk Encryption

Full Disk Encryption (FDE) does not prevent someone from copying the data in the storage subsystems through Fibre Channel host port connections when the drives are unlocked and operating. FDE also does not prevent unauthorized management access. A security capable drive encrypts data during writes and decrypts data during reads. FDE prevents the physical removal of the disk from the DS5000 system and interpreting data it contained. The FDE drive with Drive Security enabled will be locked on power up and will only unlock after successful authentication with the DS5000 system.

The Encryption Key is generated by the drive and never leaves the drive, so it always stays secure. It is stored in encrypted form performing symmetric encryption and decryption of data at full disk speed with no impact on disk performance. Each FDE drive uses its own unique encryption key which is generated when the disk is manufactured and regenerated when required by the storage administrator using the DS5000 Disk Encryption Manager.

The security enabled drives can be used as normal drives and intermixed in an array with drives of equal type and capacity when this feature is not enabled. This new feature is detailed in *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

2.2.5 DS5000 arrays and RAID levels

In this section, we introduce disk arrays, logical drives, and associated terminology, and then describe the various RAID levels that are supported by the IBM System Storage DS5000 storage subsystem. RAID is an acronym for Redundant Array of Independent Disks. It is a storage solution in which part of the total storage capacity is used to store redundant information about user data stored on the remainder of the storage capacity.

RAID relies on a series of configurations, called levels, to determine how user data and redundancy data are written to and retrieved from the drives. RAID Level 1, RAID Level 10, RAID Level 3, RAID Level 5, and RAID Level 6 write redundancy data to the drive media for fault tolerance. The redundancy data might be an exact copy of the data (mirrored) or an error correcting code derived from the data. If a drive fails, you can use the redundancy data to quickly reconstruct information on a replacement drive.

RAID levels

An array is a set of drives that the system logically groups together to provide one or more logical drives to an application host or cluster. The DS5000 storage subsystems support RAID levels 0, 1, 10, 3, 5, and 6. Each of these RAID levels offers a compromise between capacity, performance, and data redundancy. The attributes of each of these RAID levels is described in more detail over the following pages of this book.

The DS5000 family of storage subsystem is able to dynamically change the RAID level without requiring downtime. This feature is called Dynamic RAID Migration (DRM).

RAID 0: Data striping

RAID 0 (Figure 2-2) provides the highest performance of all the RAID types, but has no parity protection for the data. It is well-suited for test and development environments with program libraries requiring rapid loading of large tables, or more generally, applications requiring fast access to read-only data or fast writing.

RAID 0 is only designed to increase performance. With no redundancy, any disk failures require reloading from backups. Select the RAID 0 for applications that will benefit from the increased performance capabilities of this RAID level, but can afford data loss, and recovery from backup is acceptable in terms of recovery point objective (RPO). Never use this level for critical applications that require high availability with a tough RPO.

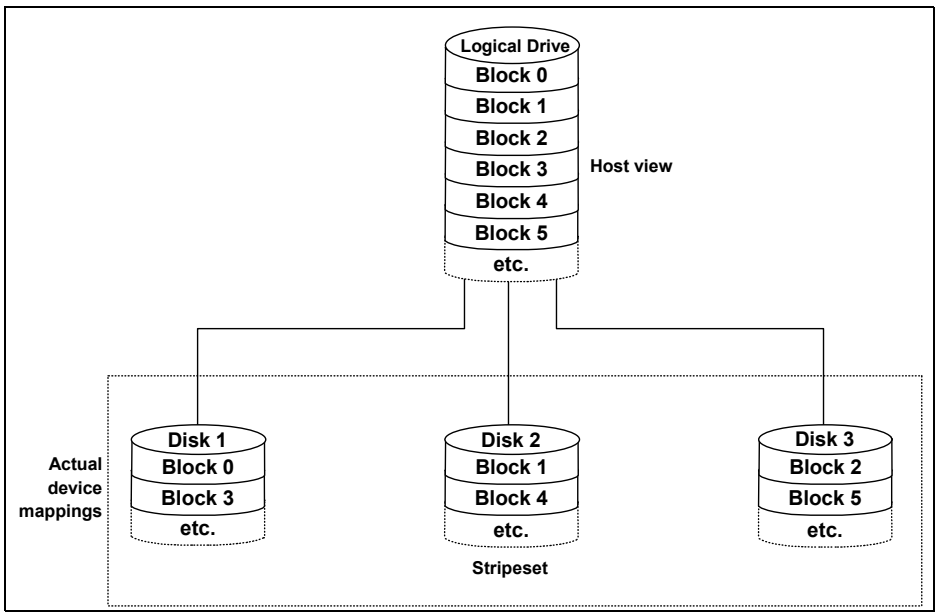


Figure 2-2 RAID 0

Important: The failure of a single disk in a RAID 0 array will cause the failure of the entire array and all of the associated logical drives, and access to all data on the array will be lost.

RAID 1 and RAID 10: Disk mirroring and disk mirroring with striping

RAID 1 is also known as *disk mirroring* and is a mirrored pair of drives without parity. RAID Level 1 uses exactly two drives to mirror the data between them. A RAID 10 (Figure 2-3) array is automatically created when you create a RAID 1 array with four or more drives (two pairs of drives). RAID 10 is also known as RAID 1+0 and *disk mirroring with striping*, and it implements block interleave data striping and mirroring. In RAID 10, data is striped across the physical disk drives, and each of those drives is then mirrored to a second drive.

RAID Levels 1/10 provide good redundancy; in the case of a single disk failure in each mirrored pair, the array and associated logical drives become degraded but all the data is still available and accessible from the second drive of the mirrored pair.

For each pair of mirrored drives, read operations can be performed from either physical disk of the mirrored pair. Write operations are performed by writing to both physical disks of the mirrored pair. In this manner, small block size writes can be completed quickly, making this RAID type a great solution for a high write-intensive application when data protection is desired. For this type of application, this RAID type is generally preferred by database administrators.

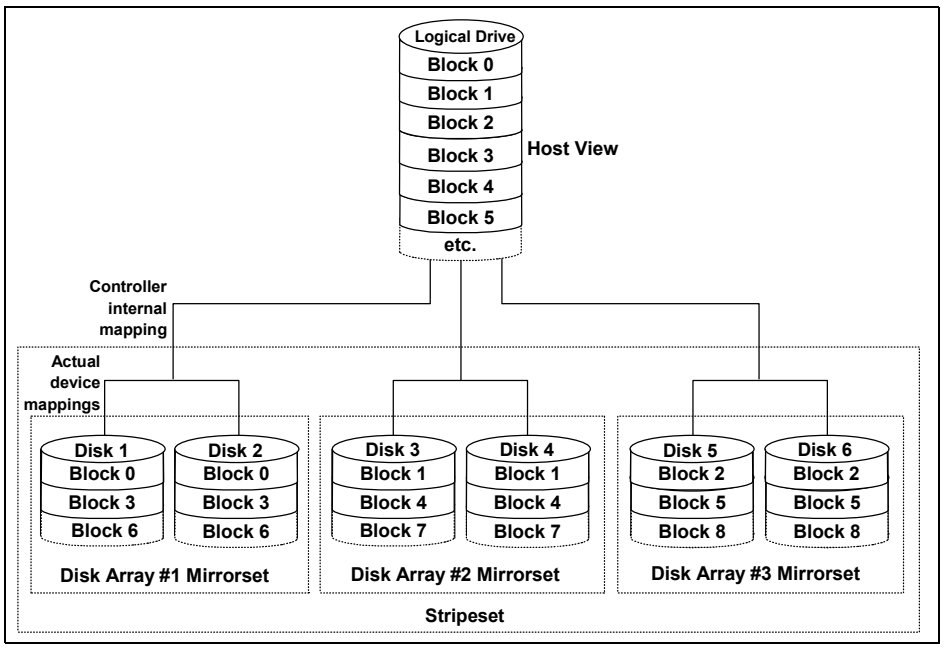


Figure 2-3 RAID 10; RAID 1 consist of one mirrored pair only (#1 Mirrorset)

However, because the data is mirrored, the capacity of the associated logical drives on a RAID 1 or RAID 10 array is 50% of the physical capacity of the hard disk drives in the array.

Here are some suggestions when using RAID 1/10:

- Use RAID 1 for the disks that contain your operating system. It is a good choice, because the operating system can usually fit on one disk.
- Use RAID 1 for transaction logs. Typically, the database server transaction log can fit on one disk drive. In addition, the transaction log performs mostly sequential writes. Only rollback operations cause reads from the transaction logs. Therefore, we can achieve a high rate of performance by isolating the transaction log on its own RAID 1/10 array.
- Use write caching on RAID 1/10 arrays. Because a RAID 1/10 write will not complete until both writes have been done (two disks), performance of writes can be improved by using a write cache. When using a write cache, be sure it is battery-backed up.
- The performance of RAID 10 is comparable to RAID 0 for sequential I/Os, but RAID 10 provides data redundancy through disk mirroring.

Tip: There are no guaranteed choices as to which type of RAID to use because it is dependent on the workload read and write activity. A good general guideline might be to consider using RAID 1 if random writes exceed about 25%, with a peak sustained I/O rate that exceeds 50% of the storage subsystem's capacity.

RAID 3: Data striping with a dedicated parity drive

A RAID 3 array uses data striping with a dedicated parity drive. Similar to RAID 0 data striping, information written to disk is split into chunks (a fixed amount of data), and each chunk is written out to the same physical position on separate disks (in parallel). This architecture requires parity information to be written for each stripe of data. RAID 3 uses a dedicated physical drive for storing parity data. If any one disk drive in the array fails, the array and associated logical drives become degraded, but all data is still accessible by the host application.

However, with RAID 3, the dedicated parity drive is a performance bottleneck during writes. Because each write operation requires the parity to be re-computed and updated, this means that the parity drive is accessed every time a block of data is written to the array. Because of this, RAID 3 is rarely used today in the industry and RAID 5 has taken its place including in the DS5000 family of storage subsystems.

RAID 5: Striped data with distributed parity

Like RAID Level 3, RAID Level 5 also uses parity for data protection but unlike RAID 3 it does not use a dedicated parity drive. Instead, the parity blocks are evenly distributed across all physical disk drives in the array, as shown in Figure 2-4. The failure of a single physical drive in a RAID 5 array will cause the array and associated logical drives to be degraded, but all the data will remain accessible to the host application. This level of data redundancy is known as n+1 redundancy because the data remains accessible after a single drive failure. When you create a RAID 5 array, the capacity of the array is reduced by the equivalent capacity of one drive (for parity storage).

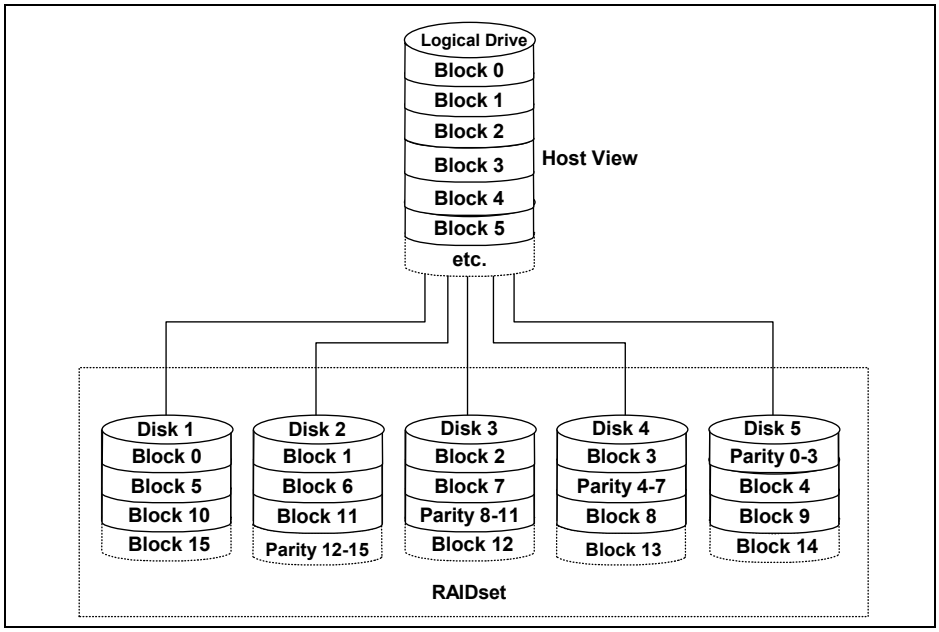


Figure 2-4 RAID 5

RAID Level 5 can be good for multi-user environments, such as database or file system storage where the typical I/O size is small and there is a high proportion of read activity. Applications with a low read percentage and a high random write percentage might not perform as well on RAID 5 logical drives because parity data must be recalculated for each write operation and then written to each drive in the array.

Use write caching on RAID 5 arrays, because each RAID 5 write will not be completed until at least two read I/Os (one data one parity) and two write I/Os (one data and one parity) have occurred. The write performance penalty can be mitigated by using battery-backed write cache. RAID 5 arrays with caching can give as good as performance as any other RAID level, and with certain workloads, the striping effect can provide better performance than RAID 1/10. Applications that require high throughput sequential write I/Os are an example of one such workload. In this situation, a RAID Level 5 array can be configured to perform just one additional parity write when using “full stripe writes” (also known as “full stride writes”) to perform a large write I/O when compared to the two writes per data drive (self and its mirror) that are needed for each write I/O with a RAID 1 array.

You need to configure the RAID 5 array with a certain number of physical drives to take advantage of full stripe writes. This rule is illustrated in Figure 2-5 for the case of a RAID 5 array with 8 total drives (7 data + 1 parity).

A: Total # of drives written	B: Total # of RAID 5 Read Data IOs	C: Total # of RAID 5 Read Parity IOs	D: Total # of RAID 5 Write Data IOs	E: Total # of RAID 5 Write Parity IOs	X: Total # of IOs for RAID 5 (= B+C+D+E)	Y: Total # of IOs for RAID 1 (= Ax2)
1	1	1	1	1	4	2
2	2	1	2	1	6	4
3	3	1	3	1	8	6
4	3	0	4	1	8	8
5	2	0	5	1	8	10
6	1	0	6	1	8	12
7	0	0	7	1	8	14

Figure 2-5 Potential performance advantages of RAID 5 full stripe writes

Column A lists the number of drives that are being written to. Column Y is the number of write I/Os for a RAID 1 and will always be twice the value of A. Columns B, C, D, and E contain the numbers of read data/parity and write data/parity I/Os required for the number of drives that are being written to. You can see that for seven drives, no read I/Os are required for RAID 5 arrays because the full stripe is being written at once. This substantially reduces the total number of I/Os (column X) required for each write operation.

The decrease in the overhead read operations with the full stripe write operation is the advantage you are looking for. You must be careful when implementing this type of layout to ensure that your data pattern does not change, which might decrease its effectiveness. However, this layout might work well for you in a large sequential write environment. Due to the small size of segments, reads might suffer, so mixed I/O environments might not fare well, which might be worth testing if your writes are high.

When the DS5000 detects that it is receiving contiguous full stripe writes, it will switch internally to an even faster write capability known as Fast Stripe Write Through. In this method of writing, the DS system storage uses the disk as the mirror device for the cache write and shortens the write process. This method of writing can increase throughput as much as 30% on the system storage. However, it requires the following rules are being met by the IO pattern:

- All write I/Os are full stripe writes (no partial stripe writes can be requested)
- Write I/Os are sequential and contiguous in nature so no seeks are required.

If any interruptions in this pattern are detected, the writes will revert back to the standard full stripe write model, which nevertheless still gives a benefit to the RAID5 for large sequential writes over the RAID1.

RAID 6: Data striping with dual distributed parity

RAID 6 (see Figure 2-6) with dual rotational parity that is distributed across the drives in the array. A RAID 6 array has n+2 redundancy, which means that data remains accessible to the host application after two concurrent disk drives failures in the array. RAID 6 achieves n+2 redundancy because it calculates two sets of parity information for each block of data (P+Q) that is striped across the disks. The DS5000 storage subsystem performs the P+Q parity calculations in hardware.

There is no performance penalty for read operations from a RAID 6 array, but there is a performance penalty for write operations because two sets of parity information (P+Q) must be calculated for each write operation. The write penalty in RAID Level 6 can be mitigated by using battery-backed write caching.

The calculation of q is complex. In the case of the DS5000 storage subsystem, this calculation is made by the hardware and thus more performant than the software-based implementation found in other storage subsystems.

By storing two sets of distributed parities, RAID 6 is designed to tolerate two simultaneous disk failures. It is a good implementation for environments using SATA disks.

Due to the added impact of more parity calculations, in terms of writing data, RAID 6 is slower than RAID 5, but might be faster in random reads thanks to the spreading of data over one more disks.

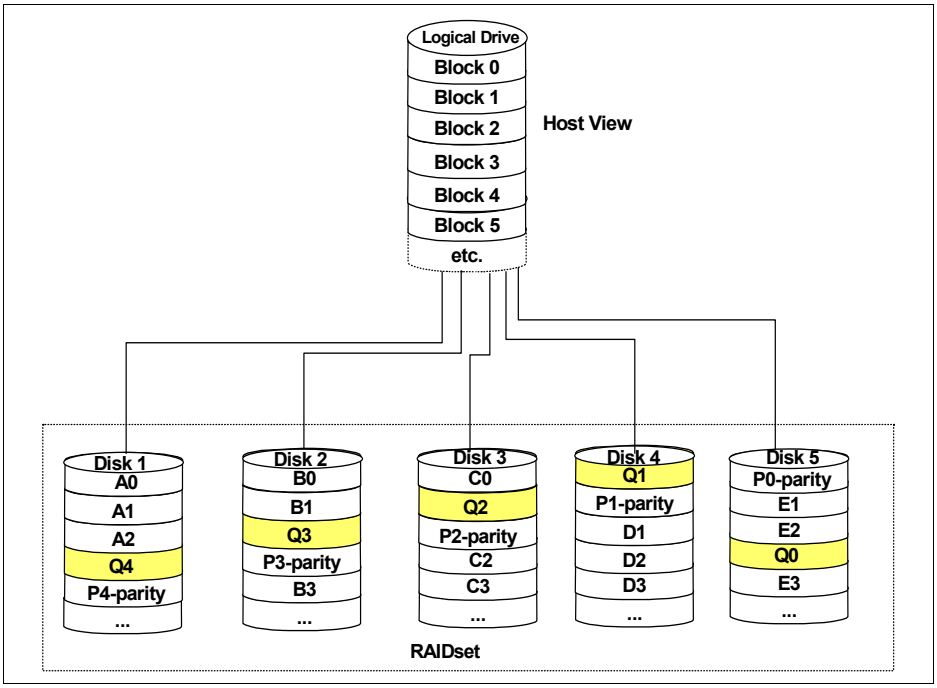


Figure 2-6 RAID 6

Summary of RAID levels

Table 2-2 represents the technical summary of all RAID level with the suggested usage.

Table 2-2 RAID levels comparison

RAID	Description	Application	Advantage	Disadvantage
0	Stripes data across multiple drives.	IOPS Mbps	Performance, due to parallel operation of the access.	No redundancy. If one drive fails, the data is lost.
1	The disk's data is mirrored to another drive.	IOPS	Performance, as multiple requests can be fulfilled simultaneously.	Storage costs are doubled.

RAID	Description	Application	Advantage	Disadvantage
10	Data is striped across multiple drives and mirrored to the same number of disks.	IOPS	Performance, as multiple requests can be fulfilled simultaneously. It is the best option for random write performance. Most reliable RAID level.	Storage costs are doubled.
3	Drives operate independently with data blocks distributed among all drives. Parity is written to a dedicated drive.	Mbps	High performance for large, sequentially accessed files (image, video, and graphics).	Degraded performance with 8-9 I/O threads, random IOPS, and smaller, more numerous IOPS.
5	Drives operate independently with data and parity blocks distributed across all drives in the group.	IOPS Mbps	Good for reads, small IOPS, many concurrent IOPS, and random I/Os. Best for sequential writes, throughput intensive applications. Best option for mixed workload and space efficiency.	Random writes are particularly demanding.
6	Stripes blocks of data and parity across an array of drives and calculates two sets of parity information for each block of data.	IOPS Mbps	Good for multi-user environments, such as database or file system storage, where typical I/O size is small, and in situations where additional fault tolerance than RAID 5 is required.	Slower in random writing data, complex RAID controller architecture.

RAID reliability considerations

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

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

This problem is exacerbated by using large arrays with RAID 5. This *stripe kill* can lead to data loss when the information to rebuild the stripe is not available. The end effect of this issue will depend on the type of data and how sensitive it is to corruption. While most storage subsystems (including the DS5000 storage subsystem) have mechanisms in place to try to prevent this from happening, they cannot work 100% of the time.

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

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

2.2.6 Array configuration

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

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

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

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

Number of drives

Small arrays make easier configuring for enclosure loss protection; however, this design has a high cost of capacity for parity when used with RAID 5 or RAID 6 An additional factor to consider is that these small arrays can make high I/O environments encounter bottlenecks which slow application workloads.

With transactional based processes, the more physical drives you have per array, the shorter the access time for read and write I/O operations is. So having many disk drives per array will benefit this environment, where a high number of host I/O operations per second are needed. However, take care not to build such a large array that it becomes unmanageable during the sparing or recovery processes. Only RAID 10 is not impacted during by the sparing and rebuild.

With sequential access applications, however, the number of drives is not the critical factor. These workloads are more dependent on bandwidth and the speed of the processor and bus of the storage system to move the high throughput.

As you can see, the number of drives to select per array need to be considered according to the application environment for which the system is being designed.

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

- Transactional Performance: By doubling the number of the physical drives, you can expect up to a 50% increase in transactions, keeping the number less than twelve.
- Flexibility: Using more physical drives gives you more flexibility to build arrays and logical drives according to your needs.
- Data capacity: When using RAID 5 and RAID 6 arrays, more data space is available with smaller physical drives because less capacity is used for parity. However, pay attention to the protection level; because with larger RAID array, the rebuild time might expose you to a second failure and lead to an offline array.

Enclosure layout and loss protection planning

Depending on the DS5000 storage subsystem and the expansions used in your configuration, you need to plan in advance how these expansions will be best interconnected with your DS5000 system controller and the drives used for your array configurations. In order to optimize performance from each controller to each expansion, you need to determine the proper of your cabling configuration. It is most important with the larger storage subsystems.

With larger configurations with multiple expansion enclosures, you can provide better controller access to the drives by spreading them across the drive ports by selecting array members from different enclosures, and thus optimize performance. Other considerations to optimize the DS5000 storage subsystem is to evenly spread disk array members from odd and even drive slots, because each controller has a preferred management for the odd or even slots.

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

By default, the automatic configuration is enabled. With the default values used with this method of configuring very few applications are suited to the design, and the overhead cost can be quite high. When building a system for a single workload tuning the default settings can make this configuration method very useful. However, in many cases, it is not the best method of creating arrays. Instead, using the manual method allows for more configuration options to be available at creation time and a better design can be built that will meet the workload and performance requirements.

Manual array configuration: Manual array configuration allows for greater control over the creation of arrays because it allows you to specify each array to meet your planned optimal configuration options.

Figure 2-7 shows an example of enclosure loss protection. If enclosure number 2 fails, the array with the enclosure loss protection would still function (in a degraded state), as the other drives are not affected by the failure.

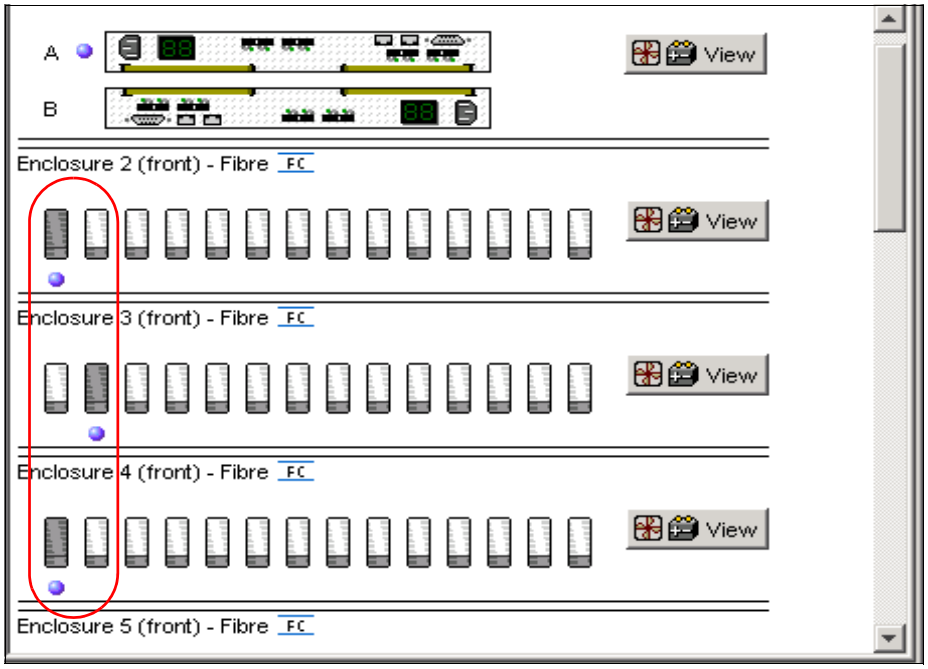


Figure 2-7 Enclosure loss protection

In the example shown in Figure 2-8, without enclosure loss protection, if enclosure number 2 fails, the entire array becomes inaccessible.

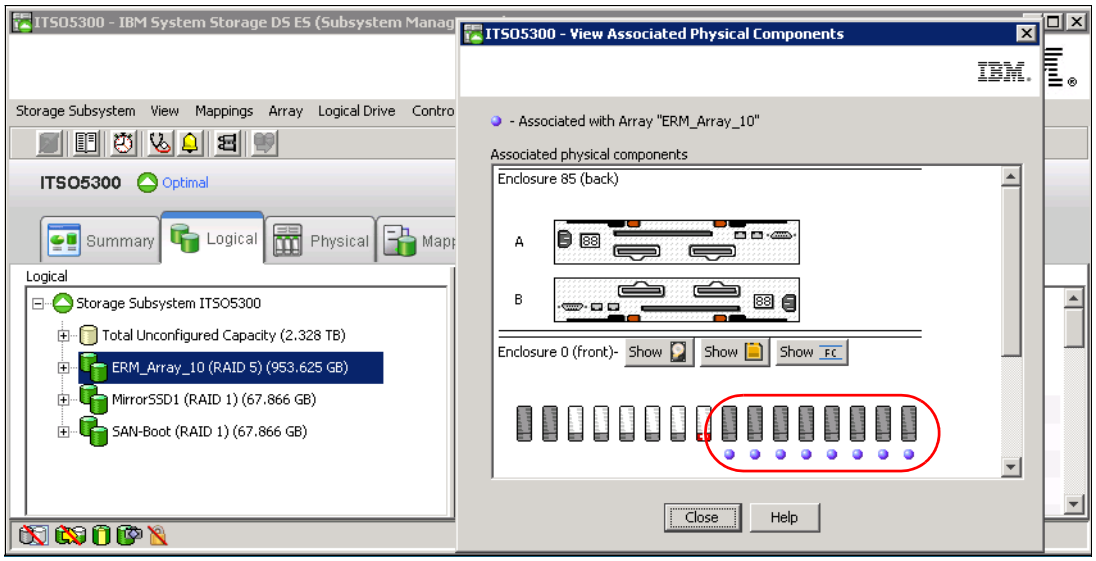


Figure 2-8 Array without enclosure loss protection

Best practice: When possible, plan to use enclosure loss protection for your arrays.

2.2.7 Segment size

A segment, in a logical drive, is the amount of data, in kilobytes, that the controller writes on a single physical drive before writing data on the next physical drive. Depending on your data structure, you can optimize the creation of your logical drives for this value.

When you create a logical drive, the default segment size choose is 128 KB, unless you are creating logical drives under a RAID 3, or choosing the multimedia type for the logical drive, in which case a segment size of 256 KB is used.

The choice of a segment size can have a major influence on performance in both IOPS and throughput. You can choose the defaults, which is a good choice for general usage, or consider the following to properly select a the logical drive segment size customized for your needs:

- The characteristics of your application, IOPS demanding, or throughput demanding, whether random or sequential.
- The I/O block size of the host that will use the logical drive:
  - Small IO sizes allow for greater transaction (IOPS) rates.
  - Large IO sizes allow for a better transfer rate in MBps.
- The RAID protection and number of disk drives that are participating in the logical drive array.

  To calculate the best segment size in this environment, consider the RAID protection and number of drives participating in the array.

Segment size > I/O block size request

This option is best for high IOPs and random requests:

- Several transactions are satisfied in a single operation to a disk.
- There is higher IOPS.
- It is ideal for random I/O requests, such as the database file system.
- The best option is start with segment size >= 2 x block size for high IOPS. Segment size = I/O block size request.

It is very difficult to align segment size with block size, so this option is not practical to implement. You can use the other two scenarios:

- Every request transaction uses exactly one disk operation.
- There is high IOPS.
- This option is ideal for random I/O requests, such as the database file system.

Segment size < I/O block size request

This option is best to obtain MBps performance and low IOPs, which is the norm for multimedia applications.

- More disks are used or requested.
- There is higher throughput (Mbps).
- It is ideal for sequential writes, such as a large multimedia application.
- It is optimized when a single I/O request can be serviced with a single or exact multiple data stripes (the segment size multiplied by the number of drives in the array that are used for I/O). In this case, multiple disks are used for the same request, but each disk is only accessed once, or the exact number of times if the block size is too big and you need to use multiple stripes.

To calculate the best segment size in this environment, consider the RAID protection and number of drives participating in the array by using the following formula:

Segment size = (block size / X)

Where X is the number of drives used in the array for I/O, as follows:

- RAID 0 = number of drives
- RAID 1 or 10 = number of drives / 2
- RAID 5 = number of drives - 1
- RAID 6 = number of drives - 2

If the resulting segment size is >512 KB, then divide it by an integer number staring with 2 to obtain multiple data stripes of each block requested.

Tip: The possible segment sizes available are 8 KB, 16 KB, 32 KB, 64 KB, 128 KB, 256 KB, and 512 KB.

- Storage Manager sets a default block size of 128 KB for every logical volume, except for RAID 3 volumes, which are set to 256 KB.
- For database applications, block sizes between 32–128 KB have shown to be more effective.
- In a large file environment, such as media streaming or CAD, 128 KB or more are suggested.
- For a Web server or file and print server, the range needs to be between 16–64 KB.

The performance monitor and collected support data can be used to evaluate how a given segment size affects the workload.

Segment sizes: Undertake a performance testing schedule in the environment before going into production with a given segment size. Segment size can be dynamically changed, but only by rewriting the data, which consumes bandwidth and impacts performance. Plan this configuration carefully to avoid having to reconfigure the options chosen.

2.2.8 Logical drives and controller ownership

Logical drives, sometimes simply referred to as volumes or LUNs (LUN stands for Logical Unit Number and represents the number a host uses to access the logical drive), are the logical portions of arrays. A logical drive is a logical structure you create on a storage subsystem for data storage. A logical drive is defined across a set of drives (called an array) and has a defined RAID level and capacity (see 2.2.5, “DS5000 arrays and RAID levels” on page 26). The drive boundaries of the array are hidden from the host computer.

General considerations

The IBM System Storage DS5000 storage subsystem provides great flexibility in terms of configuring arrays and logical drives. However, when assigning logical volumes to the systems, it is very important to remember that the DS5000 storage subsystem uses a preferred controller ownership approach for communicating with LUNs. This means that every LUN is owned by only one controller. It is, therefore, important at the system level to make sure that traffic is correctly balanced between the two controllers. It is a fundamental principle for a correct setting of the storage subsystem.

Balancing traffic is unfortunately not always a trivial task. For example, if an application requires a large disk space to be located and accessed in one chunk, it becomes harder to balance traffic by spreading the smaller volumes among controllers. To this end many system's implement an internal volume management method to help with this balancing effort.

In addition, typically, the load across controllers and logical drives is constantly changing. The logical drives and data accessed at any given time depend on which applications and users are active during that time period, which is why it is important to monitor the system.

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

- Assign LUNs across all controllers to balance controller utilization.
  - With multiple arrays, assign all LUNs on an array to the same controller.
- Use the manual method of creating logical drives. This allows greater flexibility for configuration settings, such as enclosure loss protection and utilizing both drive loops.
- If you have highly used LUNs, where possible, spread them out on their own separate arrays. This will reduce disk contention for the arrays.

Enhanced Remote Mirror (ERM) considerations

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

For more information about ERM, see the *IBM System Storage DS Storage Manager Copy Services Guide*, SG24-7822.

2.2.9 Hot spares

The DS5000 storage subsystem provides global hot spare drives for use when a disk failure occurs. A hot spare drive is like a replacement drive installed in advance. Hot spare disk drives provide additional protection that might prove to be essential if there is a disk drive failure in a fault tolerant array.

The number and location of these spares is completely determine by you when you configure the system. Hot spares can cover any drive of the same type (FC, FC-SSD, SATA, SAS, SAS-SSD, and FDE or non-FDE capable) and can spare for any drive of equal or smaller size.

Therefore; it is suggested to consider using the largest size of each drive type in your configuration as a hot spare. This allows for all hot spares of a type to be completely global for their drive type, and can decrease the number of spares required for protection of the storage solution.

Here are some general guidelines to plan your hot spare coverage properly:

- Hot spare disk drives must be of the same media type and interface type as the disk drives that they are protecting.
- Hot spare disk drives must have capacities equal to or larger than the used capacity on the disk drives that they are protecting. The DS5000 storage subsystem can use a larger drive to recover a smaller failed drive to it. It will not use smaller drives to recover a larger failed drive. If a larger drive is used, the remaining excess capacity is blocked from use.
- FDE disk drives provide coverage for both security capable and non-security capable disk drives. Non-security capable disk drives can provide coverage only for other non-security capable disk drives:
  - For an array that has secured FDE drives, the hot-spare drive ought to be an unsecured FDE drive of the same or greater capacity.
  - For an array that has FDE drives that are not secured, the hot-spare drive can be either an unsecured FDE drive or a non-FDE drive.

For example, in a mixed disk environment that includes non-security capable SATA drives, non-security-capable Fibre Channel drives, and FDE Fibre Channel drives (with security enabled or not enabled), use at least one type of global hot-spare drive (FDE Fibre Channel and a SATA drive) at the largest capacity within the array. With a secure-capable FDE Fibre Channel drive of the largest size used, and a SATA hot-spare drive of the largest size in the system used for hot spares, all arrays can be protected.

Hot spares locations

Distribute the hot spare drives evenly across the different expansions of your storage subsystem, but avoid having multiple ones in a single enclosure. Because hot spare drives are in standby, without traffic or I/O until a drive fails, then you want to maximize the overall performance of your system by evenly distributing your production drives across the different expansions. At the same time, this avoids the risk of a single disk drive channel, or expansion enclosure failure, causing loss of access to all hot spare drives in the storage subsystem.

However, in some configurations, for example, a DS5000 storage subsystem with five expansions evenly distributed across the four different drive channels to maximize the traffic to each enclosure, you can choose to maximize performance over availability by having all the spares defined in the fifth expansion. This way, the channel with two expansions will not be penalized for excessive traffic, because the spares expansion will not contribute to the traffic load in that channel.

Important: Distribute your spare drives evenly across the different expansion to avoid the possibility of a general enclosure failure.

Quantity of drives as hot spares

The is no fixed rule about the quantity of disk drives to assign as hot spares. But as a general rule about disk usage and availability, we suggest defining a minimum of one of every 30 drives of a particular media and interface type, or one for every two fully populated enclosures. In large configurations with arrays containing numerous drives, the reconstruction of a failed drive to a hot spare drive can take a long time, proportional to the quantity of drives in the array and the size of the disks. If in addition to that time, you need to wait to have a new disk available onsite to replace the failed drive, then the probability of having another disk failure increases. Having multiple spare drives will not mitigate the reconstruction time for an array, but will provide protection for another array in case of a second disk failure.

Tip: There is no definitive suggestion about how many hot spares to install, but is best that a hotspare be configured for every 20 - 30 drives of a specific drive type. Be aware that Fiber Channel, SSD, SATA, SAS, and SAS SSD cannot spare for each other.

When possible, split the hot spares so that they are in separate enclosures and are not on the same drive loops (see Figure 2-9).

Best practice: When assigning disks as hot spares, make sure that they have enough storage capacity. If the failed disk drive is larger than the hot spare, reconstruction is not possible. Using the largest drive size in your subsystem for each type as your hot spares can allow for greater coverage of disk with fewer spares. Ensure that you have at least one of each size or all larger drives configured as hot spares.

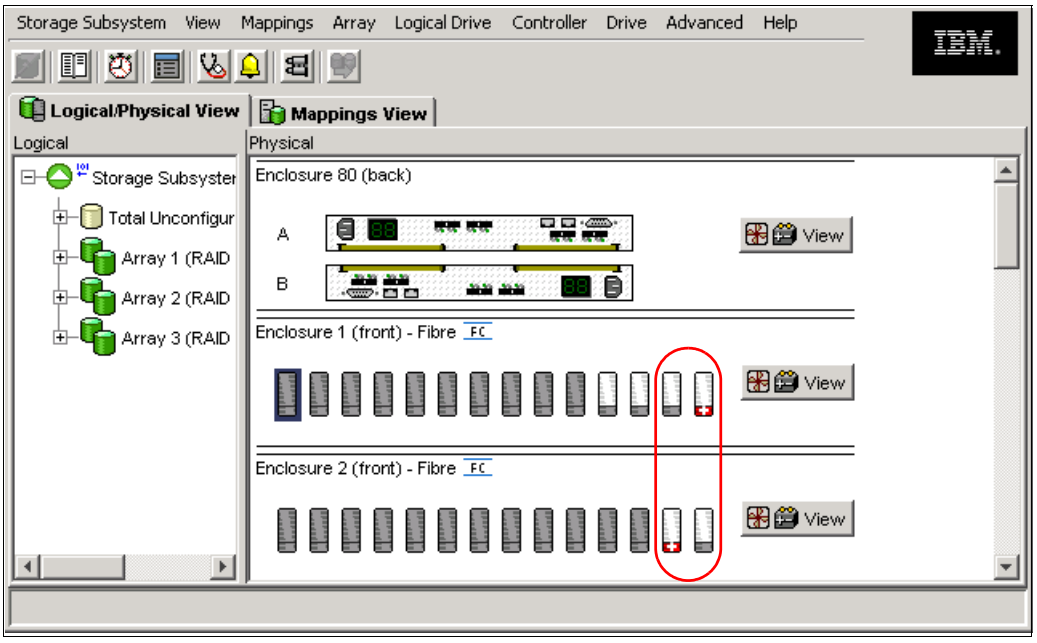


Figure 2-9 Hot spare coverage with alternating loops

2.2.10 Media scan

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

Media scan continuously runs in the background, using spare cycles to complete its work. The default media scan is for a scan every 30 days, that is, the maximum time media scan will need to complete the task. During the scan process, the DS5000 calculates how much longer the scan process will take to complete, and adjusts the priority of the scan to ensure that the scan completes within the time setting allocated. After the media scan has completed, it will start over again and reset its time for completion to the current setting. This media scan setting can be reduced, however if the setting is too low, priority will be given to media scan over host activity to ensure that the scan completes in the allocated time. This scan can impact on performance, but improve data integrity.

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

A media scan can be considered a surface scan of the hard drives, whereas a redundancy check scans the blocks of a RAID 3, 5, or 6 logical drive and compares it against the redundancy data. In the case of a RAID 1 logical drive, then the redundancy scan compares blocks between copies on mirrored drives.

We have seen no effect on I/O with a 30 day setting unless the processor is utilized in excess of 95%. The length of time that it will take to scan the LUNs depends on the capacity of all the LUNs on the system and the utilization of the controller. Figure 2-10 shows an example of default media scan settings.

Tip: If you change the media scan duration setting, the changes will not take effect until the current media scan cycle completes or the controller is reset.

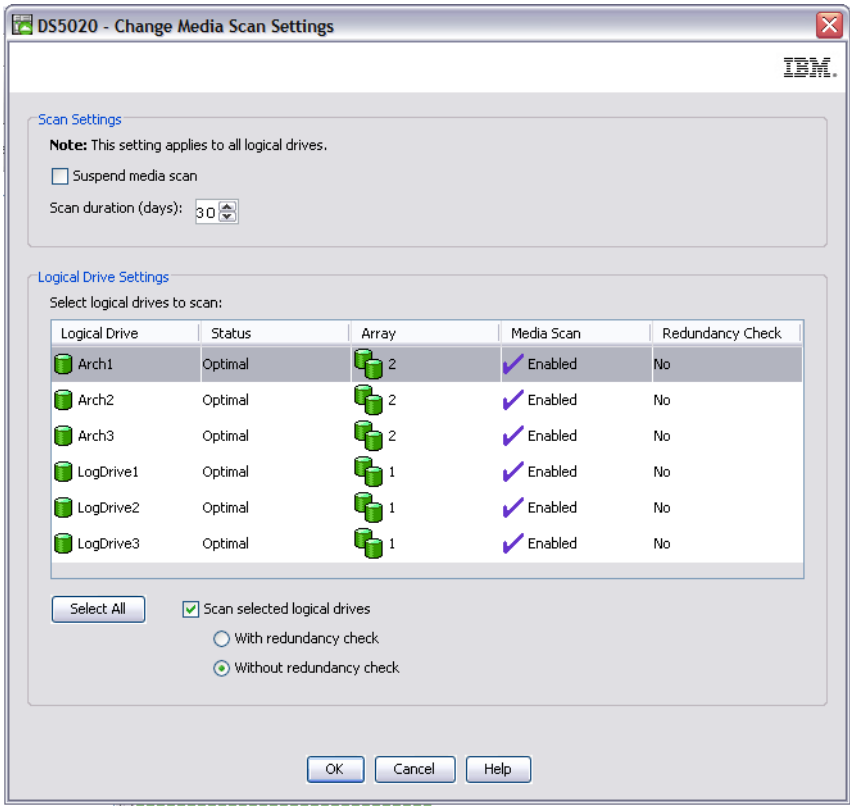


Figure 2-10 Default media scan settings at DS5000 storage subsystem

An example of logical drive changes to the media scan settings is shown in Figure 2-11.

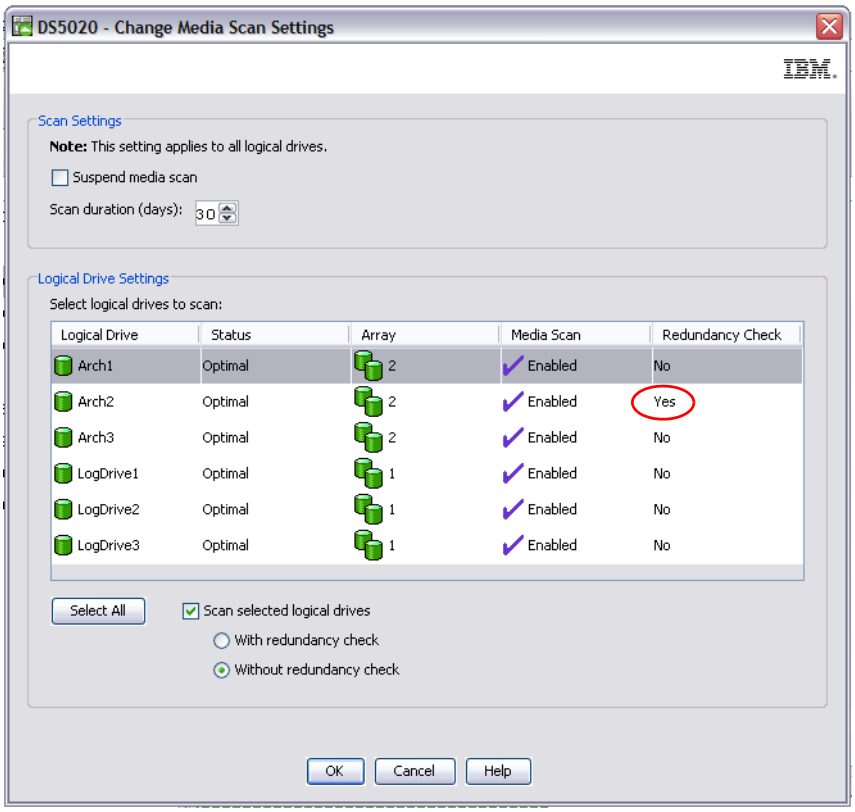


Figure 2-11 Logical drive changes to media scan settings

Tip: You cannot enable background media scans on a logical drive comprised of Solid State Disks (SSDs).

2.2.11 Cache parameters

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

Efficient use of the RAID controller cache is essential for good performance of the DS5000 storage subsystem.

Figure 2-12 shows a schematic model of the major elements of a disk storage subsystem, which are elements through which data moves (as opposed to other elements, such as power supplies). In the model, these elements are organized into eight vertical layers: four layers of electronic components shown inside the dotted ovals and four layers of paths (that is, wires) connecting adjacent layers of components to each other. Starting at the top in this model, there are some number of host servers (not shown) that connect (over some number of paths) to host adapters. The host adapters connect to cache components. The cache components, in turn, connect to disk adapters that, in turn, connect to disk drives.

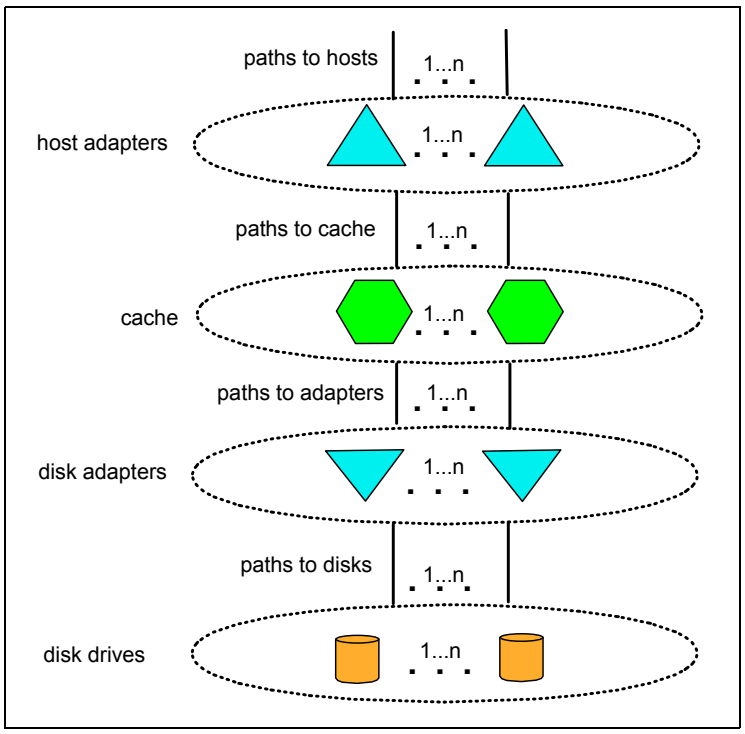


Figure 2-12 Conceptual model of disk caching

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

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

There are many different settings (related to caching) that come into play. The IBM System Storage DS® Storage Manager utility enables the following settings to be configured:

- DS5000 system wide settings (these settings will affect all host IO to arrays and logical drives created on the system)
- Start and stop cache flush levels: Percentage levels are 80/80, but are generally suggested to be set to 50/50:
  - Cache block size: Choices of 4 KB, 8 KB, 16 KB, and 32 KB, depending on host IO block size, which is prevalent with the critical applications.
  - The default level is 8 KB, and it works well for most small IO applications. Larger sizes might be needed for big host block size applications for high throughputs.

- Logical drive specific settings:
  - Read caching: Can be enabled or disabled; try to use cache for host read IO.
  - Dynamic read cache prefetch: Can be enabled or disabled; will read extra data into cache when sequential reads are detected from the host IO.
  - Write caching: Can be enabled or disabled; will write host IO to cache or write through to disks before sending an acknowledgement to the host.
  - Write cache mirroring: Can be enabled or disabled; will write host IO to cache and mirror the write to the cache of the second controller before it sends an acknowledgement to the host.

The default parameters configured in the Storage Manager can be used for many installations; however, you might want to modify these values for your specific environment. Select the defaults, or for specific configuration details, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023. As all of these settings are dynamically changeable, you can adjust as needed while testing your actual host operations with the storage.

When implementing a DS5000 storage subsystem as part of a whole solution, performance testing and monitoring to adjust the settings ought to be planned over at least a one week period.

2.3 Planning for premium features

The premium features that come with the DS5000 Storage Manager software are enabled by purchasing either a premium feature pack, or a premium feature license key for each feature capability. Not all premium features are available on all DS5000 storage subsystems. When planning for any of the premium features, it is a good idea to document what are the goals and rationale for purchasing the feature, which clearly defines from the outset what you want to achieve and why. See Table 2-3 for available premium features and their supported subsystem models.

Standard Included by default with the model of DS storage subsystem
Keyed Requires a premium feature key
N/A Not applicable to the model of the DS storage subsystem

Table 2-3 Premium Features list

Premium feature	DS5020	DS5100	DS5300
Performance Tier	Standard	Keyed	Standard
RAID-6	Standard	Standard	Standard
FDE Internal Key Management	Keyed	Keyed	Keyed
FDE External Key Management	Keyed	Keyed	Keyed
T10PI	Standard	Standard	Standard
SSD	Standard[a]	Standard[a]	Standard[a]
Max drive slot 64	Keyed[b]	N/A	N/A
Max drive slot 112	Keyed	N/A	N/A
Max drive slot 448	N/A	Keyed[c]	Standard

Premium feature	DS5020	DS5100	DS5300
Eight EXP5060's	N/A	Keyed[d]	Keyed
Storage Partitions[e] 4, 8, 16, 32, 64, 128	Keyed[f]	Keyed[g]	Keyed[g]
Storage Partitions 256, 512	N/A	Keyed	Keyed
Remote Mirroring 8, 16, 64	Keyed	Keyed	Keyed
Remote Mirroring 128	N/A	Keyed	Keyed
FlashCopy 4, 8	Keyed	Keyed	Keyed
FlashCopy 16	N/A	Keyed	Keyed
Volume copy	Keyed[h]	Keyed[h]	Keyed[h]

a. Currently maximum of 20 SSDs per subsystem.
b. Base is 32 drive slots.
c. Base is 256 drive slots.
d. Requires "Max drive slot 448" premium feature as a prerequisite.
e. 2 Storage partitions are included by default with the firmware
f. Minimum Storage Partition level is 4.
g. Minimum Storage Partition level is 8.
h. Must be ordered with one of the FlashCopy license keys.

2.3.1 Storage partitioning

Storage partitioning adds a high level of flexibility to the DS5000 storage subsystem. It enables you to connect multiple and heterogeneous host systems to the same storage server, either in stand-alone or clustered mode. The term storage partitioning is somewhat misleading, because it actually represents a host or a group of hosts and the logical drives they are assigned to access.

Without mapping to storage partitions, the logical drives configured on a DS5000 storage subsystem can only be accessed by a single host system or by a single cluster, which can lead to inefficient use of storage subsystem's hardware unless the use of the DS5000 storage subsystem is dedicated to a single host (for example, SVC attachment, where it is seen as a single host).

Storage partitioning, on the other hand, allows the creation of "sets", containing the hosts with their host bus adapters and the logical drives. We call these sets *storage partitions*. The host systems can only access their assigned logical drives, just as though these logical drives were locally attached to them. Storage partitioning adapts the SAN idea of globally accessible storage to the local-storage-minded operating systems.

Storage partitioning allows mapping and masks the logical drive or LUN (that is why it is also referred to as "LUN masking"), which means that after the logical drive is assigned to a host, it is hidden from all other hosts connected to the same storage server. Therefore, access to that logical drive is exclusively reserved for that host.

It is a good practice to configure storage partitioning prior to connecting multiple hosts. Operating systems such as Windows will write their signatures to any device it can access.

Heterogeneous host support means that the host systems can run various operating systems. But be aware that all host systems within a particular storage partition have unlimited access to all logical drives assigned to the partition. Therefore, file systems or disk structure on these logical drives must be compatible with host systems. To ensure this, it is best to run the same operating system on all hosts within the same partition. Certain operating systems might be able to mount foreign file systems.

Storage partition topology is a collection of topological elements (default group, host groups, hosts, and host ports) shown as nodes in the topology view of the mappings view. To map a logical drive or LUN to a specific host server or group of hosts, each component of the storage partition must be defined.

A storage partition contains several components:

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

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

A *host* is a single system that can be mapped to certain *logical drive(s)*.

The number of storage partitions that can be used is defined by the premium feature key that has been enabled on the DS5000 storage subsystem. See Table 2-3 on page 44 for the supported numbers of partitions for each of the DS5000 models.

A *host port* is the FC port of the host bus adapter (HBA) on the host system. The host port is identified by its world-wide name (WWN). A single host can contain more than one host port. If the servers are attached using full redundancy, each server will have two host bus adapters, that is, it needs two host ports within the same host system. It is possible to have a host with a single HBA, but for redundancy, it must be able to access both DS5000 controllers, which can be achieved by SAN zoning.

The DS5000 storage subsystem only communicates through the use of the WWN. The DS5000 storage subsystem is not aware of which host bus adapters are in the same server or in servers that have a certain relationship, such as a cluster. The host groups, the hosts, and their host ports reflect a logical view of the physical connections of the SAN, as well as the logical connection between servers, such as clusters.

With the logical setup defined as previously described, mappings are specific assignments of logical drives to particular host groups or hosts.

The storage partition is the combination of all these components. It ensures correct access to the various logical drives even if there are several hosts or clusters connected.

The default host group is a placeholder for hosts that are defined but have not been mapped. The default host group is also normally used only when storage partitioning is not enabled. If it is the case, then only one type of operating system must be sharing the logical drives.

Every unassigned logical drive is mapped to the undefined mappings group, which means that no host (or host port, to be precise) can access these logical drives until they are mapped.

With Storage Manager, it is possible to have up to 512 storage partitions on some models of DS5000 storage subsystems, which allows these storage subsystems to have storage capacity that is available to a great number of heterogeneous hosts, allowing for a great deal of flexibility and scalability.

For the maximum number of logical drives a host can have mapped in a storage partition, see Table 2-4.

Table 2-4 Maximum logical drives per host type

Operating system	Maximum number of logical drives supported per partition
Windows Server 2003, 2008	255
HP-UX	127
AIX	255
Linux	255

Also, for different DS5000 storage subsystem models, the number of storage partitions allowed depends on the premium feature licence that is purchased; and the fact that various models can support different maximum numbers.

Every mapping of a logical drive to a new host or host group creates a new storage partition. If additional logical drives are required for an existing host or host group, a new storage partition is not required. For example, a cluster with two nodes with redundant I/O paths gets configured as one host group with two hosts. Each host then has two host ports for redundancy, and several logical drives are mapped to this host group. All these components represent one storage partition. If another single host system is attached to the same DS5000 storage subsystem and other logical drives are mapped to that host, then another storage partition must be created for it. If a new logical drive is created and mapped to either the cluster or the single host, it uses an existing storage partition.

For a step-by-step guide, see *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Logical drives: There are limitations as to how many logical drives you can map per host. DS5000 series storage servers will support up to 256 logical drives (including the “access” logical drive) per partition (although there are also restrictions, depending on the host operating system) and a maximum of two partitions per host. Keep these limitations in mind when planning the DS5000 storage subsystem installation.

Storage partitioning considerations

In a heterogeneous environment, the “access” logical drive is mapped only to the default host group. If in-band management is being used, then the access logical drive must be mapped to the storage partition for the managing host.

In a security-sensitive environment, you can also assign the access logical drive to a particular storage partition and ensure host-based management access only through the servers in this storage partition. In this environment, you can assign a password to the DS5000 storage subsystem as well.

Tip: Each host with separately assigned storage will use a storage partition. Each host group with separately assigned storage will also use an additional storage partition.

In order to configure the storage partitioning correctly, you need the WWN of your host HBAs. Mapping is done on a WWN basis. Depending on your HBA, you can obtain the WWN either from the BIOS or QLogic SANsurfer tool if you have QLogic cards. Emulex adapters and IBM adapters for IBM System p and IBM System i® servers have a sticker on the back of the card. The WWN is also usually printed on the adapter itself or the box in which the adapter was shipped.

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

When planning your partitioning, keep the following considerations in mind:

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

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

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

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

Table 2-5 shows an example of a storage partitioning plan, which clearly shows the host groups, hosts, port names, WWN of the ports, and the operating systems used in that environment. Other columns can be added to the table for future references, such as HBA BIOS levels, driver revisions, and switch ports used, all of which can then form the basis of a change control log.

Table 2-5 Sample plan for storage partitioning

Host group	Host name	Port name	WWN	OS type
Windows 2003	Windows Host	MailAdp_A	200000E08B28773C	Windows 2003 Non-Clustered
		MailAdp_B	200000E08B08773C	
Linux	Linux_Host	LinAdp_A	200100E08B27986D	Linux
		LinAdp_B	200000E08B07986D	
IBM POWER6®	AIX_Host	AIXAdp_A	20000000C926B6D2	AIX
		AIXAdp_B	20000000C926B08	

Heterogeneous hosts

When implementing a DS5000 storage subsystem solution, a mixture of host servers with various operating systems can be used (clustered and non-clustered variants of the same operating systems). However, all logical drives in a single storage partition must be configured for the same operating system. Also, all hosts in that same storage partition must run the same defined operating system.

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

Delete the access logical drive (ID 31)

The DS5000 storage subsystem will automatically create a logical drive for each host attached (logical drive ID 31). This drive is used for in-band management, so if you do not plan to manage the DS5000 storage subsystem from that host, you can delete this logical drive, which will give you one more logical drive to use per host.

If you attached a Linux or AIX to the DS5000 storage subsystem, you need to delete the mapping of this access logical drive.

2.3.2 DS5000 copy services premium features

For a detailed understanding of the copy services premium features of the DS5000 storage subsystem, see the *IBM System Storage DS Storage Manager Copy Services Guide*, SG24-7822.

Consider the following requirements:

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

Then document all of these needs and requirements.

2.3.3 FlashCopy

A FlashCopy logical drive is a point-in-time image of a logical drive. It is the logical equivalent of a complete physical copy, but you can create it much more quickly than a physical copy. Additionally, it requires less disk space. In DS5000 Storage Manager, the logical drive from which you are basing the FlashCopy, called the base logical drive, must be a standard logical drive in the storage server. Typically, you create a FlashCopy so that an application (for example, an application to take backups) can access the FlashCopy and read the data while the base logical drive remains online and user-accessible.

Plan carefully with regard to the space available to make a FlashCopy of a logical drive, even though FlashCopy takes only a small amount of space compared to the base image.

2.3.4 VolumeCopy

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

A VolumeCopy creates a complete physical replication of one logical drive (source) to another (target) within the same storage subsystem. The target logical drive is an exact copy or clone of the source logical drive. VolumeCopy can be used to clone logical drives to other arrays inside the DS5000 storage subsystem.

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

2.3.5 Enhanced Remote Mirroring

The Enhanced Remote Mirroring (ERM) option is used for online, real-time replication of data between storage subsystems over a remote distance.

Operating modes for Enhanced Remote Mirroring

Enhanced Remote Mirroring (ERM), which was formerly named “Remote Volume Mirroring,” offers three operating modes:

- Metro mirroring:

  Metro mirroring is a synchronous mirroring mode. Any host write requests are written to the primary (local) storage subsystem and then to the secondary (remote) storage subsystem. The remote storage controller acknowledges the write request operation to the local storage controller, which reports a write completion to the host. This mode is called synchronous. The host application does not get the write request result until the write request has been executed on both (local and remote) storage controllers.

- Global copy:

  This mode copies a non-synchronous, remote copy function designed to complete write operations on the primary storage subsystem before they are received by the secondary storage subsystem. This capability is designed to prevent primary performance from being affected by wait time from writes on the secondary subsystem. Therefore, the primary and secondary copies can be separated by long distances. This function is appropriate for remote data migration, offsite backups, and transmission of inactive database logs at virtually unlimited distances.

- Global mirroring:

  This mode is a two-site remote data mirroring function designed to maintain a complete and consistent remote mirror of data asynchronously at virtually unlimited distances with virtually no degradation of application response time. Separating data centers by longer distances helps provide protection from regional outages. This asynchronous technique can help achieve better performance for unlimited distances by allowing the secondary site to trail in data currency a few seconds behind the primary site. With Global Mirror, currency can be configured to be as little as three to five seconds with respect to host I/O. This two-site data mirroring function is designed to provide a high-performance, cost-effective global distance data replication and disaster recovery solution.

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

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

Planning considerations for ERM

Keep in mind the following planning considerations:

- DS5000 storage subsystems (minimum of two)
- Type of links between sites
- Distances between sites (ensure that it is supported)
- Switches or directors used
- Redundancy
- Additional storage space requirements

Attention: ERM requires a dedicated *switched fabric* connection per controller to be attached to Host port 4 on both A and B controllers. This dedication is required at both ends of the ERM solution.

2.3.6 Obtaining the premium feature key

When you have purchased a premium feature, you are provided with the information that you will require in order to generate the required premium feature key for your DS5000 storage subsystem. You can generate the feature key file by using the premium feature activation tool that is located at the following website:

`http://www-912.ibm.com/PremiumFeatures/jsp/keyInput.jsp`

The key can then be added to your DS5000 system as detailed in *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

2.4 Planning your host attachment method

In this section, we review the different attachment methods available for the IBM Midrange System Storage DS5000 storage subsystem so you can evaluate the requirements needed in each case. You have these options to attach your hosts to the DS5000 storage subsystems:

- Fibre Channel:
  - Storage Area Network (SAN) attached
  - Direct attach
- iSCSI attach

2.4.1 Fibre Channel: SAN or direct attach

When planning the setup of a Storage Area Network (SAN), you want the solution to answer your current requirements and allow for expansion for your future needs.

First, the SAN fabric must be able to accommodate a growing demand in storage (it is estimated that storage needs double every two years). Second, the SAN must be able to keep up with the constant evolution of technology and resulting hardware upgrades and improvements. It is estimated that a storage installation needs to be upgraded every 2 to 3 years.

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

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

1. Produce a statement outlining the solution requirements that can be used to determine the type of configuration you need. Then use this statement to cross-check that the solution design delivers the basic requirements. The statement must have easily defined bullet points covering the requirements, for example:
   - New installation or upgrade of existing infrastructure
   - Host Bus Adapter (HBA) selection
   - HBA driver type selection: SCSIPort or StorPort for example.
   - Multipath Driver selection: RDAC, MPIO, DMMP, or SDDPCM

- Types of applications accessing the SAN (whether I/O intensive or high throughput)
- Performance measurements to be met
- Required capacity
- Required redundancy levels
- Type of data protection needed
- Current data growth patterns for your environment
- Whether application workload is more read or write based
- Backup strategies in use: Network, LAN-free, or Server-less
- Premium features required
- Number of host connections required, and types of connections
- Types of hosts and operating systems that will connect to the SAN
- Zoning required
- Distances between equipment and sites (if there is there more than one site)

2. Produce a hardware checklist. It must cover such items that require you to:
   - Make an inventory of existing hardware infrastructure. Ensure that any existing hardware meets the minimum hardware requirements and is supported with the DS5000 storage subsystem.
   - Make a complete list of the planned hardware requirements.
   - Ensure that you have enough rack space for future capacity expansion.
   - Ensure that the power and environmental requirements are met.
   - Ensure that your existing Fibre Channel switches and cables are properly configured.
3. Produce a software checklist to cover all the required items that need to be certified and checked. It must include such items that require you to:
   - Ensure that the existing versions of firmware and storage management software are up to date.
   - Ensure that host operating systems are supported with the DS5000 storage subsystem. For the latest in supported host information, see the IBM System Storage Interoperation Center website:

     http://www.ibm.com/systems/support/storage/config/ssic/index.jsp
   - These lists are not exhaustive, but the creation of the statements is an exercise in information gathering and planning; it gives you a greater understanding of what your needs are in your current environment and creates a clearer picture of your future requirements. The goal ought to be quality rather than quantity of information.

Use this chapter as a reference to help you gather the information for the statements.

Understanding the applications is another important consideration in planning for your DS5000 storage subsystem setup. Applications can typically either be I/O intensive, such as high number of I/O per second (IOPS); or characterized by large I/O requests, that is, high throughput or MBps.

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

By understanding your data and applications, you can also better understand growth patterns. Being able to estimate an expected growth is vital for the capacity planning of your DS5000 storage subsystem installation. Clearly indicate the expected growth in the planning documents: The actual patterns might differ from the plan according to the dynamics of your environment.

Selecting the right DS5000 storage subsystem model for your current and perceived future needs is one of the most crucial decisions you will make. The good side, however, is that the DS5000 offers scalability and expansion flexibility. Premium features can be purchased and installed at a later time to add functionality to the storage server.

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

2.4.2 Fibre Channel adapters

We now review topics related to Fibre Channel adapters:

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

Host system bus

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

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

It is always a best practice to distribute high-speed adapters across several buses. When you use PCI adapters, make sure that you first review your system specifications. Certain systems include a PCI adapter placement guide.

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

Failover

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

In general, all operating systems support two paths to the DS5000 storage subsystem. Microsoft Windows 2003 and 2008 and Linux support up to four paths to the storage controller. AIX can also support four paths to the controller, provided that there are two partitions accessed within the DS5000 storage subsystem. You can configure up to two HBAs per partition and up to two partitions per DS5000 storage subsystem.

Load balancing

Load balancing or load sharing means distributing I/O requests from the hosts between multiple adapters, which can be done by assigning LUNs to both the DS5000 controllers A and B alternatively (see also 2.2.8, “Logical drives and controller ownership” on page 37).

Figure 2-13 shows the principle for a load-sharing setup. A multipath driver checks all available paths to the controller. In Figure 2-13, that is two paths (red and blue). The driver forces the data down all paths in a *round-robin* scheme, which means that it does not really check for the workload on a single path, but moves the data down in a *rotational manner* (round-robin).

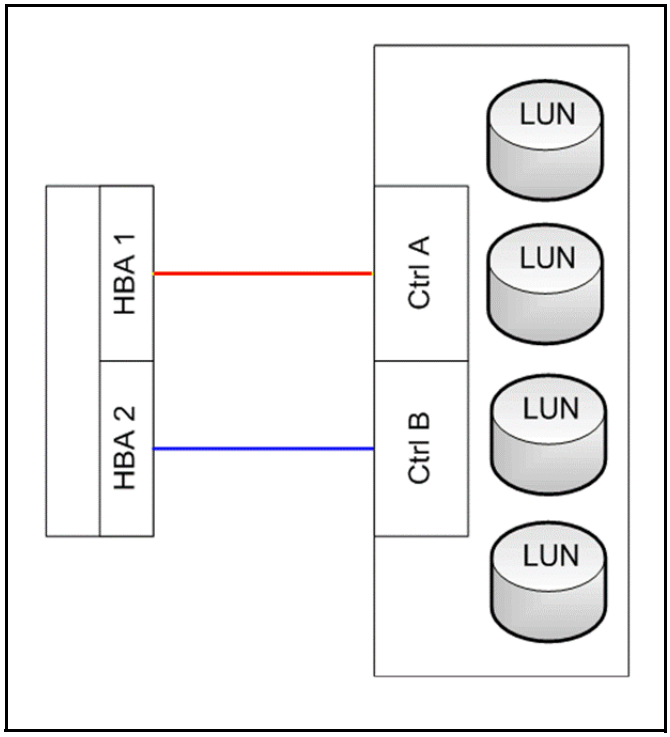


Figure 2-13 Load sharing approach for multiple HBAs

The RDAC drivers for Linux supports round-robin load balancing.

Tip: In a cluster environment, you need a single path to the each of the controllers (A and B) of the DS5000 storage subsystem. However, if the cluster software and host application can do persistent reservations, you can keep multiple paths and the multipath driver will route the I/O request using the appropriate path to the reserved logical drive.

Queue depth

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

For the latest firmware for the DS5000 controller, see the following website:

`http://www-03.ibm.com/systems/storage/disk/`

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

2.4.3 SAN zoning for the DS5000 storage subsystem

Zoning is an important part of integrating a DS5000 storage subsystem into a SAN. When done correctly, it can eliminate many common problems.

General considerations

Depending on your host operating system and the device driver you are planning to use, you need to zone your SAN differently. In AIX or Solaris using old RDAC, only one path per controller is supported, so you need to create a zone for the connection between the host bus adapter (HBA1) and controller A and a separate zone that contains the other HBA2 to controller B.

If using a multipath driver such as MPIO, Veritas DMP, and MPP or DMM in Linux, then you need to connect and zone your SAN to have multiple paths to the same controller. It is done by creating the additional target paths in the zone with a single initiator (HBA). This creates each zone in its best form, by keeping all the initiators isolated from each other while allowing for multiple paths to be available to the storage from each initiator.

Best practices:

- Connect your storage and create zones by considering the number of paths allowed to the same controller by your host operating system and the multipath driver being used.
- Make sure to have only one initiator per zone.

Important: Disk and tape should always be on separate HBAs, following the best practice for zoning; then the disk and tape access will also be in separate zones. With certain UNIX systems, the sharing of the HBA might be supported by the DS5000 storage subsystem, but hardware limitations might exist with the overall solution. Therefore, *do not share HBAs between disk storage and tape.*

Enhanced Remote Mirroring considerations

The Enhanced Remote Mirroring (ERM) Fibre Channel connection must be dedicated for data replication between the subsystems. The ports used for ERM cannot receive I/Os from any host. These requirements are addressed by defining special SAN zones. The zones must separate the host access ports from the DS5000 storage subsystem mirroring ports as well as separate the mirroring ports of the redundant controllers from each other.

When using Enhanced Remote Mirroring (ERM), create two separate dedicated zones:

- The first zone contains the ERM source DS5000 controller A and ERM target DS5000 controller A.
- The second zone contains the ERM source DS5000 controller B and ERM target DS5000 controller B.

Restriction: When using Enhanced Remote Mirroring, the last Fiber Channel host port on each controller is dedicated to only remote mirroring traffic, and must be zoned away from all host ports.

ERM is detailed further in *IBM System Storage DS Storage Manager Copy Services Guide*, SG24-7822.

2.4.4 iSCSI connection to the DS5000 storage subsystem

In addition to using Fibre Channel as the interface connection method, iSCSI host interfaces allows servers with specialized hardware iSCSI cards, or with software running over regular Ethernet cards, run the iSCSI protocol as the connection method to attach the DS5000 storage subsystems. As with FC, before beginning an iSCSI deployment, plan in advance and understand and document the network topology that will be used.

Connecting the DS5000 storage subsystem to host by iSCSI is supported over both 1 Gbps and 10 Gbps networks.

With the level of training in the IP environment being high, many users are using this technology to simplify their storage networking needs for their environments. Therefore, there are many host types that are supported and with a variety of host bus adapters to chose from; and the list is continuously growing. For the latest in host types, and bus adapters supported see the IBM System Storage Interoperation Center website:

http://www.ibm.com/systems/support/storage/config/ssic/

When planning for a network configuration involving QLogic HBA of both Fiber Channel or iSCSI, it is suggested that you use SANsurfer to assist with the installation and management of your connections. This utility supports the QLA4xxx HBA for iSCSI along with the QLA2xxx for Fiber Channel. For greater detail on the supported HBAs for this tool see:

http://driverdownloads.qlogic.com/QLogicDriverDownloads_UI/Product_detail_new.aspx?oemid=344&companyid=6

For more detailed information about setting up the host with iSCSI, see Chapter 4, “Host configuration guide” on page 151.

Partitioning challenges

As described in 2.3.1, “Storage partitioning” on page 45, because the FC and iSCSI protocols provide radically different latency and throughput capabilities, and this mixture within a server might cause failover driver conflict, performance degradation, or potential data loss, you need to plan ahead by using the following suggestions:

- Define separated partitions for FC-based hosts from iSCSI-based hosts, avoid mixing them in same storage partition.
- A single host should not be configured for both iSCSI connections and FC connections to the DS5000 storage subsystem.

In order to define hosts and partitions in iSCSI, remember to plan for iSCSI addresses for the host ports, and use the iSCSI qualified name (IQN) of the host you want to map.

Network settings

Unlike traditional Fibre Channel, which requires special cabling and SAN switches, iSCSI can be run over an existing network infrastructure. However, in complex network environments, and in order to protect the integrity of the data in your DS5000 storage subsystem, and its continuous access, we suggest that, whenever possible, to try to isolate the iSCSI traffic in a dedicated network. The iSCSI multipathing architecture provides failover to the alternate controller in the event of an outage situation. Also with MPIO, IBM provides the DSM, which also offers load-balancing algorithms.

For better redundancy, you can increase the availability of your connections using redundant networks, so a failure in one does not interrupt the remaining redundant connection.

Aside from the basic iSCSI connectivity parameters, such as IP address per target Ethernet port and associated iSCSI qualified names, you need to plan in advance several optional configuration parameters, including ennoblement of jumbo frames, configuration of a VLAN, and setting a specific Ethernet priority:

Jumbo Frames Jumbo frames are created when the MTU is adjusted above 1500 bytes per frame; they are set by port. The frame sizes supported are of 1500 and 9000 bytes. When using jumbo frames, ensure that all of the devices on your iSCSI network, including switches, initiators, and targets, are configured to use the same maximum jumbo frame size.

VLAN As previously mentioned, we suggest, for performance and availability reasons, having separated networks for redundant interfaces. If it is not possible to segregate an iSCSI storage subsystem onto a physically separate LAN, with the IBM DS5000 storage subsystems that are connected by iSCSI, you can use VLANs to maximize the potential performance.

Ethernet priority Ethernet priority, sometimes referred to as quality of service or class of service, is supported in the DS5000 storage subsystems. You can set the Ethernet priority of the target iSCSI interfaces to increase the class of service received within the network itself.

For more details on how to set up a host in an iSCSI network, see Chapter 4, “Host configuration guide” on page 151.

Security

Unlike FC SANs, Ethernet networks can be more open, so in order to provide additional security, you can configure the following additional authentication protocols on the DS5000 storage subsystems:

- The Internet Storage Name Service (iSNS) protocol allows for automated discovery, management, and configuration of iSCSI devices on a TCP/IP network. iSNS servers offer additional security services through explicitly defined initiator-to-target mappings and simplified asset locators, similar to that provided by DNS and WINS for IP address lookup facilities
- Challenge Handshake Authentication Protocol (CHAP) provides an additional security layer within the iSCSI SAN on the IBM DS5000 storage subsystems.

2.5 Host support and multipathing

The intent of this section is to list the most popular supported operating system platforms and topologies used on the DS5000 storage subsystem. See the IBM Storage System Interoperation Center (SSIC) for a complete up to date list of compatible hosts, host bus adapters, and their driver version levels:

http://www-01.ibm.com/systems/support/storage/config/ssic/index.jsp

Here we describe the available multipathing drivers and their supported operating systems.

2.5.1 Supported server platforms

The following server platforms are supported:

- IBM System x
- IBM System p
- IBM System i
- IBM Power Systems™
- IBM BladeCenter®
- HP Compatible servers
- AMD and Intel Compatible servers
- Sun Compatible servers

2.5.2 Supported operating systems

At the time of publication, the following operating systems are supported:

- Microsoft Windows Server 2003 SP2
- Microsoft Windows Server 2008 R2, R2SP1 SP1, SP1R1, SP2
- Red Hat Enterprise Linux 4.7, and 4.8
- Red Hat Enterprise Linux 5.3, 5.4, 5.5, and 5.6
- Red Hat Enterprise Linux 6
- Novell SUSE SLES 9 SP4
- Novell SUSE SLES 10 SP2 - SP4
- Novell SUSE SLES 11 and 11 SP1
- VMware ESX 3.5 U3, U4, and U5
- VMware vSphere/ESX 4.0, 4.0 U1, 4.1, 4.1 U1, and 5.0
- IBM AIX 5L™ V5.1(ML9), V5.2 and 5.2(TL9 and TL10), V5.3 and V5.3 (TL5 - TL12), V6.1 and V6.1 (TL1 - TL6), and V7.1
- IBM i V6.1.1, and V7.1
- By VIOS guest client: DS5020, DS5100, and DS5300
- Native IBM i attach: DS5100 and DS5300
- IBM VIOS V1.5.2, V2.1, V2.1.1, V2.1.2, V2.1.3, and V2.2
- HP-UX 11iv2 (11.23) and HP-UX 11iv3 (11.31)
- Sun Solaris 10 u8 and 10 u9storage subsystem
- MacOS 10.6.x and later

2.5.3 Clustering support

At the time of publication, the following clustering services are supported:

- IBM PowerHA® (formerly HACMP)
- Microsoft Cluster Services
- Microsoft Windows Failover Clustering
- Novell Cluster Services
- Red Hat Cluster Suite
- Oracle Sun Cluster
- SIOS Lifekeeper
- HP Service Guard
- Symantec VERITAS Cluster
- Veritas Storage Foundation for Oracle RAC
- VMware Fault Tolerance (FT)
- VMWare High Availability (HA)

Tip: Make sure to check the System Storage Interoperation Center, found at the following website, for the current supported operating systems and clustering environments:

`http://www.ibm.com/systems/support/storage/config/ssic/index.jsp`

2.5.4 Multipathing

IBM offers various multipath drivers that you can use with your DS5000 storage subsystem. Only one of these drivers is required. Each driver offers multipath support, I/O load balancing, and automatic path failover.

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

When you create a logical drive, you assign one of the two active controllers to own the logical drive (called *preferred controller ownership,* as described in 2.2.8, “Logical drives and controller ownership” on page 37) and to control the I/O between the logical drive and the application host along the I/O path. The preferred controller normally receives the I/O requests from the logical drive. If a problem along the data path (such as a component failure) causes an I/O to fail, the multipath driver issues the I/O to the alternate controller.

A multipath device driver is not required when the host operating system has its own mechanism to handle multiple I/O paths, but if not, you need to install the supplied multipath device driver, even if your server does not has multiple paths.

Table 2-6 shows the DS5000 storage subsystem interoperability of multipathing drivers.

Table 2-6 DS5000 storage subsystem interoperability of multipathing drivers

	Windows 2003/2008	**RHEL**	**SLES**	**AIX 5L V5.3**	**AIX V6.1**	**HP-UX 11.23**	**HP-UX 11.31**	**MacOS 10.6.x**
Linux RDAC[a]		X	X					
Linux DMM[b]		X	X					
Windows MPIO	X							
AIX MPIO				X[c]	X[d]			
AIX SDDPCM				X	X			
LVM (HP-UX)						X	X	
SDD for HP-UX						X		
ATTO Multipath Director								X

a. Supported with Red Hat 5 or earlier and SLES 10 or earlier
b. Supported with Red Hat 6 or later and SLES 11 or later
c. Included in the operating system
d. Included in the operating system

Tip: See the IBM Storage System Interoperation Center (SSIC) for a complete up to date list of compatible drivers and their version levels:

`http://www-01.ibm.com/systems/support/storage/config/ssic/index.jsp`

Load balancing policy

When you have multiple paths to the DS5000 storage subsystem, the driver can select between different options about how to manage the traffic between the host and storage. The options available for the different operating systems are shown in Table 2-7.

Table 2-7 Load Balancing Policies per OS

Operating system	Multipath driver	Load balancing policy
AIX	MPIO	Round robin Selectable path priority
Red Hat Enterprise Linux 4 Update 7	MPP	Round robin Least queue depth
Solaris	MPxIO	Round robin
SUSE Linux Enterprise 9 Service Pack 4	MPP	Round robin Least queue depth
Windows	MPIO	Round robin Least queue depth Least path weight

One path per controller: Failover

Failover is not a balancing algorithm, but we include it here because it is an option that is presented with the actual balancing policies in some operating systems, such as Windows. The driver uses failover only when there is one path to each controller of your system. Using failover, one of the device paths is active, the one corresponding to the controller who owns the logical volume, and the other is in a standby state. It is the default for RDAC drivers (AIX and Solaris), because it only supports two paths, that is, one to each controller. Make sure to zone accordingly when working with RDAC so there is not more than one path per controller.

Multiple paths per controller

The multi-path driver transparently balances I/O workload without administrator intervention, across multiple paths to the same controller, but not across both controllers. To make use of the feature, you need to cable and zone your SAN properly. The load balancing policy uses one of three following algorithms:

- Round robin with subset:

  The round robin with subset I/O load balance policy routes I/O requests, in rotation, to each available data path to the controller that owns the volumes. This policy treats all paths to the controller that owns the volume equally for I/O activity. Paths to the secondary controller are ignored until ownership changes. The basic assumption for the round robin policy is that the data paths are equal. With mixed host support, the data paths might have different bandwidths or different data transfer speeds.

- Least queue depth with subset:

  The least queue depth with subset policy is also known as the least I/Os or least requests policy. This policy routes the next I/O request to a data path that has the least outstanding I/O requests queued. For this policy, an I/O request is simply a command in the queue. The type of command or the number of blocks that are associated with the command are not considered. The least queue depth with subset policy treats large block requests and small block requests equally. The data path selected is one of the paths in the path group of the controller that owns the volume.

- Least path weight with subset:

  The least path weight with subset policy assigns a weight factor to each data path to a volume. An I/O request is routed to the path with the lowest weight value to the controller that owns the volume. If more than one data path to the volume has the same weight value, the round-robin with subset path selection policy is used to route I/ O requests between the paths with the same weight value.

2.5.5 Microsoft Windows MPIO

This section describes the available Windows multipath options. MPIO is a Driver Development Kit (DDK) from Microsoft for developing code that manages multipath devices. It contains a core set of binary drivers, which are installed with the DS5000 Device Specific Module (DSM) to provide a transparent system architecture that relies on Microsoft Plug and Play to provide LUN multipath functionality while maintaining compatibility with existing Microsoft Windows device driver stacks.

Tip: The MPIO Driver is included in the Storage Manager software package for Windows and supports Microsoft Windows 2003 and 2008 on 32-bit and x64 systems. In Windows 2008, MPIO is already part of the operating system.

The MPIO driver performs the following tasks:

- Detects and claims the physical disk devices presented by the DS5000 storage subsystems based on vendor/product ID strings and manages the logical paths to the physical devices.
- Presents a single instance of each LUN to the rest of the Windows operating system.
- Provides an optional interface through WMI for use by user-mode applications.
- Relies on the vendor's (IBM) customized Device Specific Module (DSM) for information about the behavior of storage subsystem devices for the following items:
  - I/O routing information
  - Conditions requiring a request to be retried, failed, failed over, or failed back (for example, vendor-unique errors)
  - Handles miscellaneous functions, such as release/reservation commands
- Multiple Device Specific Modules (DSMs) for different disk storage subsystems can be installed in the same host server.

See 2.5.10, "Auto Logical Drive Transfer feature" on page 65 for more details about multipathing and failover considerations.

2.5.6 AIX MPIO

With Multiple Path I/O (MPIO), a device can be uniquely detected through one or more physical connections, or paths. A path-control module (PCM) provides the path management functions.

An MPIO-capable device driver can control more than one type of target device. A PCM can support one or more specific devices. Therefore, one device driver can be interfaced to multiple PCMs that control the I/O across the paths to each of the target devices.

The AIX PCM has a health-check capability that can be used to do the following actions:

- Check the paths and determine which paths are currently usable for sending I/O.
- Enable a path that was previously marked failed because of a temporary path fault (for example, when a cable to a device was removed and then reconnected).
- Check currently unused paths that will be used if a failover occurred (for example, when the algorithm attribute value is failover, the health check can test the alternate paths).

MPIO is part of the AIX operating system and does not need to be installed separately. You can find more information about MPIO in 4.4, "AIX configuration" on page 190.

2.5.7 AIX Subsystem Device Driver Path Control Module (SDDPCM)

SDDPCM is a loadable path control module for supported storage devices to supply path management functions and error recovery algorithms. When the supported storage devices are configured as Multipath I/O (MPIO) devices, SDDPCM is loaded as part of the AIX MPIO Fibre Channel Protocol (FCP) device driver during the configuration. The AIX MPIO-capable device driver with the supported storage devices SDDPCM module enhances data availability and I/O load balancing.

2.5.8 Linux: RHEL/SLES

With the Linux operating systems, there are now two drivers that you can choose. With earlier releases of the operating system, only the Redundant Disk Array Controller (RDAC) was supported. With the newer operating system releases, a native DMM was added and has been added to the supported multipathing driver list to choose from.

Linux: RHEL/SLES RDAC

The Redundant Disk Array Controller (RDAC), also known as Multi-Path Proxy (MPP), is the suggested multipathing driver for Linux based operating systems like Red Hat Enterprise Linux (RHEL) or SUSE Linux Enterprise Server (SLES).

The current RDAC driver implementation performs the following tasks:

- Detects and claims the physical devices (LUNs) presented from the DS5000 storage subsystems (*hides* them) based on vendor/product ID strings and manages all of the paths to the physical devices.
- Presents a single instance of each LUN to the rest of the Linux operating system components.
- Manages all of the plug and play interactions.
- Provides I/O routing information.
- Identifies conditions requiring a request to be retried, failed, or failed over.
- Automatically fails over the LUNs to their alternate controller when detecting problems in sending I/Os to the LUNs in their preferred controller, and fails back the LUNs to their preferred controller when detecting the problems in the preferred path fixed.
- Handles miscellaneous functions, such as persistent reservation translation.
- Uses a round robin (load distribution or load balancing) model.

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

- It is much more transparent to the OS and applications.
- I/O controls at the HBA driver level are not as tightly coupled to the OS as those at the disk driver level. Consequently, it is easier to implement I/O control functionality in the MPP-based RDAC driver for routing functions.

As the driver is positioned at the HBA level (see Figure 2-14), it has access to the SCSI command and sense data interface of the HBA driver and therefore can make more informed decisions about what to do in the case of path failures.

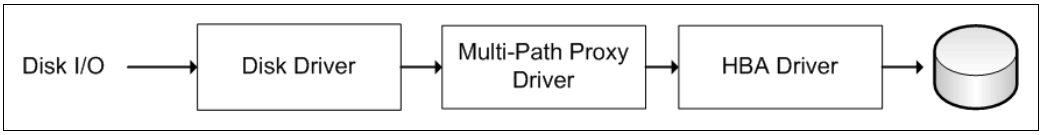


Figure 2-14 Linux RDAC/MPP driver

Tip: In Linux, RDAC cannot be installed with the Installation Wizard. If you need RDAC, you need to download and install it separately.

Linux RHEL/SLES Device Mapper Multipath (DMM)

Starting with Red Hat 6 and SLES 11, the IBM DS5000 storage subsystems now support the use of the Linux internal DMM driver to provide multipath capabilities. This driver includes the DS5000 DSM driver to provide the interface to the DS5000 storage subsystem. With this driver, multiple paths can be used to provide access to the defined logical drives for the host allowing for the use of load balancing policies. For details on how to set up and use this multipath driver, see the *IBM System Storage DS Storage Manager Version 10 Installation and Host Support Guide*, GA32-0963, available at the following website:

http://www-947.ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5075652&brandind=5000028

2.5.9 Apple MacOS

With the new release of the DS5000 storage subsystem's firmware 7.77.x, there is now support available for the Apple MacOS v 10.6. This support requires one of the following host bus adapters be used to connect the host to the DS5000 though a Fiber Channel connection:

- FC-81EN (8Gb single channel adapter)
- FC-82EN (8Gb dual channel adapter)
- FC-84EN (8Gb quad channel adapter)

All of these adapters require the use of the *ATTO Multipath Director* to be able to support the DS5000 storage subsystem's logical drives on the MacOS host. This driver supports the DS5000 in both failover and load balancing operations.

When setting up the HBA and driver parameters, you need to use the ATTO Configuration Tool. This tool also provides monitoring and management capabilities for multiple host connections to the storage.

There are a number of configurations for use with the MacOS connections to the DS5000 storage subsystems as shown next. The main focus is to ensure that you have the level of redundancy that your environment needs:

- Direct attached MacOS configuration: This configuration, shown in Figure 2-15, offers path failover through the dual port HBA or the single port HBA cards in the MacOS host.

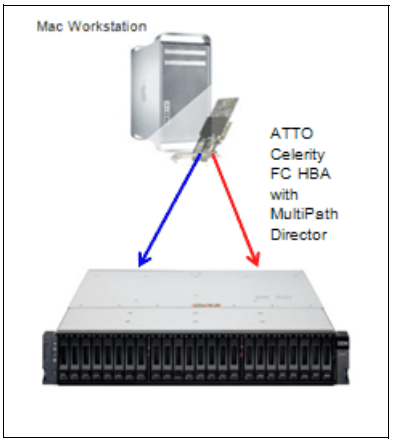


Figure 2-15 MacOS support with direct attached DS5000

Basic SAN fabric support: This configuration, shown in Figure 2-16, offers redundant paths for access and failover through use of a dual HBA to two fabrics connecting the MacOS host and DS5000 storage subsystem.

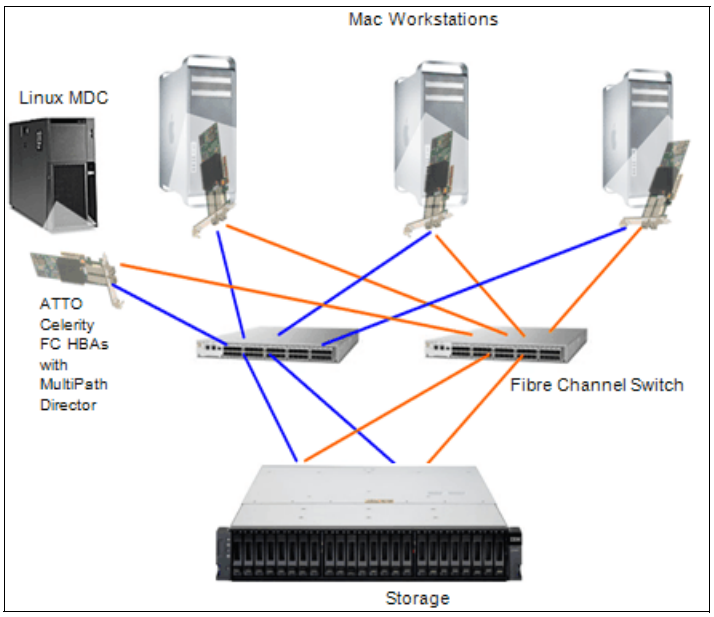


Figure 2-16 Basic SAN fabric with dual failover paths using dual ported HBA for MacOS host

- Redundant SAN fabric support: This configuration, shown in Figure 2-17, offers redundant paths for failover through dual fabrics and using single port HBA cards in the MacOS host.

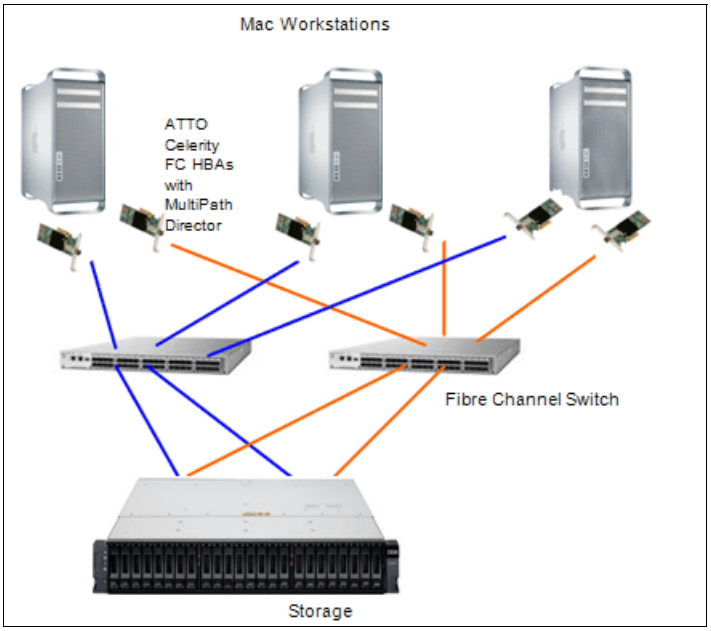


Figure 2-17 MacOS support with redundant SAN fabric

- High performance and redundant path protection: This feature provides a high number of access paths as well as failover protection by using the four port HBA with two ports to each of the fabrics, allowing greater load balancing as well as failover protection. See Figure 2-18 for an example configuration.

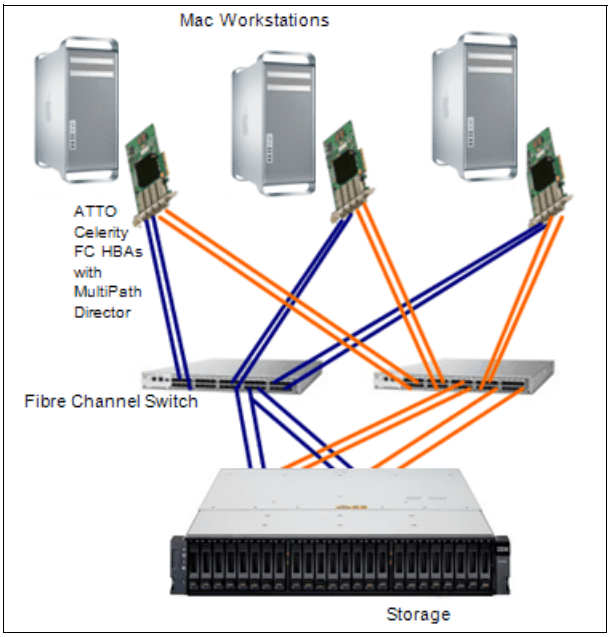


Figure 2-18 High Performance and redundant configuration for MacOS

To use this host kit, you need to download the following files from the locations given here:

- MultiPath Director Driver for MacOS 10.5x - 10.6.x: Driver for Celerity HBA listed above
- MacOS Flash Bundle: Firmware file for the HBA to work with MacOS (10.5.x - 10.6.x)
- MacOS Configuration Tool: ATTO Configuration Tool for MacOS 10.5.x - 10.6.x

For details on the setup and use of this environment, see this website:

`http://www.attotech.com/solutions/IBM/`

There is no support for the DS Storage Manager software to be installed on the MacOS host. Therefore, when using the DS5000 storage subsystem with a MacOS host, you need to have either a Windows or a Linux server available to be the management server for the storage subsystem.

2.5.10 Auto Logical Drive Transfer feature

In a DS5000 storage subsystem, you can provide redundant I/O paths with the host systems. There are two different components that provide this redundancy:

- A multipath driver
- Auto Logical Drive Transfer (ADT)

Auto-Logical Drive Transfer (ADT) is a built-in feature of controller firmware that allows logical drive-level failover rather than controller-level failover. Depending on the attached host, the feature will allow auto-volume transfer based on the characteristics of the operating system and the driver used. You can also enable or disable this feature.

Tip: ADT is not a failover driver. ADT provides storage subsystems with the flexibility to work with some third-party failover software.

Next, we describe the two modes.

- ADT-disabled failover:

  The multi-path software sends a SCSI Mode Select command to cause a change in volume ownership before using the alternate path. All logical drives on the preferred controller are transferred to the alternate controller. It is the configuration setting for Microsoft Windows, IBM AIX, IBM i5, Solaris, VMWare, and Linux (non-AVT) systems. When ADT is disabled, the I/O data path is still protected as long as you use a multi-path driver. After the I/O data path problem is corrected, the preferred controller does not automatically reestablish ownership of the logical drive. You must open a storage management window, select **Redistribute Logical Drives** from the Advanced menu, and perform the Redistribute Logical Drives task.

  Figure 2-19 shows the ADT-disabled failover mode phases.

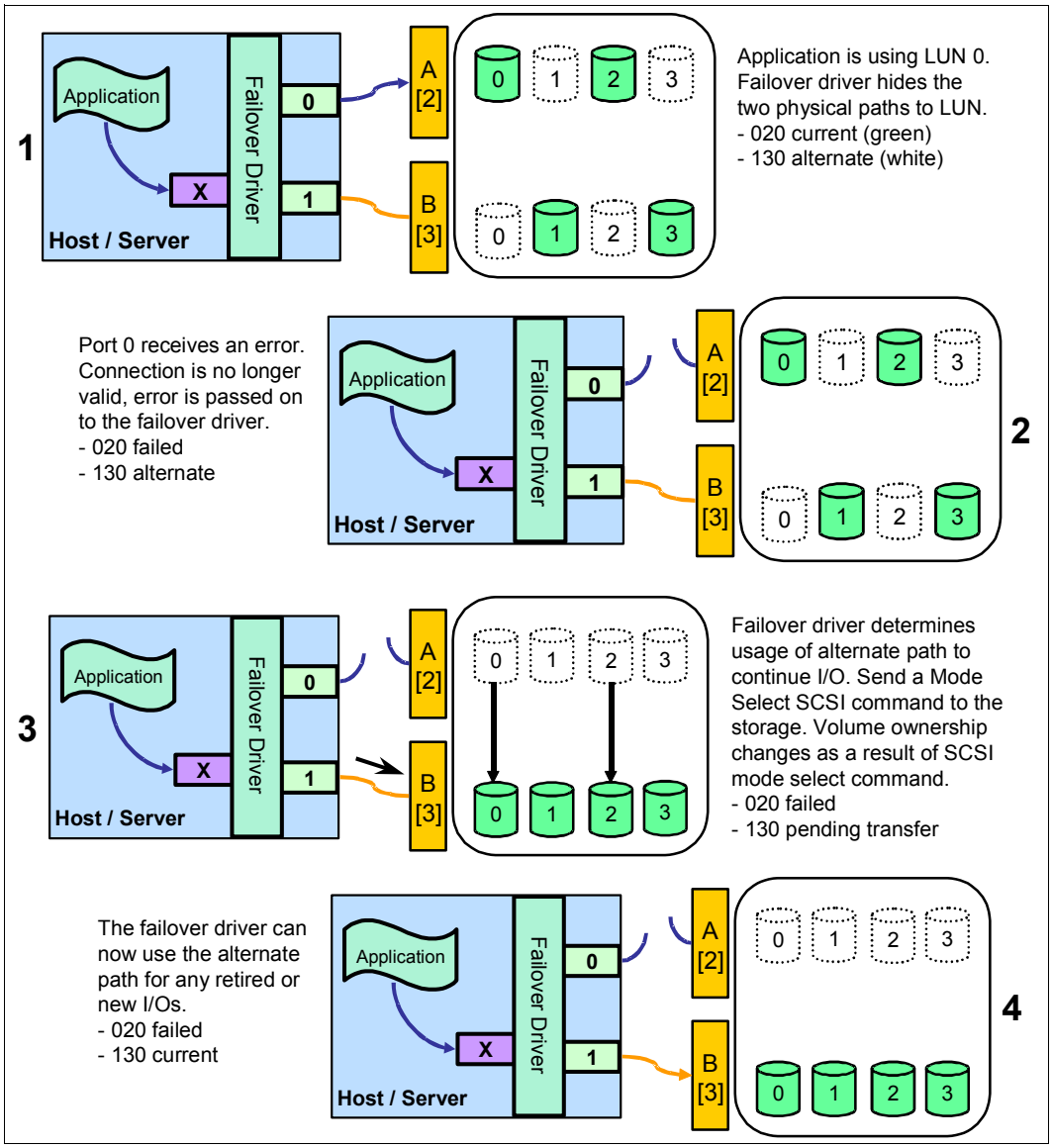


Figure 2-19 ADT-disabled mode path failover

Attention: In ADT-disabled mode, you are required to issue a redistribution command manually to balance the LUNs across the controllers.

- ADT-enabled failover:

  The multi-path driver starts using the alternate path by sending the I/O down the path it chooses and lets the ADT react. It is the normal configuration setting for host with Veritas DMP, and HP-UX systems. After the I/O data path problem is corrected, the preferred controller automatically reestablishes ownership of the logical drive as soon as the multipath driver detects that the path is normal again.

  Figure 2-20 shows the phases of failover in the ADT-enabled case.

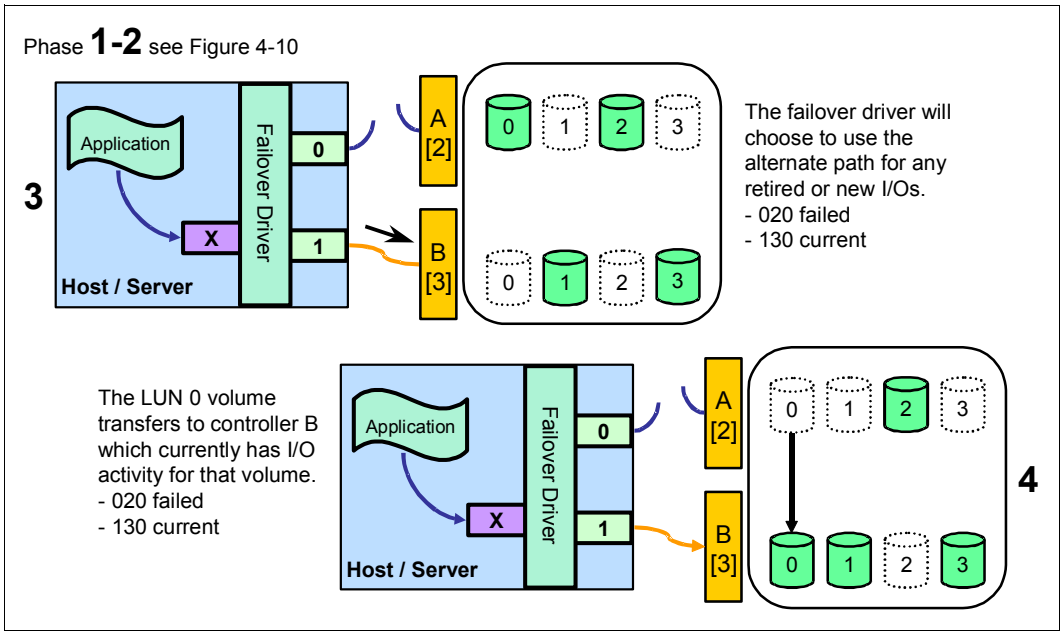


Figure 2-20 ADT-enabled mode path failover

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

2.5.11 Virtualization

With the growth and popularity of storage area networks, storage environments are getting more and more complex. Storage virtualization reduces the complexity and costs of managing storage environments and optimizes storage utilization in a heterogeneous environment.

The IBM Storage Area Network Volume Controller (SVC) and the new IBM Storwize V7000 storage subsystem can be used with IBM TotalStorage Productivity Center products to address these virtualization needs.

Important: In this section, we describe the SAN Volume Controller (SVC) as an example. However, the Storwize V7000 is capable of performing these functions in the same manner and can be considered as an equal replacement in this description.

IBM System Storage SAN Volume Controller and Storwize V7000

The IBM System Storage SAN Volume Controller is a scalable hardware and software solution to allow aggregation of storage from various disk subsystems. It provides storage virtualization and a consistent view of storage across a Storage Area Network (SAN).

The IBM Storwize V7000 is a new virtual storage subsystem with the abilities to perform as a SVC but with internal storage available with it as well. When the Storwize V7000 is used to virtualize the DS5000 storage subsystem as a backend storage, it behaves in the same manner as the SVC and therefore, the procedures and configuration suggestions are the same as described here for the SVC.

For details on the IBM Storwize V7000, see *Implementing the IBM Storwize V7000 V6.3*, SG24-7938.

The SAN Volume Controller provides in-band storage virtualization by creating a pool of managed disks from attached back-end disk storage subsystems. These managed disks are then mapped to a set of virtual disks for use by various host systems.

In conjunction with the DS5000 storage subsystem family, the SAN Volume Controller (SVC) can increase the storage copy services functionality and also the flexibility of SAN-based storage. The SVC is very flexible in its use. It can manage all host storage requirements or just part of them. A DS5000 storage subsystem can still be used to allocate storage to hosts or use the SVC, which is dependent upon various needs and requirements.

The SVC also offers an alternative to FlashCopy, VolumeCopy, and Enhanced Remote Mirroring for disaster recovery, high availability, and maintenance. If the use of SVC with a DS5000 storage subsystem is planned, then these premium features will not be required.

The SVC can also reduce the requirement for additional partitions. The SVC only consumes one storage partition. If you plan to use the SVC for all of your hosts, then a storage partition upgrade might not be required.

SVC is licensed by the capacity that is being managed. This capacity also includes the capacity used by the copy services.

Virtualization references

For detailed information about SVC implementation, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

For more information about Storage Virtualization, see the IBM TotalStorage Virtualization website:

`http://www.ibm.com/servers/storage/software/virtualization/index.html`

For more information about SAN Volume Controller, see its home page at this website:

`http://www.ibm.com/servers/storage/software/virtualization/svc/index.html`

For more information about IBM Storwize V7000, see its home page at this website:

`http://www-03.ibm.com/systems/storage/disk/storwize_v7000/index.html`

2.6 Additional host planning considerations

In this section, we review additional elements to consider when planning your DS5000 storage subsystems for use with a Logical Volume Manager (LVM) or virtualization options.

2.6.1 Planning for systems with LVM: AIX example

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

The Logical Volume Manager controls disk resources by mapping data between a simple and flexible logical view of storage space and the actual physical disks. The Logical Volume Manager does this by using a layer of device driver code that runs above the traditional physical device drivers. This logical view of the disk storage is provided to applications and is independent of the underlying physical disk structure. Using the Logical Volume Manager to create logical volumes that are striped across multiple physical devices can aid in performance and provide higher I/O handling capabilities for heavy transaction workloads. It is accomplished by the spreading of the request across multiple sets of disks through each of the physical device members.

Figure 2-21 illustrates the layout of the components in a volume manager using AIX Logical Volume Manager as our example.

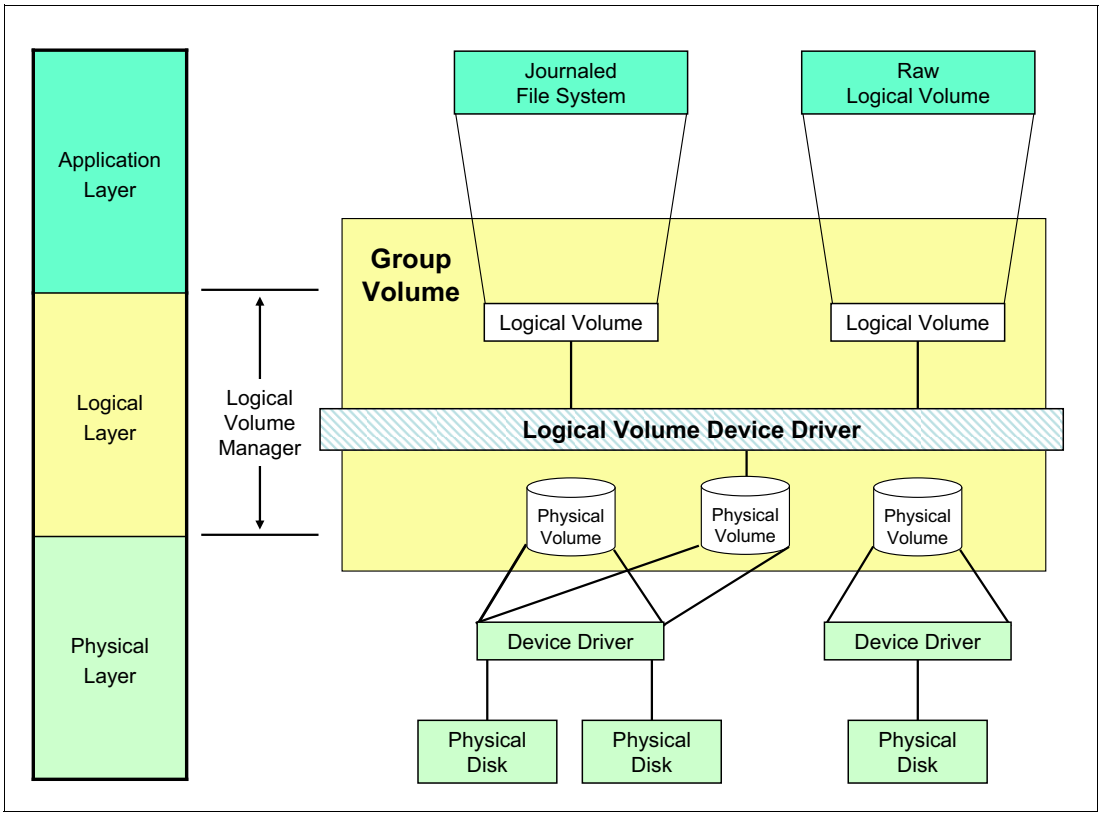


Figure 2-21 AIX Logical Volume Manager

Hierarchy of structures in disk storage

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

In AIX, each individual disk drive is called a physical volume (PV) and has a name, usually /dev/hdiskx (where *x* is a unique integer on the system). In the case of the DS5000 storage subsystem, such physical volumes correspond to a LUN.

- Every physical volume in use belongs to a volume group (VG) unless it is being used as a raw storage device.
- Each physical volume is divided into physical partitions (PPs) of a fixed size for that physical volume.
- Within each volume group, one or more logical volumes (LVs) are defined. Logical volumes are groups of information located on physical volumes. Data on logical volumes appear contiguous to the user, but can be spread (striped) on multiple physical volumes.
- Each logical volume consists of one or more logical partitions (LPs). Each logical partition corresponds to at least one physical partition (see Figure 2-22). If mirroring is specified for the logical volume, additional physical partitions are allocated to store the additional copies of each logical partition.
- Logical volumes can serve a number of system purposes (paging, for example), but each logical volume that holds ordinary systems, user data, or programs, contains a single journaled file system (JFS or JFS2). Each file system consists of a pool of page-size blocks. In AIX Version 4.1 and later, a given file system can be defined as having a fragment size of less than 4 KB (512 bytes, 1 KB, or 2 KB).

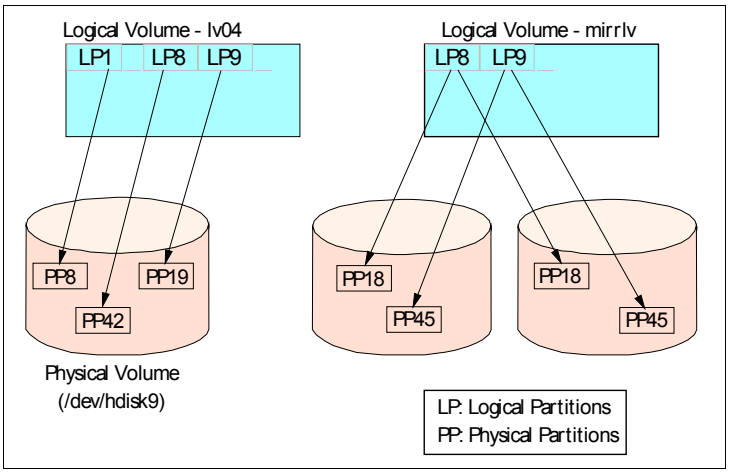


Figure 2-22 Relationships between LP and PP

The Logical Volume Manager controls disk resources by mapping data between a simple and flexible logical view of storage space and the actual physical disks. The Logical Volume Manager does this by using a layer of device driver code that runs above the traditional physical device drivers. This logical view of the disk storage is provided to applications and is independent of the underlying physical disk structure (Figure 2-23).

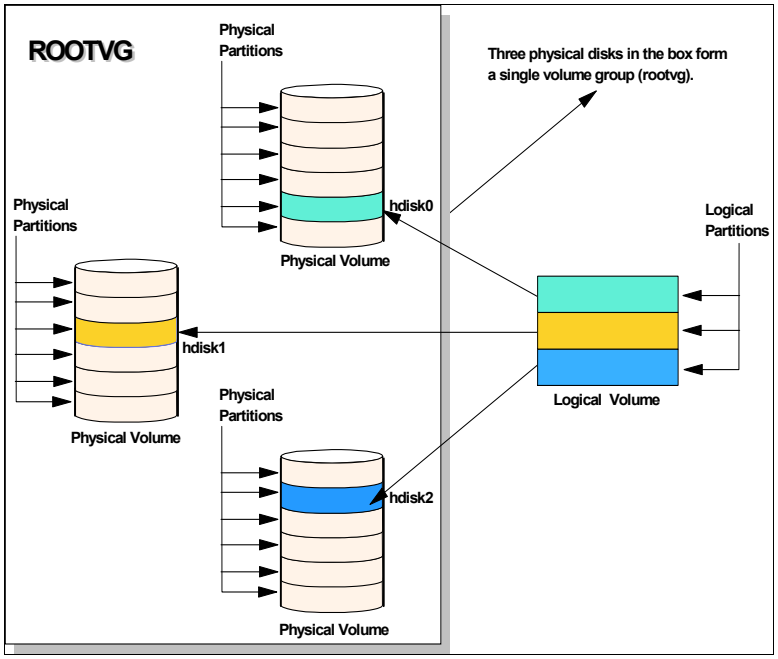


Figure 2-23 AIX LVM conceptual view

Best practice: When using a DS5000 storage subsystem with operating systems that have a built-in LVM, or if a LVM is available, plan to make use of the LVM.

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

Because a DS5000 storage subsystem has its own RAID arrays and logical volumes, we do not work with real physical disks in the system. Functions, such as intra-disk allocation, write scheduling, and write-verify policies, do not help much, and it is hard to determine the performance benefits when using them. They must only be used after additional testing, and it is not unusual that trying to use these functions will lead to worse results.

On the other hand, do not forget about the important inter-disk allocation policy, as described in the following section.

Inter-disk allocation policy

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

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

By setting the inter-physical volume allocation policy to maximum, you also ensure that the reads and writes are shared among PVs, and in systems such as a DS5000 storage subsystem, also among controllers and communication paths.

Best practice: For random I/O, the best practice is to create arrays of the same type and size. For applications that do not spread I/Os equally across containers, create VGs comprised of one logical drive from every array, use a maximum inter-disk allocation policy for all LVs in the VG, and use a random disk order for each LV. Check that the ownership of logical drives selected are spread evenly across DS5000 controllers. Applications that spread their I/Os equally across containers, such as DB2, use another layout.

If systems are using only one big volume, it is owned by one controller, and all the traffic goes through one path only, which happens because of the static load balancing that DS5000 controllers use.

2.6.2 Planning for systems without LVM: Windows example

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

With Windows 2003 and 2008, there are two types of disks: basic disks and dynamic disks. By default, when a Windows system is installed, the basic disk system is used.

Basic disks and basic volumes are the storage types most often used with Microsoft Windows operating systems. A basic disk refers to a disk that contains basic volumes, such as primary partitions and logical drives. A basic volume refers to a partition on a basic disk. For Windows 2003 and 2008, a primary partition on a basic disk can be extended using the **`extend`** command in the diskpart.exe utility.

Dynamic disks provide features that basic disks do not, such as the ability to create volumes that span multiple disks (spanned and striped volumes), as well as the ability to create software level fault tolerant volumes (mirrored and RAID 5 volumes). All volumes on dynamic disks are known as *dynamic volume*s.

With the DS5000 storage subsystem, you can use either basic or dynamic disks, depending upon your needs and requirements (certain features might not be supported when using dynamic disks). There are cases for both disk types; this depends on your individual circumstances. In certain large installations, where you might have the requirement to span or stripe logical drives and controllers to balance the work load, dynamic disk might be your only choice. For smaller to mid-size installations, you might be able to simplify and just use basic disks, which is entirely dependent upon your environment and your choice of disk system needs to be made on those circumstances.

When using the DS5000 storage subsystem, the use of software mirroring and software RAID 5 is not required. Instead, configure the storage on the DS5000 storage subsystem for the redundancy level required.

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

- If you want to have the UNIX-like capabilities of LVM, you can purchase and use the Veritas Storage Foundation (from Symantec) suite or a similar product. With this product, you get several features that go beyond LDM. Volume Manager does not just replace the Microsoft Management Console (MMC) snap-in; it adds a much more sophisticated set of storage services to Windows 2003 and 2008. After Windows is upgraded with Volume Manager, you are able to manage better multidisk direct server-attached (DAS) storage, JBODs (just a bunch of disks), Storage Area Networks (SANs), and RAID.

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

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

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

  Using Veritas Volume Manager and tuning the databases and applications goes beyond the scope of this guide. Browse the application vendor sites and vendor documentation for more information.

  For Veritas Volume Manager (VxVM), see this website:

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

- You can use the DS5000 storage subsystem and Windows dynamic disks to spread the workload between multiple logical drives and controllers, which can be achieved with spanned, striped, mirrored, or RAID 5:

  - Spanned volumes combine areas of un-allocated space from multiple disks into one logical volume. The areas of un-allocated space can be various sizes. Spanned volumes require two disks, and you can use up to 32 disks. If one of the disks containing a spanned volume fails, the entire volume fails, and all data on the spanned volume becomes inaccessible.
  - Striped volumes can be used to distribute I/O requests across multiple disks. Striped volumes are composed of stripes of data of equal size written across each disk in the volume. They are created from equally sized, un-allocated areas on two or more disks. The size of each stripe is 64 KB and cannot be changed. Striped volumes cannot be extended and do not offer fault tolerance. If one of the disks containing a striped volume fails, the entire volume fails, and all data on the striped volume becomes inaccessible.
  - Mirrored and RAID 5 options are software implementations that have an additional impact on top of the existing underlying fault tolerance level configured on the DS5000 storage subsystem. They can be employed to spread the workload between multiple disks, but then there are two lots of redundancy happening at two separate levels.

  Text these possibilities in your environment to ensure that the solution chosen suits your needs and requirements.

2.7 Software and microcode upgrades

Periodically, IBM releases new firmware (which is posted on the support the website) that will need to be installed. Occasionally, IBM might remove old firmware versions from support. Upgrades from unsupported levels are mandatory to receive warranty support.

Upgrades to the DS5000 Storage System firmware must generally be preceded by an upgrade to the latest available version of the Storage Manager client software because this might be required to access the DS5000 Storage System when the firmware upgrade completes. In most cases, it is possible to manage a DS5000 Storage System running down-level firmware with the latest SMclient, but not possible to manage a storage server running the latest version of firmware with a down-level client. In certain cases the only management capability provided might be to upgrade the firmware to newer level; which is the desired goal.

Tip: The version number of the Storage Manager firmware and the Storage Manager client are not completely connected. For example, Storage Manager 10.77 can manage storage Systems that are running storage System firmware 06.12 or later on them.

Always check the readme file for details of the latest Storage Manager release and special usage.

2.7.1 Staying up-to-date with your drivers and firmware using My support

My support registration provides email notification when new firmware levels have been updated and are available for download and installation. To register for My support, visit:

http://www.ibm.com/support/mysupport/us/en

Here we describe the registration process:

1. The Sign In window displays:
   - If you have a valid IBM ID and password, then sign in.
   - If you are not currently registered with the site, click **Register now** and register your details.
2. If it is your first time on the Notifications website, or subscribing to any Disk Systems product, click **Subscribe**, then select **Disk Systems** under Storage, and click each of the products you want to receive notifications.

   If you are already subscribed for any Disk Systems products, click **My Subscriptions**, then click your Subscription folder for Disks Products, and then click your subscription name to edit the **Disk Systems** category.
3. Select all the products of your interest to receive notifications and then click Continue
4. Complete the remaining fields for your new subscription, select the frequency of the notifications, daily or weekly, and click Submit
5. Click **Sign Out** to log out of My Support.

You will be notified whenever there is new firmware available for the products you selected during registration

Also, explore and customize to your needs the other options available under My support. If you need more details on the My Support configuration, see the following presentation:

ftp://ftp.software.ibm.com/systems/support/tools/mynotifications/overview.pdf

2.7.2 Compatibility matrix

In order to stay at a supported level for all your components of the solution, any time you need to upgrade your OS version, driver, firmware, and so on, plan ahead for your changes. Also check in the IBM System Storage Interoperation Center (SSIC) website for the latest levels:

http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss

Tip: Use the IBM System Interoperation Center website to check latest compatibility information before doing any change in your storage subsystem environment.

2.7.3 DS5000 firmware components and prerequisites

The microcode of the DS5000 Storage System is separated into the following packages:

- The Storage Manager package
- The Controller Firmware package, including NVSRAM and scripts for specific usage
- The environmental service module (ESM) and Hard disk drive (HDD) package

The Storage Manager package includes a set of client and host tools to manage your Storage subsystem from your management workstation.

The firmware and the NVSRAM are closely tied to each other and are therefore *not* independent. Be sure to install the correct combination of the two packages. The NVSRAM is similar to the settings in the BIOS of a host system. You can change certain specific settings using the scripts provided, with this package.

Upgrading the firmware and management software for the DS5000 Storage System is a relatively simple procedure. While it is not a requirement, when possible, schedule a maintenance window for all firmware updates, thus minimizing any possible exposure. The times for upgrading all the associated firmware and software are given in Table 2-8. These times are only approximate and can vary from system to system.

Table 2-8 Upgrade times

Element being upgraded	Approximate time of upgrade	Requirements
Storage Manager software	35 minutes	Concurrent
DS5000 Controller and NVSRAM firmware	5 to 35 minutes	Concurrent with redundant drivers Downtime if no dual paths, or executed scripts to change NVSRAM setting
DS5000 environmental service module (ESM) canister	5-10 minutes per ESM	Low I/O
Hard drives	3-5 minutes per drive. Parallel firmware upgrade is possible for multiple drive types	Downtime, NO I/O

Scheduling the firmware update

If you have a maintenance window with downtime, schedule the ESM firmware update for that time. You should stop all I/Os and update all the ESMs at one time. If it is not possible to schedule a maintenance window, apply ESM firmware one EXP at a time, and this needs to be done during non-peak utilization periods.

Drive firmware update requires downtime, however, the DS5000 storage subsystems can update all the drives in a configuration, up to four separate drive models at the same time.

It is critical that if you update one part of the firmware, you update all the firmware and software to the same level. You must *not* run a mismatched set.

Firmware installation sequence

Normally, the DS5000 storage subsystem firmware download sequence starts with controller firmware, followed or together with the NVSRAM, then ESM firmware, and concludes with the drive firmware. However, it is not necessarily the case, because many times there is a minimum prerequisite level for drive or ESM firmware for a certain level of controller firmware. Because it is dependent on each specific release level, you need to check the firmware readme file of the level you are trying to install for specific dependencies or guidelines.

2.7.4 Updating the DS5000 subsystem firmware

It is always a best practice to run your DS5000 Storage System at the latest level of microcode. Occasionally, IBM will withdraw older levels of microcode from support. In this case, an upgrade to the microcode is mandatory. In general, you need to plan on upgrading all drivers, microcode, and management software in your SAN on a periodic basis. New code levels can contain important fixes to problems that you might not have encountered yet.

Important: Before upgrading the storage System firmware and NVSRAM, make sure that the system is in an optimal state. If not, run the Recovery Guru to diagnose and fix the problem before you proceed with the upgrade.

The upgrade procedure needs two independent connections to the DS5000 Storage System, one for each controller. It is not possible to perform a microcode update with only one controller connected. Therefore, both controllers must be accessible either through Fibre Channel or Ethernet. Both controllers must also be in the active state.

Considerations for the upgrade

If you plan to upgrade doing in-band management, make sure that you have a multipath I/O driver installed on your management host. This precaution is necessary because access logical drive moves from one controller to the other during this procedure and the DS5000 Storage System must be manageable during the entire time.

Important: Here are various considerations for your upgrade:

- See the readme file to determine whether the ESM or the controllers must be upgraded first. In certain cases, the expansion enclosure ESMs must be updated to the latest firmware level before starting the controller update (outdated ESM firmware can make your expansion enclosures inaccessible after the DS5000 Storage System firmware update). In certain cases it is just the opposite.
- Ensure that all hosts attached to the DS5000 Storage System have a multipath I/O driver installed and paths active.
- Any power or Network/SAN interruption during the update process might lead to configuration corruption. Therefore, do not power off the DS5000 Storage System or the management station during the update. If you are using in-band management and have Fibre Channel hubs or managed hubs, then make sure no SAN connected devices are powered up during the update. Otherwise, it can cause a loop initialization process and interrupt the process.

Controller microcode upgrade

As mentioned before, DS5000 Storage subsystem, controller firmware can be updated concurrently as long as you have dual paths to all the hosts using the storage.

You can load the controller firmware and NVSRAM to a designated flash area on the DS5000 controllers and activate it at a later time, staged microcode upgrade. Of course, you can still transfer the controller microcode to the storage subsystem and activate it in one step, which is the preferable method if you are doing in a maintenance window or during a non peak hours as an additional precaution.

Download the controller firmware and NVSRAM at the same time by selecting the option check box in the controller firmware download window. However if you have executed any script to change the host parameters NVSRAM settings (like Enable AVT on Windows), downloading NVSRAM will overwrite those changes, changing the current behavior of your storage subsystem. In that case, schedule the controller firmware and NVSRAM during a maintenance window, so you can reapply the modifications after loading the new NVSRAM file.

Tip: If you applied any changes to the NVSRAM settings, for example, running a script, you must re-apply them after the download of the new NVSRAM completes. The NVSRAM update resets all settings stored in the NVSRAM to their defaults.

The firmware is transferred to one of the controllers. This controller copies the image to the other controller. The image is verified through a CRC check on both controllers. If the checksum is OK, the uploaded firmware is marked ready and available for activation. If one of the two controllers fails to validate the CRC, the image is marked invalid on both controllers and not available for activation. An error is returned to the management station as well.

The activation procedure is similar to previous firmware versions. The first controller moves all logical drives to the second one, then it reboots and activates new firmware. After that it takes ownership of all logical drives, and the second controller is rebooted in order to have its new firmware activated. When both controllers are up again, the logical drives are redistributed to the preferred paths. Because the logical drives move between the controllers during the procedure and are all handled by just one controller at a certain point, the new firmware activation ought to be done when the disk I/O is relatively low.

A normal reboot of a controller or a power cycle of the DS5000 does not activate the new firmware. It is only activated after the user has chosen to activate the firmware.

For specific details and guided step by step procedure updating DS5000 Storage System firmware, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

2.7.5 Updating DS5000 Storage Manager software

This section describes how to update the DS5000 software in Windows and Linux environments. This procedure is also covered in detail in Chapter 4, “Host configuration guide” on page 151 for various operating systems.

Updating in a Windows environment

To update the host software in a Windows environment, proceed as follows:

1. Verify that IBM HBA firmware and device driver versions are current, and at the supported level of compatibility from the host Storage Manager software and firmware you are installing (see the SSIC website). If they are not current, download the latest versions and study the `readme` file located with the device driver, then upgrade the device drivers.
2. Go to the IBM Support website:

   `http://www-947.ibm.com/support/entry/portal/`

   Select your storage system. From the Support home page, click **Downloads**, then select the Storage Manager package for your operating system.
3. Begin the installation of your downloaded package in your server. Accept the license terms, and select **Typical** as the installation type.
4. Because you are updating your installation, without uninstalling the previous version, the installation program detects existing versions of software components and presents the warning panel shown in Figure 2-24.

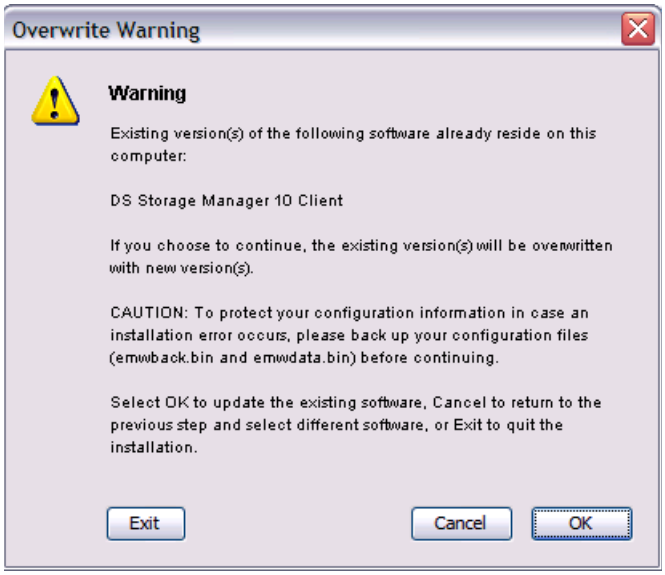


Figure 2-24 Updating host software warning

5. As noticed, back up the file emwdata.bin in case of failures, copying it to other directory outside the installation folder of the Storage Manager. This file contains the alerts notifications setup.
6. Click **OK** to continue installation, confirming the remaining windows. You are normally requested to restart your machine after the installation finishes.

> **Attention:** Windows MPIO/DSM is no longer available in the Storage Manager software package. It needs to be installed separately. The DSM software and Storage Manager Client software are available in the same folder.

Updating in a Linux environment

To update the host software in a Linux environment, proceed as follows:

1. If there is a previous version of the IBM DS4000 Storage Manager host software (such as SMruntime, SMclient, RDAC, SMutil and SMagent packages) installed in the server, you need to uninstall it first before installing the new version of the Storage Manager host software.

> **Tip:** Starting with the DS4000 Storage Manager version 9.16, all of the host software packages (SMruntime, SMclient, SMesm, SMutil and SMagent) are included in a single DS4000 Storage Manager host software installer wizard.

2. Verify that IBM host adapter device driver versions are current. If they are not current, see the `readme` file located with the device driver and then upgrade the device drivers.
3. Install the Storage Manager components in the following order:
   a. SMruntime - always first
   b. SMesm - required by client
   c. SMclient
   d. SMagent
   e. SMutil

2.8 Planning for physical components

In this section, we review elements related to physical characteristics of an installation, such as rack considerations, fiber cables, Fibre Channel adapters, and other elements related to the structure of the DS5000 storage subsystem and disks, including enclosures, arrays, controller ownership, segment size, storage partitioning, caching, hot spare drives, and Enhanced Remote Mirroring.

2.8.1 Rack considerations

The DS5000 storage subsystem and possible expansions are mounted in rack enclosures.

General planning

Consider the following general planning guidelines; determine:

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

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

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

Rack layout

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

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

The following considerations apply when you make a layout:

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

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

You need at least 152 cm (60 in.) of clearance at the front and at least 76 cm (30 in.) at the rear of the 42U rack suites. This space is necessary for opening the front and rear doors and for installing and servicing the rack. It also allows air circulation for cooling the equipment in the rack. All vertical rack measurements are given in rack units (U). One U is equal to 4.45 cm (1.75 in.). The U levels are marked on labels on one front mounting rail and one rear mounting rail.

Figure 2-25 shows an example of the required service clearances for a 9306-900 42U rack. Check with the manufacturer of the rack for the statement on clearances.

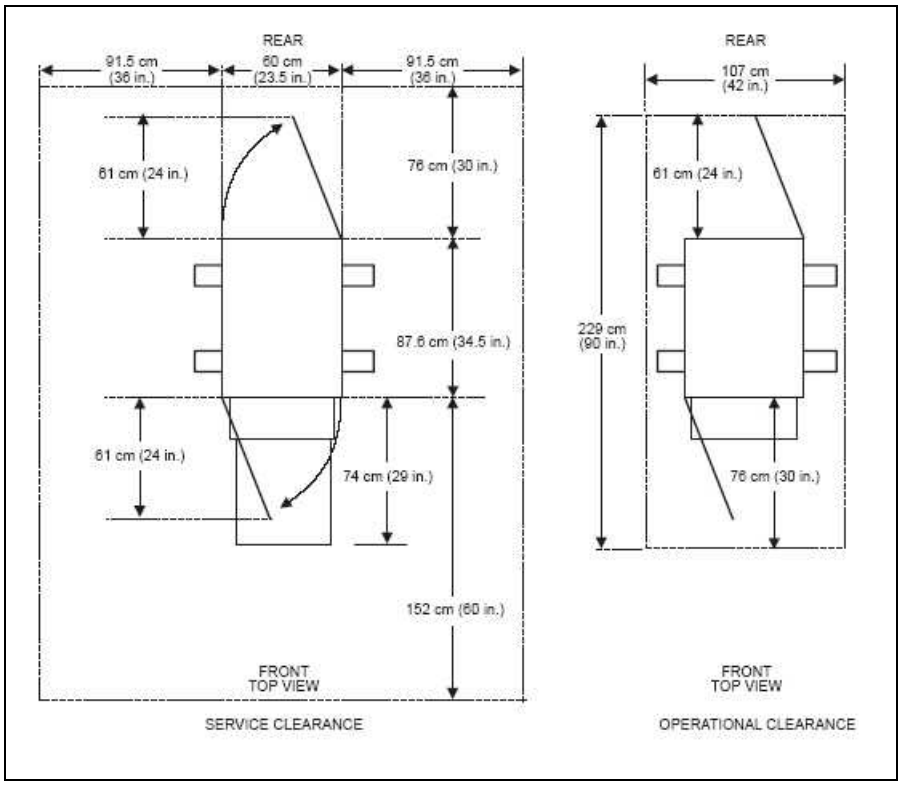


Figure 2-25 9306 enterprise rack space requirements

2.8.2 Cables and connectors

In this section, we describe various essential characteristics of fiber cables and connectors. This information can help you understand the options you have for connecting and cabling the DS5000 storage subsystem.

Cable types: Either shortwave or longwave

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

Multi-mode fiber allows light to disperse in the fiber so that it takes many paths, bouncing off the edge of the fiber repeatedly to finally get to the other end (multi-mode means multiple paths for the light). The light taking these various paths gets to the other end of the cable at slightly separate times (separate paths, separate distances, and separate times). The receiver has to determine which incoming signals go together.

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

There are two core sizes of multi-mode cabling available: 50 micron and 62.5 micron. The intermixing of the two core sizes can produce unpredictable and unreliable operation.

Therefore, core size mixing is not supported by IBM. Users with an existing optical fiber infrastructure are advised to ensure that it meets Fibre Channel specifications and is a consistent size between pairs of FC transceivers.

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

Table 2-9 Cable type overview

Fiber type	Speed	Maximum distance
9 micron SMF (longwave)	1 Gbps	10 km
9 micron SMF (longwave)	2 Gbps	2 km
50 micron MMF (shortwave)	1 Gbps	500 m
50 micron MMF (shortwave)	2 Gbps	300 m
50 micron MMF (shortwave)	4 Gbps	150 m
50 micron MMF (shortwave)	8 Gbps	50 m
62.5 micron MMF (shortwave)	1 Gbps	300 m
62.5 micron MMF (shortwave)	2 Gbps	150 m
62.5 micron MMF (shortwave)	4 Gbps	70 m
62.5 micron MMF (shortwave)	8 Gbps	21 m

Note that the “maximum distance” shown in Table 2-9 is just that, a maximum. Low quality fiber, poor terminations, excessive numbers of patch panels, and so on, can cause these maximums to be far shorter.

All IBM fiber feature codes that are orderable with the DS5000 storage subsystem will meet the standards.

Interfaces, connectors, and adapters

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

The interfaces that are used to convert the internal communication transport of gigabit transport are Small Form Factor Transceivers (SFF), also often called Small Form Pluggable (SFP). See Figure 2-26. Gigabit Interface Converters (GBIC) are no longer used on current models although the term GBIC is still sometimes incorrectly used to describe these connections.

Figure 2-26 Small Form Pluggable (SFP) with LC connector fiber cable

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

LC connector

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

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

All DS5000 series products use SFP transceivers and LC fiber cables. See Figure 2-27.

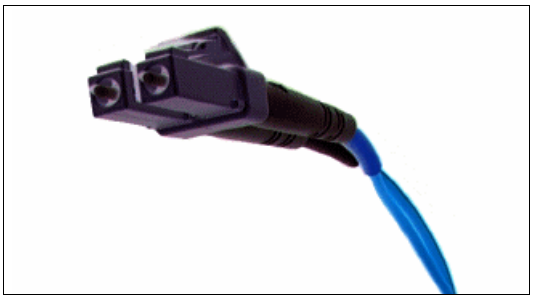


Figure 2-27 LC fiber cable connector

Best practice: When you are not using an SFP, it is best to remove it from the port on the DS5000 storage controller and replace it with a cover. Similarly, unused cables must be stored with ends covered, which will help eliminate risk of dirt or particles contaminating the connection while not in use.

Interoperability of 2 Gbps, 4 Gbps, and 8 Gbps devices

The Fibre Channel standard specifies a procedure for speedy auto-detection. Therefore, if a 4 Gbps port on a switch or device is connected to a 2 Gbps port, it must negotiate down and the link will run at 2 Gbps. If there are two 8 Gbps ports on either end of a link, the negotiation runs the link at 8 Gbps if the link is up to specifications. A link that is too long or “dirty” can end up running at 4 Gbps, even with 8 Gbps ports at either end, so care must be taken with cable lengths distances and connector quality is sound.

The same rules apply to 8 Gbps devices relative to 4 Gbps and 2 Gbps environments. The 8 Gbps and 4 Gbps devices have the ability to automatically negotiate back down to either 4 Gbps, 2 Gbps or 1 Gbps, depending upon the attached device and the link quality. If the link does unexpectedly negotiate to a slower speed than expected, then the causes or reasons for this ought to be investigated and remedied.

The DS5000 storage subsystem now has 8 Gbps functionality; there are several switches and directors that operate at this speed.

Tip: On certain fiber switch vendor models, it might be necessary to configure the port to a specific speed of 2, 4, or 8 Gbps to obtain the required speed instead of leaving "auto-detection" on the port.

2.8.3 Cable management and labeling

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

Why more detailed cable management is required

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

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

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

Cable planning

Successful cable management planning includes three basic activities:

- Site planning (before your solution is installed)
- Cable routing
- Cable labeling

Site planning

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

Cable routing

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

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

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

  - Take particular care when attaching fiber optic cables to the rack. See the instructions included with your fiber optic cables for guidance on minimum radius, handling, and care of fiber optic cables.
  - Run the cables neatly along the rack rear corner posts.
  - Use cable ties to secure the cables to the corner posts.
  - Make sure the cables cannot be pinched or cut by the rack rear door.
  - Run internal cables that connect devices in adjoining racks through the open rack sides.
  - Run external cables through the open rack bottom.
  - Leave enough slack so that the device can be fully extended without putting a strain on the cables.
  - Tie the cables so that the device can be retracted without pinching or cutting the cables.
- To avoid damage to your fiber optic cables, follow these guidelines:
  - Use great care when utilizing cable management arms.
  - When attaching to a device on slides, leave enough slack in the cable so that it does not bend to a radius smaller than that as advised by your fiber optic cable guide when extended or become pinched when retracted.
  - Route the cable away from places where it can be snagged by other devices in the rack.
  - Do not overtighten the cable straps or bend the cables to a radius smaller than that as advised by your fiber optic cable guide.
  - Do not put excess weight on the cable at the connection point and be sure that it is well supported. For example, a cable that goes from the top of the rack to the bottom *must* have a method of support other than the strain relief boots built into the cable.
  - For long cable runs, ensure that enough slack is made for rack movement in accordance with your computer room standards for earthquake proofing.

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

Cable labeling

When labeling your solution, follow these tips:

- As you install cables in the rack, label each cable with the appropriate identification.
- Remember to attach labels to any cables that you replace.
- Document deviations from the label scheme you use. Keep a copy with your Change Control Log book.
- Comply with an existing cable naming convention or define and adhere to a simple logical naming convention

An example of label naming convention might include these attributes:

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

Other cabling mistakes

Avoid making these common mistakes in cabling:

- Leaving cables hanging from connections with no support.
- Not using dust caps.
- Not keeping connectors clean. (Certain cable manufacturers require the use of lint-free alcohol wipes in order to maintain the cable warranty.)
- Leaving cables on the floor where people might kick or trip over them.
- Not removing old cables when they are no longer needed or planned for future use.

Tip: Collect all SFP, HBA, and cable dust caps, store in a dust free container, to be used for future cabling work. Do not re-use dust caps which have been left loose in the rack or computer room.

3

Configuring the IBM DS5000 Storage System

In this chapter, we provide a sequence of tasks and guidelines to properly set up, install, and configure the IBM System Storage DS Storage System:

- Initial setup of DS5000 Storage System
- Configuring the DS5000 Storage System with the Storage Manager (SM) Client
- Defining logical drives and hot spare drives
- Increasing array and logical volume capacity
- Setting up storage partitioning
- Configuring iSCSi with specific card or software initiators
- Event monitoring and alerts
- Software and firmware upgrades
- Capacity upgrades

3.1 Configuring the DS5000 Storage System

For the course of this chapter, we assume that you have installed the operating system on the host server; and have all the necessary device drivers and host software installed and configured. We also assume that you have a good understanding and working knowledge of the DS5000 Storage System product. If you require detailed information about how to perform the installation, setup, and configuration of this product, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

After you have set up the DS5000 Storage System that is connected to a server or the SAN, you can proceed with additional configuration and storage setting tasks. If there is previous configuration data on the DS5000 Storage System that you want to be able to reference, then first save a copy of the *storage subsystem profile* to a file. After you have completed your changes, you must save the profile to a separate file as well, which will be of great value when you are describing questions or problems with support; or reviewing your configuration with your performance data. For more information about the storage subsystem profile, see the maintenance and troubleshooting section of the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Best practice: You need to save a new profile, Collect of All Support Data (CASD), each time that you change the configuration of the DS5000 storage subsystem, which applies to all changes regardless of how minor they might be. The profile must be stored in a location where it is available even after a complete configuration loss, for example, after a site loss.

The configuration changes that you want can be done using the Subsystem Management Setup view of the DS5000 Storage Manager, or by following the steps outlined here.

Before defining arrays or logical drives, you need to perform some basic configuration steps, which also applies when you reset the configuration of your DS5000 Storage System:

1. Add your DS5000 storage subsystem to the DS Storage Manager Client. To add a new system to the SM Client Enterprise Management Window (EMW), click **Edit** → **Add Storage Subsystem...** and the Add New System manual window opens as shown in Figure 3-1. You need to enter the controller IP addresses. Here we use default IPs.

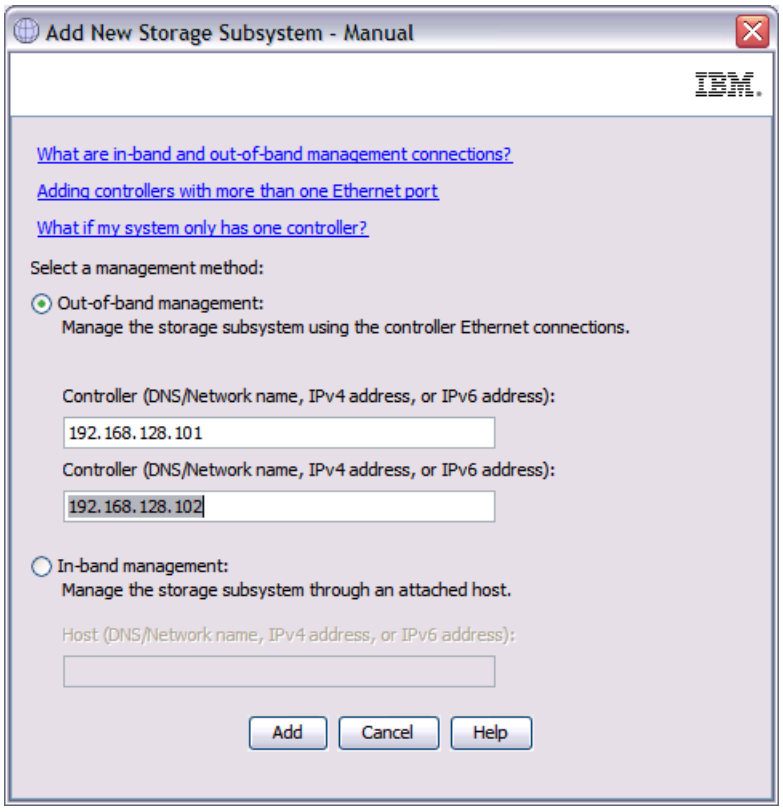


Figure 3-1 Add new storage subsystem manually

Click the **Add** button. The window opens, saying that a new system has been added. Click **NO** to complete if you have no more systems to add. Double-click **DS5000** to open the Storage Manager SMW, as shown in Figure 3-2.

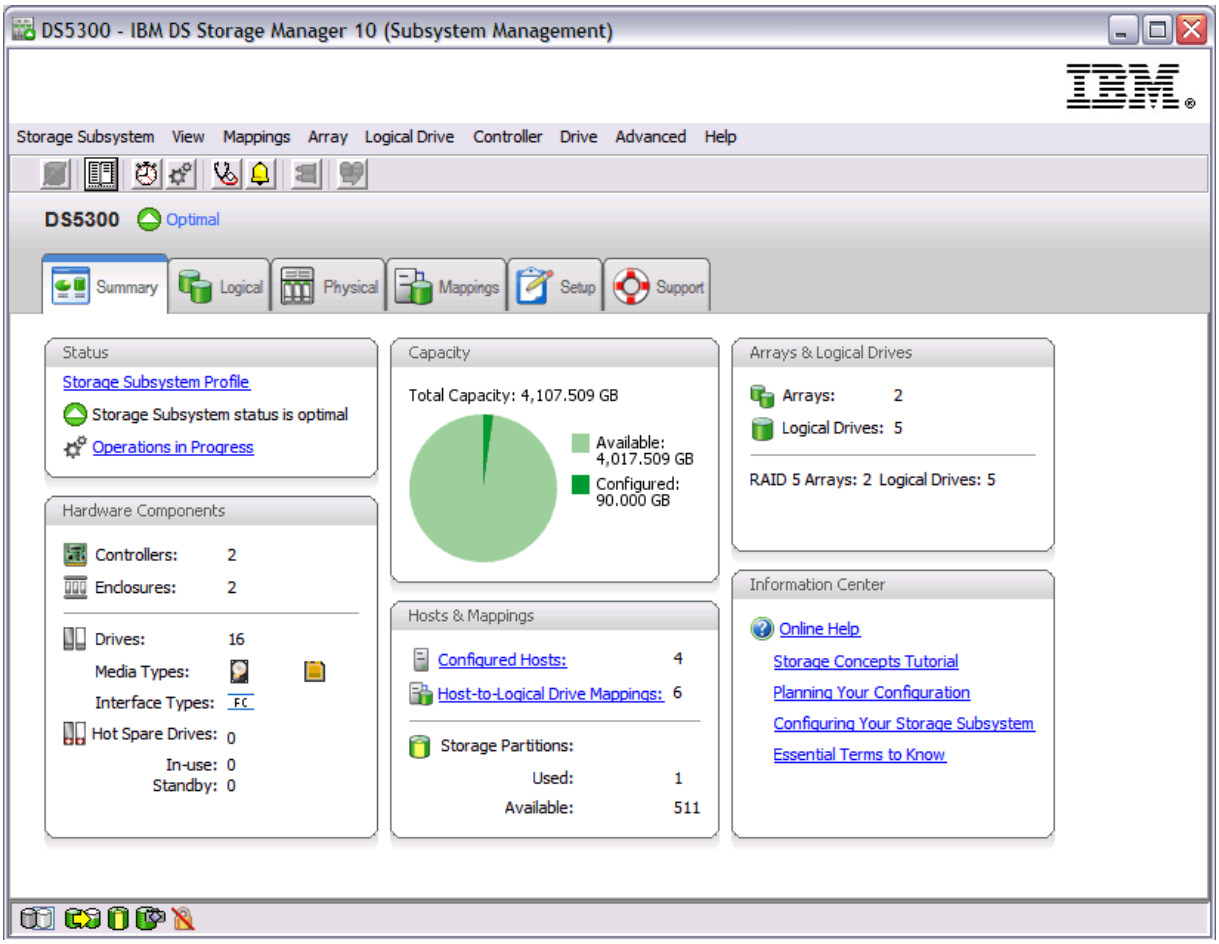


Figure 3-2 DS Storage Manager subsystem management window

2. Give literal names to the DS5000 if you install more than one storage subsystem. To name or rename the DS5000, open the Subsystem Management window, select the **Setup** tab, and click **Rename Storage Subsystem**, or select **Storage Subsystem** → **Rename** from the top menu bar.

   Make sure that you are renaming the right subsystem, if you have more than one. Select the option **Locate Storage Subsystem** in the Setup view to verify the storage subsystem. This action turns on a light in the front of the selected subsystem, so you can identify it.

3. Since the DS5000 Storage stores its own event log, synchronize the controller clocks with the time of the host system used to manage the DS5000. If you have not already set the clocks on the DS5000 Storage Systems, set them now. Be sure that your local host system is working using the correct time. Then, click **Storage Subsystem** → **Synchronize Controller Clock...**

   Best practice: Synchronize the DS5000 Storage subsystem controller clocks with the Management station in order to simplify error determination when you start comparing the various event logs. A network time server can be useful for this purpose.

4. Set a password to prevent unauthorized users from making configuration changes, for security reasons, especially if the DS5000 has Full Disk Encryption (FDE) security-capable drives, or it is connected to the Public network. This password is required for all actions on the DS5000 that change or update the configuration in any way.

The reminder window shown in Figure 3-3 prompts you to set the password each time you start managing your subsystem.

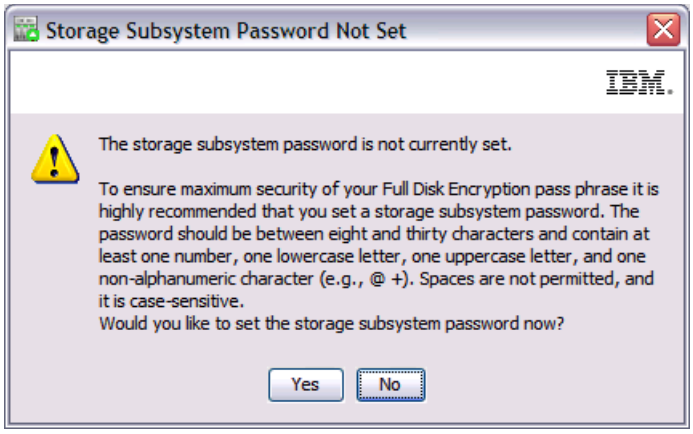


Figure 3-3 Set security password

To set a password, select **Yes** in the window show in Figure 3-3, or select the option **Set a Storage Subsystem Password** from the Subsystem management window setup view. This password is then stored on the DS5000. You need to provide it no matter which management station you are connecting from, whether you are using in-band or out-of-band management.

Best practice: Set a password to the storage subsystem to prevent unauthorized access and modifications, even if you are not using Full Disk Encryption (FDE) security-capable drives.

You are asked for the password only once during a single management session.

Important: There is no way to reset the password by the Storage Manager Client after it is set if you lose it. Ensure that the password information is kept in a safe and accessible place. Contact IBM technical support for help if you forget the password to the storage subsystem.

3.1.1 Defining hot spare drives

Hot spare drives are reserved drives that are *not* normally used to store data. You can use a hot spare drive for additional data protection from drive failures that occur in a RAID Level 1, RAID Level 3, RAID Level 5, or RAID Level 6 array. When a drive in a RAID array fails, the hot spare drive takes on the function of the failed drive and the data is rebuilt on the hot spare drive, which becomes part of the array. After this rebuild procedure, your data is again fully protected. A hot spare drive is like a replacement drive installed in advance.

A hot spare drive defined on the DS5000 Storage System is always used as a *global hot spare*. That is, a hot spare drive is defined at storage subsystem level and it cans always be used for a failed drive within the DS5000. The expansion or storage System enclosure in which it is located is not important.

Considerations

There are various considerations to evaluate when setting up hot spare drives, regarding disk types, usage, and size.

Follow these guidelines to protect your storage subsystem setting hot spares drives:

- Hot spare disk drives must be of the same media type and interface type as the disk drives that they are protecting.
- Similarly, hard disk drives can only be hot spares for other hard disk drives, not Solid State Disks (SSDs).
- Hot spare disk drives must have capacities equal to or larger than the used capacity on the disk drives that they are protecting. The DS5000 can use a larger drive to recover a smaller failed drive to it. It will not use smaller drives to recover a larger failed drive. If a larger drive is used, the remaining excess capacity is blocked from use.
- FDE disk drives provide coverage for both security capable and non-security capable disk drives. Non-security capable disk drives can provide coverage only for other non-security capable disk drives:
  - For an array that has *secured* FDE drives, the hot spare drive must be an unsecured FDE drive of the same or greater capacity.
  - For an array that has *unsecured* FDE drives, the hot spare drive can be either an unsecured FDE drive or a non-FDE drive.

  In a mixed disk environment that includes SATA drives, Fibre Channel drives, and FDE Fibre Channel drives (with security enabled or not enabled), use at least one type of global hot spare drive for each one (FDE Fibre Channel drive and a SATA drive) at the largest capacity within the array. If a secure-capable FDE Fibre Channel and SATA hot spare drives are included, all arrays are protected.

Hot spare best practices

Here is a summary of best practices to assign a Hot spare Drive:

- For maximum data protection, you must use only the largest capacity drives for hot spare drives in mixed capacity hard drive configurations, and highest speed drives.
- Hot spare drives distributed across separate disk expansions provide maximized performance and availability, preventing unlikely but possible failures such as the loss of a complete expansion or a drive channel.
- Use a minimum ratio of one hot spare drive for every 30 drives of a particular media type and interface type, or one for every two fully populated enclosures.
- Use the option **Drive** → **Hot Spare Coverage...** and then select **View/change current hot spare coverage** to ensure that all the defined arrays are hot spare protected.

Drive capacity and speed

When a drive failure occurs on a Storage System configured with multiple hot spare drives, the DS5000 will attempt to find a hot spare drive in the enclosure with the failed drive first. It will find a drive that is at least the same size as the failed drive, but not necessarily giving preference to one the exact same size as the failed drive. If a match does not exist in the same enclosure, it will look for spares in the other enclosures that contain sufficient capacity to handle the task.

The controller uses a free hot spare drive as soon as it finds one, even if there is another one that might be closer to the failed drive. If various speed drives (10 K or 15 K) are used, the fastest hot spares are used so as to not slow down an array if the hot spare is required.

Hot spare locations

Distribute the hot spare drives evenly across the various expansions of your storage subsystem and avoid having multiple drives in a single enclosure. Because hot spare drives are in standby, without traffic or I/O until a drive fails, then you want to maximize the overall performance of your system by evenly distributing your production drives across the various expansions. At the same time, this avoids the risk of a single disk drive channel, or expansion enclosure failure, causing loss of access to all hot spare drives in the storage subsystem.

However, in certain configurations, for example, a DS5000 with 5 expansions, evenly distributed across the 4 separate drive channels to maximize the traffic to each enclosure, then you can choose to maximize performance over availability, having all the spares defined in the fifth expansion. In this way, the channel having two expansions will not be penalized for excessive traffic, because the spares expansion will not contribute to load the traffic in that channel.

Quantity and type of hot spare drives

There is no fixed rule for the quantity of disk drives to assign as hot spares, but as a general rule, considering disk usage and availability, it is a best practice to define one for each 30 drives of a particular media type and interface type, or one for every 2 fully populated enclosures. Because of disk sizes, and especially in large configurations with arrays containing numerous drives, the reconstruction of a failed drive to a hot spare drive can take a long time, proportional to the quantity of drives in the array and the size of the disks. If in addition to that time, you need to wait to have a new disk available onsite to replace the failed drive, then the probabilities of having another disk failure increases. Having multiple spare drives will not mitigate the reconstruction time, but at least an array will be prepared sooner for a second disk failure.

Because the media and interface type matter when assigning disk drives, make sure to cover all the various disk types. Use the push buttons in the Physical view to show each of the types of drives for better identification of the drive types in your Storage subsystem (FC, FDE, SATA, SSD), as in Figure 3-4.

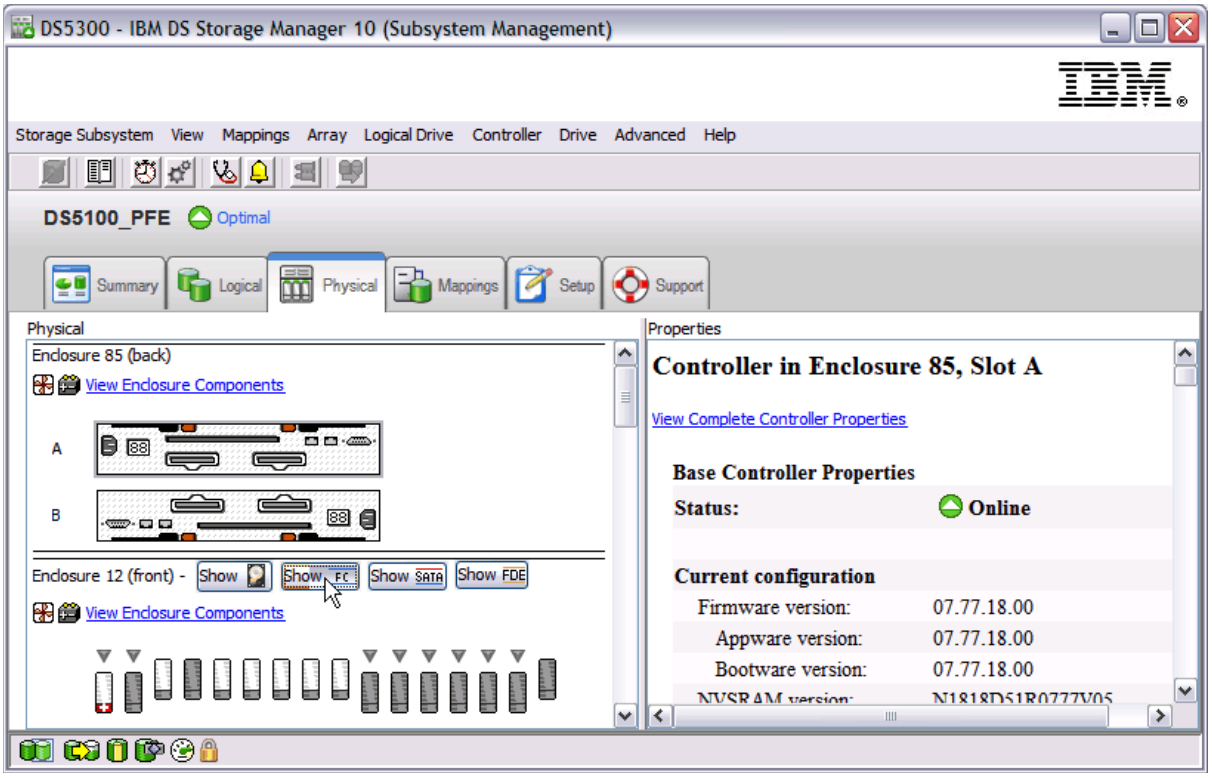


Figure 3-4 FC drive types

The newly defined hot spare drives are identified with a small red cross in the lower part of the drive icon. Available Hot Spare drive status are shown in Figure 3-5.

Optimal - Hot Spare - Standby	
Optimal - Hot Spare - In Use	
Failed - Hot Spare - Standby	

Figure 3-5 Hot Spare drive status

Verifying complete hot spare protection

Hot spare protection verification can be done from the Subsystem management window by selecting **Drive** → **Hot Spare Coverage** → **View/change current hot spare coverage**. When selecting this option, not only can you assign hot spares, you can also check if all of the defined arrays are protected by hot spare drives, for example, as shown in Figure 3-6.

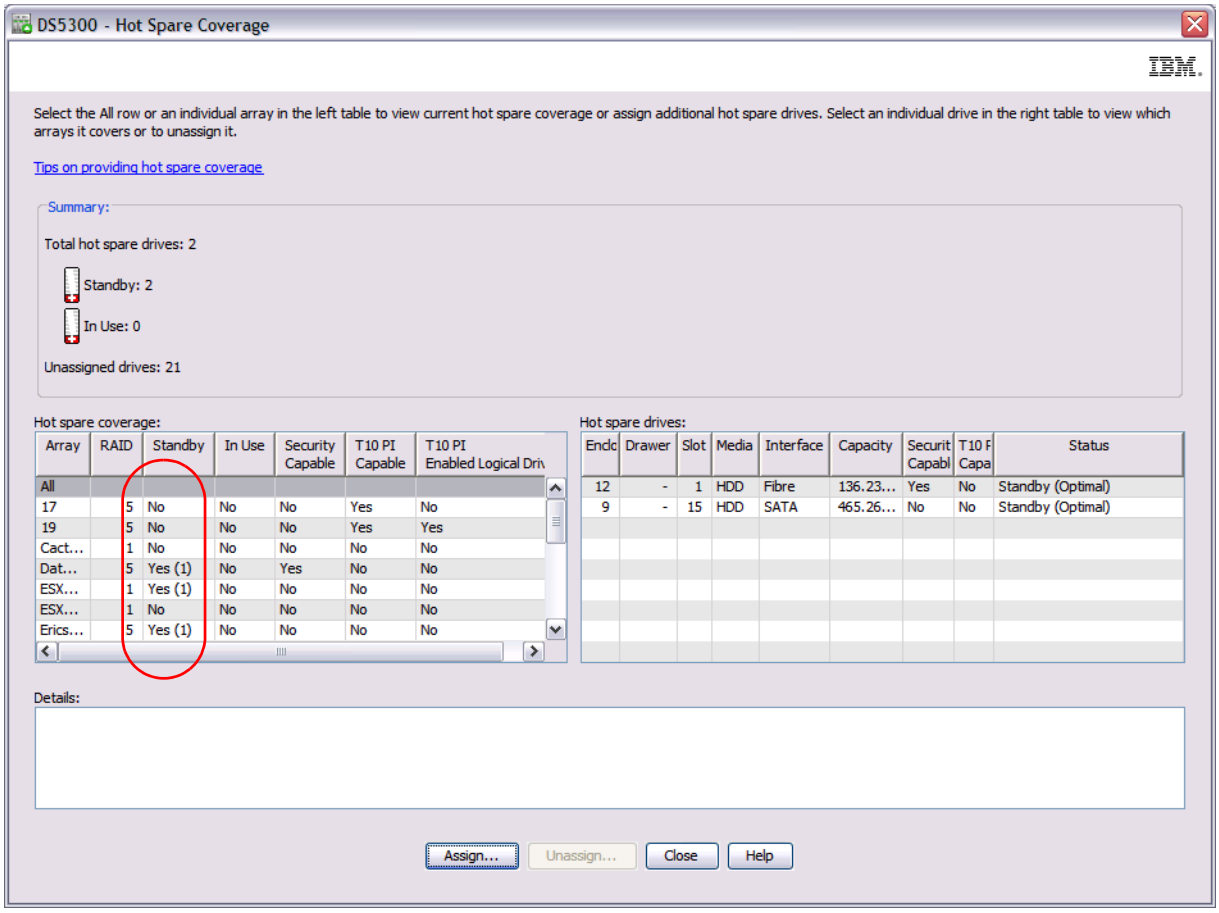


Figure 3-6 Hot spare coverage verification

If you need any further assistance with how to define the hot spare drives in the DS5000 Storage System, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Managing hot spare drives

Now that we have reviewed the considerations and best practices to define hot spare drives, let us see what happens after they take over the data of a failed disk.

If a hot spare is available when a drive fails, the controller uses redundancy data to reconstruct the data onto the hot spare. Then, you can replace the failed drive with a new one, or with an unassigned drive that is available. Next we describe the three alternatives:

1. After a failed drive is replaced with a new drive, the new replacement drive copyback process starts automatically, from the hot spare drive in use to the new replaced drive. After the copyback completed, the original hot spare drive becomes a free hot spare drive again.

This option is particularly useful when you have customized the hot spare slot assignments for better availability or performance, and for arrays of few disks. The penalty of this option is high, because it adds a second reconstruction to the process, duplicating the performance impact.

2. However, if you have free unassigned drives, you do not need to wait for the new replacement drive. You can select the option to replace the failed drive with one of the unassigned drives. Doing it starts a reconstruction process from the hot spare drive to the free unassigned drive, which then becomes a new member of the array.
3. Before a failed drive is replaced with a new drive, the hot spare drive can be selected as replacement drive, and no data copyback occurs, as shown in Figure 3-7.

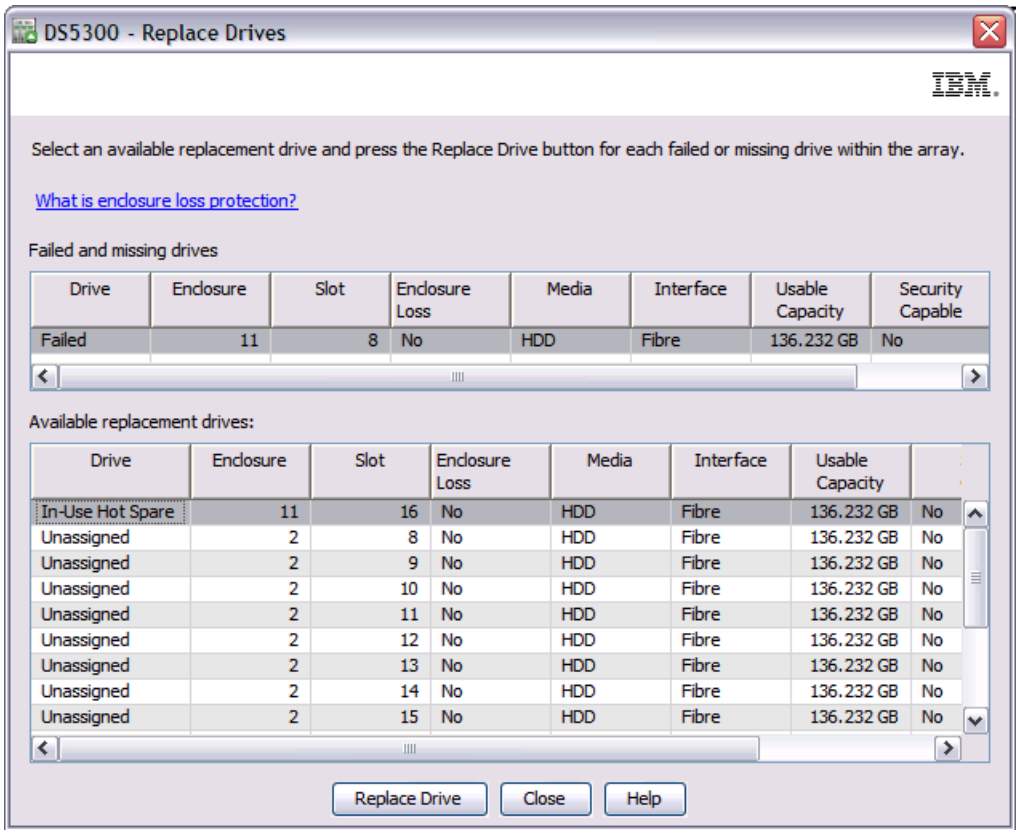


Figure 3-7 Replace drive

The original hot spare drive becomes a permanent member of the array. The location of the hot spare drive in this case is not fixed. A new hot spare drive needs to be *manually assigned* from the unassigned disk drive pool.

Tip: Select this option to avoid the performance penalty of having another copyback process. Make sure to replace the failed drive later, and assign a new hot spare.

Make sure that the new spare drive location does not compromise your enclosure loss protection availability, or performance configuration, and that you are not using a bigger drive or an FDE drive when it is not necessarily needed.

FDE enabled drives

If a global hot spare drive is an unsecured FDE drive, it can be used as a spare drive in secured or unsecured arrays with FDE drives, or as a spare drive in arrays with non-FDE drives, as shown in Figure 3-8.

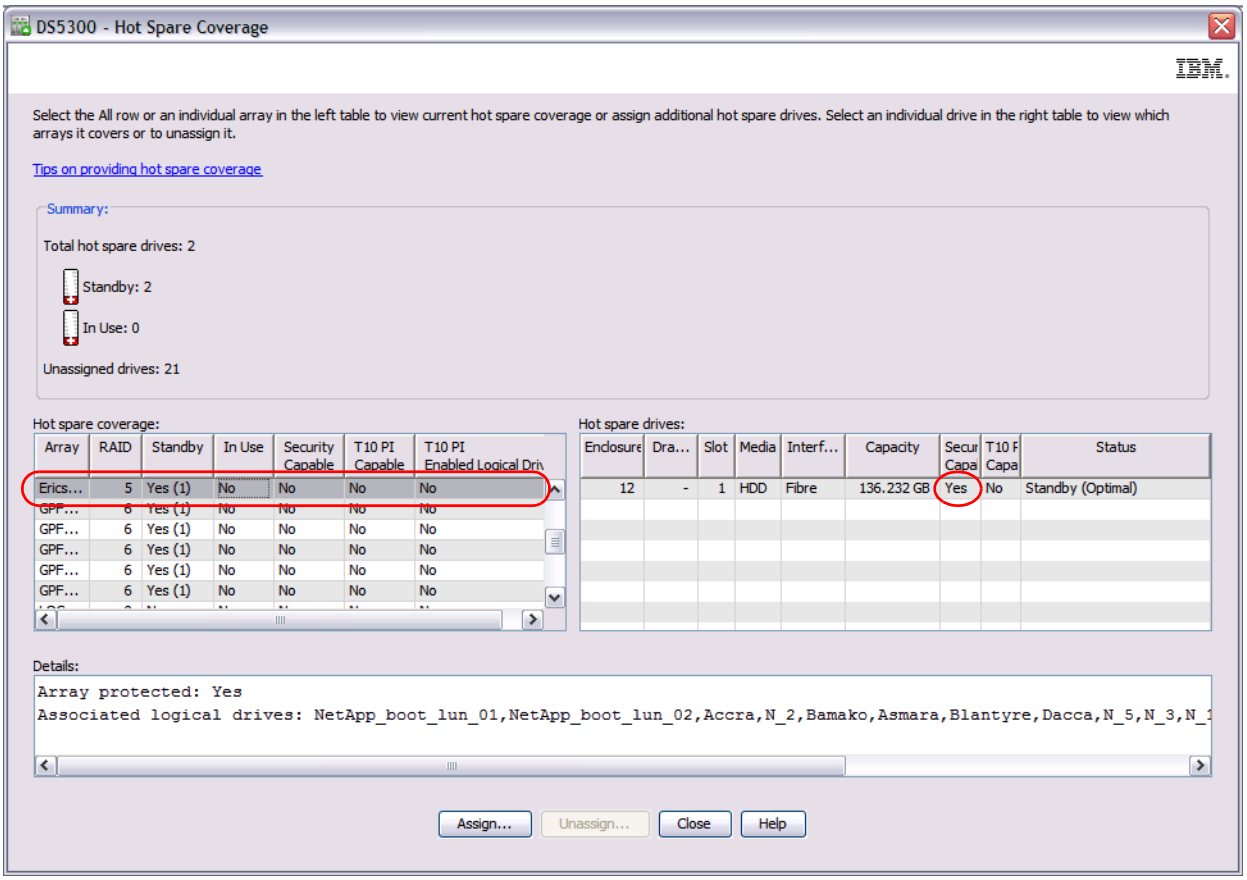


Figure 3-8 FDE capability hot spare drive protecting an unsecure array

When a global hot spare drive is used as a spare for a failed drive in a secure array, it becomes a secure FDE drive and remains secure as long as it is a spare in the secure array. After the failed drive in the secure array is replaced and the data in the global hot spare drive is copied back to the replaced drive, the global hot spare drive is automatically reprovisioned by the controllers to become an unsecured FDE global hot spare drive.

If instead of doing the copyback, you select the spare drive as a permanent drive of the array, then you must secure erase the replaced FDE drive to change it to unsecured state before it can be used as a global hot spare drive.

3.1.2 Creating arrays and logical drives

At this stage, the storage subsystem has been installed and hot spares drives defined. You can now configure the arrays and logical drives according to your requirements. With the SMclient, you can use an automatic default configuration mode to configure the arrays and LUNs for the sizes you want and the RAID type you select. If you have planned your configuration for maximum performance and reliability as described in 2.2, “Planning your DS5000 storage layout” on page 21, you will need to define them manually to enter your specific configuration and needs.

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

For more information about defining arrays, logical drives, and the restrictions that apply, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

To create logical drives, you can define them from unconfigured-capacity, un-assigned drives, or from free capacity in an already existing array:

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

While creating your logical drives, select **Customize settings** to be able to set specific values for the cache settings, and segment size for the logical drive options. In the following section, we list the steps you can use to create a new array.

Creating an array (automatic mode)

Follow these steps:

1. Right-click the unconfigured capacity in the Subsystem Management window and select **Create Logical Drive**. Click **Yes** on Array required (you get it when create first logical drive of the array only) and the wizard for creating the logical drives starts. The first window of the wizard is an introduction to the process, as shown in Figure 3-9. Read the introduction and then click **Next** in order to proceed. We present the best practices during the array creation process.

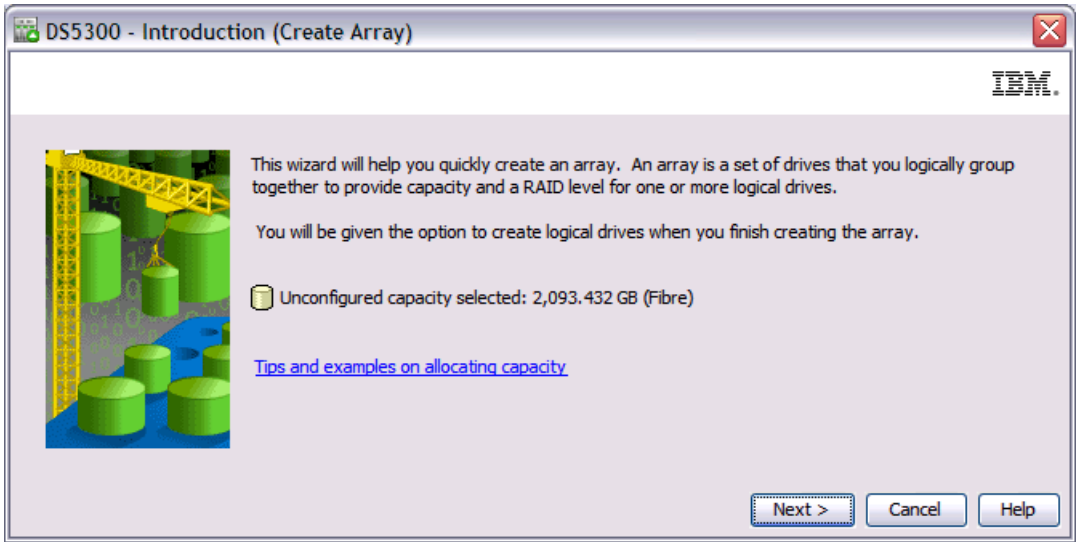


Figure 3-9 Create logical drive

2. After you specify the array name, you have two choices for drive selection, automatic and manual, as shown in Figure 3-10. The default is the automatic method. In automatic mode, the RAID level is used to create a list of available array sizes.

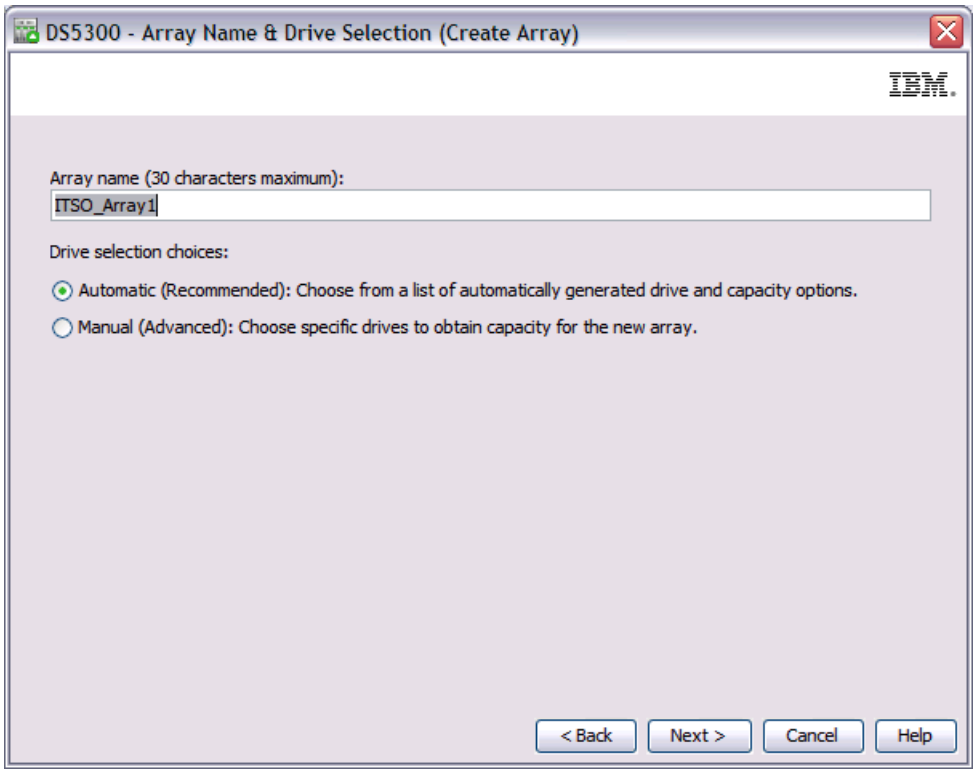


Figure 3-10 Create array with automatic configuration

3. Click **Next >** to select the RAID level and amount capacity of the array. You see a window as shown in Figure 3-11.

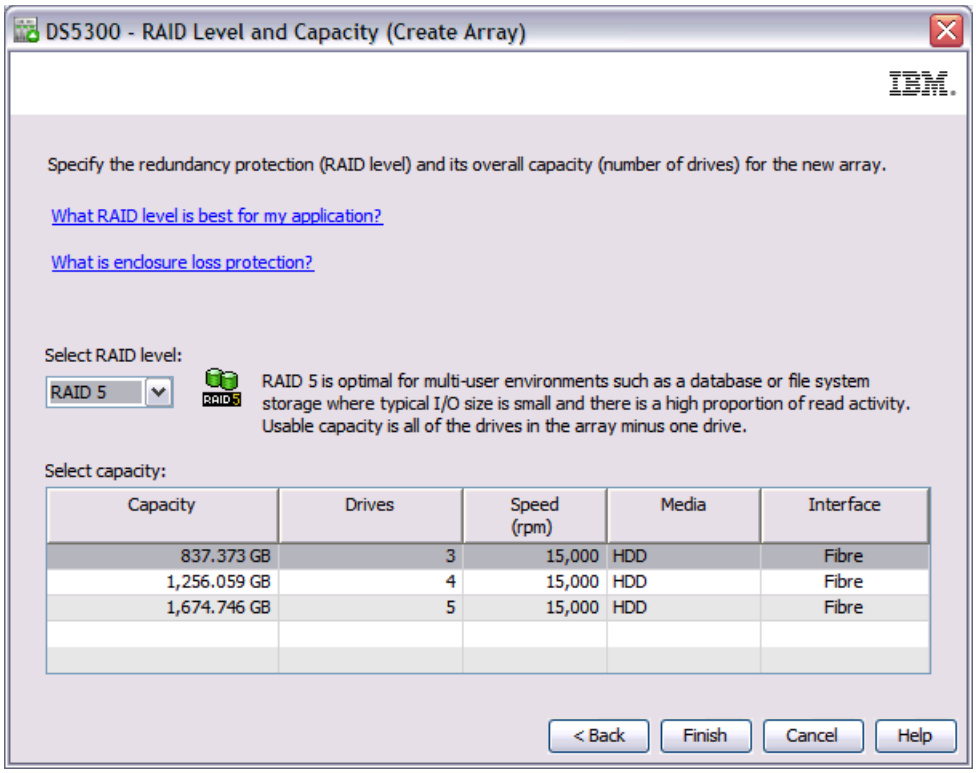


Figure 3-11 Select RAID level and capacity with automatic configuration

4. Select the desired RAID level and the capacity, and click **Finish** to complete array creation.

Manual method: To define your specific layout as planned to meet performance and availability requirements, use the manual configuration method as described in the following section. It allows for more configuration options to be available at creation time, as described in 2.2, “Planning your DS5000 storage layout” on page 21.

Creating an array (manual mode)

Using the manual configuration, you have more choices to select. As we did in the previous example for automatically configuring an array with RAID 5 protection, now let us consider an example to define a similar array manually, but taking care of each particular selection.

When using the manual method, we can create an array as follows (Figure 3-12).

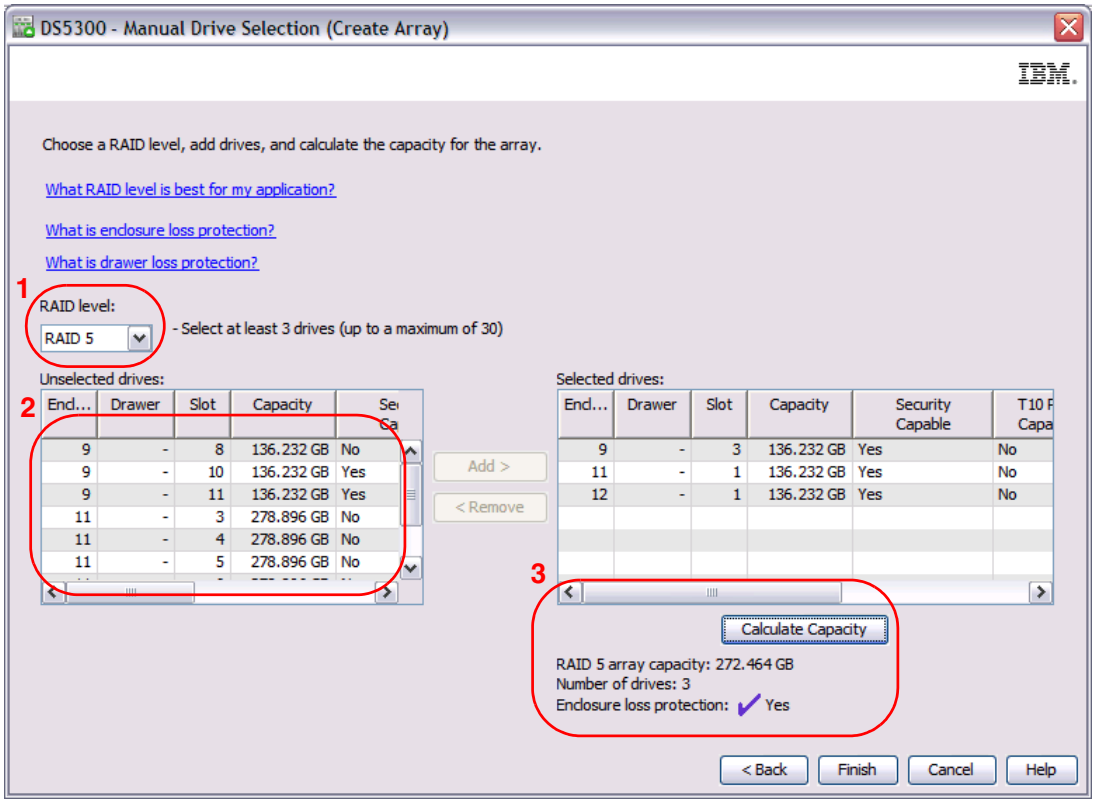


Figure 3-12 Manual configuration for enclosure loss protection

Let us review the three steps needed:

1. First, select the protection RAID protection level desired for your array. For most of the cases, you need a degree of availability protection, which eliminates RAID 0, and then you need to choose the best RAID level according to your applications and business needs. To simplify the selection, Table 3-1 provides a generic guideline.

Table 3-1 Quick guidelines for selecting RAID protection

	RAID 5	RAID 6	RAID 10
Random write performance	2	3	1
Sequential write performance	1	2	3
Availability	3	1	2
Space efficiency	1	2	3

2. Select the drives that you want to assign to the arrays. Ensure that the drives are staggered between the enclosures so that the drives are evenly distributed on the loops. No matter what your RAID level, attempt to use drives from separate enclosures, to optimize performance, and to provide enclosure loss protection if possible.

Tips for choosing array drives:

- Select the same size and speed of drives to create an array.
- To maximize IOPS:
  - More drives are better. Reasonable: 10 drives for RAID 5, 16 drives for RAID 10.
  - Select SSD disk types, then FC, avoiding SATA drives.
- To maximize MBps:
  - Create relatively compact arrays, 4+P or 8+P.
  - Select either FC or SATA drives, not SSD.
- Attempt to select drives from separate enclosures, to optimize controller access performance and array availability, in case of a complete enclosure problem.
- Select drives evenly across both even and odd slots. Because Controller A has a preference to manage odd slots, and Controller B to manage even slots, this approach can result in better performance.
- In configurations with multiple enclosures and high performance requirements, select drives from only odd loop pairs, or only even loop pairs, because controller A is the preferred path to odd loop pairs, and Controller B to even loop pairs.

3. Click **Calculate Capacity** to proceed, after selecting the drives. When enclosure loss protection is achieved, then a checkmark and the word *Yes* will appear in the bottom right side of the window, otherwise, it will have a circle with a line through it and the word *No*. After you are satisfied that capacity and protection have been obtained, click **Finish** to create the array. See Figure 3-12.

After creating the array, the logical drive creating wizard windows are displayed as described in the following sections.

Creating logical drives

After creating an array, you are prompted to continue creating logical drives as shown in Figure 3-13:

1. To alternatively select a free capacity space in any array, right-click it, and select **Create Logical Drive**.

Figure 3-13 Create logical drive wizard

2. Click **Next** to define logical drive properties as shown in Figure 3-14.

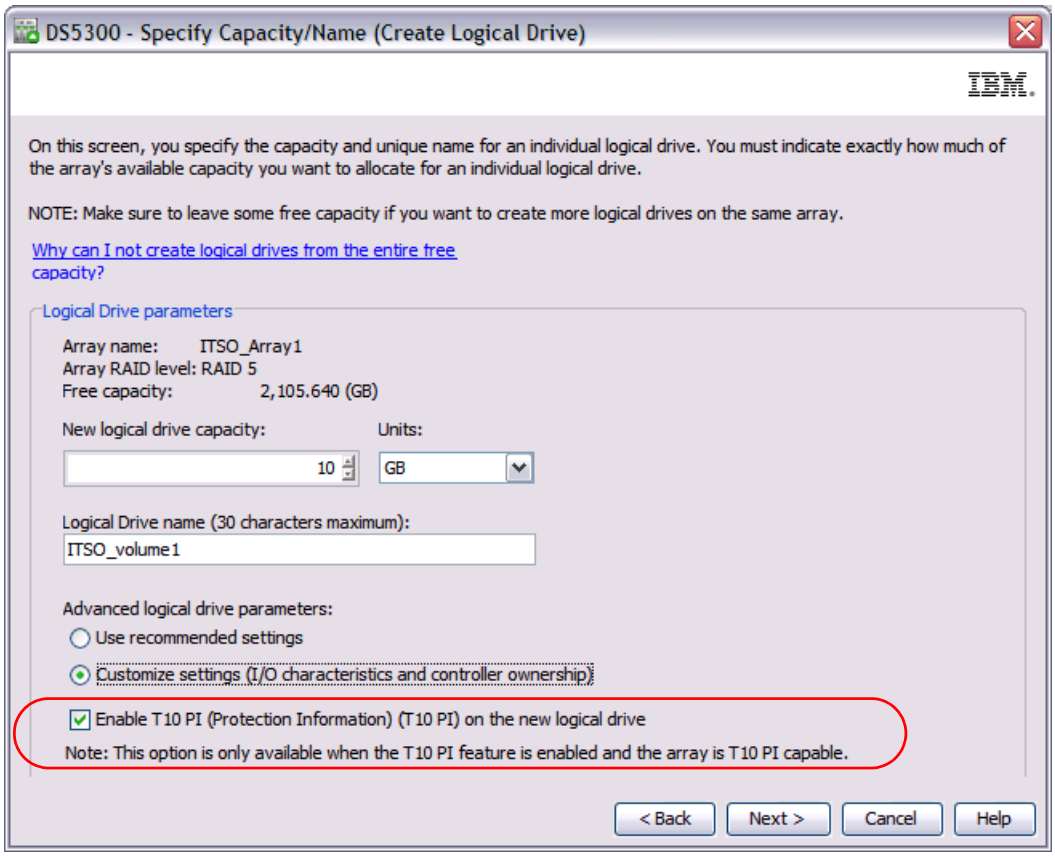


Figure 3-14 Create logical drive wizard

3. Define the logical drive capacity size, assign a name to the logical drive. Then you can either **Use recommended settings**, or **Customize Settings**. Select **Customize Settings** to optimize the new logical drive, by specifying segment size or cache settings option, and click **Next.**

> **Tip:** The feature industry standard extension T10 Protection Information (T10PI) features is now available from firmware version 7.77. To create a T10PI-capable RAID array, all of the drives in the RAID array must be T10PI capable. If you create the logical drive without T10PI functionality, it cannot be converted into a T10PI-enabled logical drive later.

To enable T10PI capability, check the box Enable T10 PI (Protection Information) while creating the new logical drive as shown in Figure 3-14.

Customizing logical drives

The customize parameters window is displayed as shown in Figure 3-15.

Best practice: Select **Customize Settings** in the creation of the logical drive, so you can set your planned specific values according to your configuration. The defaults using the **Use Recommended Settings** option sets a default segment size of 128 KB for every logical volume, except for RAID 3 volumes, which are set to 256 KB.

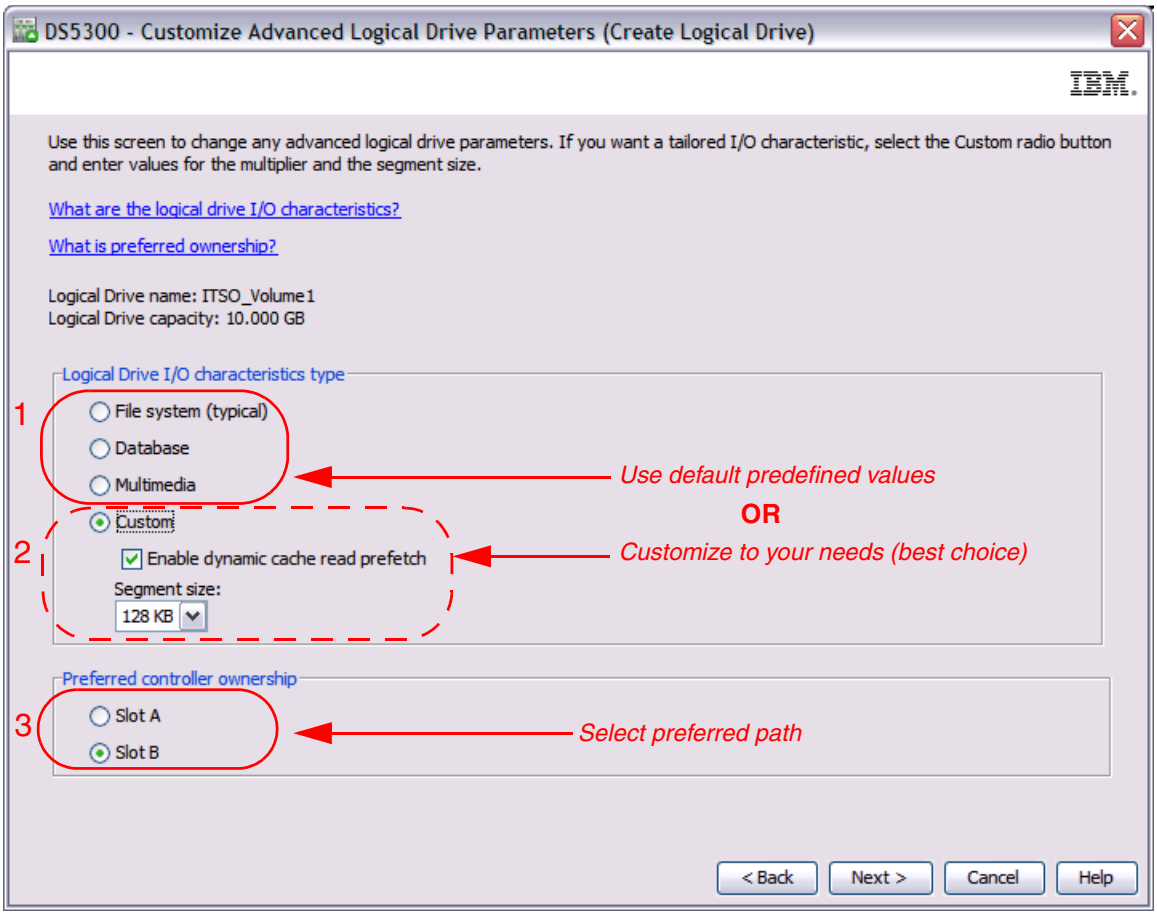


Figure 3-15 Customizing logical drive parameters

You can customize the logical drive using predefined defaults for typical applications or manually, enabling or not, dynamic cache read prefetch, and specifying the best segment size for your configuration. Follow these steps:

1. Specify the logical drive I/O characteristics. You can specify file system, database, or multimedia defaults.
2. Setting a segment size of 128 K for all but for Multimedia and RAID 3 volumes, which are set to 256 KB. The Custom option allows you to disable the dynamic cache read prefetch, and specify the segment size to one of the following values: 8, 16, 32, 64, 128, 256, 512 KB.

   As a best practice, select the **Custom** option and choose the segment size according to the usage pattern as follows.

Segment size:

- For database applications, block sizes between 32–128 KB have shown to be more effective.
- In a large file environment, such as media streaming, CAD, or file system, 128 KB or more are preferred.
- For a Web server or file and print server, the range must be between 16–64 KB.

- The read ahead setting is really an on or off decision because the read ahead multiplier is dynamically adjusted in the firmware. After it is enabled, the dynamic cache read-ahead function will automatically load in advance as many blocks as the current usage pattern.
- If the reads are mostly sequential, then many blocks are read in advance; if it is random, then it will not read ahead any blocks at all. However, in order to analyze the traffic pattern, and decide how many blocks to read in advance, the controller is using processing cycles.
- If you know in advance that the traffic pattern is random, as in database environments, then it is a good choice to disable this option, so you can save this additional processing made by the controllers.

Dynamic cache read prefetch:

- Disable this option for databases, or all traffic patterns where you are sure that the reads are random, to save controller processing time analyzing the traffic to calculate how many blocks are read in advance.
- Enable this option for file system, multimedia, or any other applications that use sequential I/O traffic to increase the read throughput.

3. Choose the preferred controller ownership option as follows.

   You can distribute your logical drives between both controllers to provide better load balancing between them:

   - The default is to alternate the logical drives on the two controllers, so after creating one logical drive, the next one is assigned automatically to the other controller. The problem with this option is that because not all the logical drives are normally used in the same way, you can end with one controller being much more used than the other.
   - Obviously it is better to spread the logical drives by the load they cause on the controller. It is possible to monitor the load of each logical drive on the controllers with the Performance Monitor and change the preferred controller in case of need, so you can split the traffic load evenly across both.
   - The best option, however, which is suitable for configurations with multiple enclosures, is to assign all the logical drives within an enclosure to the specific controller that has the preferred path to the enclosure channel, based on architecture design. Figure 3-16 shows the order of expansion enclosures installation (first column) and the preferred path for each enclosure according to the channel, port loop (first row).

Number of EXPs	A		B		A		B		<----PREFERRED PATH-----|	
	Ch1		Ch2		Ch3		Ch4		Channel #	Controller A
	8	7	6	5	4	3	2	1	Port #	
1	1	-	-	-	-	-	-	-		
2	1	-	1	-	-	-	-	-		
3	1	-	1	-	1	-	-	-		
4	1	-	1	-	1	-	1	-		
5	1	1	1	-	1	-	1	-		
6	1	1	1	1	1	-	1	-		
7	1	1	1	1	1	1	1	-		
8	1	1	1	1	1	1	1	1		
9	2	1	1	1	1	1	1	1		
10	2	1	2	1	1	1	1	1		
11	2	1	2	1	2	1	1	1		
12	2	1	2	1	2	1	2	1		
13	2	2	2	1	2	1	2	1		
14	2	2	2	2	2	1	2	1		
15	2	2	2	2	2	2	2	1		
16	2	2	2	2	2	2	2	2		
	1	2	3	4	5	6	7	8	Channel #	Controller B
	Ch5		Ch6		Ch7		Ch8		Port #	

Figure 3-16 Cabling EXPs and controller preferred paths

Even if it is the best configuration for assigning volume ownership, the counterpart of it is that you need to plan in advance to balance the traffic across both controllers, knowing the expected traffic for each volume.

Plan in advance for this configuration if you want to optimize your DS5000 storage subsystem performance, and use the Performance Monitor to make adjustments to balance traffic across both controllers.

Controller ownership: Assign the controller ownership to each logical drive, considering the preferred path based on the hardware architecture (for configurations for multiple expansions) and the overall traffic balance between each controller.

Use the Performance Monitor to measure the balance across both controllers, and make adjustments as necessary.

When you have completed setting your values as desired in Figure 3-15 on page 107 (either default or custom), click **Next** to complete the creation of the logical drive.

4. Choose between default mapping and storage partitioning, when the Specify Logical Drive-to-LUN Mapping dialog is displayed.

   Storage partitioning is a separate licensing option, but depending in your DS5000 Storage System, the basic configuration might allow more than one partition.

   Unless you are using your DS5000 Storage System with a single host attached, avoid selecting the option for Default Mapping. If you choose **Default mapping**, then the physical volume is mapped to the default host group and is available to any host zoned to the DS5000 Storage System, so it is not a desirable choice if your DS5000 Storage System supports more than a single partition. You can map the logical volumes more accurately later, by selecting **Map later using the Mappings View**.

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

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

In the past, various enhancements provided in new code versions encountered difficulties in implementing certain changes when all the space in the array was used. Therefore, as a good practice, keep extra space free in each array whenever you are able to afford having empty space.

Tip: You cannot create a logical drive with a capacity equal to the entire free capacity remaining in the array. When you create a logical drive, some additional capacity is pre-allocated for Dynamic Segment Size (DSS) migration. DSS migration is a feature of the software that allows you to change the segment size of a logical drive.

3.1.3 Adding free capacity to an array

Add Free Capacity option lets you to expand the capacity of a selected array by adding unassigned drives. It adds additional free capacity to the array while the system remains fully operational. You can use this free capacity to create additional logical drives. This action is called also *Dynamic Capacity Expansion* (DCE).

This procedure might have a performance impact, because the expansion process competes with normal disk access. We suggest that, where possible, that this type of activity be performed when I/O activity is at a minimum.

Adding free capacity: Procedure

To add capacity to an array, follow these steps:

1. Select the array by clicking and choose **Array** → **Add Free Capacity (Drives)...** . As an alternative way, you can also right-click **Array** and select **Add Free capacity**, as shown in Figure 3-17.

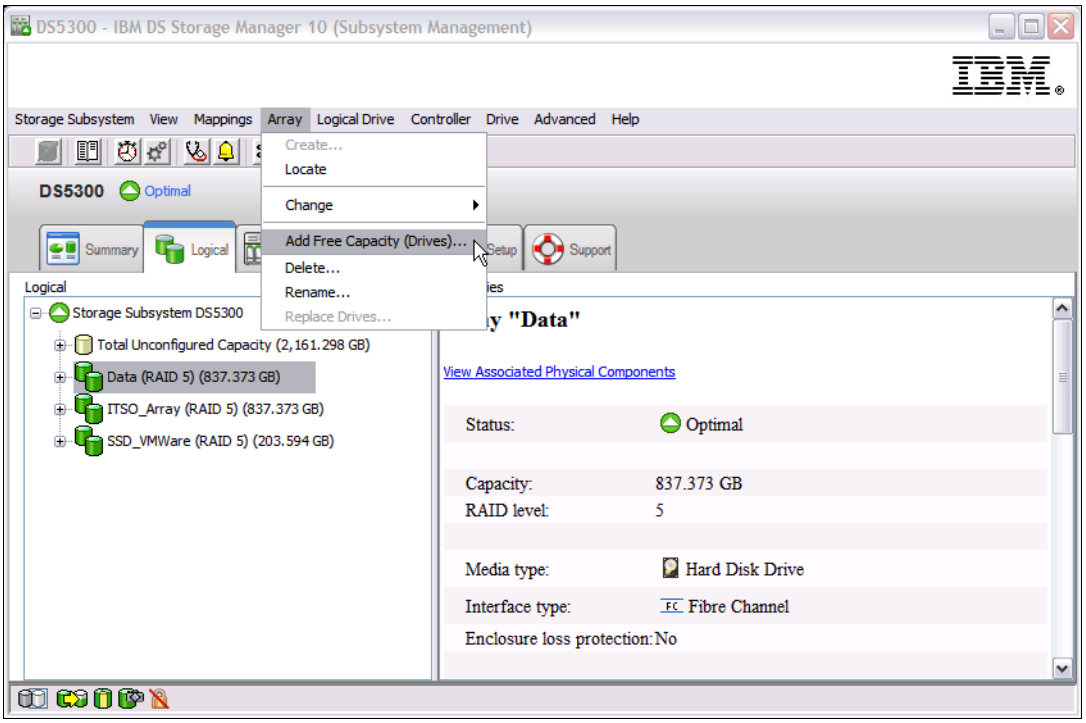


Figure 3-17 Adding free capacity to array

2. The add capacity drives window lists the available drives that you can add to your array, as shown in Figure 3-18.

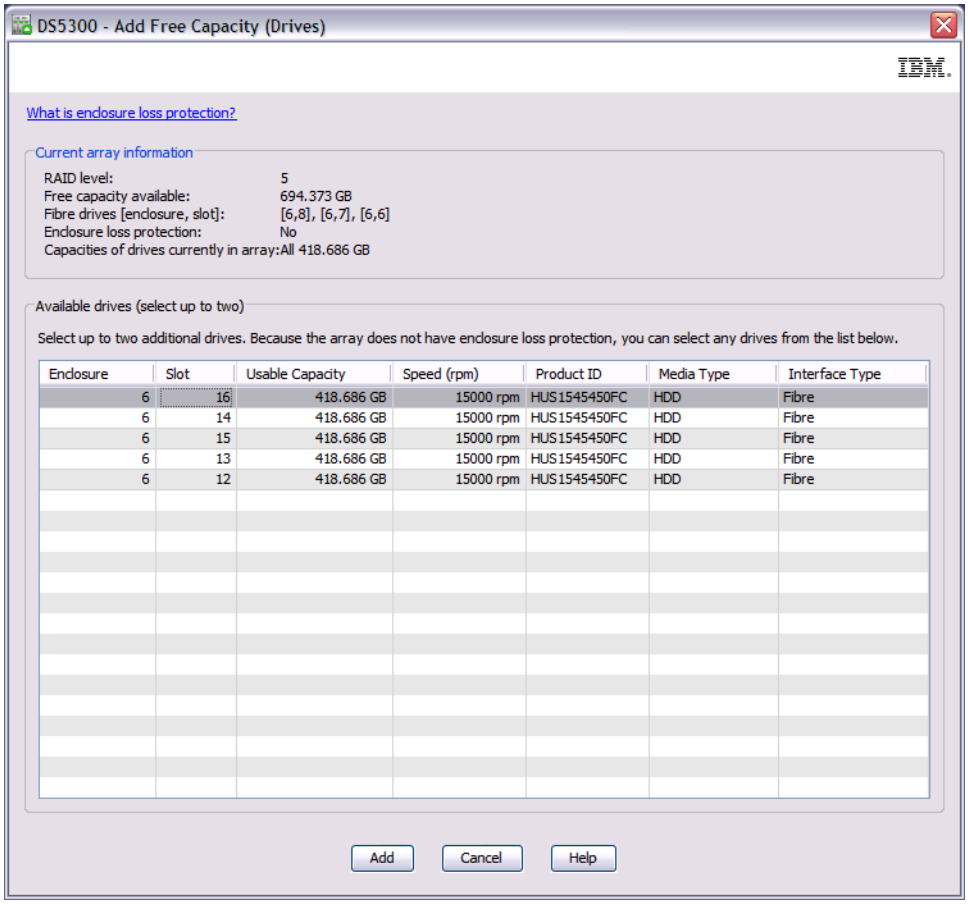


Figure 3-18 Add capacity window

3. Select the drive that you want to add and click the **Add** button, then click **OK** on the confirmation dialog to add drives to your array. A clock on the array's volumes means that the reconfiguration process is ongoing. You can see the operation progress bar just by clicking the volume, as shown in Figure 3-19.

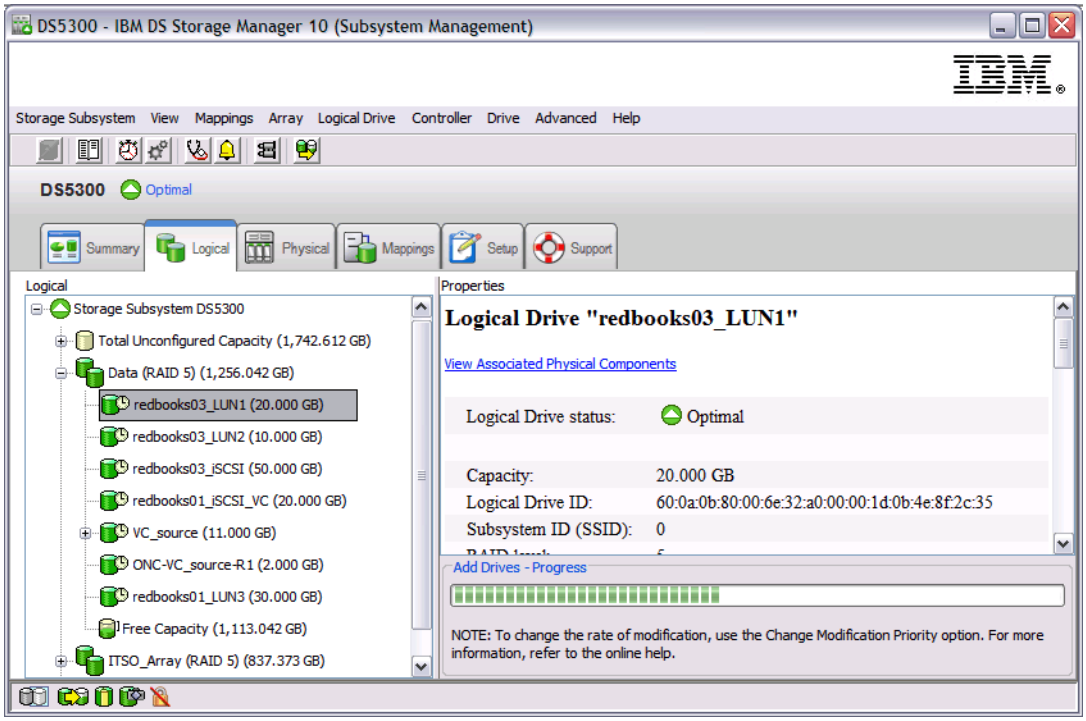


Figure 3-19 Add capacity progress bar

Adding free capacity: Guidelines

Keep these guidelines in mind when you add free capacity to an array:

- You cannot cancel this operation after it begins.
- You cannot mix different media types or different interface types within a single array.
- Data remains available during this operation.
- The array must be in Optimal status before you can perform this operation.
- This option works only if there are unassigned drives in the storage subsystem.
- The existing logical drives in the array do not increase in size. This operation redistributes existing logical drive capacity over the larger number of drives.
- If you add drives with a smaller capacity, be aware that the usable capacity of each drive currently in the array is reduced. Therefore, the drive capacity is the same across the array. If data exists on the current drives in the array that could be lost when the usable capacity is reduced.
- If you add drives with a larger capacity, be aware that the usable capacity of the new drives will be reduced so that they match the current capacities of the drives in the array.
- Only security-capable drives can be added to a security-enabled or security-capable array.

Tip: It is not possible to use more than 30 drives in RAID 3, 5, and 6 arrays. After the maximum number of drives is reached, you obviously cannot add new drives anymore.

3.1.4 Increasing logical drive capacity

It is also possible to increase the size of logical drives by adding free capacity available within the array. This action is called Dynamic Volume Expansion (DVE).

The following operating systems support an increase of capacity for a standard logical drive:

- AIX
- Linux
- NetWare
- Solaris
- Windows Dynamic Disks

Important: Before increasing capacity, be sure that the running operating system supports it. If the logical drive capacity is increased on a host operating system that does *not* support it, the expanded capacity will be unusable and the original logical drive capacity cannot be restored.

Increasing logical drive capacity: Procedure

To increase the logical drive capacity, follow these steps:

1. Highlight the logical drive to be expanded, right-click it, and select **Logical Drive** → **Increase Capacity**... as shown in Figure 3-20.

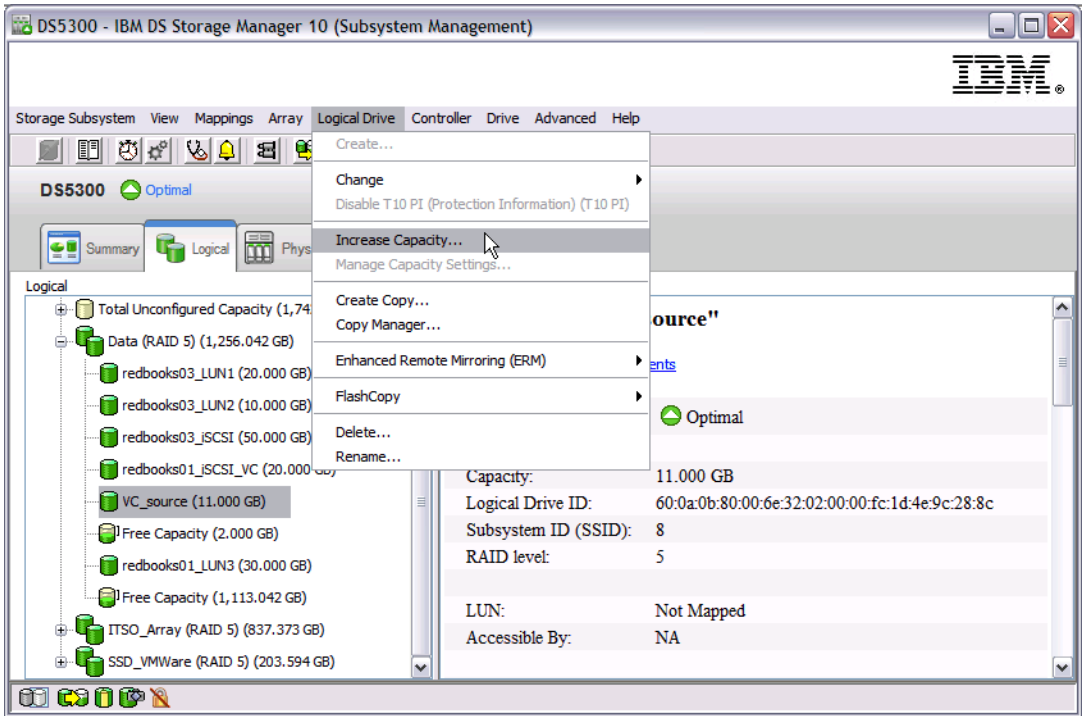


Figure 3-20 Increasing logical drive capacity

2. Click **OK** on the Additional Instruction dialog. An Increase Logical Drive Capacity window appears as shown in Figure 3-21.

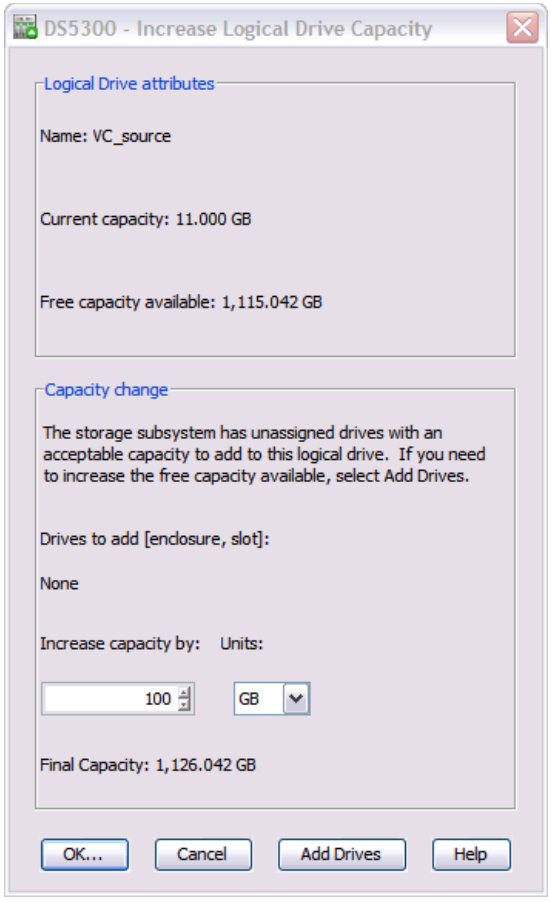


Figure 3-21 Increase logical drive capacity

3. Click **Yes** on the confirmation dialog to increase the logical drive capacity. You can see the operation progress bar just by clicking the volume, as shown Figure 3-22.

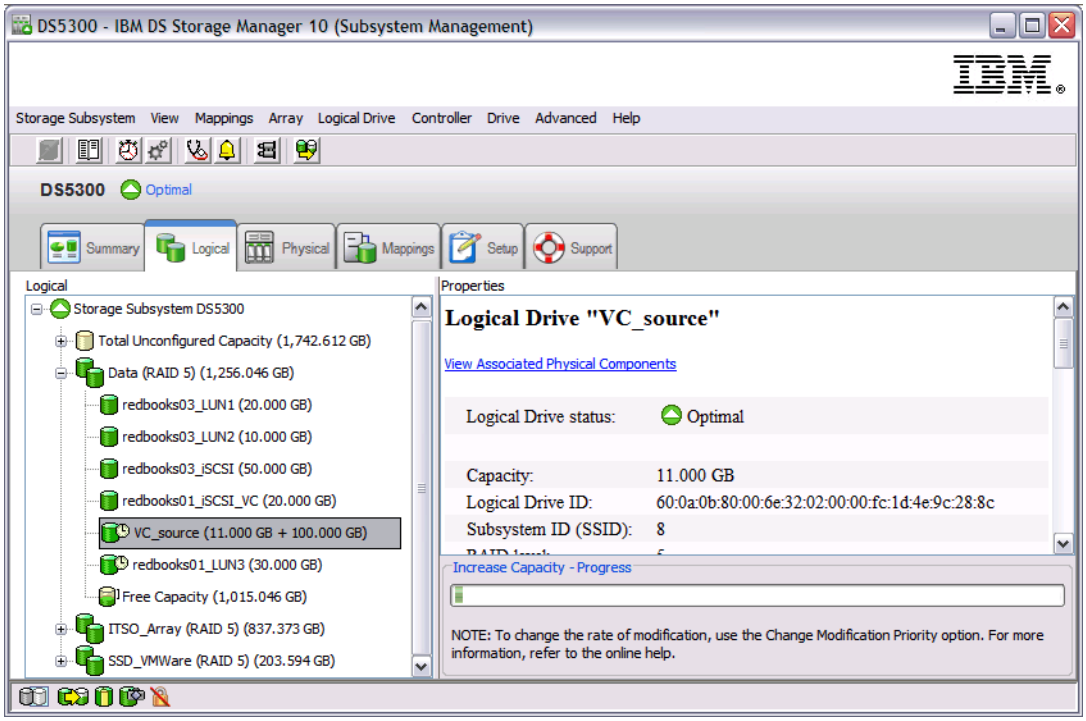


Figure 3-22 Increase logical drive progress bar

Increasing logical drive capacity: Guidelines

Keep these guidelines in mind when you increase logical drive capacity:

- You can increase storage capacity by performing one of these tasks:
  - Use the free capacity that is available on the array of the standard logical drive.
  - Add unconfigured capacity (in the form of unused drives) to the array of the standard logical drive. Use this option when no free capacity exists on the array.
- You cannot increase the storage capacity of a standard logical drive if any of these conditions exist:
  - One or more hot spare drives are in use in the logical drive.
  - The logical drive has a non-Optimal status or does not have an Optimal status.
  - Any logical drive in the array is in any state of modification.
  - The controller that owns this logical drive is in the process of adding capacity to another logical drive (each controller can add capacity to only one logical drive at a time).
  - No free capacity exists in the array and no unconfigured capacity (in the form of drives) is available to add to the array.

3.1.5 Configuring storage partitioning

Because the DS5000 Storage System is capable of having heterogeneous hosts attached, a way is needed to define which hosts are able to access which logical drives on the DS5000 Storage System. Therefore, you need to configure storage partitioning, for two reasons:

- Each host operating system has particular settings that vary slightly, required for proper operation on the DS5000 Storage System. For that reason, you need to tell the storage subsystem the host type that is attached.
- There is interference between the hosts if every host has access to every logical drive. By using storage partitioning, you mask logical drives from hosts which are not to use them (also known as LUN masking), and you ensure that each host or host group only has access to its assigned logical drives. You can have a maximum of 256 logical drives assigned to a single storage partition. A maximum of 2048 logical drives (LUNs) per storage System is possible, depending on the model.

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

The overall process of defining the storage partitions is as follows:

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

Defining the partitions

Follow these detailed steps to define the partitions:

1. First, select the **Mappings View** in the Subsystem Management window. All functions that are performed with regard to partitioning are performed from within this view.

   If you have not defined any storage partitions yet, the Mapping Start-Up Help window is displayed. You are advised to create only the host groups that you intend to use. For example, if you want to attach a cluster of host servers, then you need to create a host group for them. On the other hand, if you want to attach a host that is not a part of the cluster, it is not necessary to put it into a particular host group; however, it might help you to have them grouped together.

   Best practice: Although not required, mapping all hosts to a host group for their specific purpose can prevent confusion and mistakes.

   For detailed information about each of the process steps, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

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

   – Before starting to map logical volumes, create hosts, or host groups, make sure to activate the premium feature for partitions, to expand the current limit to the new limit purchased.

   – When creating logical volumes, avoid mapping them to the default group. Whenever possible, select the option to map them later, so you can assign them individually as needed.

– All information, such as host ports and logical drive mappings, is shown and configured in the **Mappings View**. The right side of the window lists all mappings that are owned by the object you choose in the left side, as shown in Figure 3-23.

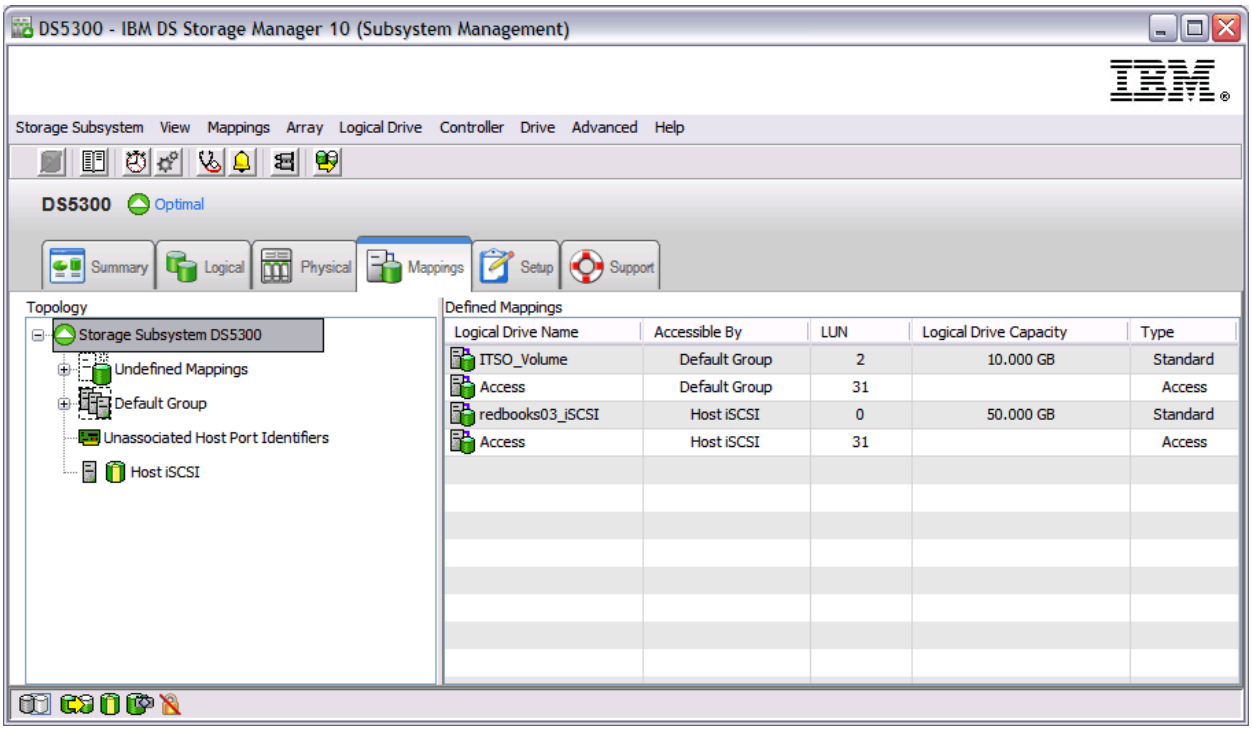


Figure 3-23 Mappings view

If you highlight the storage subsystem, you see a list of all defined mappings. If you highlight a specific host group or host, only its mappings are listed.

2. Move the host if necessary. If you accidentally assigned a host to the wrong host group, you can move the host to another group. Simply right-click the host name and select **Move**. A pop-up window opens so you can specify the host group name.
3. Storage partitioning of the DS5000 Storage System is based on the World Wide Names of the host ports for FC, and the iSCSI Qualified Name (IQN) for iSCSI. The definitions for the host groups and the hosts only represent a view of the physical and logical setup of your fabric. Having this structure available makes it much easier to identify which host ports are allowed to see the same logical drives, and which are in separate storage partitions.

Best practice: Before creating the host and assigning the host ports, it is better have connectivity to the host so you can pick up the host ID or WWPNs from a known unassociated host port identifier instead of typing it manually. The purpose is to prevent possible mistakes.

4. For iSCSI connections, define the iSCSI host ports of each of the ports of both controllers planned to connect iSCSI hosts (see 3.1.6, “iSCSI configuration and management” on page 123).
5. Create your host definitions according to your connection type. Select the connection method first (FC or iSCSI), and assign the host port identifier (WWPN for FC, or IQN for iSCSI. If the host is properly configured and attached to the DS5000 Storage System of the Storage Are Network, you can select the port identifier from a list.

   If for any reason the hosts are not yet configured, or the SAN not zoned, or the network switches not available, then you cannot pick the port identifiers from the dialog box to select it. However, you can still type in the IQN or WWPN of the host adapter manually, as in Figure 3-24, to avoid further delays with your implementation.

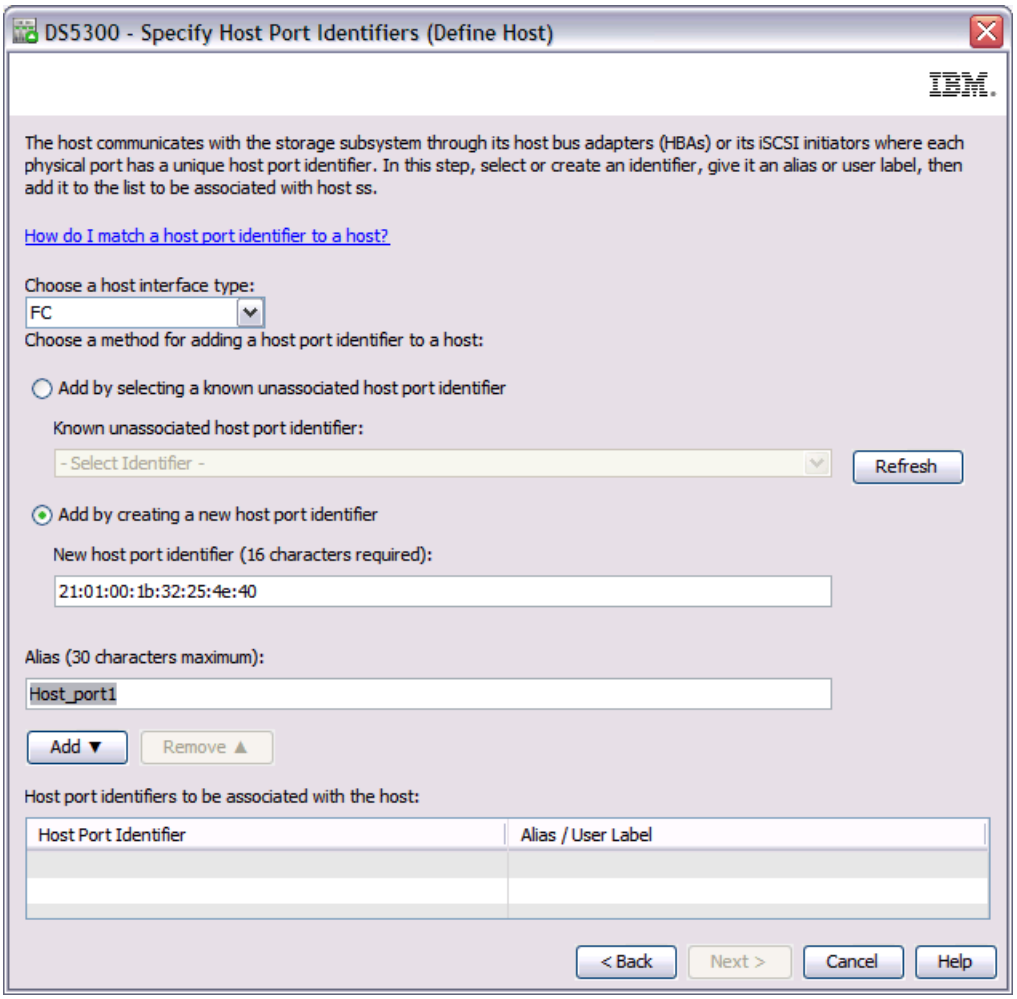


Figure 3-24 Host port identifiers

In order to determine the WWPN, you can read it from the external side of the HBA in certain cases. If not, from Windows or Linux, use the utility provided by the HBA manufacturer, as SANsurfer, or equivalent. In AIX, you can query the adapter VPD using the **lscfg** command, as shown in Example 3-1 on page 120.

Example 3-1 Determine WWPN in AIX

```
# lscfg -vl fcs0 |grep Network
        Network Address.............10000000C955E566
#
```

6. Ensure that if your DS5000 Storage System has mixed host interface cards for both FC and iSCSI, a given partition does not contain hosts that are FC-based as well as hosts that are iSCSI-based. Also, a single host must not be configured for both iSCSI connections and FC connections to the storage system.

   Storage partitioning is not the only function of the storage System that uses the definition of the host ports. When you define the host port, the host type of the attached host is defined as well. Through this information, the DS5000 Storage System modifies the NVSRAM settings to behave according to the host type defined. The most common of the specifications that differ according to the host type setting is the Auto Volume Transfer function, or AVT, which is set as *enable* for certain host types, and *disable* for others.

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

7. Assign the mapping of a logical volume to a host group, only if you need to share the volume across more than one host. If you have a single server in a host group that has one or more logical drives assigned to it, and no other host has to use the same volumes, assign the mapping to the host, and not the host group. Numerous servers can share a common host group; but might not necessarily share drives. Only place drives on the host group mapping that you truly want *all* hosts in the group to be able to share.

8. If you have a cluster, assign the logical drives that are to be shared across all, to the host group, so that all of the host nodes on the host group have access to them.

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

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

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

Follow the aforementioned best practices to configure your mappings, hosts, and partitions. Remember that after it has been done, the host mapped can already start using the logical volumes mapped.

Managing host port identifiers

With the hosts already defined, now let us see how you can get additional help using the Storage Manager to view and make changes to the assignments already done,

Because your DS5000 Storage subsystem can support many hosts, you can use the option **Mappings** → **Manage Host Port Identifiers** from the Subsystem management window to list all the host port identifiers defined in your configuration as shown in Figure 3-25.

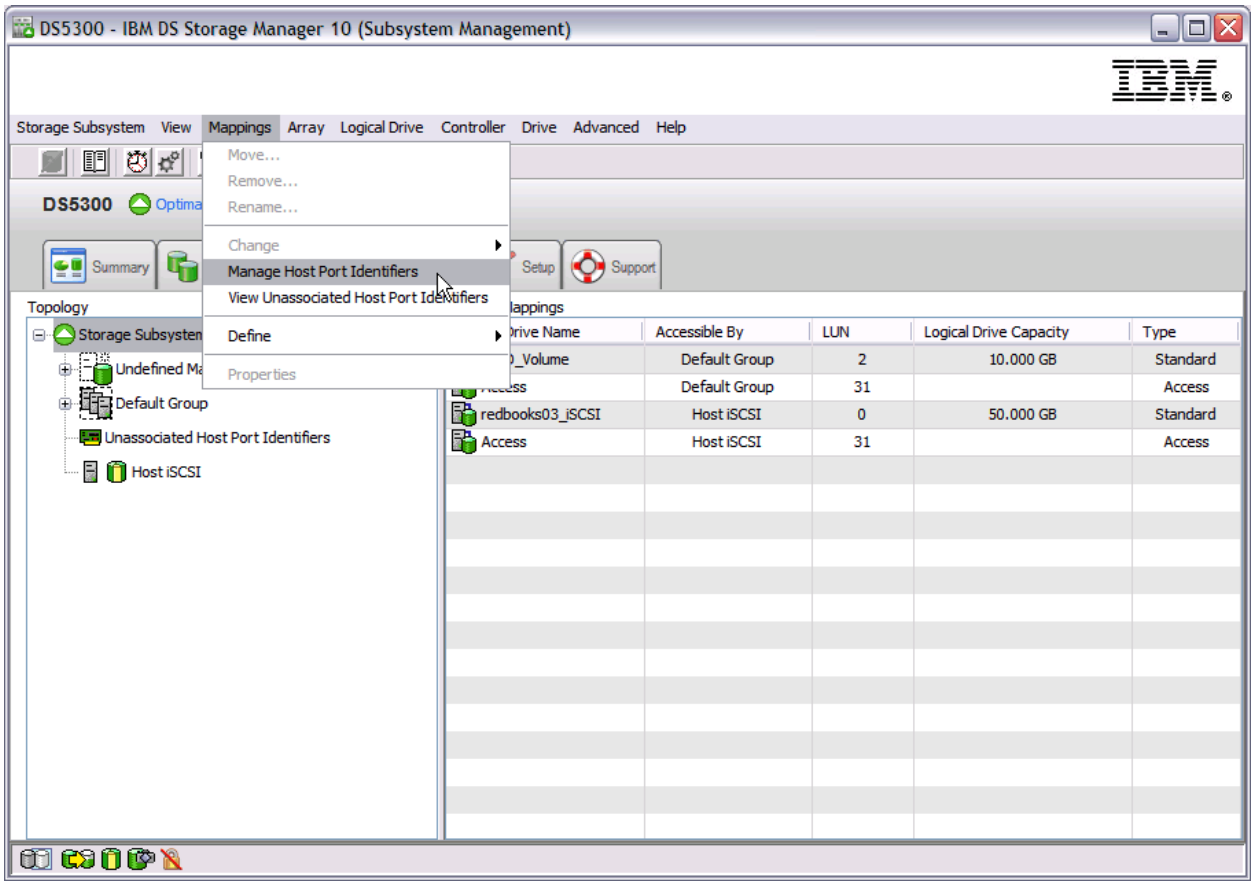


Figure 3-25 Manage host port identifiers

The Manage Host Port Identifiers window opens, from where you have the facility of listing all hosts or selecting a specific one in a single window. See Figure 3-26.

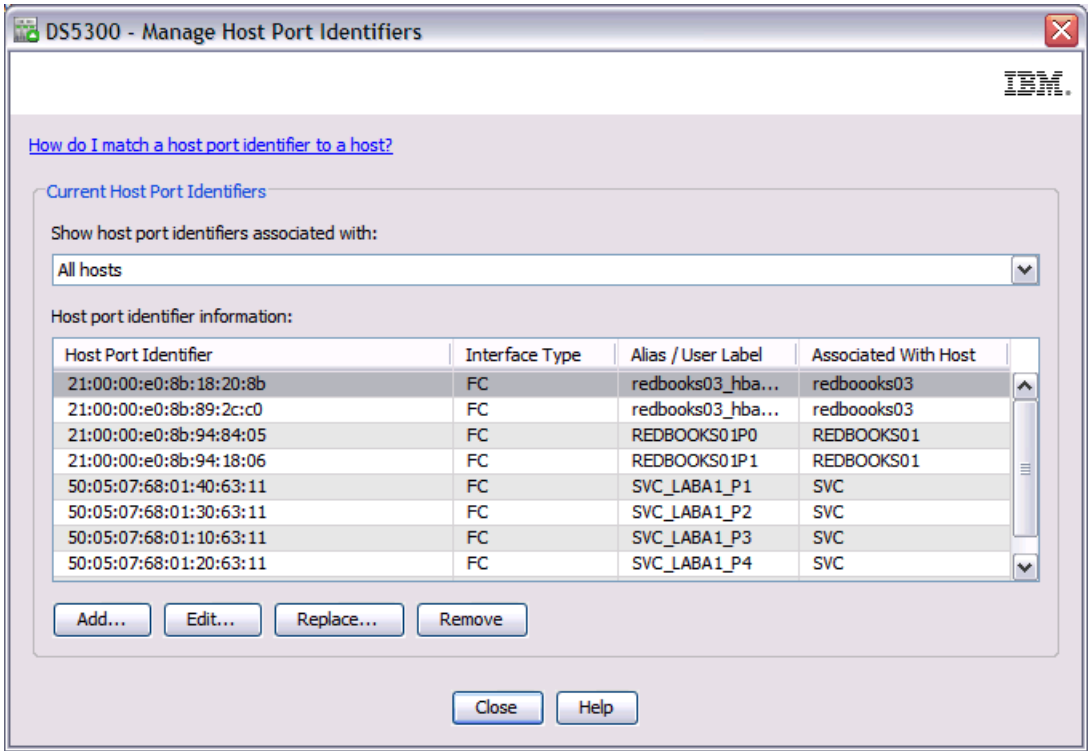


Figure 3-26 Manage host port identifiers

Use this window if you want to modify a host attachment setting, such as replacing a host bus adapter. You can also use it to add new host mappings, however, remember that this first view only shows defined hosts.

To check what other host port identifiers are already recognized by your storage subsystem, but are not already assigned to a Host, select the option **Mappings → View Unassociated Host Port Identifiers** from the Subsystem management window. If there are any, they are shown in a window as in Figure 3-27.

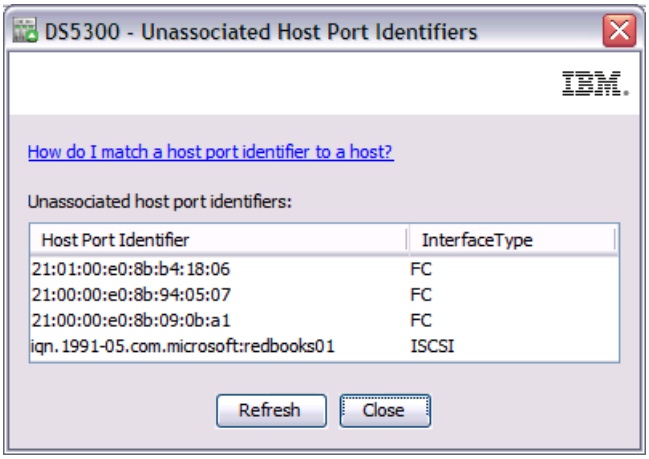


Figure 3-27 Unassociated host port identifiers

Depending on your operating system, you can use a utility to dynamically discover the mapped logical volume, avoiding a complete rescan or reboot. The utility, installed with the DS5000 Utilities package, provides the **`hot_add`** command line tool for certain operating systems. You simply run **`hot_add`**, and all host bus adapters are re-scanned for new devices, and the devices must be accessible to the operating system.

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

3.1.6 iSCSI configuration and management

The DS5000 Storage System has two separate host interface card options to connect your hosts, iSCSI and FC. Configuring the arrays and logical drives is the same for both connection types. The only difference to configure iSCSI attachment from your DS5000 Storage Manager is how the hosts are defined, by its iSCSI Qualified Name (IQN), instead of the WWPN as in Fibre Channel, as presented in 3.1.5, “Configuring storage partitioning” on page 117.

Next we show how to set up a host for iSCSI, and the specific options available in the Storage Manager for setup and management of iSCSI connections.

For additional information about iSCSI, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Configuring iSCSI

To configure iSCSI, follow these steps:

1. Start the iSCSI attachment configuration defining the iSCSI host ports. Do it for each of the ports of both controllers that you are planning to use for connecting iSCSI hosts. From the Setup tab of the Subsystem management window, select the option **Configure iSCSI Host Ports,** as shown in Figure 3-28.

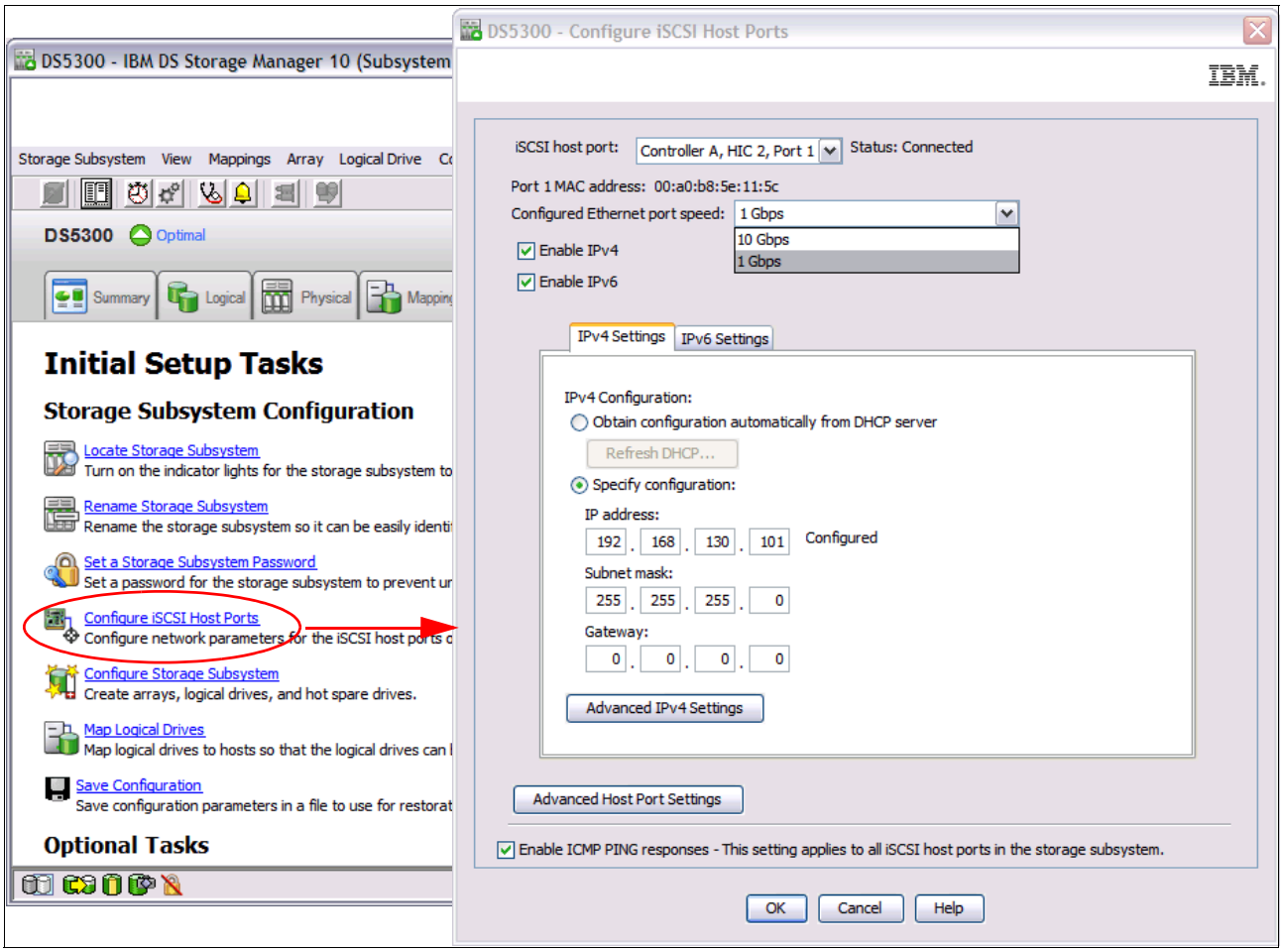


Figure 3-28 Configure iSCSI host ports

2. Select an iSCSI host port, select Ethernet port speed, and assign the IP address, mask, and gateway as usual with all Ethernet cards, in either of the IPv4 or IPv6 tabs. If you are planning to use VLAN Tags (802.1Q), or set Ethernet priority (802.1P), click **Advanced IP Settings** to define them, as shown in Figure 3-29. Repeat the process for at least one port in controller A, and one port in controller B.

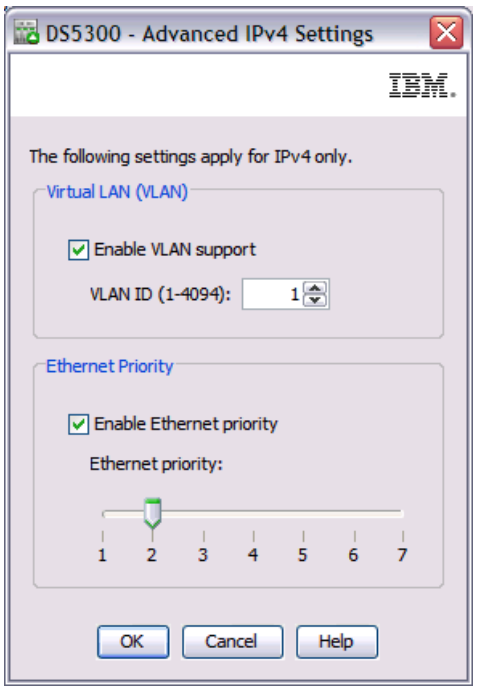


Figure 3-29 VLAN and priority

3. If you plan to use Jumbo frames, click **Advanced Host Port Settings** to enable it and select the frame size to either 1500 or 9000 Bytes. Here you can change the default TCP port 3260 for any other port. See Figure 3-30.

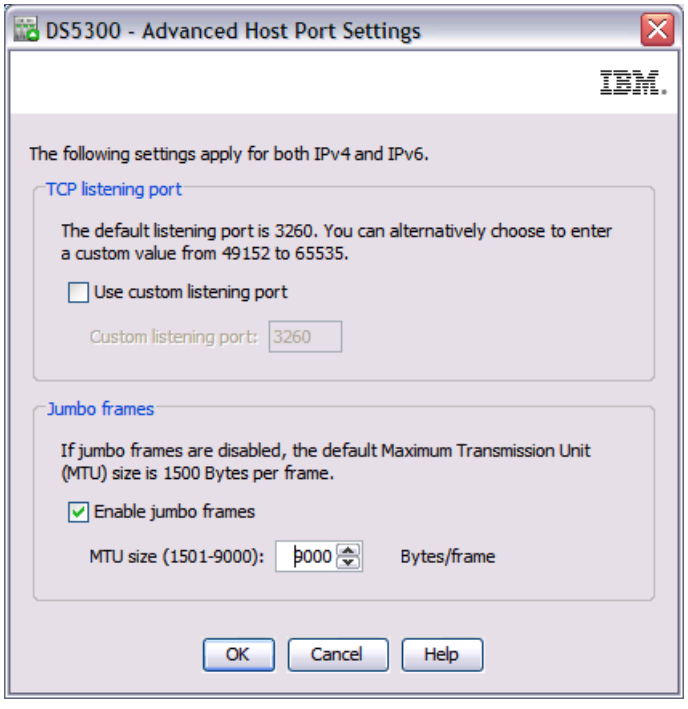


Figure 3-30 Jumbo frames

Now you can create your host definitions according to your connection type. However, the host server needs to be already configured, so you can easily map its port identifiers to the DS5000 Storage Manager host definition.

The storage subsystem iSCSI ports support the Internet Protocol version 6 (IPv6) TCP/IP. Note that only the final four octets can be configured if you are manually assigning the local link address. The leading four octets are fe80:0:0:0. The full IPv6 address is required when you are attempting to connect to the target from an initiator. If you do not provide the full IPv6 address, the initiator might fail to be connected.

Configuring using an iSCSI host adapter

Use your iSCSI adapter management software to configure the HBA. In our example, we show how to configure a QLogic iSCSI adapter using iSCSI SANsurfer. You can also use the BIOS setup menu. If you are planning to boot from your iSCSI attached disk, then you need to use the BIOS adapter menu to enable it. For more information, see *SAN Boot Implementation and Best Practices Guide for IBM System Storage*, SG24-7958.

1. Begin by starting the iSCSI SANsurfer, and check the adapter drivers, BIOS, and firmware levels. Make sure that they are at the required support level. See Figure 3-31.

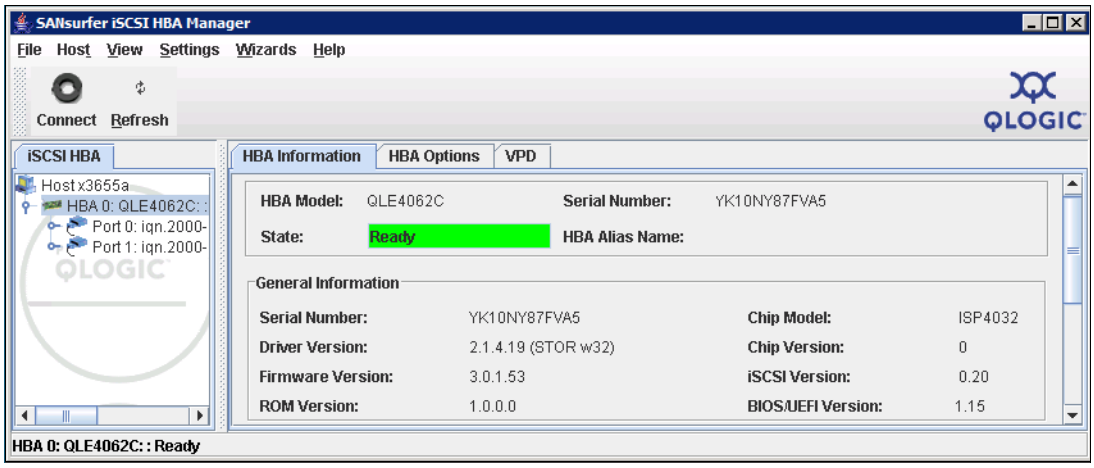


Figure 3-31 Displaying iSCSI card levels

2. Configure your iSCSI adapter port with the network settings already planned. On the SANsurfer iSCSI HBA Manager main window HBA tree, select the HBA port of the card to configure. On the Port Options page, click the **Network** tab. See Figure 3-32.

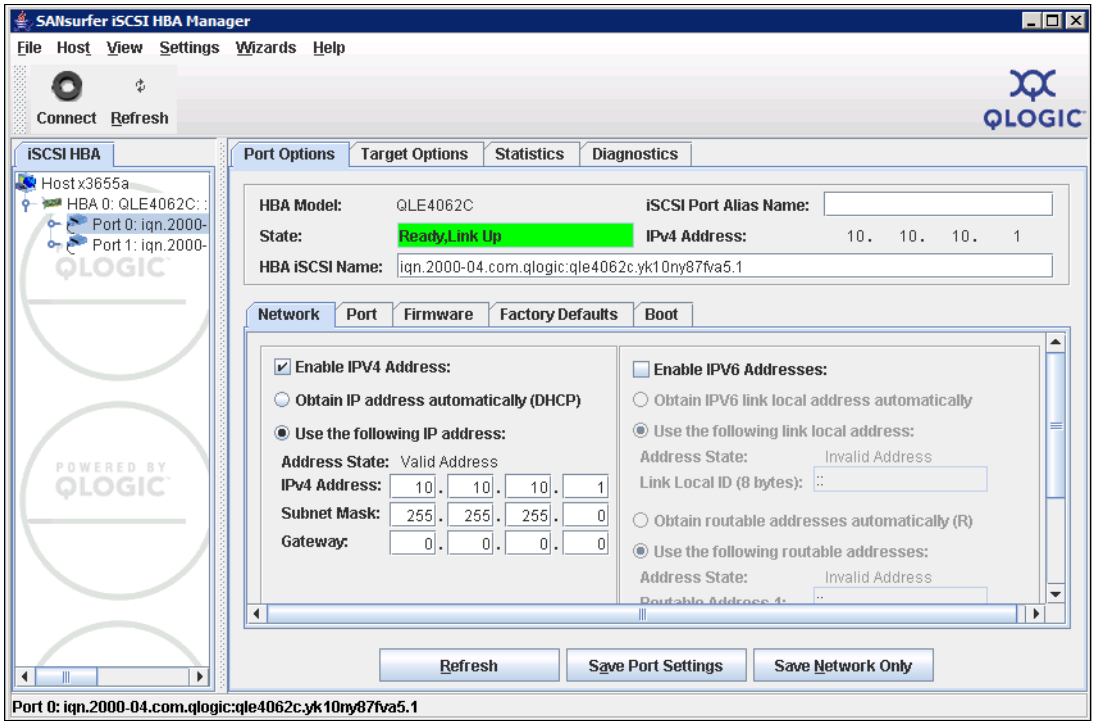


Figure 3-32 Setting iSCSI port

Enable IPv4 or IPv6, and select DHCP or manually assign the IP address, mask, and gateway. You can also assign an alias name for the port. Save the setting when done.

3. Set Jumbo frames, VLAN, and Ethernet priority settings according your previous configuration settings defined using the Storage Manager. Move to the **Firmware Tab** and edit the values as shown in Figure 3-33.

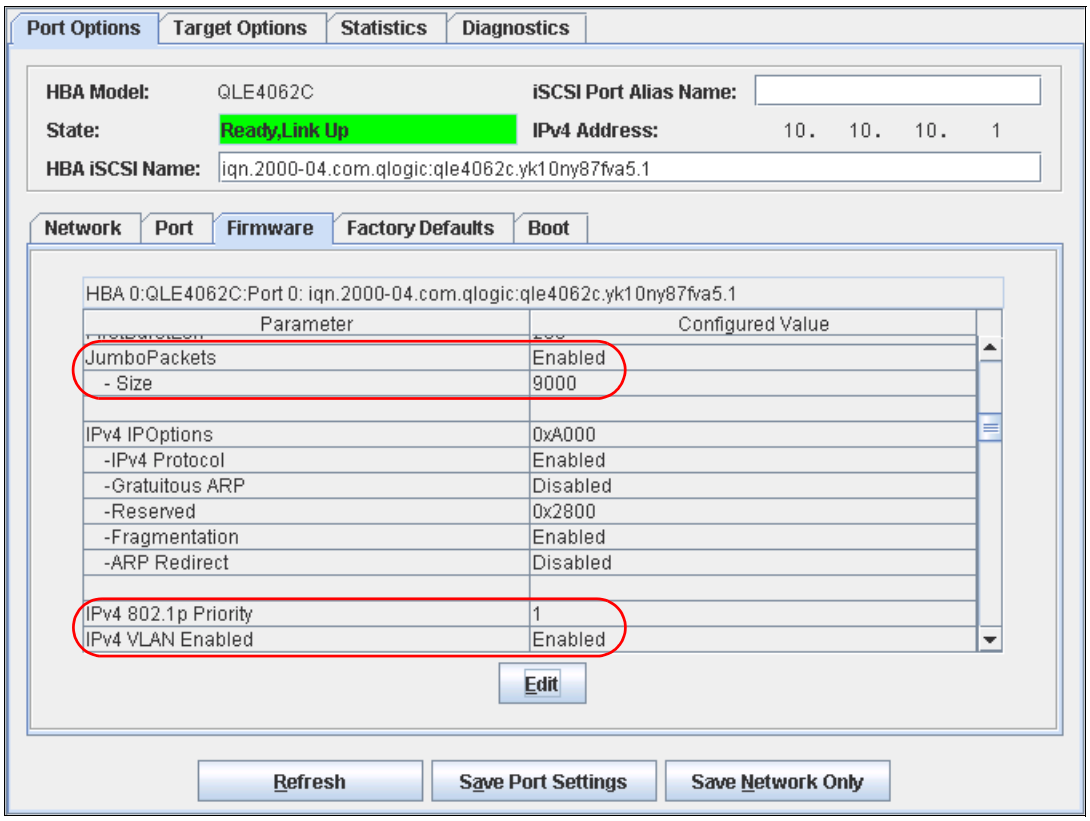


Figure 3-33 iSCSI additional customization

Set Jumbo Frames, VLANs, and Ethernet priority in the iSCSI adapter in accordance to the settings specified in the DS5000 Storage Manager in Figure 3-29, “VLAN and priority” and in Figure 3-30, “Jumbo frames” on page 125.

4. Now move to the **Target Settings** tab, and add the IP address of the controller A of your subsystem as shown in Figure 3-34.

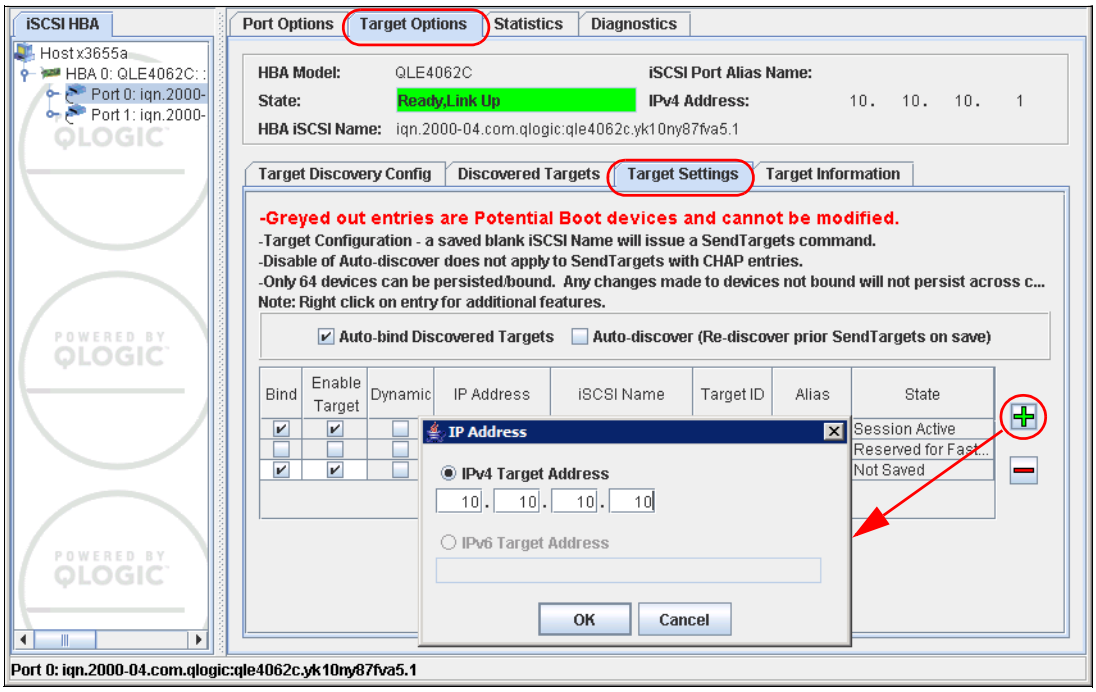


Figure 3-34 Target settings

5. Save the configuration of the HBA selecting the option, **Save Target Settings** (after adding the Target of the DS5000 Storage Controller to map to this iSCSI HBA port). The adapter resets and refreshes its configuration, querying the added target server, and showing any devices discovered.
6. Repeat the same process to add the controller B IP address. For redundancy, use dual iSCSI HBA cards. In our example, we used a dual port iSCSI HBA.

7. Check the iSCSI names of the adapters to ensure that they are recognized, as follows. After both controller IP addresses are configured as targets in the iSCSI HBA ports, and because we already defined the iSCSI host ports as in Figure 3-28, “Configure iSCSI host ports” on page 124, then the iSCSI names of the adapters must be recognized by the DS5000 storage subsystem.

   Select the option **Mappings** → **View Unassociated Host Port Identifiers**. Confirm the iSCSI names of the recently configured HBA ports are shown, as in Figure 3-35, for both iSCSI HBA ports.

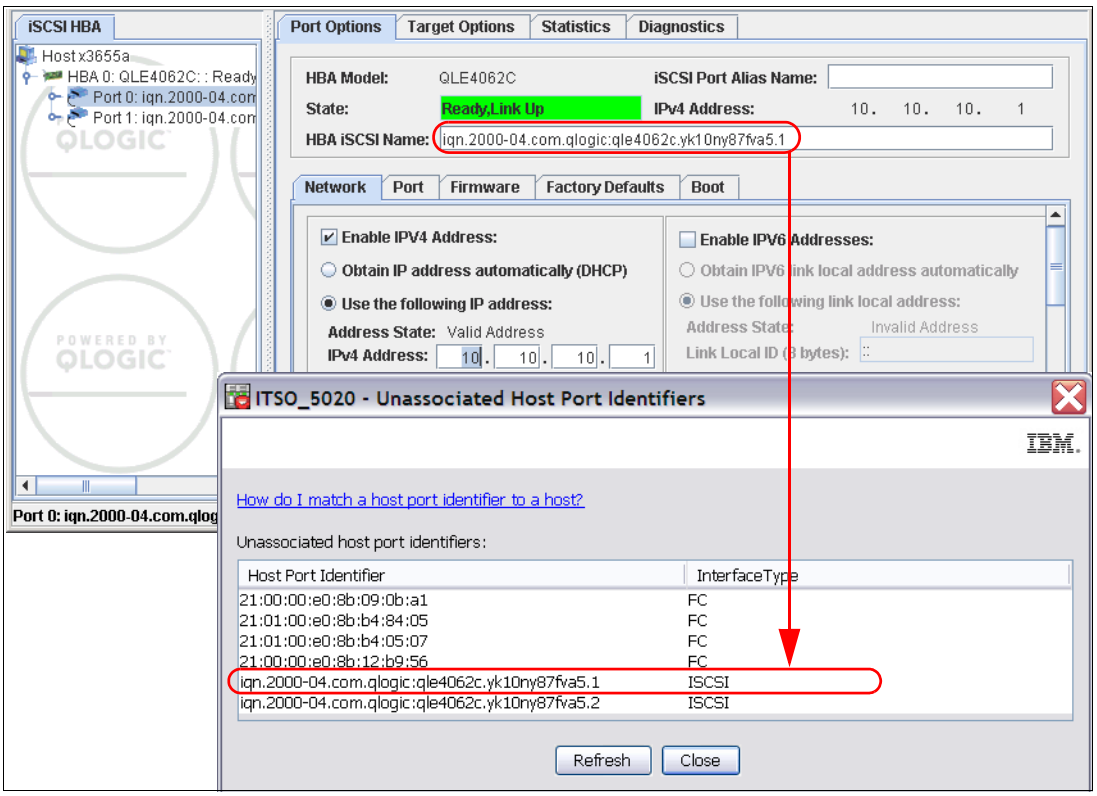


Figure 3-35 Checking host ports identifiers

The configuration of your iSCSI host adapter is now finished.

8. Define the hosts with their corresponding host port identifiers in the DS5000 Storage Manager (later we can map the logical drives desired).

   Select from the DS5000 Subsystem Management window mappings view tab, the option **Mappings** → **Define** → **Hosts.** Provide a meaningful host name, and click **Next**.

   Select **iSCSI** as the host interface type. Pick one of the identifiers from the known list for the first port, assign a label for the port, and click **Add** as shown in Figure 3-36.

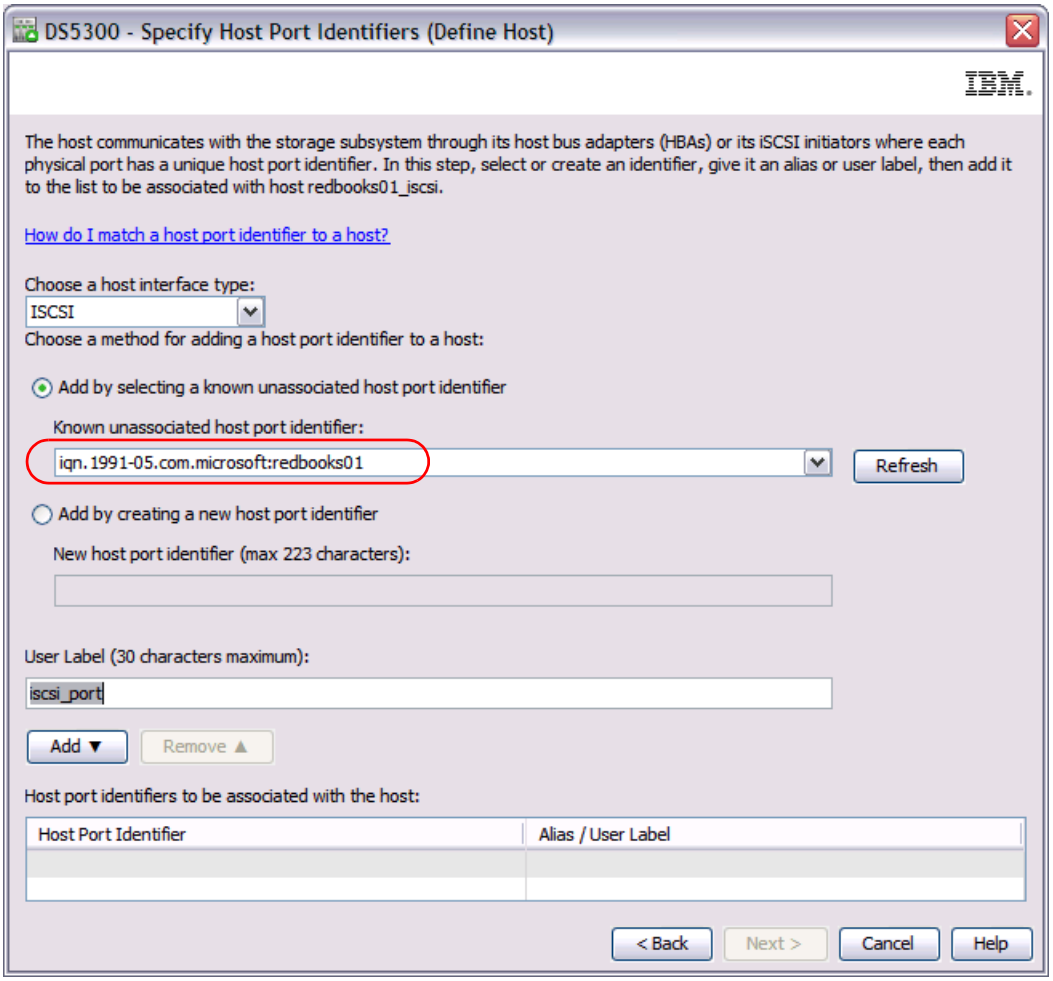


Figure 3-36 Adding iSCSI port identifier

9. Repeat the same steps to add the remaining host port identifier. Because we defined two iSCSI HBA ports, we need to map both to the definition of the host in the Storage Manager.
10. Continue assigning the host type and specifying if the host will be in a partition to finish the Host Creation wizard.
11. Now that the host is defined with both iSCSI ports, map the desired logical volumes to this host as usual. Remember the partitioning guidelines described in 3.1.5, “Configuring storage partitioning” on page 117.

Considerations for iSCSI partitioning:

- If your Storage System has mixed host interface cards for both FC and iSCSI, a given partition must not contain hosts that are FC-based as well as hosts that are iSCSI-based.
- A single host must not be configured for both iSCSI connections and FC connections to the storage system.

12. Confirm that the logical volumes mapped are recognized by the host, as shown in Figure 3-37.

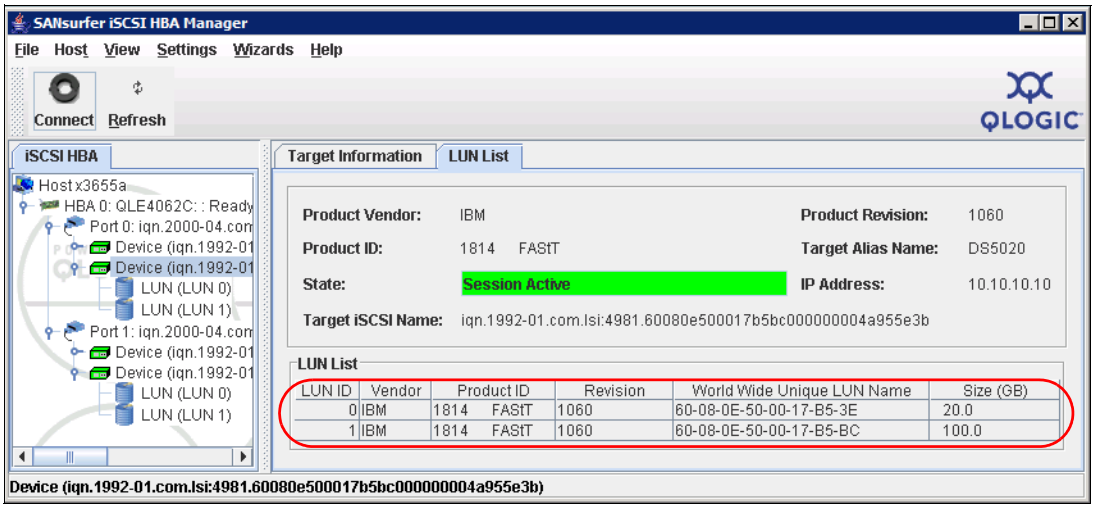


Figure 3-37 Viewing mapped LUNs at the host

Additional information

For additional information about using the mapped hosts and verifying multipath driver configuration, as well as how to configure iSCSI Software Initiator, see 4.3.8, “iSCSI Software Initiator implementation” on page 184.

We cover additional iSCSI settings such as security authentication and iSNS target discovery in “Managing iSCSI settings” on page 133.

When host configuration is completed, we proceed to define the host with the corresponding host port identifiers in the DS5000 Storage Manager, so that later we can map the logical drives desired (see 3.1.5, “Configuring storage partitioning” on page 117).

Managing iSCSI settings

For a quick review of your current iSCSI configuration, you can select the **Logical** tab of the Subsystem Management window, and go down in the right frame to display iSCSI properties, as shown Figure 3-38.

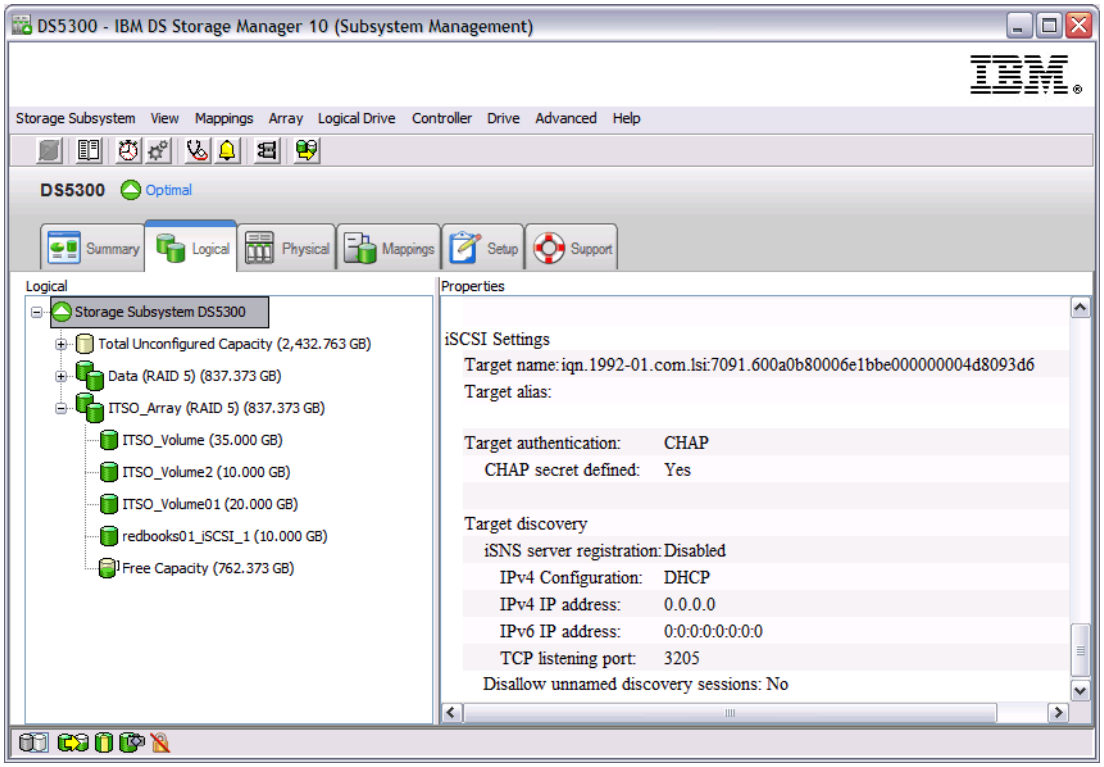


Figure 3-38 Viewing iSCSI settings

In order to manage the iSCSI connections and configuration, select from Subsystem Management in the Storage Manager, **Storage Subsystem** → **iSCSI** → **Manage Settings**, as in Figure 3-39. Or, select the option **Manage iSCSI Settings** from the Setup tab.

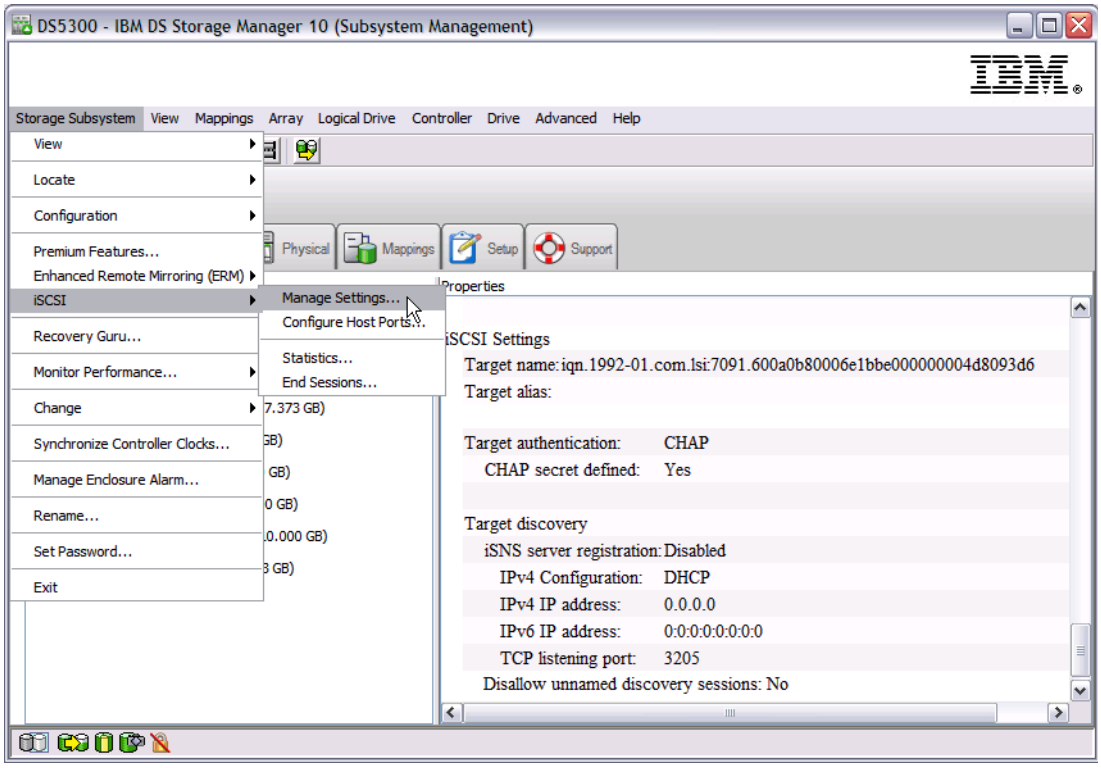


Figure 3-39 Managing iSCSI settings

The iSCSI Manage Settings window can be used for the following tasks:

- Configuring iSCSI security authentication:
  - For Target Device, DS5000 Subsystem
  - For Initiator Device, iSCSi Host Bus Adapter
- Identification and alias creation for both Target and Initiator
- Target discovery configuring ISNS

iSCSI Security Authentication

Challenge Handshake Authentication Protocol (CHAP) is the authentication scheme implemented by the DS5000 Storage System. CHAP validates the identity of remote clients, when you establish the initial link. The authentication is based on a shared secret and you can enable it in only one-way, where the initiator needs to authenticate with the target, or two-way, where the initiator needs to request identification from the target too.

If an initiator and a target negotiate during login to use authentication, by default, the initiator authenticates to the target that it is who it says it is, given that the initiator knows the target Challenge Handshake Authentication Protocol secret (CHAP secret). If the initiator optionally requests mutual authentication, then the target authenticates to the initiator that the target is who it says it is, given that the target knows the initiator's CHAP secret.

Target authentication

Use the **Target Authentication** tab to specify the target challenge handshake authentication protocol (CHAP) secret that the initiator must use during the security negotiation phase of the iSCSI login. By default, **None** is selected. To change the selection, click the check box for **CHAP**, and click **CHAP secret** to enter the CHAP secret. You can also select the option to generate a random secret, which enables 1-way CHAP. See Figure 3-40.

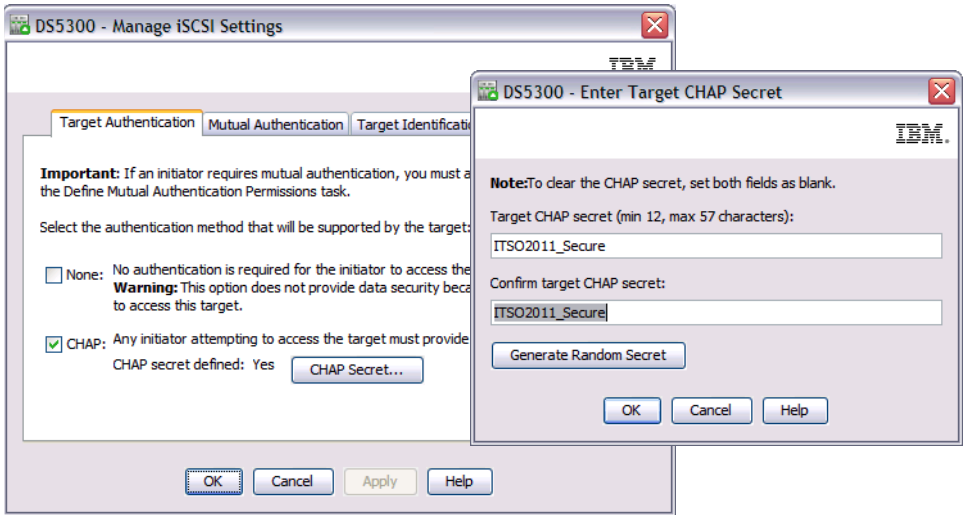


Figure 3-40 iSCSI security

Mutual authentication

To define two-way or mutual authentication select the **Mutual Authentication** tab, which enables the target (or DS5000 Storage System) to authenticate to the initiator (or server iSCSI HBA) so it can confirm that the target is who it says. If an initiator is already configured to only allow connections from targets that know the CHAP secret, then you can check and set the authentication by selecting the **Authentication Tab**. See Figure 3-41.

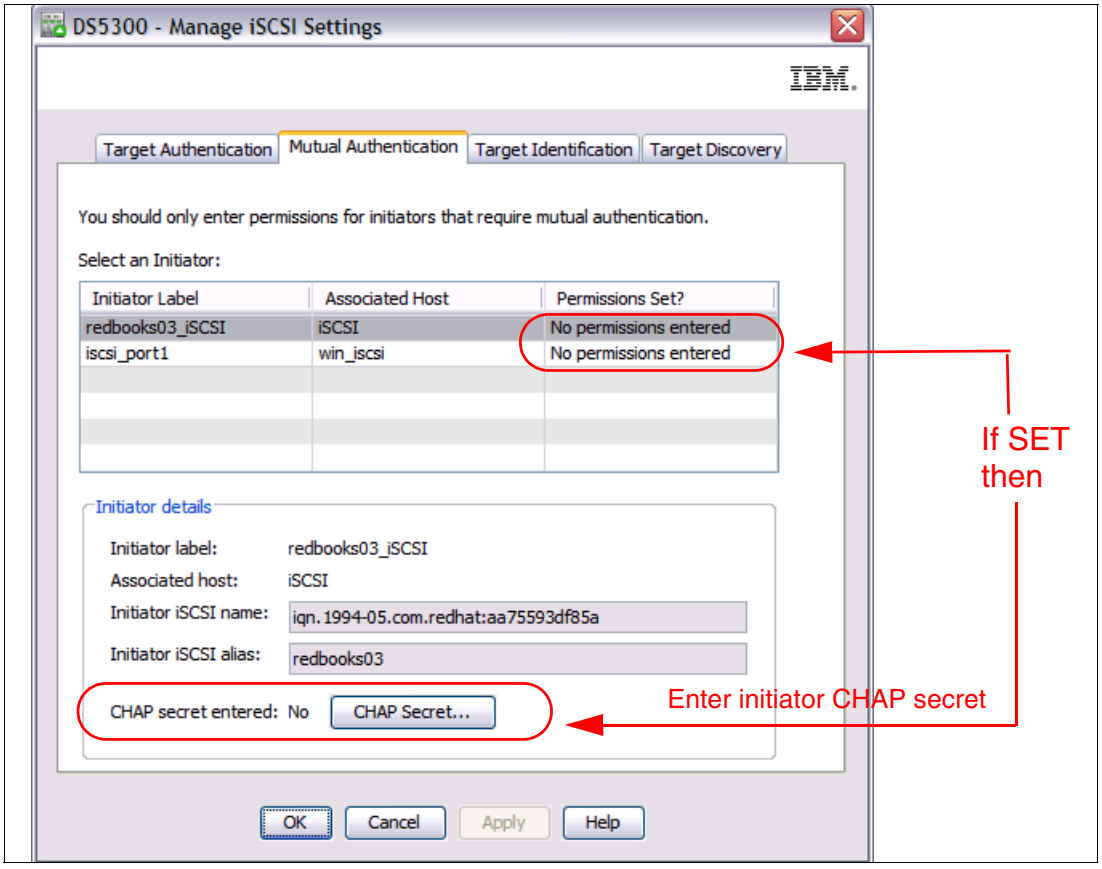


Figure 3-41 iSCSI mutual authentication

Notice that this window lists the host port identifiers already defined. You need to define a host previously to be allowed to set the mutual authentication.

Click the **Chap Secret** button to specify the secret that the target passes to the initiator to authenticate itself. After being set, the initiator will only be able to log on to targets that know the initiator secret.

You can also use this window to view defined host ports and the authorization level.

Target identification

The **Target Identification** option shows the iSCSI Qualified Name (IQN), of the DS5000 Storage System. In iSCSI protocol, the Storage is considered the Target, whereas the host server is the Initiator. Assign an alias here so you can identify it easily from the host. See Figure 3-42.

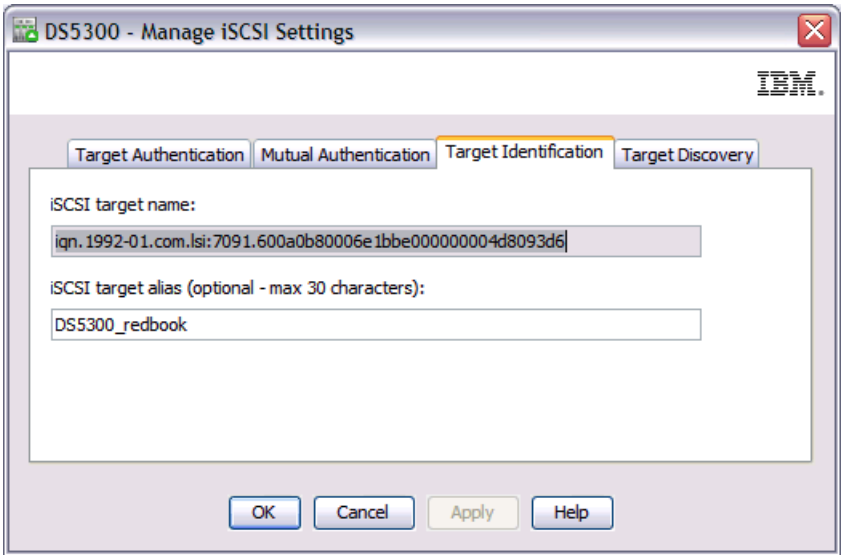


Figure 3-42 DS5000 storage system target name

Target discovery

Select the **Target Discovery** tab to perform a device discovery using the iSCSI simple naming service (iSNS). Enable the use of iSNS by selecting the **Use iSNS Server** check box, and then use DHCP to discover the iSNS server on your network, or type in manually an Internet Protocol version 4 (IPv4) or IPv6 address. See Figure 3-43.

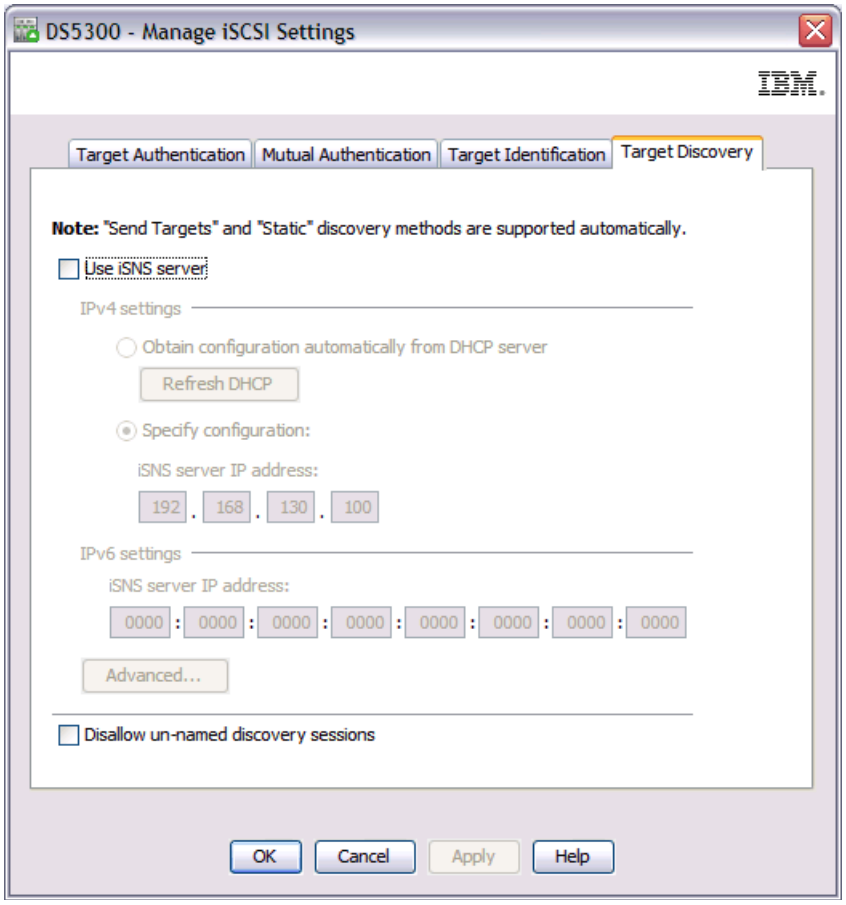


Figure 3-43 iSCSI setting iSNS discovery

When iSNS service is enabled, the **Advanced** tab lets you assign a separate TCP/IP port for your iSNS server for additional security, as shown in Figure 3-44.

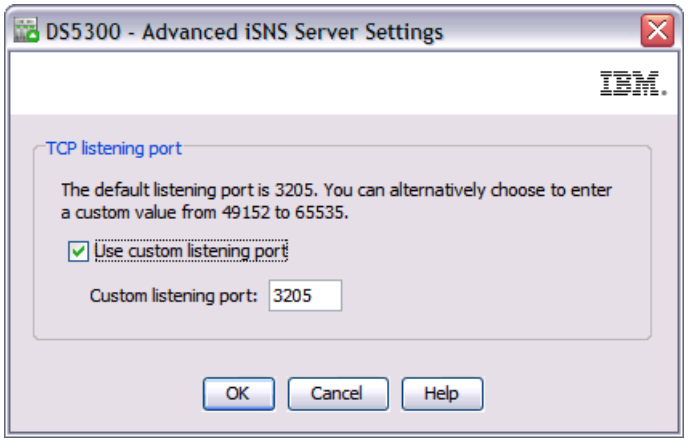


Figure 3-44 Advanced iSNS server settings

Managing iSCSI sessions

From your Subsystem Management window, select **Storage Subsystem** → **iSCSI** → **End Sessions** (Figure 3-45). Doing it will not terminate any session unless you later select it.

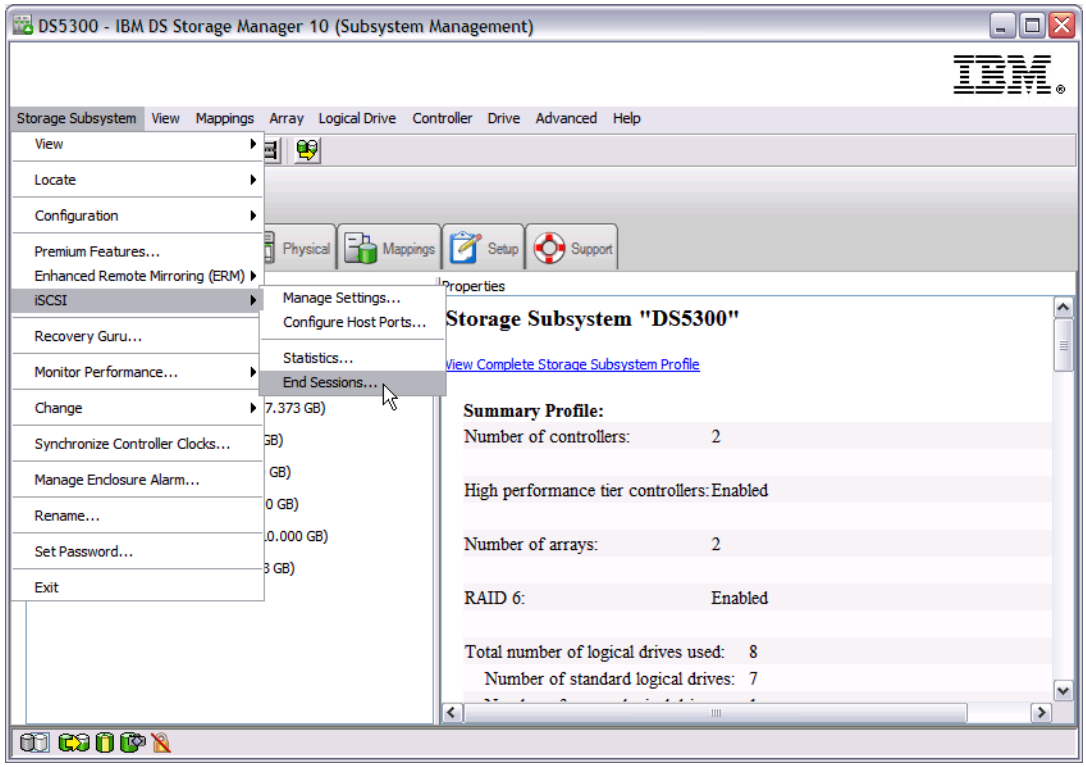


Figure 3-45 Manage iSCSI sessions

After selecting **End Sessions**, the following window opens. See Figure 3-46.

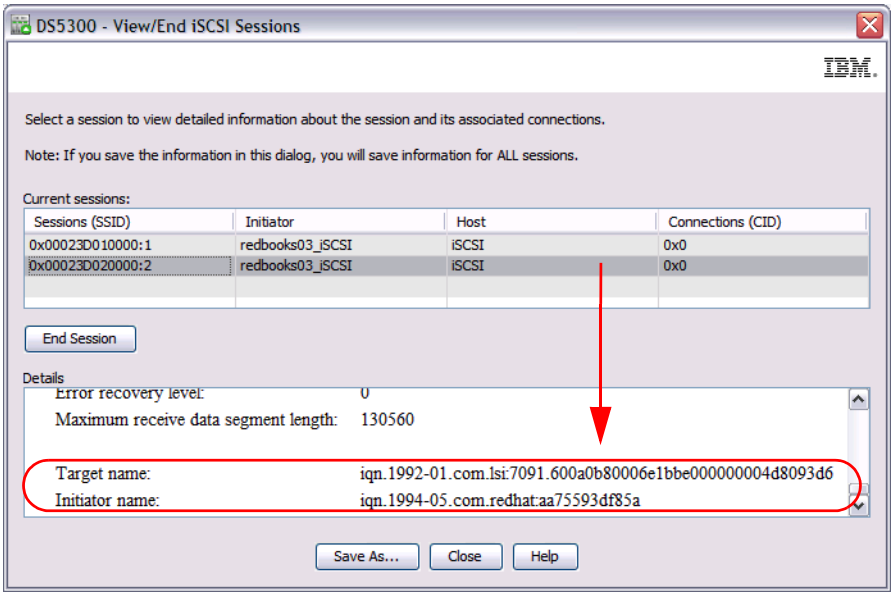


Figure 3-46 View/end iSCSI sessions

From here, you can check what connections are established to the Storage System, and from which initiator. You can end a particular session if you want to, but make sure to check the session initiator name to avoid ending one not desired. Remember that each session in the current implementation might not have more than one connection. So if you have multiple paths between host and controller, you must have multiple sessions established.

Viewing iSCSI statistics

You can use the performance monitor to display the traffic statistics from the entire DS5000 Storage System. In addition, you can view a list of all iSCSI session data, and monitor its traffic for your storage subsystem iSCSI ports by selecting **Storage Subsystem** → **iSCSI** → **Statistics** from your Subsystem Management window. The window in Figure 3-47 opens.

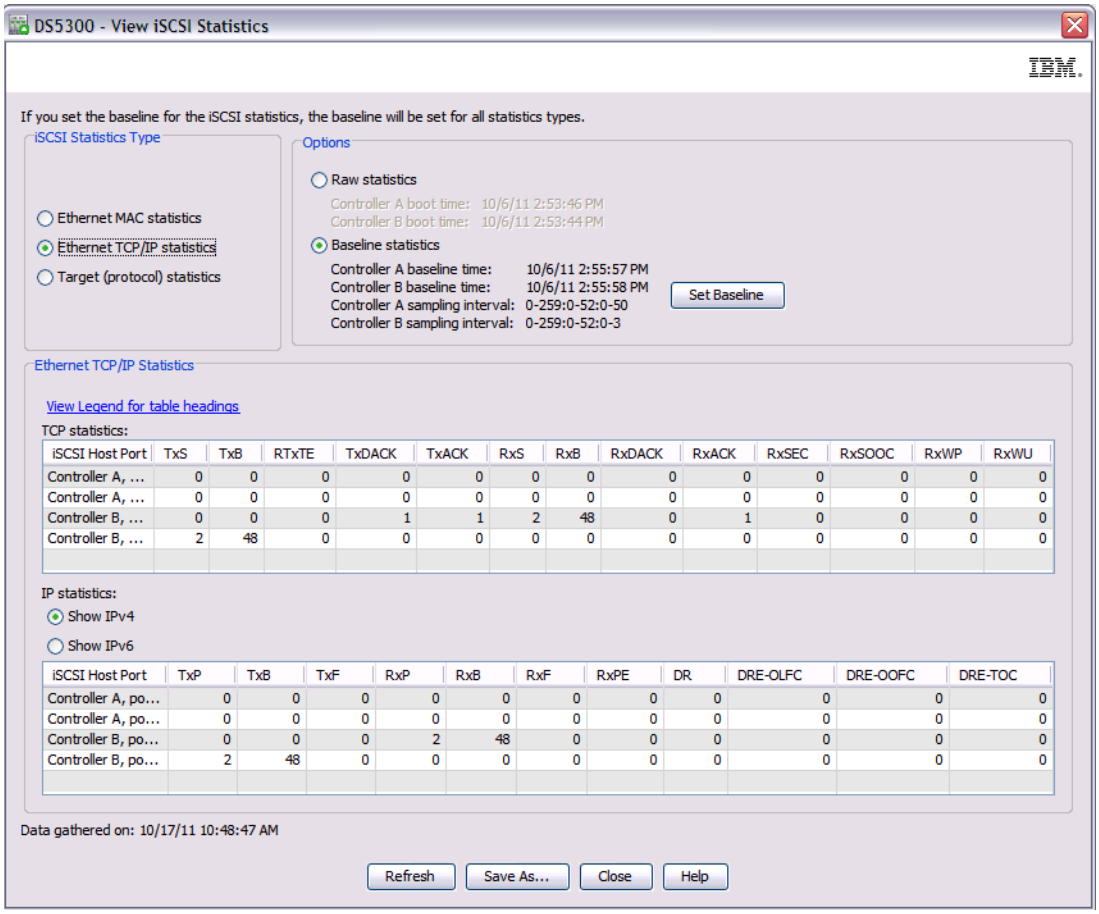


Figure 3-47 iSCSI statistics

As well as in the Performance Monitor, you can set a baseline to start monitoring the traffic from the moment the baseline is set. You can use this data to troubleshoot connection problems, for example, to track the number of header digest errors, number of data digest errors, and successful protocol data unit counts. After a corrective action, set a baseline again to determine whether the problem is solved, by displaying the next statistics.

3.1.7 Configuring for Copy Services functions

The DS5000 Storage System has a complete set of Copy Services functions that can be added to it. These features are all enabled by premium feature keys, which come in the following types:

- FlashCopy:

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

- VolumeCopy:

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

  You can perform either an offline VolumeCopy or an online VolumeCopy. To ensure data integrity, all I/O to the target logical drive is suspended during either VolumeCopy operation.

  - Offline Copy:

    An offline copy reads data from the source logical drive and copies it to a target logical drive, while suspending all updates to the source logical drive with the copy in progress. All updates to the source logical drive are suspended to prevent chronological inconsistencies from being created on the target logical drive.

  - Online Copy:

    An online copy creates a point-in-time FlashCopy copy of a source logical drive within a storage subsystem, while still being able to write to the logical drive with the copy in progress, using the FlashCopy as the actual source logical drive for the copy. The online VolumeCopy relationship is between a FlashCopy logical drive and a target logical drive.

Tip: If the logical drive that you want to copy is used in a production environment, the FlashCopy feature must be enabled. A FlashCopy of the logical drive must be created, and then specified as the VolumeCopy source logical drive, instead of using the actual logical drive itself. This requirement allows the original logical drive to continue to be accessible during the VolumeCopy operation.

- Enhanced Remote Mirroring:

  Enhanced Remote Mirroring (ERM) is used to allow mirroring to another DS5000 Storage System either co-located, or situated at another site. The main usage of this premium feature is for to enable business continuity in the event of a disaster or unrecoverable error at the primary storage System. It achieves this by maintaining two copies of a data set in two separate locations, on two or more separate storage Systems, and enabling a second storage subsystem to take over responsibility. Methods available for use with this feature are: synchronous, asynchronous, and asynchronous with write order consistency (WOC).

  Tip: Enhanced Remote Mirror is only allowed using Fibre Channel Host Interface Cards, not with the iSCSI host interfaces. After the mirror ports are presented in the SAN, then you can use FC-IP routers to reach the destination DS5000 storage subsystem.

  To connect both DS5000 Storage Systems, only FC is allowed, not iSCSi host interface cards.

Configurations of all of these features are documented in great detail in *IBM System Storage DS Storage Manager Copy Services Guide*, SG24-7822 and *IBM System Storage DS Copy Services Guide*, GA32-0964.

3.2 Event monitoring and alerts

Included in the DS5000 Client package is the Event Monitor service. It enables the workstation running this monitor to send out alerts by email (SMTP) or traps (SNMP). The Event Monitor can be used to alert you of problems in any or all of the DS5000 Storage Systems in your environment. It is also used for the Remote Support Manager Service as described in 3.2.3, “IBM Remote Support Manager (RSM)” on page 145.

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

Depending on the setup you choose, various storage subsystems are monitored by the Event Monitor. If you select the **Setup** tab of the Enterprise Management window, and then **Configure Alerts**, you have the option to enable it for all the storage subsystems managed by this station, or to select a specific one.

If you right-click your local system in the Enterprise Management window (at the top of the tree) and select **Configure Alerts**, this applies to all storage subsystems listed in the Enterprise Management window, without mattering are in-band or out of band managed.

If you right-click a specific storage subsystem, you only define the alerting for this particular DS5000 Storage System.

An icon at the right of the managing host, or at the right of specified Storage subsystem in the Enterprise Management window indicates wether all the DS5000 subsystems managed by this host, or only an specific one have Alerts configured. In the example shown in Figure 3-48, both DS5000 subsystems are configured to send alerts, because it is enabled at the host level. The subsystem ITSO5300 is also set at the subsystem level, although it is not necessary in this case; you will receive only one alert if this subsystem has a problem.

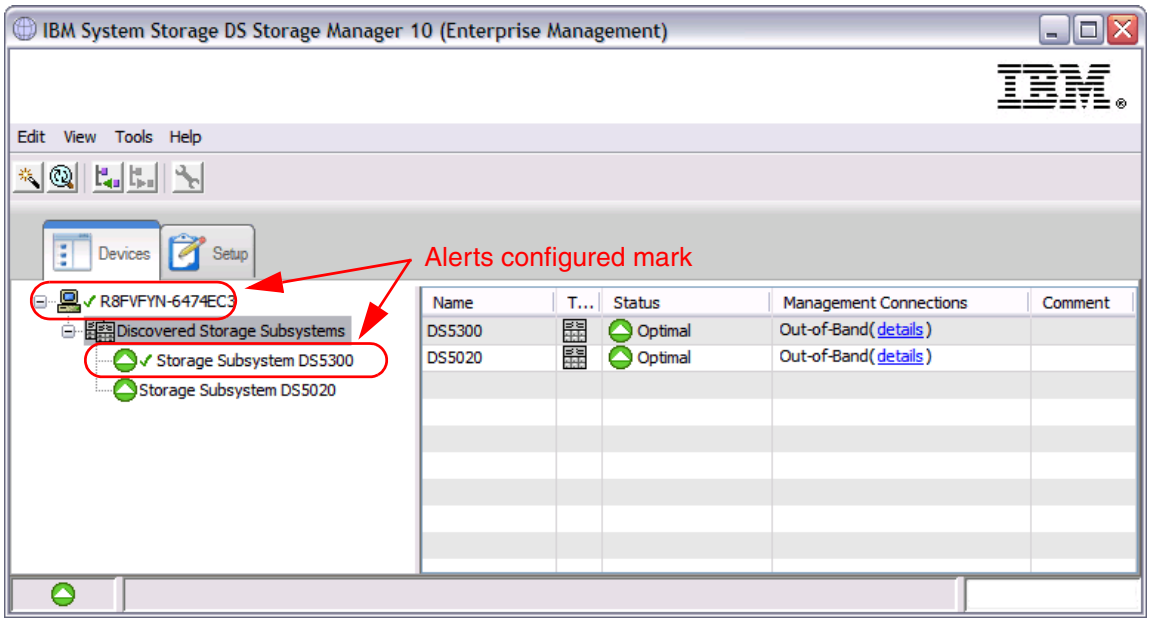


Figure 3-48 Alerts configured

See the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023, for specific details about configuring event notifications, such as the following examples:

- You can configure alert notifications for all your DS5000 storage subsystems, or at least, for your most critical DS5000 Storage Systems.
- As a best practice, configure your critical system with both in-band and out of band management capabilities, so in case of problems with one of the management methods, you are still notified.
- Make sure that the Event Monitor service is installed and running as an automatic service in at least two systems attached to your DS5000 subsystems, and allow in-band management, running 24 hours a day. This practice ensures proper alerting, even if one server is down, although you will receive duplicate messages.
- You can replicate the alert settings defined in the first Storage Management station by copying the emwdata.bin generated in the file to every storage management station from which you want to receive alerts.
- If you do not install the Event Monitor, or if the active process or service is stopped, then alerts are not sent unless the Storage Management station is open.
- If you want to send email alerts, you need to define an SMTP server first, and then the email addresses to which alerts are sent. If you do not define an address, no SMTP alerts are sent.
- Make sure to take the option to test a notification to the added email addresses to ensure a correct delivery and test your setup.
- If you choose the SNMP tab, you can define the settings for SNMP alerts: the IP address of your SNMP console and the community name. As with the email addresses, you can define several trap destinations, and test each one for correct operation.
- You need an SNMP console for receiving and handling the traps sent by the service. There is an MIB file included in the Storage Manager software, which must be compiled into the SNMP console to allow proper display of the traps. See the documentation for the SNMP console that you are using to learn how to compile a new MIB.
- With these notifications set, and the appropriate software, you can forward these alerts to your pager or cell phone.

Best practice: Configure alert notifications through SNMP, email, or both, to receive instant notifications of all critical events of your storage subsystem. Such alerts can help you take the appropriate corrective action as soon as the problem is logged.

3.2.1 ADT alert notification

Auto Drive Transfer (ADT) is a function that provides automatic failover in case of controller failure on a storage subsystem. ADT alert notification, which is provided with Storage Manager, accomplishes three things:

- It provides notifications for persistent “Logical drive not on preferred controller” conditions that resulted from ADT.
- It guards against spurious alerts by giving the host a “delay period” after a preferred controller change, so it can get reoriented to the new preferred controller.
- It minimizes the potential for the user or administrator to receive a flood of alerts when many logical drives fail over at nearly the same point in time due to a single upstream event, such as an HBA failure.

Upon an ADT event or an induced logical drive ownership change, the DS5000 controller firmware waits for a configurable time interval, called the *alert delay period*, after which it reassesses the logical drive distribution among the arrays.

If, after the delay period, certain logical drives are not on their preferred controllers, the controller that owns the not-on-preferred-logical drive logs a critical Major Event Log (MEL) event. This event triggers an alert notification, called the *logical drive transfer alert*. The critical event logged on behalf of this feature is in addition to any informational or critical events that are already logged in the RDAC, as seen in Figure 3-49.

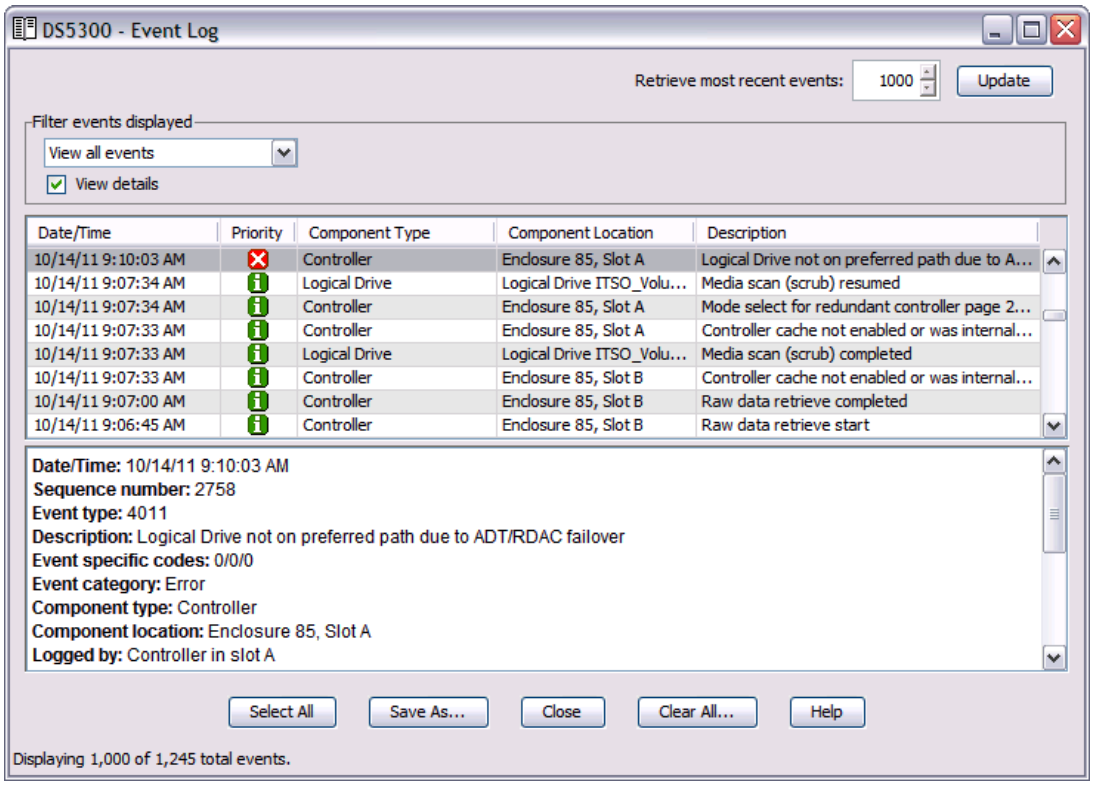


Figure 3-49 Example of alert notification in MEL of an ADT/RDAC logical drive failover

Tip: Logical drive controller ownership changes occur as a normal step of a controller firmware download process. However, the logical-drive-not-on-preferred-path events that occur in this situation will *not* result in an alert notification.

3.2.2 Failover alert delay

The failover alert delay lets you delay the logging of a critical event if the multipath driver transfers logical drives to the non-preferred controller. If the multipath driver transfers the logical drives back to the preferred controller within the specified delay period, no critical event is logged. If the transfer exceeds this delay period, a logical drive-not-on-preferred-path alert is issued as a critical event. This option also can be used to minimize multiple alerts when many logical drives failover because of a system error, such as a failed host adapter.

The logical drive-not-on-preferred-path alert is issued for any instance of a logical drive owned by a non-preferred controller and is in addition to any other informational or critical failover events. Whenever a logical drive-not-on-preferred-path condition occurs, only the alert notification is delayed; a needs attention condition is raised immediately.

To make the best use of this feature, set the failover alert delay period such that the host driver failback monitor runs at least once during the alert delay period. Note that a logical drive ownership change might persist through the alert delay period, but correct itself before you can inspect the situation. In such a case, a logical drive-not-on-preferred-path alert is issued as a critical event, but the array will no longer be in a needs-attention state. If a logical drive ownership change persists through the failover alert delay period, see the Recovery Guru for recovery procedures.

Considerations for failover alerts

Important: Here are several considerations regarding failover alerts:

- The failover alert delay option operates at the storage subsystem level, so one setting applies to all logical drives.
- The failover alert delay option is reported in minutes in the storage subsystem profile as a storage subsystem property.
- The default failover alert delay interval is five minutes. The delay period can be set within a range of 0 to 60 minutes. Setting the alert delay to a value of zero results in instant notification of a logical drive not on the preferred path. A value of zero does not mean that alert notification is disabled.
- The failover alert delay is activated after controller start-of-day completes to determine if all logical drives were restored during the start-of-day operation. Thus, the earliest that the not-on-preferred-path alert will be generated is after boot up and the configured failover alert delay.

Changing the failover alert delay

To change the failover alert delay, follow these steps:

1. Select the storage subsystem from the Subsystem Management window, and then select either the **Storage Subsystem** → **Change** → **Failover Alert Delay** menu option, or right-click and select **Change** → **Failover Alert Delay**. See Figure 3-50.

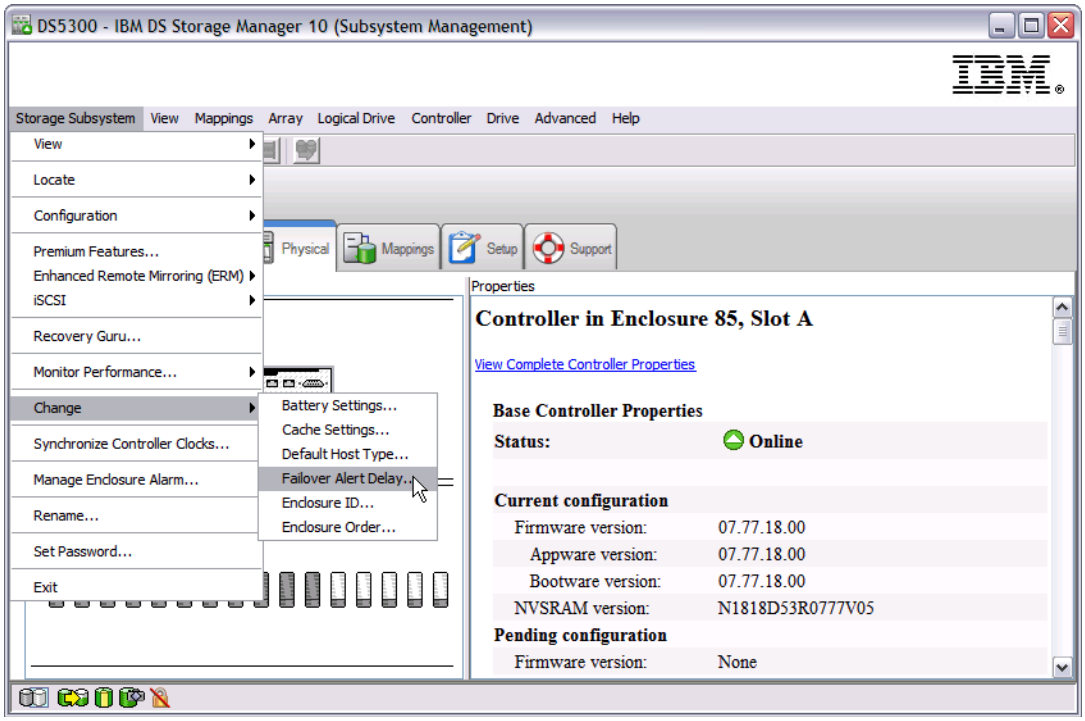


Figure 3-50 Changing the failover alert delay

2. In the Failover Alert Delay dialog box opens, as shown in Figure 3-51. Enter the desired delay interval in minutes and click **OK**.

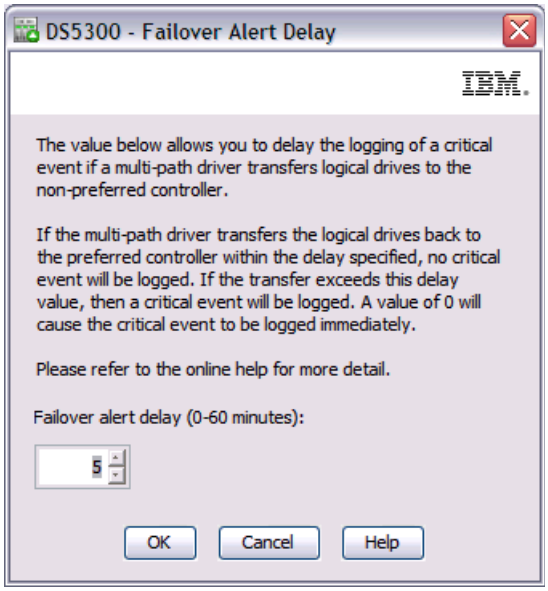


Figure 3-51 Failover alert delay window

3.2.3 IBM Remote Support Manager (RSM)

The challenges facing your business in regards to storage are very demanding. Faster response times are needed for alerts, and more detailed information is needed from these errors in order to resolve the issue, and to resolve it within a timely manner.

The IBM Remote Support Manager for Storage is designed to address these challenges by providing the following capabilities:

- Fast response time to alerts. The problem reporting provided by the RSM application automatically creates an entry in the IBM call management system for each storage subsystem that reports a problem.
- Detailed information on each alert for error analysis and accuracy. Files and logs needed for problem determination are sent to IBM using email or an FTP connection using the RSM server Ethernet interface.
- Low IT cost. The RSM for Storage is a non-charge software package designed to manages up to 50 storage systems per implementation and with no annual fees.
- Security designed to give control of remote access and notifications when remote users connect. Isolation of remote and local users is performed by an internal firewall that is managed by the RSM for Storage application. Remote users do not have the ability to change any security features of the application.
- Works with your existing IBM Storage Manager client application to help detect events.
- Sends logs and status along with the alert to IBM for fast problem resolution.
- Allows IBM service to dial in to obtain additional information and logs to aid problem determination and speed problem resolution.

Tip: For more details on how RSM works and how to install and configure RSM, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

3.3 Capacity upgrades and system upgrades

The DS5000 Storage subsystem has the ability to accept new disks or disk expansion units dynamically, with no downtime to the DS5000 unit. In fact, the DS5000 *must* remain powered on when adding new hardware.

3.3.1 Capacity upgrades

With the DS5000 Storage System, you can add capacity by adding expansion enclosures or disks to the enclosure being used. Care must be taken when performing these tasks to avoid damaging the configuration currently in place. For this reason, you must follow the detailed steps laid out for each part.

Important: Prior to physically installing new hardware, we advise that you collect all support data and follow the instructions in the *IBM System Storage DS4000/DS5000 Hard Disk Drive and Storage Enclosure Installation and Migration Guide GA32-0962*, at the following website:

https://www-947.ibm.com/support/entry/myportal/docdisplay?lndocid=MIGR-57818&brandind=5000028

Failure to consult this documentation might result in data loss, corruption, or loss of availability to your storage.

For more information, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

After physical installation, use Storage Manager to create new arrays/LUNs, or extend existing arrays/LUNs. (Note that certain operating systems might not support dynamic LUN expansion.)

Guidelines when adding new disks

Observe these guidelines to add new disks to your configuration:

1. Make additions to the DS5000 storage subsystem only while it is powered on and in optimal state.
2. To add new expansion enclosures, schedule the addition at a time period of low I/O between the DS5000 storage subsystem and host servers. You can complete this process while the DS5000 storage subsystem is receiving I/O from the hosts, however, you might have a performance problem while the drive loops are interrupted momentarily during the addition.
3. If you are adding new expansion enclosures model or new disks type on your DS5000 configuration, check following:
   a. Review the minimum requirements of firmware for your DS5000 storage subsystem, and its corresponding Storage Manager software. Also check the System Storage Interoperation Center (SSIC) website.
   b. Check if your DS5000 storage subsystem supports the new disk or expansion to add, and if the expansion supports the new disk type.
   c. Check if the new hardware is compatible to operate in the current speed of the current drive side Fibre Channel speed.
4. Temporarily disable the alert monitoring, because it will report unnecessary drive enclosure lost redundancy path errors while the new expansion is connected.

Other considerations when adding expansion enclosures and drives

Here are various considerations to keep in mind:

- If new enclosures have been added to the DS5000 Storage System, and you want to optimize your resulting configuration for maximum availability or performance, plan for the necessary downtime to re-distribute the physical drives of arrays between enclosures.

  Drives can be moved from one slot (or enclosure, or loop) to another with no effect on the data contained on the drive. This operation must, however, be done following a specific procedure: exporting the array, removing array disks, reinserting all of them in their final destination, and importing back the array. For more information about migrating arrays between DS storage subsystems, see the advanced maintenance section of the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023. All must be done while the subsystem is online, as explained in the *IBM System Storage DS4000/DS5000 Hard Disk Drive and Storage Enclosure Installation and Migration Guide*, GA32-0962.
- When adding drives to an expansion unit, do not add more than two drives at a time.
- For maximum resiliency in the case of failure, arrays must be spread out among as many expansion units as possible. If you merely create a multi drive array in a new drawer every time you add an EXP, all of the traffic for that array will be going to that one tray, which can affect performance and redundancy (see “Enclosure layout and loss protection planning” on page 34).
- For best balance of LUNs and I/O traffic, drives must be added into expansion units in pairs. In other words, every expansion must contain an even number of drives, not an odd number such as 5.

- If you are utilizing two drive loop pairs, approximately half of the drives in a given array must be on each loop pair. In addition, for performance reasons, half of the drives in an array must be in even numbered slots, and half in odd-numbered slots within the expansion units. (The slot number affects the default loop for traffic to a drive.)
- To balance load among the two power supplies in an expansion, there must also be a roughly equal number of drives on the left and right hand halves of any given expansion. In other words, when adding pairs of drives to an expansion, add one drive to each end of the expansion.

5. The complete procedure for drive migration is given in *IBM System Storage DS4000/DS5000 Hard Disk Drive and Storage Enclosure Installation and Migration Guide*, GA32-0962.

3.3.2 Storage System upgrades

The procedures to migrate disks and enclosures or upgrade to a newer DS5000 Storage System controller are not particularly difficult, but care must be taken to ensure that data is not lost. The checklist for ensuring data integrity and the complete procedure for performing capacity upgrades or disk migration is beyond the scope of this book. For more information about upgrading to DS5000, see the advanced maintenance section of the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Here we explain the DS5000 feature that makes it easy for upgrading subsystems and moving disk enclosures. This feature is known as *DACstore*.

DACstore

DACstore is an area on each drive in a storage subsystem or expansion enclosure where configuration information is stored. This 512 MB reservation (as pictured in Figure 3-52) is invisible to a user and contains information about the DS5000 configuration.

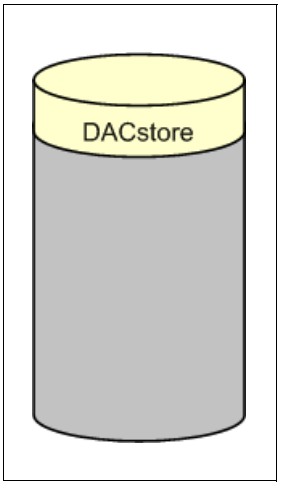


Figure 3-52 The DACstore area of a DS5000 disk drive

The standard DACstore on every drive stores the following information:

- Drive state and status
- WWN of the DS5000 controller (A or B) behind which the disk resides
- Logical drives contained on the disk

Certain drives also store extra global controller and subsystem level information; they are called *sundry drives*. The DS5000 controllers will assign one drive in each array as a sundry drive, although there will always be a minimum of three sundry drives even if only one or two arrays exist.

Additional information stored in the DACstore region of the sundry drive includes:

- Failed drive information
- Global Hot Spare state/status
- Storage subsystem identifier (SAI or SA Identifier)
- SAFE premium feature identifier (SAFE ID)
- Storage subsystem password
- Media scan rate
- Cache configuration of the storage subsystem
- Storage user label
- MEL logs
- LUN mappings, host types, and so on.
- Copy of the controller NVSRAM

Why DACstore is used

This particular feature of DS5000 storage Systems offers a number of benefits:

- Storage system level reconfiguration: Drives can be rearranged within a storage system to maximize performance and availability through channel optimization.
- Low risk maintenance: If drives or disk expansion units are relocated, there is no risk of data being lost. Even if a whole DS5000 subsystem needed to be replaced, all of the data and the subsystem configuration can be imported from the disks.
- Data intact upgrades and migrations: All DS5000 subsystem recognize configuration and data from other DS5000 subsystems so that migrations can be for the entire disk subsystem as shown in Figure 3-53, or for array-group physical relocation as illustrated in Figure 3-54.

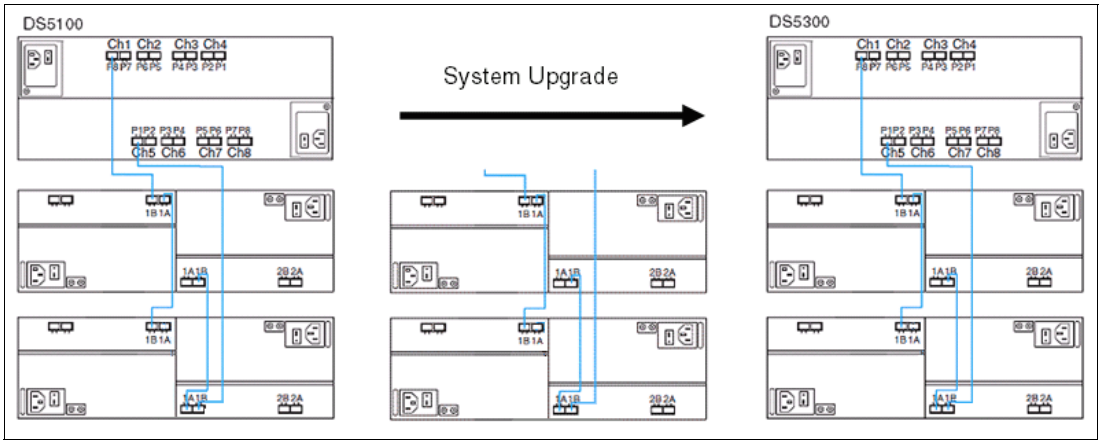


Figure 3-53 Upgrading DS5000 controllers

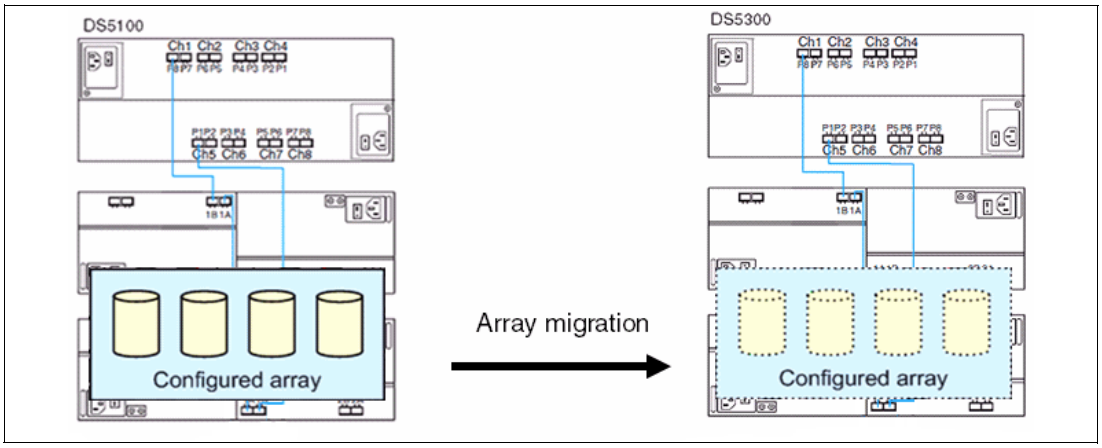


Figure 3-54 Relocating arrays

3.3.3 Increasing bandwidth

You can increase bandwidth by moving expansion enclosures to a new or unused channels pair (this doubles the drive-side bandwidth).

Let us assume that the initial configuration is the one depicted on the left in Figure 3-55. We are going to move EXP2 to the unused channels pair on the DS5000.

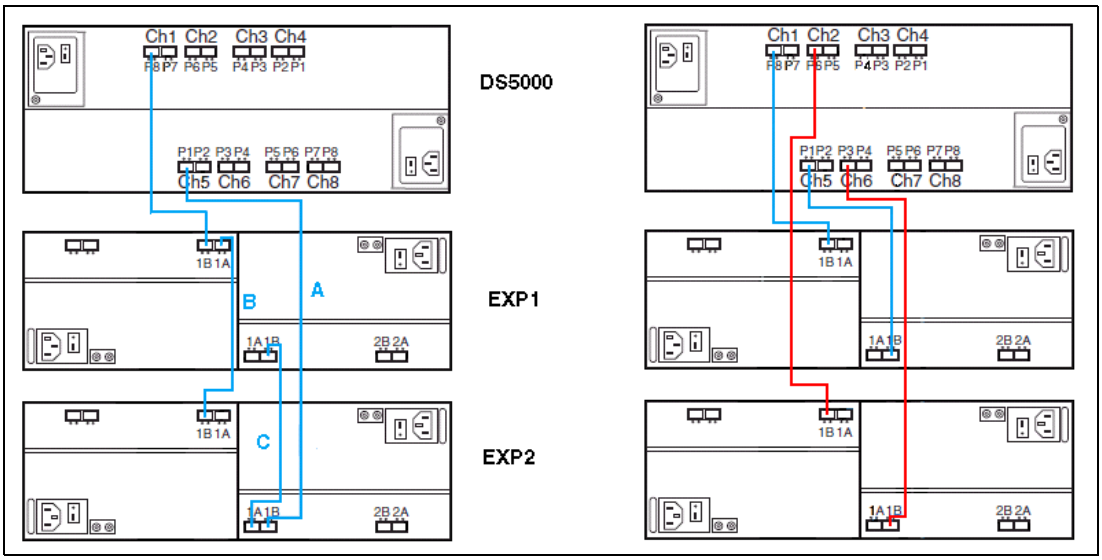


Figure 3-55 Increasing bandwidth

To move EXP2 to the unused channels pair, proceed as follows:

1. Disconnect the cable A from Channel 5 Port 1 and connect to Channel 6 Port 3.
2. Disconnect the cable B from EXP1 and connect to Channel 2 Port 6.
3. Disconnect the cable C from EXP2 and connect to Channel 5 Port 1.

4

Host configuration guide

In this chapter, we explain how to manage your IBM System Storage DS5000 from the most common host servers and operating system environments. After describing the installation of the Storage Manager (SM) software, we explain how to verify a correct installation and show you how to view your storage logical drives from your hosts. We describe specific tools and give you the commands necessary to use the tools from various operating systems such as Microsoft Windows Server 2008, Red Hat Linux 6, and AIX. Finally, we refer you to other sources of information for VMware vSphere5, IBM i5/OS™, and Microsoft Hyper-V Server 2008 R2.

We assume that you have already performed the necessary planning, and have set up and configured the DS Storage System.

Whatever operating system you need to connect to your DS Storage System, make sure to use the guidelines for levels of firmware and device driver levels of each component of the solution. For the supported levels, see the IBM System Storage Interoperation Center (SSIC) website:

`http://www-03.ibm.com/systems/support/storage/config/ssic`

If you need to update any components, you can find details at the IBM System Storage support website:

`http://www.ibm.com/servers/storage/support/disk`

4.1 Planning your host attachment method

Throughout this unit, we go over the main procedures needed to manage our IBM System Storage DS5000 from the operating systems most commonly used. This requirement is the reason why, in this section, we detail the design used as good practice.

4.1.1 Fibre Channel SAN attach (FC SAN)

The configuration in Figure 4-1 shows three physical servers running different operating system, connected to the same DS storage subsystem (IBM DS5300). They have two HBAs each, connected to different SAN Fabric Switches. Each server has his own zone defined on both Fabric Switch to separate the traffic for stability and improve the management. The IBM DS 5300 have two controllers defined as Controller A and Controller B, both controllers are physically connected to different SAN Fabric Switches. Based on the Multipath Driver implemented at the OS level and the proposed cabling connections, the OS will be able to access to the SAN attach storage using alternatives paths for redundancy.

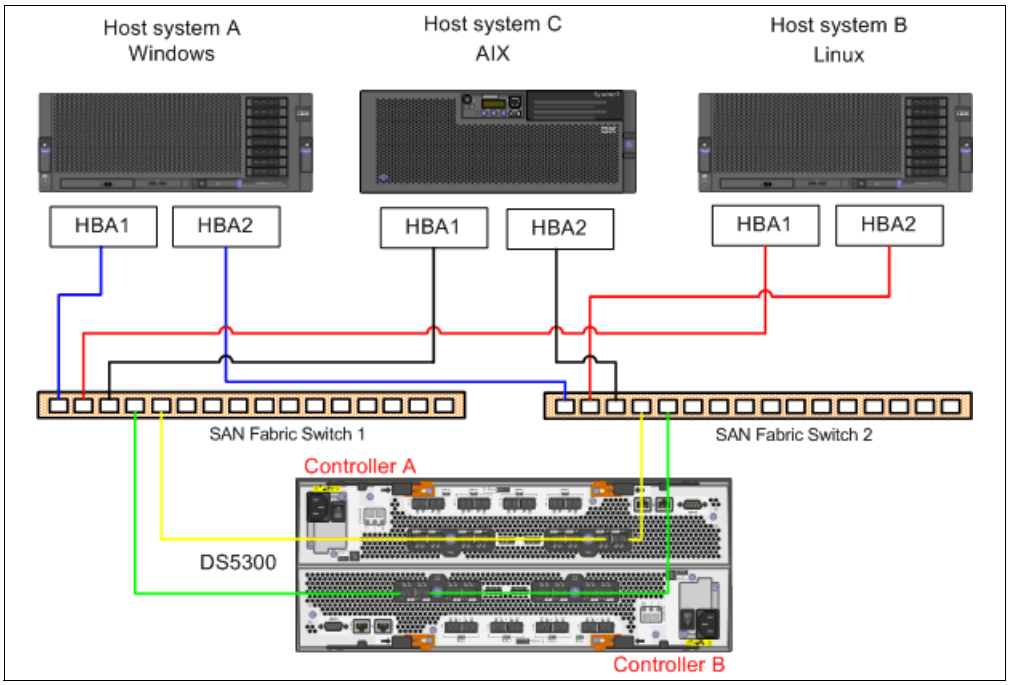


Figure 4-1 Fibre Channel SAN attach layout

For further information about cabling guidelines and zoning the SAN switches, see *Implementing an IBM/Brocade SAN with 8 Gbps Directors and Switches*, SG24-6116, or *Implementing an IBM/Cisco SAN*, SG24-7545.

4.1.2 iSCSI SAN attach: Using iSCSI Software Initiator

The configuration in Figure 4-2 shows three physical servers running different operating systems and configured for using iSCSI Software Initiator. These servers are connected to the same DS storage subsystem (IBM DS5300) where the iSCSI protocol is configured and defined as SAN provider. They have two physical network adapters (NICs) each, connected to different Ethernet Switches for redundancy. Each server will have the iSCSI Software Initiator installed and configured.

The IBM DS5300 has two controllers defined as Controller A and Controller B; both controllers are physically connected to different Ethernet Switches. Based on the iSCSI Software Initiator version and the multipath driver implemented at the OS level, the host servers will be able to access to the iSCSI targets disks using alternative paths for redundancy or for improving performance.

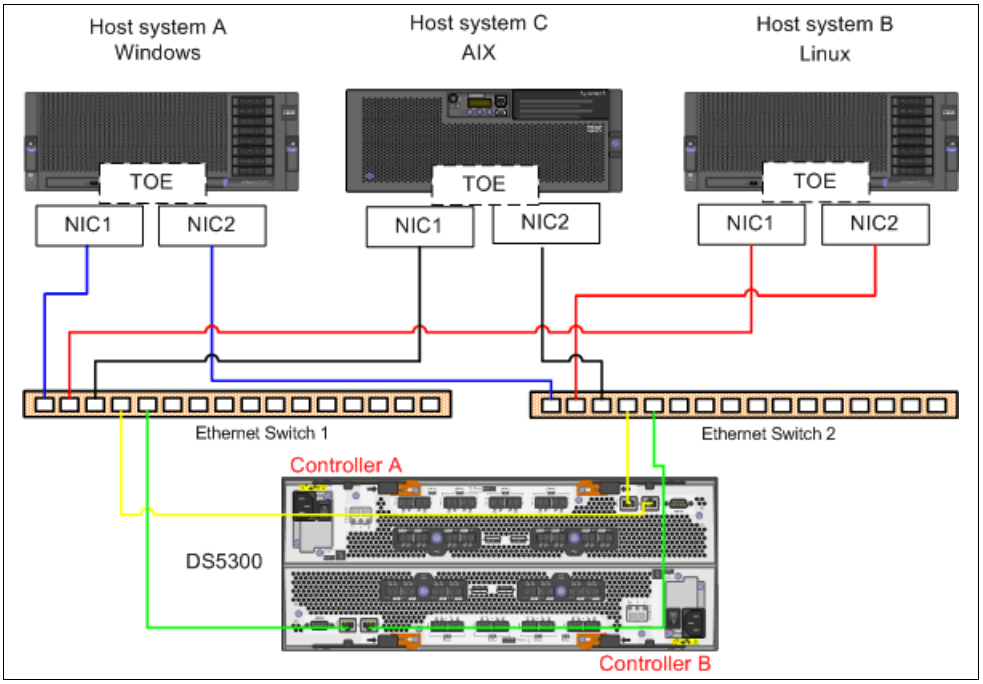


Figure 4-2 iSCSI SAN attach layout

4.2 Intermixing device drivers

With the variety of host types supported by the DS5000 storage subsytems, there are a number of different drivers that can be encountered on the hosts. In many cases, these drivers might be used to support other devices and can interfere with the functions of the DS5000 storage subsystem. Understanding how to manage and configure these drivers so as to permit co-existence is an important part of planning a successful implementation. The following are examples of some common driver intermix scenarios seen by customers.

4.2.1 AIX MPIO and fcp_array drivers

With AIX there are two drivers available with the operating system: the MPIO driver and the earlier **`fcp_array`** (RDAC) driver. Because earlier storage subsystems still require the RDAC driver, AIX allows for both of these drivers to co-exist and be used depending on the storage subsystem needs and configuration selections. With the MPIO, AIX offers much better multipathing support rather than the simple failover support previously provided with the RDAC driver.

Best practice: It is a best practice to use the AIX MPIO driver whenever it is supported.

See 4.4, “AIX configuration” for details on how to configure the MPIO for DS5000 storage subsystems.

4.2.2 Windows 2003 and 2008

With the supported Windows operating systems, all supported IBM DS storage subsystems are supported with the Windows MPIO driver and the Device Specific Module (DSM) device specific module provided with the host kit for Windows. In this environment, only one driver is supported and intermix is not permitted. Customers can either choose the internal MPIO or, if they want, they can purchase Veritas Volume Manager and use its DMP multipath driver.

4.2.3 Red Hat and SLES Linux operating systems

With the Red Hat 5 and SLES 10 and later, a new internal Device Mapper Multipath (DMM) was released. As of Red Hat 6 and SLES 11, these new drivers are now supported with DS5000 storage subsystems. However, with the earlier operating system levels, the internal drivers are not supported and the Linux MPP (RDAC) driver is required. With earlier storage subsystems, the DMM is also not supported, so situations can arise either where these two drivers must co-exist, or where the DMM might need to be disabled. The procedure for disabling the DMM driver is as follows:

1. Verify that the current settings of HBAs are the suggested settings. See Table 4-1 and Table 4-2 for examples of settings for the QMI2582 Qlogic host bus adapter.

Table 4-1 Common Linux settings for the QMI2582 adapter

HBA parameters	Suggested value
Host Adapter BIOS	Disabled
Frame Size	2048
Loop Reset Delay	8
Spin-up Delay	Disabled
Connect Options	2
Fibre Channel Tape Support	Disabled

Table 4-2 Advanced HBA settings

HBA parameters	Suggested value
Execution Throttle	256
LUNs per Target	0
Enable LIP Reset	No
Enable LIP Full Login	Yes
Enable Target Reset	Yes
Login Retry Count	30
Port Down Retry Count	35
Link Down Timeout	60
Extended Error Logging	Disable
RIO Operation Mode	0
Interrupt Delay Timer	0
IOCB Allocation	256

Check with the specific HBA vendor for the settings suggested for various adapters.

2. Remove duplication of multi path drivers:

```
service multipathd stop
mppUpdate
reboot
```

3. If pdisk thrashes, then run the following command:

```
mppUtil -o DebugLevel=0x3
```

4. Verify that the setting changes have been effective by checking the DebugLevel as shown in Table 4-3. Use the following command syntax to view this information:

```
root@p5top ~]# mppUtil -o
```

Table 4-3 Standard default versus new changed values

Variable Option	Value	Value	Value	Value
DebugLevel	0x0	0x0	0x0	0xffffffff
DebugLevel	0x3	0x0	0x0	0xffffffff

In certain situations, it can be necessary to have both drivers running simultaneously on the same host. It can occur when different storage subsystems are being ran on the same host and one is compatible with the DMM and one is requiring another driver (for example, the MPP (RDAC). In these cases, the device that is not compatible with the DMM needs to be excluded from the DMM trying to manage it. It can be done by using the blacklist option of the multipath.conf file being configured. Example 4-1 shows a DS5000 being excluded from the DMM.

Example 4-1 Exclude DS5000 from DMM

```
#==[ Configuration File ]===========================#
# /etc/multipath.conf
defaults {
user_friendly_names yes
}
blacklist {
devnode "sda"
devnode "^(ram|raw|loop|fd|md|dm-|sr|scd|st)[0-9]*"
devnode "^hd[a-z]"
devnode "^cciss!c[0-9]d[0-9]*"
device {
vendor "IBM"
product "VirtualDisk"
}
}
devices {
device {
vendor "NETAPP"
product "LUN"
path_grouping_policy group_by_prio
getuid_callout "/sbin/scsi_id -g -u -s /block/%n"
features "1 queue_if_no_path"
path_checker readsector0
failback immediate
}
}
```

In Example 4-1 on page 155, the IBM DS5000 storage subsystem is being blacklisted, while the NetApp fileserver is being allowed to used the DMM.

With newer Linux releases, we support the DS5000 with the DMM. Therefore, if intermixing the DS5000 with earlier storage subsystems that must use the MPP(RDAC), we would want to blacklist using the WWN of the storage subsystem that we want to exclude.

You can use the WWN of the LUNs to be blacklisted, as done before. For example, we can actually run with a configuration like this one:

```
blacklist {
wwid "*"
}
blacklist_exceptions {
wwid "3600[0-9a-f]+"
}
```

Take care when setting up these types of blacklists to ensure that the desired results are obtained. Testing the solution will ensure that you have the results you are seeking.

4.3 Microsoft Windows Server 2008 configuration

In the following section, we describe the steps that you need to perform from your Windows 2008 host in order to install and manage your DS Storage System from this operating system.

4.3.1 Installing Storage Manager software

This section covers a guided installation of the DS Storage Manager v10.77 software in a Windows Server 2008 R2 host environment.

Tip: You can also use this information as a base reference for installing in environments using Windows Server 2003 and DS4000.

The host software for Windows includes the following components:

- SMclient
- Multipath driver (MPIO Device Specific Module - DSM)
- SMagent
- SMutil

Tip: On DS Storage Manager v10.77, the MPIO Device Specific Module has a separate installation package but still provided on the IBM DS Storage Manager installation source.

Follow these steps to install the software:

1. Log on with administrator rights for installing the new software, including new drivers.
2. Locate and run the installation executable file, either in the appropriate CD-ROM directory, or the file that you have downloaded from the IBM support website. After it has been executed, select your language of choice. After presenting the introduction and copyright statement windows, you are asked to accept the terms of the license agreement, which is required to proceed with the installation.
3. Select the installation target directory of your choice. Here is the default installation path:

   ```
   32Bit OS -> C:\Program Files\IBM_DS\
   64Bit OS -> C:\Program Files (x86)\IBM_DS\
   ```

4. Select the installation type, as shown in Figure 4-3.

Figure 4-3 InstallAnywhere: Select Installation Type

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

b. Decide what type of installation you want. Two additional choices are offered: Typical (Full Installation) and Custom. As the name indicates, Typical (Full Installation) installs all components, and Custom installation lets you choose the components, as you can see in Figure 4-4.

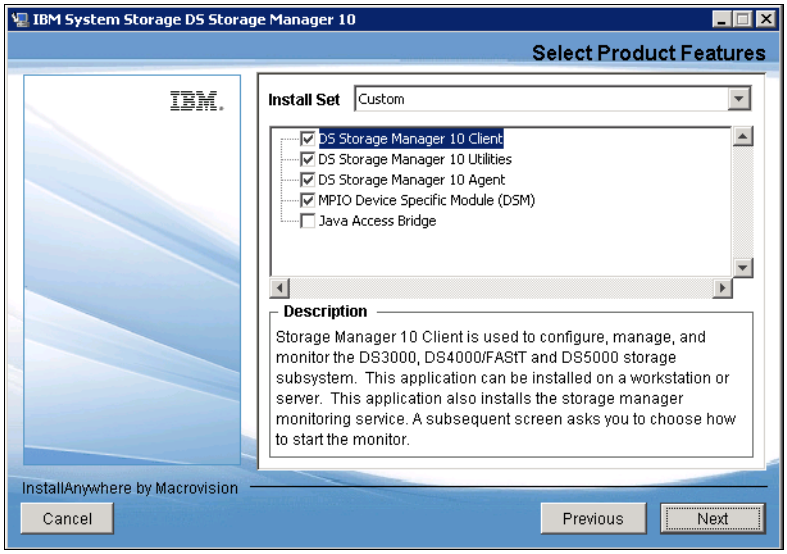


Figure 4-4 InstallAnywhere: Select Storage Manager components

Tip: Remember that the SM Failover drivers (MPIO/DSM) are only supported for connection to DS5000 Storage Servers with controller firmware V6.19 and later.

In addition to the usual Storage Manager components, you can choose to install Java Access Bridge. This selection enables support for the window reader (such as JAWS from Freedom Scientific, Inc.) for blind or visually impaired users.

5. Decide whether you want to automatically start the Storage Manager Event Monitor, as shown in Figure 4-5, which depends on your particular management setup. In case there are several management machines, the Event Monitor must only run on one. If you want to use the Event Monitor with SNMP, you need to install the Microsoft SNMP service first, because the Event Monitor uses its functionality.

Tip: The Event Monitor must be enabled for both the automatic ESM synchronization and the automatic support bundle collection on critical events.

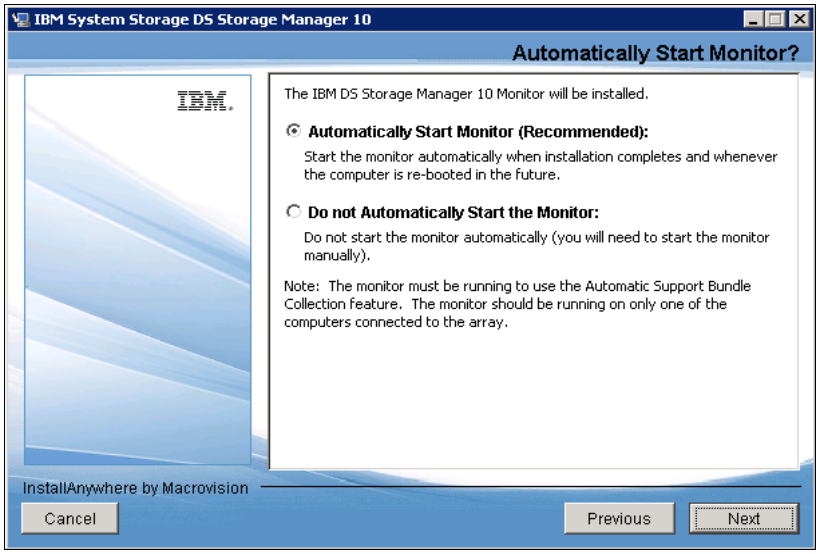


Figure 4-5 InstallAnywhere: Automatically Start Monitor

6. Verify that you have selected the correct installation options by examining the Pre-Installation Summary window, which is presented next. Then click the **Install** button. The actual installation process starts as shown in Figure 4-6.

Figure 4-6 nstallAnywhere: Pre-Installation Summary window

Verifying the SM installation

This section provides instructions on how to verify that you have installed the SM correctly in your Windows Server 2008.

Look under your programs folder for a new program entry in your named IBM DS Storage Manager 10 Client. This name is the name of the program created after a successful installation, which also generates a log file with the details of the installation process and options selected, and places it into the installation directory. Here is the file name:

```
IBM_System_Storage_DS_Storage_Manager_10_InstallLog.log
```

In case of problems during the installation, look at this file for a possible hint about what might be wrong.

Verifying the SMagent installation

This section provides instructions on how to verify that you have installed the SMagent correctly on Windows operating systems. Follow these steps:

1. Select **Start** → **Administrative Tools** → **Services**. The Services window opens.
2. Scroll through the list of services until you find IBM DS Storage Manager 10 Agent as shown in Figure 4-7.

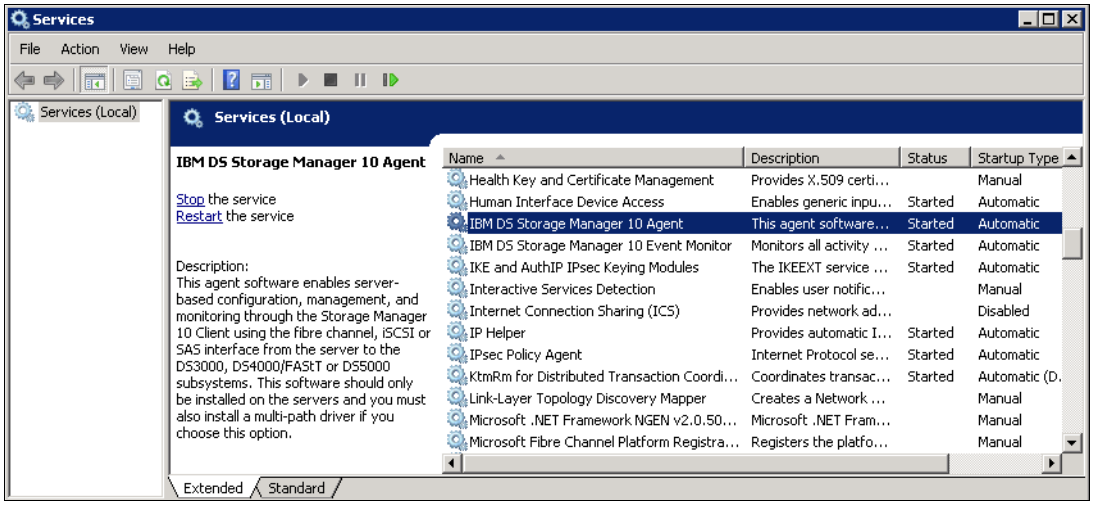


Figure 4-7 Verifying the SMagent installation

3. Determine whether the IBM DS Storage Manager 10 Agent and the Event Monitor services have been started. These services were created by the installation, which normally starts both of them by default. However, if a service has not been started, right-click it and select **Start**. Make sure the Startup Type is set to **Automatic**.

   If you are installing the host server and do not plan to use the host-agent software to manage one or more storage systems, you can set the Startup Type to **Manual**.

Verifying the SMutil installation

To verify that you have installed the SMutil correctly on Windows operating systems, follow these steps:

1. Go to the *installation_directory*\Util directory, typically:

   ```
   32Bit OS -> C:\Program Files\IBM_DS\Util
   64Bit OS -> C:\Program Files (x86)\IBM_DS\Util
   ```

2. Verify that the directory contains the following files:

   - hot_add.exe
   - SMdevices.bat
   - SMrepassist.exe

Tip: On 64-Bit OS, the installation path needs to be C:\Program Files <x86>\IBM_DS\Util

4.3.2 Updating the host software

To update the host software in a Windows environment, follow these steps:

1. Verify that IBM HBA firmware and device driver versions are current, and at the supported level of compatibility from the host Storage Manager software and firmware you are installing (see SSIC website). If they are not current, download the latest versions, study the readme file located with the device driver, and then upgrade the device drivers.
2. From the IBM Disk Support website, select your DS storage hardware and click **Download**, then select the Storage Manager package for your operating system:

   http://www.ibm.com/servers/storage/support/disk

3. Begin the installation of your downloaded package in your server. Accept the license terms, and select **Custom** as the installation type.

4. In the next window, select the components to update:
   - DS Storage Manager client
   - Utilities
   - Agent

   Because you are updating your installation, without uninstalling the previous version, the installation program detects the presence of existing software versions and presents the window shown in Figure 4-8.

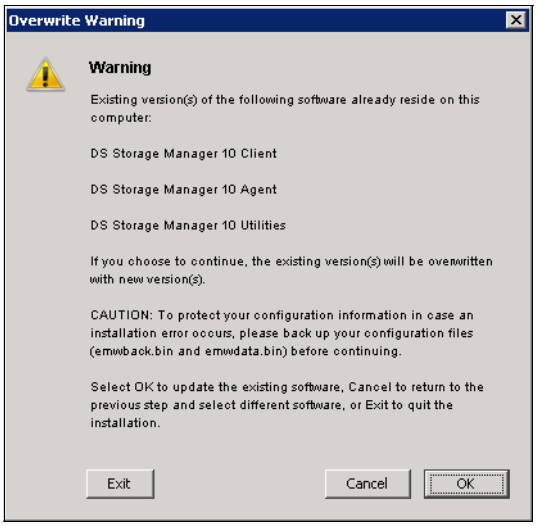


Figure 4-8 Updating host Software warning

Back up the file, emwdata_v0x.bin, to protect configuration information in case of failure by copying the file to another directory outside the installation folder of the Storage Manager. This file contains the alerts notifications setup. The default path for emwdata_v0x.bin:

```
32Bit OS -> C:\Program Files\IBM_DS\Client\Data
64Bit OS -> C:\Program Files (x86)\IBM_DS\Client\Data
```

Click **OK** to continue installation, confirming the remaining windows. You are normally requested to restart your machine after the installation finishes.

4.3.3 HBA and Multipath device drivers

For a successful implementation, and also to keep your system running at a supported level, be sure to check your host bus adapter firmware and driver levels, as well as the DS Storage System driver level, and compare them against the System Storage Interoperation Center, SCIC. For details, see the IBM System Storage Interoperation Center (SSIC) website:

http://www-03.ibm.com/systems/support/storage/config/ssic

You can download the necessary files for the DS Storage Systems, including the host bus adapter drivers, from the IBM System Storage support website:

http://www.ibm.com/servers/storage/support/disk

Verifying your Host Attachment card

Host adapter cards are tested at specified levels for interoperability with the DS Storage Systems, so very often the SSIC website output shows specific firmware levels as supported. It might also show specific preferred HBA settings.

To display or update your adapter firmware levels, and specific settings, so that they match the supported SSIC website output, you can use the management tool corresponding to your HBA manufacturer. In Figure 4-9, the QLogic SANsurfer Manager is used as an example for showing a QLogic iSCSI adapter additional information, such as firmware, BIOS, and driver versions.

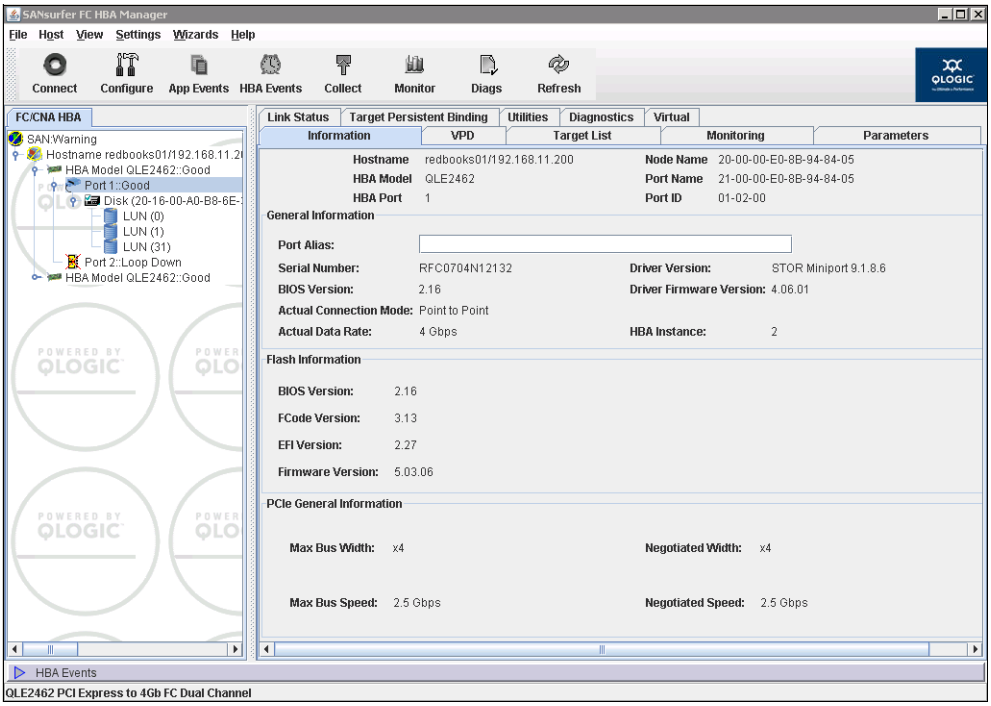


Figure 4-9 Using QLogic SANsurfer to display HBA firmware

Notice that the HBA model is specified in Figure 4-9. You need that specific model to search the supported levels of firmware and drivers to use it to attach your DS Storage System.

Emulex has another graphic utility for the same purpose, named HBAAnyware, and Brocade too. For more information about the usage of this tool, including how to update the firmware or BIOS levels, see the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Verifying your host attachment cards driver level

To check or update your host bus adapter driver level in Windows, use the Device Manager to display the status of your host bus adapter cards for both Fibre Channel or iSCSI depending on your system configuration. See Figure 4-10.

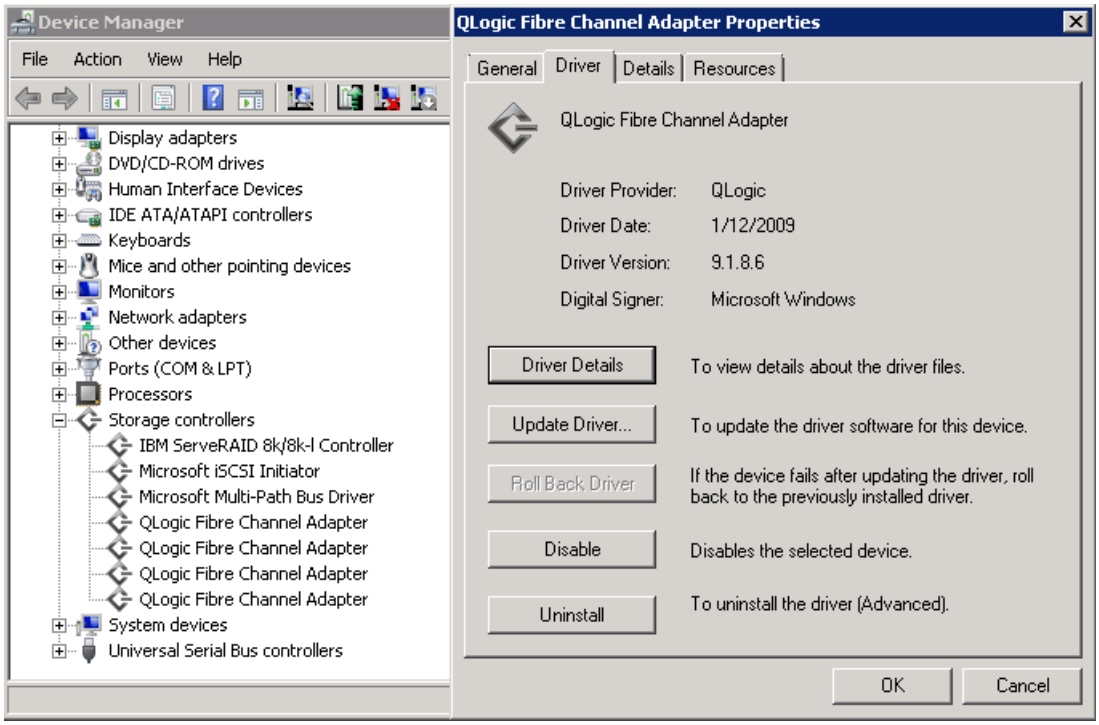


Figure 4-10 Displaying HBA drivers

From here you can display the installed driver level of your adapter card, and even update it by selecting the option **Update Driver**. This procedure is easy, however, it is important to make sure which kind of adapter you have, and its specific model, because both driver and firmware might be specifics for the adapter types. The Windows Device Manager does not provides this information, so it is a better alternative to use your adapter specific management tool to determine the exact adapter type, as seen in Figure 4-9 on page 162.

SM Failover driver (MPIO/DSM)

The following drivers are supported for the DS storage Systems in Windows:

- Microsoft MPIO:

  Included with Windows operating systems.

- MPIO Device Specific Module (DSM):

  Provided with the DS Storage Manager installation package.

Installing MPIO Device Specific Module (DSM)

On DS Storage Manager v10.77, the MPIO Device Specific Module has a separate installation package. It is provided on the IBM DS Storage Manager installation source under the Windows folder:

`\ibm_sw_ds3-5k_10.77.xx.16_windows_intl386\WS03WS08_10p77_IA32\Windows\)`

On prior versions, MPIO DSM support was provided on the same DS Storage Manager installation package and available on “Custom” installation.

Follow these steps:

1. To install MPIO DSM, log on with administrator rights for installing MPIO Device Specific Module support. Locate and run the installation executable file, click **Next** on the Welcome windows, and then accept the License Agreement, as shown in Figure 4-11 and Figure 4-12.

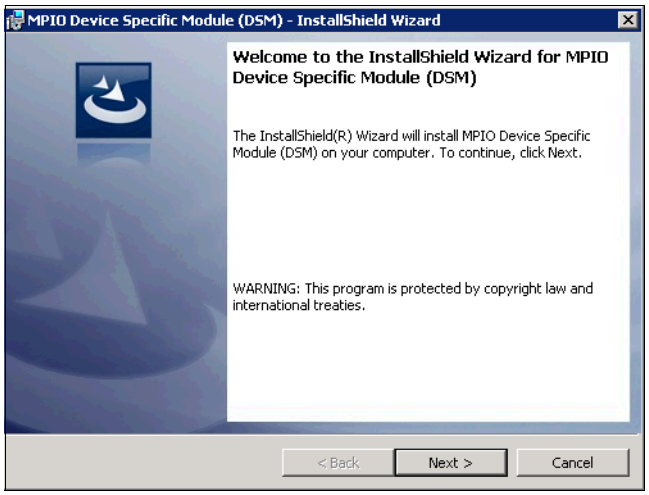


Figure 4-11 MPIO Device Specific Driver - Welcome screen

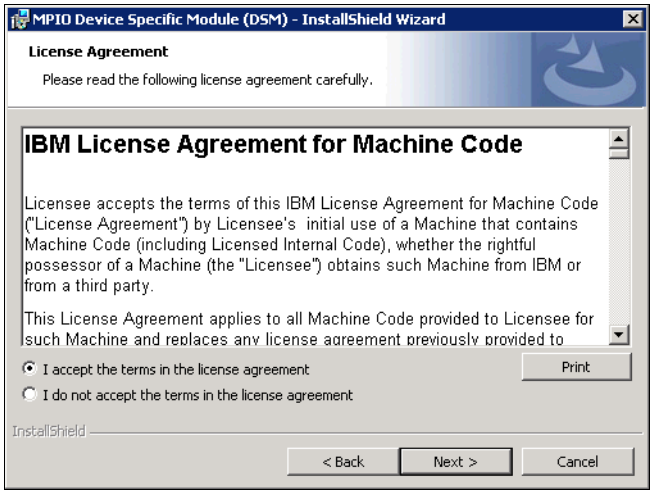


Figure 4-12 MPIO Device Specific Driver - License Agreement

2. Click **Install** to start with the installation process, as shown in Figure 4-13.

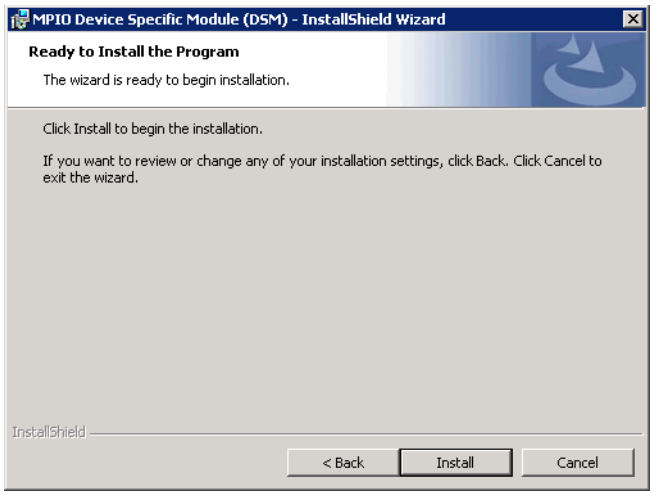


Figure 4-13 MPIO Device Specific Driver - Installation begin

3. During the installation process, a new window will be opened showing the installation process from the console. See Figure 4-14.

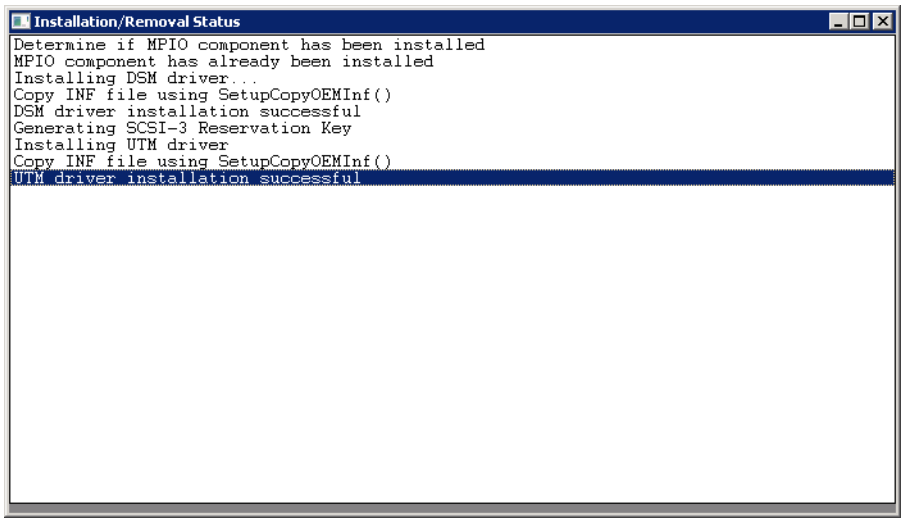


Figure 4-14 - MPIO Device Specific Driver - installation process status console

4. Finally, click **Finish** to complete the installation process. A reboot is required to apply the new driver changes. See Figure 4-15.

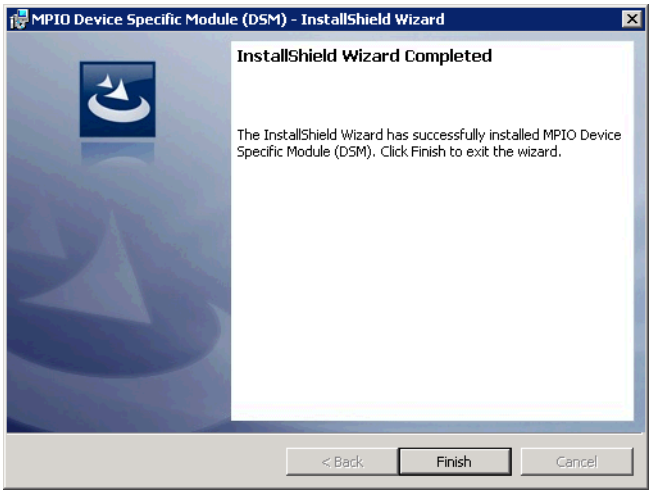


Figure 4-15 MPIO Device Specific Driver - installation completed

Verifying the MPIO Device Specific Module (DSM)

Follow these steps to verify the installation of the MPIO/DSM driver:

1. Check the install directory for the SM Failover driver. Here is the default install directory:

```
32Bit OS -> C:\Program Files\DSMDrivers\ds4dsm
64Bit OS -> C:\Program Files (x86)\DSMDrivers\ds4dsm
```

2. Open the Device Manager in Computer Management. There must be a Multi-Path Support entry under the System Devices folder, corresponding to the Device Specific Module loaded during the Storage Manager installation.
3. Right-click it, select **Properties**, and check the **Driver** tab for the version, as shown in Figure 4-16.

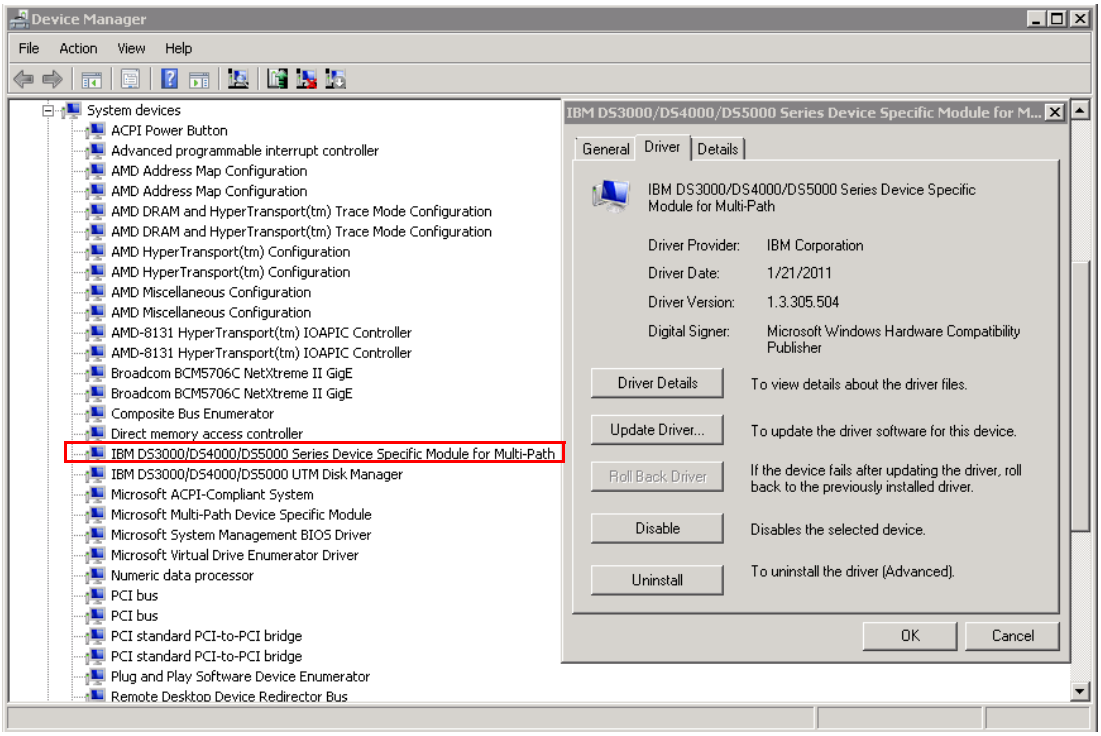


Figure 4-16 Verifying DSM driver level

Recognizing the DS Storage volumes

After the host software is installed in your Windows host, together with the host bus adapter drivers, and the DS Storage System is configured with logical volumes properly mapped, then you can recognize the DS Storage volumes in your host.

However, before starting to work on recognizing these volumes in your host, make sure that the SAN zoning is properly set up, if you are working in an FC environment, according to your planned configuration. For specific steps to configure SAN FC Zoning, see *Implementing an IBM/Brocade SAN with 8 Gbps Directors and Switches*, SG24-6116 and *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

For iSCSI attachment, make sure that the network being used is properly configured (IP, VLANs, Frame size, and so on), and has enough bandwidth to provide Storage attachment. You need to analyze and understand the impact of the network into which an iSCSI target is to be deployed prior to the actual installation and configuration of an IBM DS5000 storage system. See the “iSCSI” sections of the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Follow these steps to detect DS Storage logical volumes mapped to your Windows 2008 host:

1. Go to the Windows device manager, and highlight the **Disks** folder. After everything is configured, connected, and mapped correctly, select from the Server Manager window **Action** → **Scan for hardware changes.** Observe the results shown in Figure 4-17.

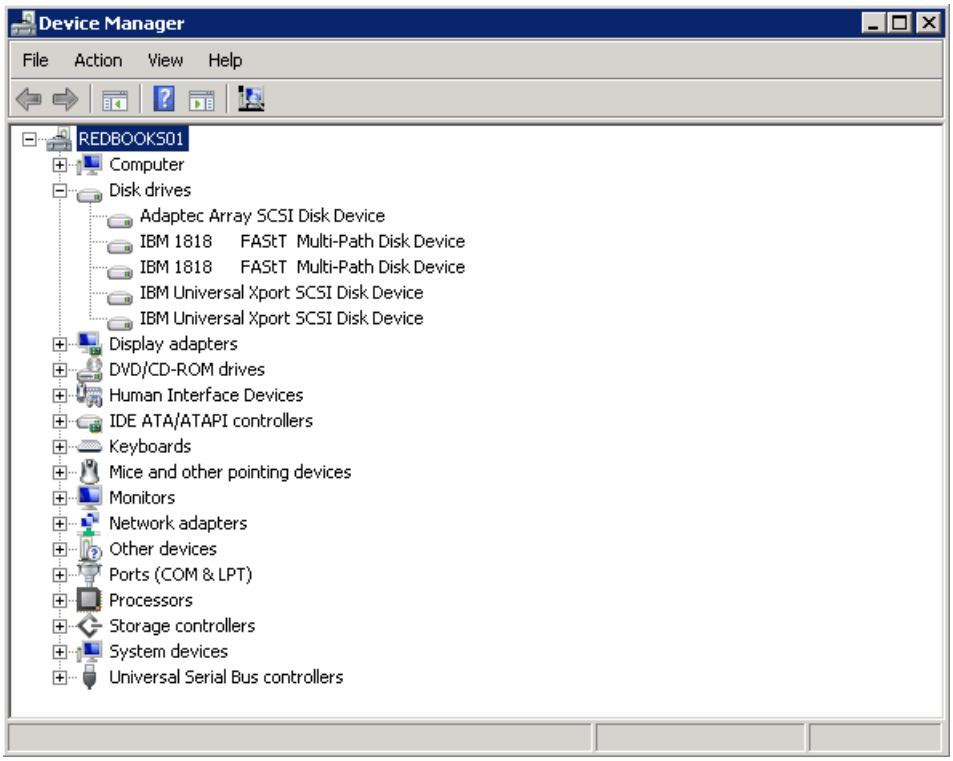


Figure 4-17 Displaying logical volumes and access LUNs

2. Each logical drive is presented as an IBM *XXX* Multipath Disk Device under the Disk drives folder, where *XXX* is the product ID of the DS Storage subsystem. Notice that in Figure 4-17, there are two disks drives, IBM 1818, mapped to this host, the logical volumes from a DS5300 Storage System.
3. Count the number of drives that you are detecting as coming from your DS Storage Systems, and make sure that this number is the same as the number of logical volumes mapped in your Storage Systems. It does not matter how many paths each volume has, it shows only as one device in the Disks folder.

   Later we explain how to individualize each disk drive presented in the device manager, to the logical drive, and how to check all the paths.

Determining device paths

From the Windows Device Manager, select a disk drive, right-click it, and select **Properties**. In the Properties windows, select the **MPIO** tab to display the configured paths to the selected disks. See Figure 4-18.

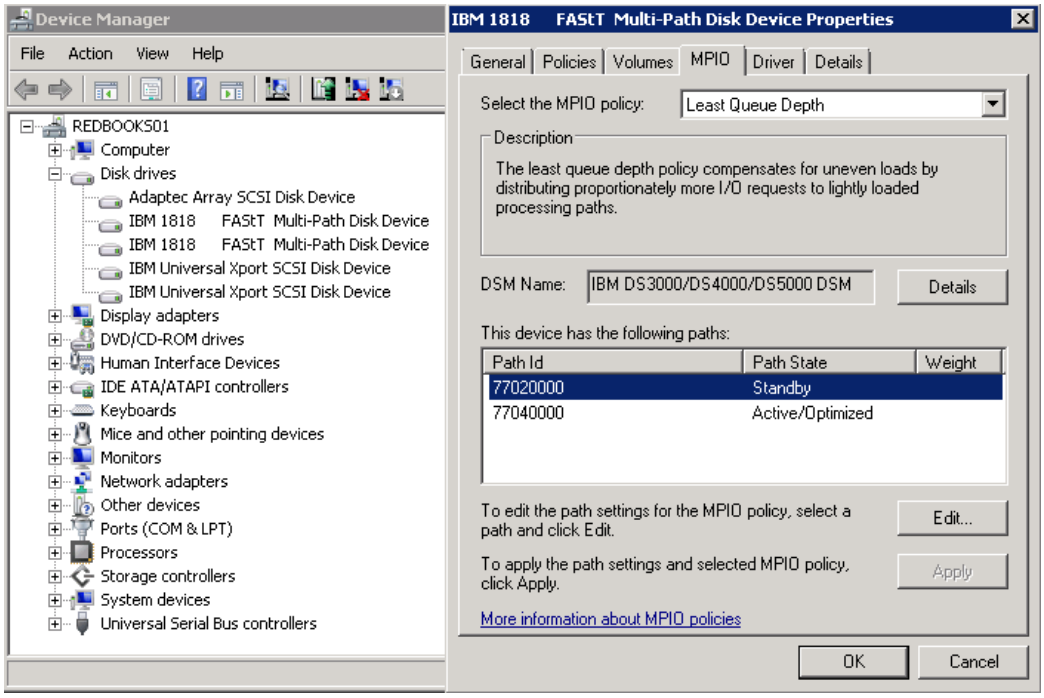


Figure 4-18 Determining device paths

Important: After a new installation, make sure to check the paths available for each detected disk, and their state. You must see only one disk per logical volume mapped in the Device Manager, but each one has to have as many paths as defined by your cabling, SAN Zoning, or iSCSI configuration.

Matching device path to DS controller

Determine which device path corresponds to your DS Storage System, controller A or controller B. From the **Device Manager** → **Disk drives** → **Properties** window (select a disk drive first), select the MPIO tab, highlight one specific path, and click **Edit**, to get the "Controller ID" as shown in Figure 4-19.

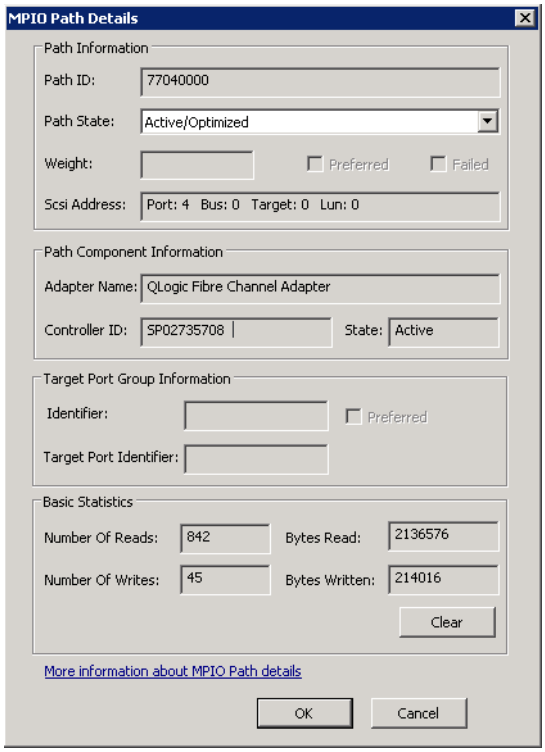


Figure 4-19 Getting the controller ID from MPIO driver

Next, we show how to match the path to the specific DS Storage System controller. Open the DS Storage Manager, select the Storage subsystem to be manager, then go to the physical tab and select the first controller. Navigate to the entire information until matching device path to DS controller by Serial Number, as shown in Figure 4-20.

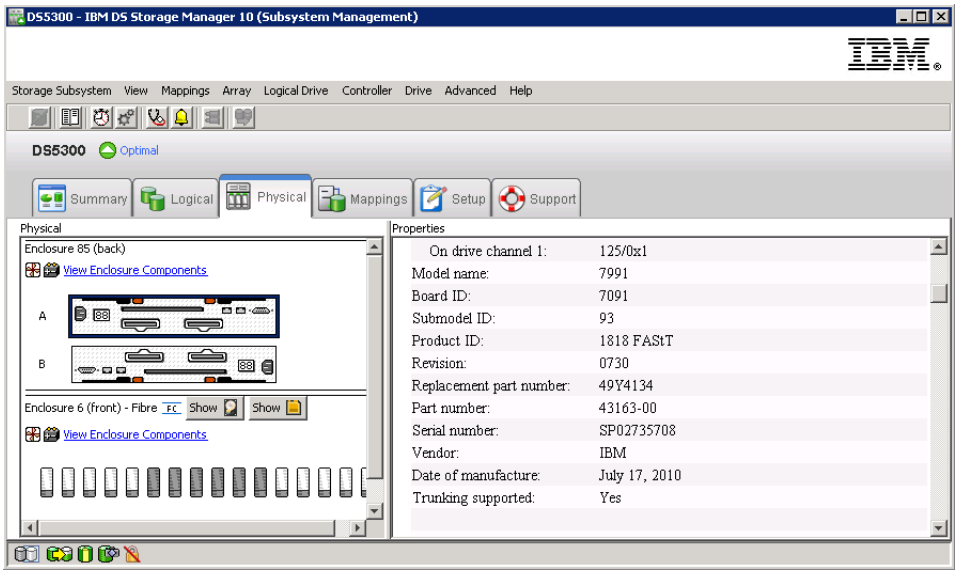


Figure 4-20 Matching device path on the DS Manager by serial number

To find which is the current path of a particular logical volume, select it from the Subsystem Management Logical View. In the right frame, scroll down until the Current Owner is displayed. You will also see the Preferred Owner information for the same logical volume.

4.3.4 Load balance policy

When you have multiple paths to the DS Storage System, the driver can select between various options specifying how to manage the traffic between the Host and the Storage. The following options are available for Windows.

One path per controller: Fail Over

The Fail Over option is not a balancing algorithm, but we include it here because it is presented to select it together with the actual balancing policies. The driver uses Fail Over only when there is one path to each controller of your system. Using Fail Over, one of the device paths is Active, and the other is in Standby State.

Multiple paths per controller

The multi-path driver transparently balances I/O workload without administrator intervention, across multiple paths to the same controller, but not across both controllers. The load balance policy uses one of three algorithms:

- Round robin with subset
- Least queue depth with subset
- Least path weight with subset

For specific details about how to select the right policy for your installation, see Chapter 2, “IBM System Storage DS5000 storage subsystem planning tasks” on page 19.

Determining load balance policy

To determine or change the load balance policy algorithm used by the device driver to redistribute the traffic between your host and the DS Storage System, follow these steps:

1. Open the Device Manager folder in your Windows host. With MPIO, each device listed here represents a Logical volume of your DS Storage System. In order to display the path failover policy, paths, and device driver version, select one of them, right-click it, and then choose **Properties**.
2. Select the **MPIO** tab to display the various paths to the DS Storage System. From the same Properties window, you can set the various **Load Balance** policies for the MPIO driver. Depending on the current physical path state and the policy selected, the paths are either both Active (round robin, weighted paths), or one is Active and the other is Standby (Fail Over, Least Queue Depth). See Figure 4-21.

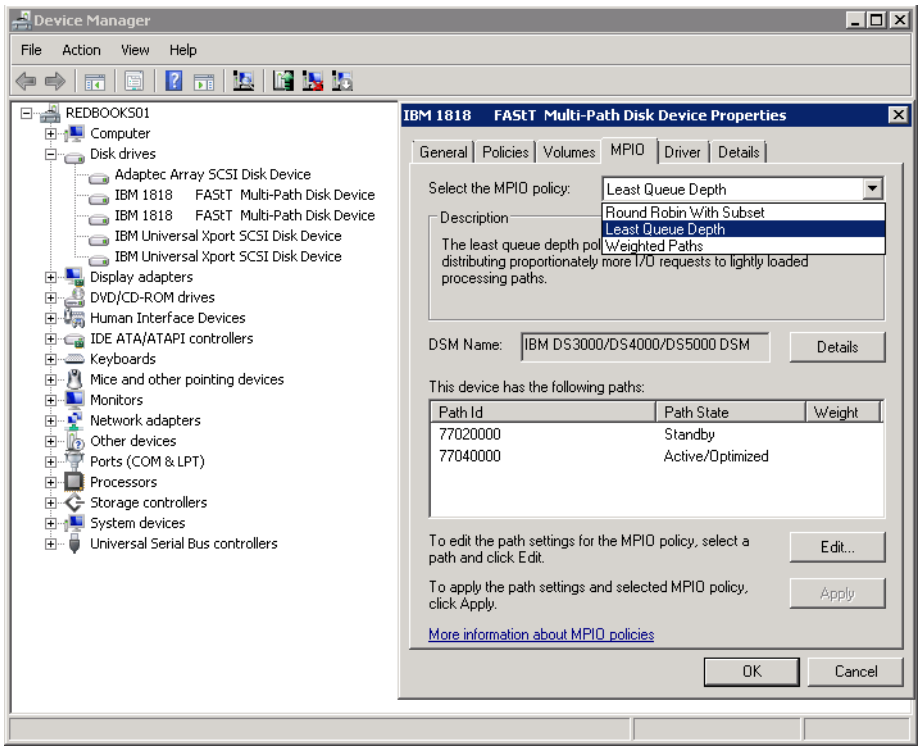


Figure 4-21 MPIO device paths load balance policies

Using the same window, you can also check the status of each path, and also the DSM driver level. In the previous example, one paths is Active, and one is Standby. The Active paths are the ones going to the current controller owning the logical volume.

4.3.5 Matching DS logical drives with Windows devices

The logical drives configured and mapped to your Windows host are represented in the Windows Device Manager under the Disk drives section. To work with them, follow these steps:

1. Open the Device Manager folder to see the detected mapped logical volumes. Each logical drive is presented as IBM xxx Multipath Disk Device under the Disk drives folder, where xxx is the product ID of the DS Storage subsystem, as shown in Figure 4-22.

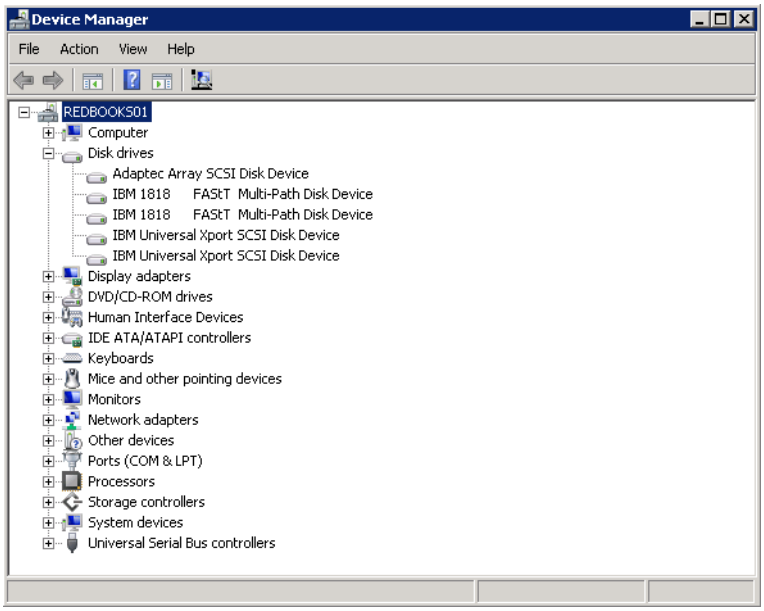


Figure 4-22 Verifying the disks in Device Manager

If the MPIO DSM driver is installed and your DS Storage System was configured properly to map each logical drive to each of the hosts adapters, then you can see in the Device Manager, the same number of devices as logical drives mapped in your DS Storage subsystem.

2. For a correct data assignment to your defined logical volumes, you need to determine which of the disks drives in the Windows Device Manager, or Disk Manager relates to the previously defined logical volumes of your DS Storage System. Select the disk drive to identify, right-click, and choose **Properties**.

3. Use the LUN number to isolate each logical drive as shown in Figure 4-23, comparing it against the Storage Subsystem Mapping tab in Figure 4-24.

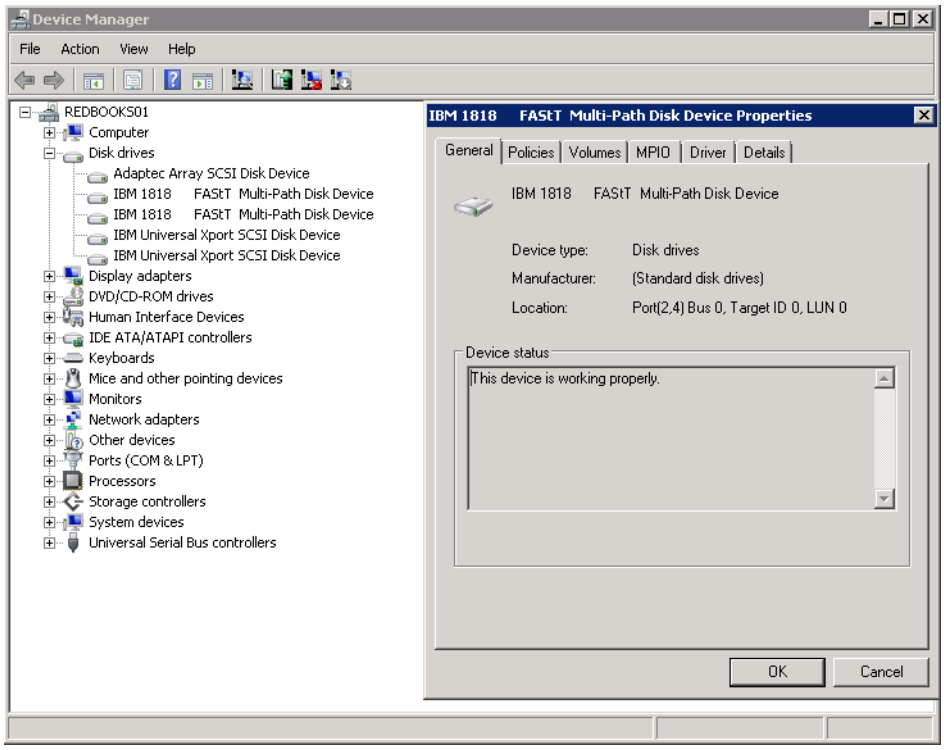


Figure 4-23 Matching LUNs between Windows host and SM

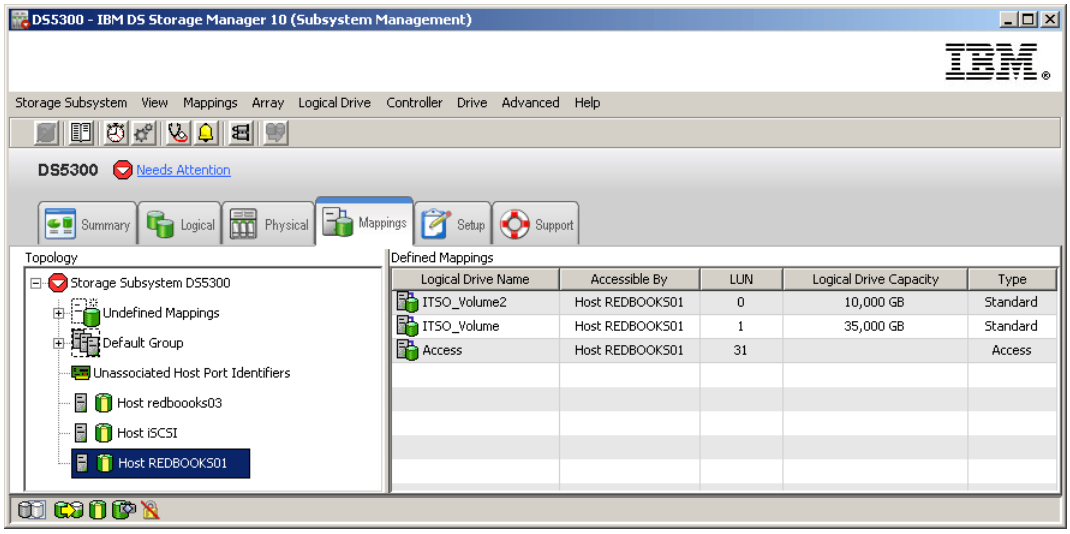


Figure 4-24 Matching LUNs on DS Storage Manager

4. Finally, use the Windows Disk Management panel to start using your newly mapped DS disks. Before making any changes to your disk structure, make sure to identify your disks correctly. You can get the same information presented previously in Figure 4-23 on page 174, by right-clicking each of the drives and selecting **Properties**, as shown in Figure 4-25.

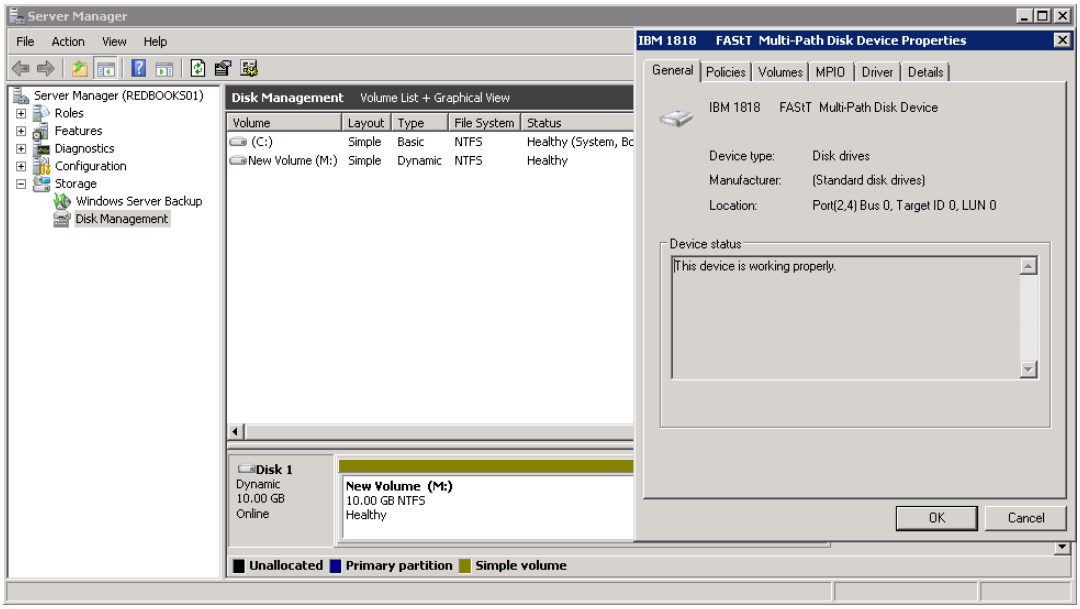


Figure 4-25 Matching LUNs on Windows Disk Management

Important: Before making any changes in the disk configuration for your server, make sure that you have correctly selected the drive that you want to configure or modify, displaying the Device Properties to check LUN, disk name, and size. Do not rely only on the LUN information, because you might have various partitions in DS Storage Systems, each with the same LUN being presented to the host. Use the device driver utilities for further help.

4.3.6 Using Windows Disk Manager

When the logical volumes are mapped to your Windows host from the DS Storage Manager, you can use the Device Manager to individualize them and adjust the driver properties.

Then, in order to start using these volumes, follow these steps:

1. Open the Windows Disk Manager to make the volumes available for data. Initially, the new disks are presented in offline/unknown state, as shown in Figure 4-26.

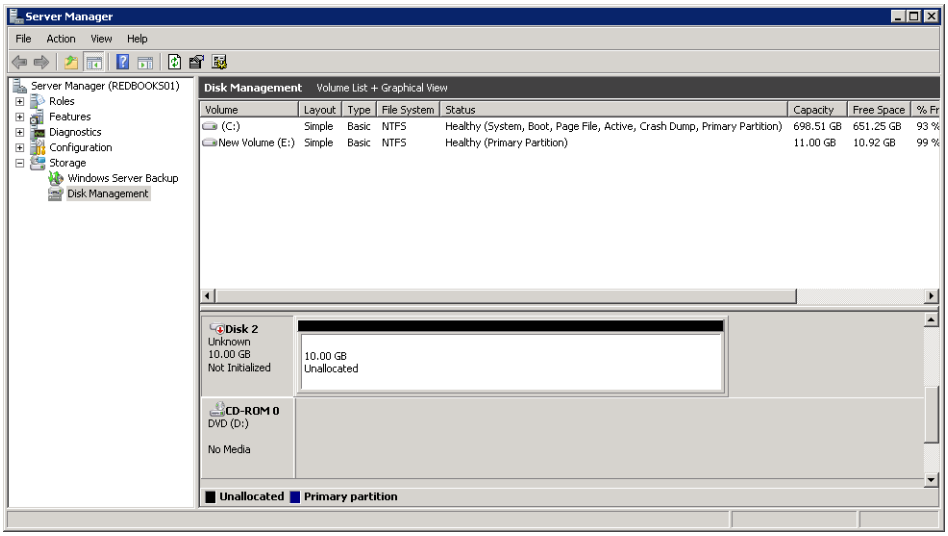


Figure 4-26 Configuring disks in online state

2. From the Disk Manager, right-click the disk to be initialized, select the disk and the partition table to be used, in this case “Disk2” & “MBR” as shown in Figure 4-27.

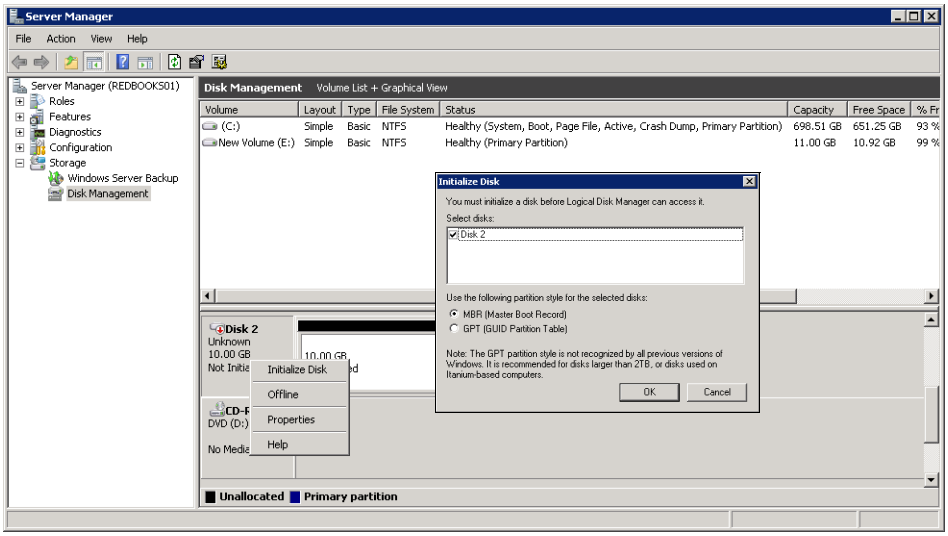


Figure 4-27 Initializing disks

3. After running the **Initialize** option, the disk is ready to be used. Create a partition by right-clicking in the **Unallocated** space area, and selecting **New Simple Volume**, thus launching the New Simple Volume Wizard.

   By default, a basic disk is created, but you can later change it to a dynamic disk, which allows added features. With the DS Storage Server, you can use either basic or dynamic disks, depending upon your needs and requirements, although certain features might not be supported when using dynamic disks. For more information, see the “Advanced maintenance and toubleshooting” section of the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

4. Select the size that you want for the new partition being created. Click **Next** to continue.

5. Select the desired drive letter to assign to this partition, or specify the mount point. Click **Next** to continue.

6. Specify the format options for the volume and other characteristics such as allocation unit size, volume label, and format type. Click **Next** and **Finish**.

The wizard finishes with a new volume available for use, according to the specifications set.

Expanding logical drives

It is possible to increase the size of logical drives after they are already created, which is called *Dynamic Volume Expansion* (DVE).

Increasing logical drive capacity from the DS Manager Console

Follow these steps to increase the capacity:

1. To increase the logical drive capacity, on the Subsystem Manager Logical Tab, highlight the logical drive to be expanded, right-click it, and select **Increase Capacity**. A new information window will be opened summarizing which OS supports logical expansion, as shown in Figure 4-28.

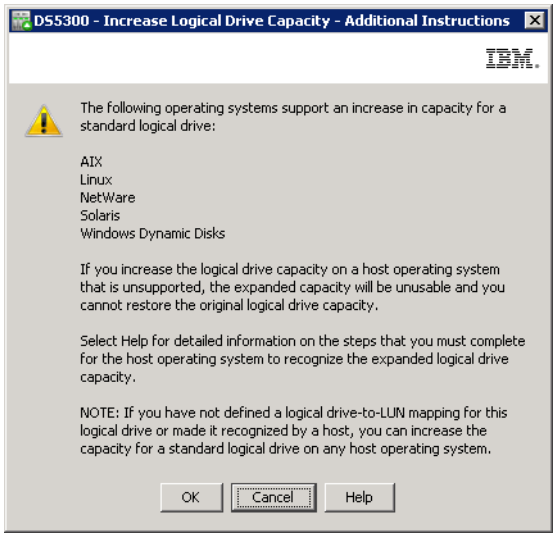


Figure 4-28 Increasing logical capacity

2. In the Increase Logical Drive Capacity window, Figure 4-29, enter the amount of space by which the logical drive will be increased.

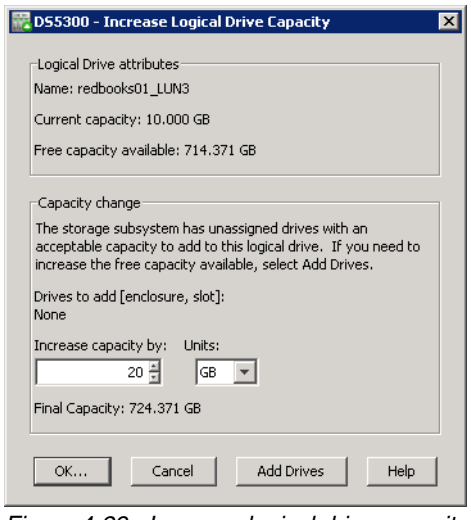


Figure 4-29 Increase logical drive capacity

In the top part of the window shown in Figure 4-29 on page 177, the current size of the logical drive and the available free capacity in the array are displayed. If no free configured space is available but there are unassigned drives, they can be added from this same window by clicking **Add Drives** before proceeding to enlarge the logical drive.

3. Before to proceed with the new space allocation, the system will confirm the task, as shown in Figure 4-30.

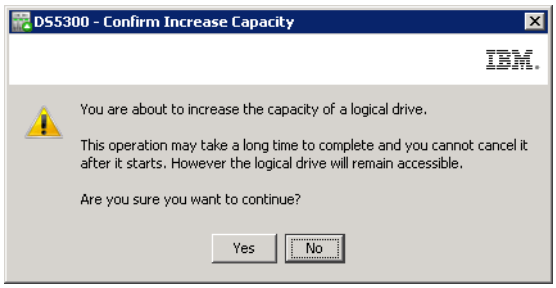


Figure 4-30 Confirm increase capacity

After the capacity of the logical drive is increased, you need to work in your host definition of the volume to make it available to use.

4. For checking the status of the task (increasing drive size) go to the **Storage Subsystem** → **View** → **Operation in progress** as shown in Figure 4-31.

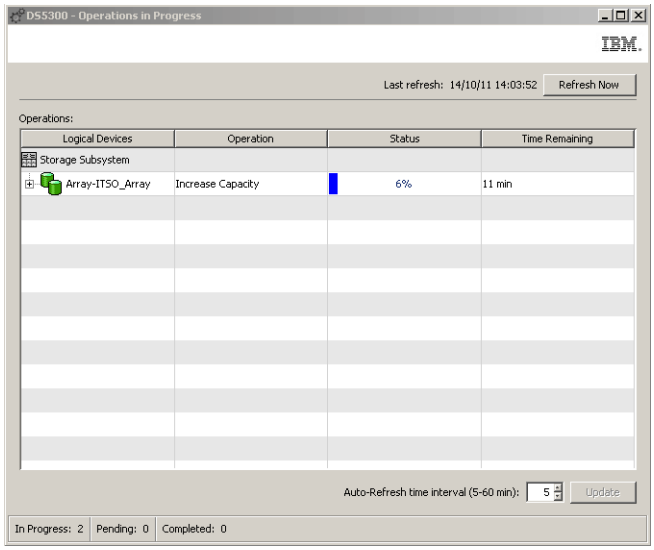


Figure 4-31 Checking Operations in progress

For more information about increasing logical disk, see 3.1.4, “Increasing logical drive capacity” on page 114.

Extending capacity using Disk Manager

In the previous section, we covered how to increase the space on a logical disk using DS Manager. Now we explain how to show/add this space to the Disk2 (F: Drive), previously created as Dynamic Volume:

1. Right-click **Disk2** and select **Extend Volume** as shown in Figure 4-32.

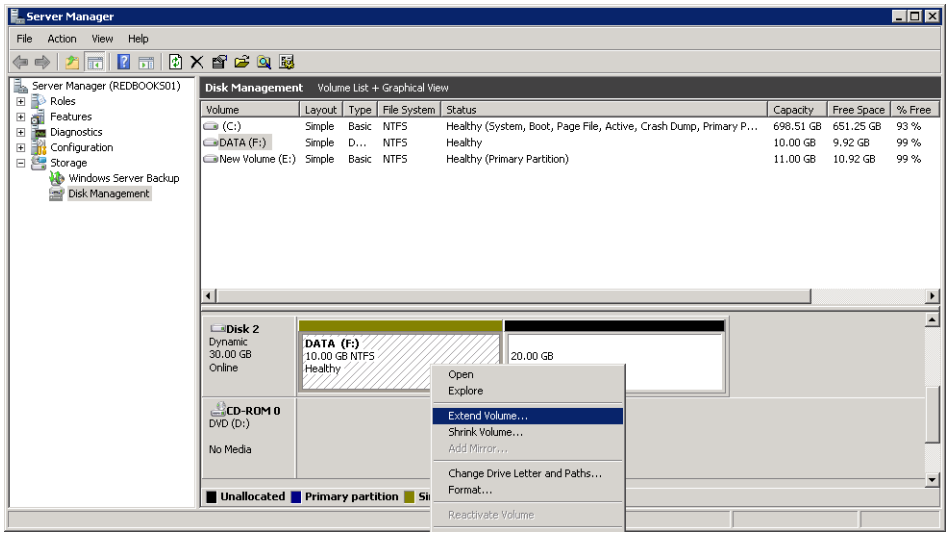


Figure 4-32 Extending Disk Volume from Disk Manager GUI

2. Select the space to be added to the Dynamic Disk and click **Next** as shown in Figure 4-33.

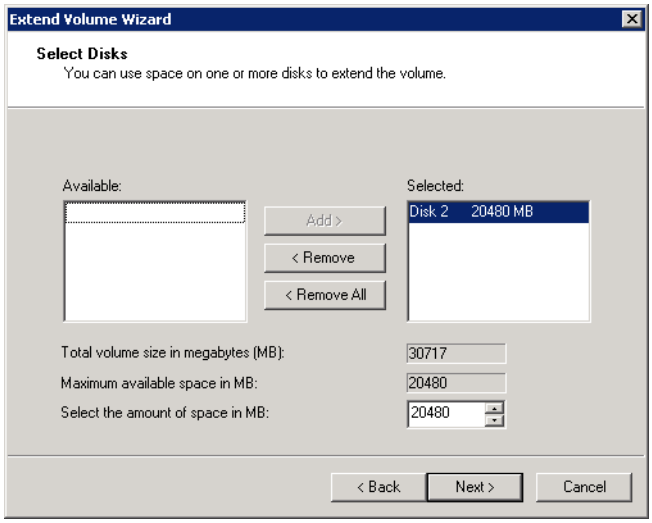


Figure 4-33 Selecting the space to be added

Finally, we have the new space allocated to the Disk2 (F: Drive) as shown in Figure 4-34.

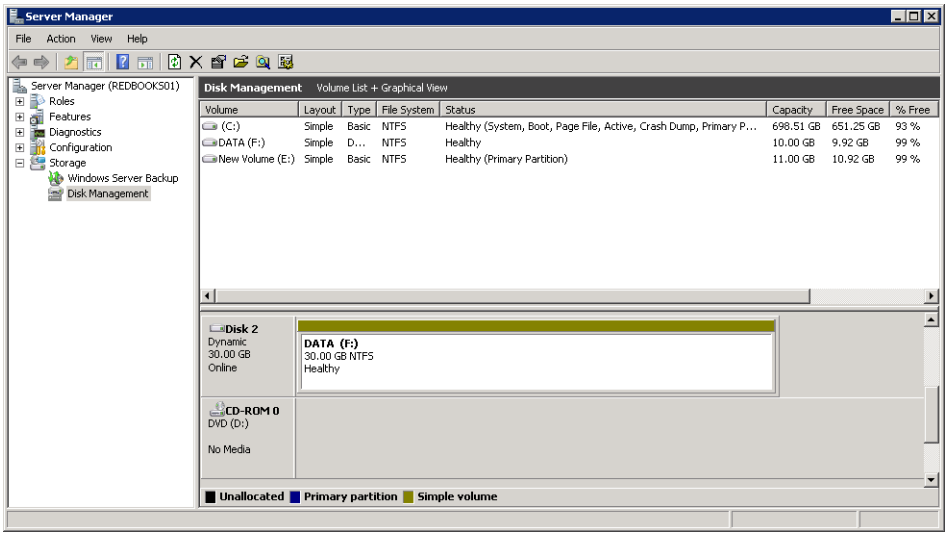


Figure 4-34 Showing the space on the Disk2 (F: Drive)

Other disk management tools

Microsoft Diskpart and third party offline utilities such as Partition Magic and Gparted can be also used to manage the disk management tasks.

Check the following tecnnical documents for more information using Microsoft disk management tools:

A description of the Diskpart Command-Line Utility:
http://support.microsoft.com/kb/300415/en-us

Using Disk Management GUI:
http://technet.microsoft.com/en-us/library/cc770943.aspx

4.3.7 Using the IBM Device Driver utilities

Storage Manager provides other utilities to display the logical drives mapped.

Storage Manager SMdevices utility

Move to the folder until under your Storage Manager installation directory:

```
32Bit OS -> C:\Program Files\IBM_DS\util
64Bit OS -> C:\Program Files (x86)\IBM_DS\util
```

Then run the SMdevices utility from the **cmd** prompt.

You can use the information provided by the SMdevices utility as follows:

- It helps you to determine the Storage Manager utility version (which is the same as the Storage Manager version).
- It provides a complete list of all the logical volumes, including these volumes:
  - Parent Storage Subsystem Name
  - Logical Drive name as assigned from Storage Manager
  - LUN ID
  - Preferred path in use

You can use this output to easily relate your Windows devices with your DS Storage Manager mappings. As shown in Figure 4-35, redbooks01_LUN3 Logical Drive is the LUN0 running on a DS5300 Storage System.

```
c:\Program Files (x86)\IBM_DS\util>SMdevices.bat
DS Storage Manager Utilities Version 10.00.35.15
Built Mon Dec 13 08:10:51 CST 2010
(C) Copyright International Business Machines Corporation, 2003-2010 Licensed Material -
Program Property of IBM All rights reserved.
US Government Users Restricted Rights - Use, Duplication, or disclosure restricted by
GSA ADP Schedule Contract with IBM Corp.

\\.\PHYSICALDRIVE2 [Storage Subsystem DS5300, Logical Drive redbooks01_LUN3,
LUN 0, Logical Drive ID <600a0b80006e32020000fcfd4e9eb175>, Preferred Path
(Controller-B): In Use]
```

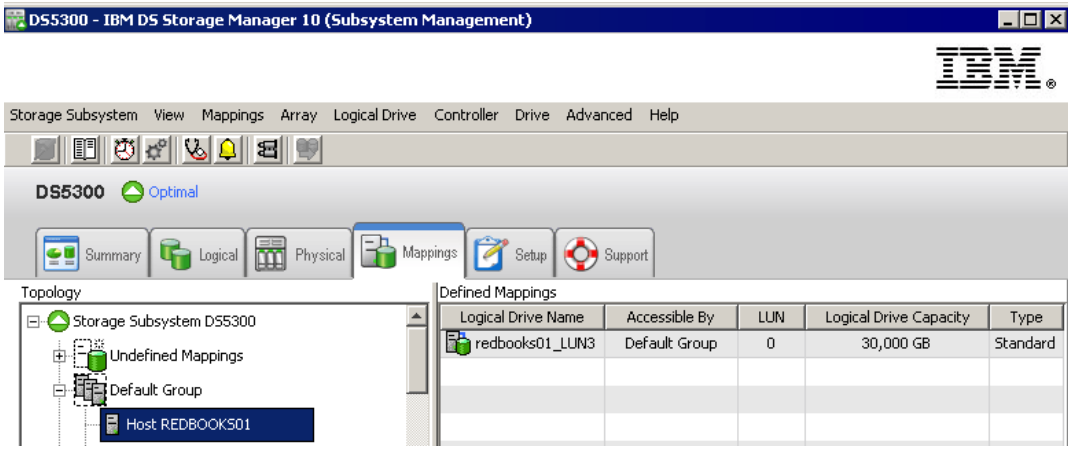


Figure 4-35 Matching LUNs using SMdevices command

MPIO DSM dsmUtil

DSMUtil is part of the IBM DSM driver installed by default in the directory:

```
32Bit OS -> C:\Program Files\DSMDrivers\ds4dsm
64Bit OS -> C:\Program Files (x86)\DSMDrivers\ds4dsm
```

You can use this utility for the following options:

- **dsmUtil -a**

  This option lists all the storage subsystems mapped to this host. See Example 4-2.

Example 4-2 dsmUtil -a

```
c:\Program Files (x86)\DSMDrivers\ds4dsm>dsmUtil.exe -a
Hostname    = REDBOOKS01
Domainname  = N/A
Time        = GMT Wed Oct 19 17:10:17 2011
------------------------------------------------------------------------
Info of Array Module's seen by this Host.
------------------------------------------------------------------------
ID              WWN                         Interface(s)    Name
------------------------------------------------------------------------
 0      600a0b80006e1bbe000000004d8093d6  iSCSI, FC       DS5300
------------------------------------------------------------------------
```

- **dsmUtil -M**

  This option lists all LUNs including System Name and Logical Name assigned from the DS Storage Manager. It is similar to the SMdevices output, so you can use it to relate your DS Logical volumes to its corresponding representation in your Windows host. It also shows the number of paths per device. See Example 4-3.

Example 4-3 dsmUtil -M

```
c:\Program Files (x86)\DSMDrivers\ds4dsm>dsmUtil.exe -m

c:\Program Files (x86)\DSMDrivers\ds4dsm>dsmUtil.exe -M
\\.\PHYSICALDRIVE1    MPIO Disk0    [Storage Array DS5300, Volume VC_source, Lun 1,
Serial Number <600a0b80006e32020000fc1d4e9c288c>][Physical Paths <1>, DSM Driver <IBM
DS3000/DS4000/DS5000 DSM>]
\\.\PHYSICALDRIVE2    MPIO Disk1    [Storage Array DS5300, Volume redbooks01_LUN3, Lun
0, Serial Number <600a0b80006e32020000fcfd4e9eb175>][Physical Paths <2>, DSM Driver <IBM
DS3000/DS4000/DS5000
DSM>]
```

In the previous example, observe there are two LUNs 0, each from a separate DS Storage System. Make sure that you select the correct one for use, especially before doing changes.

- **dsmUtil -a *SysName***

  This option lists the complete information for the given Sysname. It includes detailed information about each volume and paths, useful for troubleshooting path problems. See Example 4-4.

Example 4-4 Detailed data using dsmUtil

```
c:\Program Files (x86)\DSMDrivers\ds4dsm>dsmUtil.exe -a DS5300
Hostname    = REDBOOKS01
Domainname  = N/A
Time        = GMT Wed Oct 19 17:13:41 2011

MPP Information:
----------------
      ModuleName: DS5300                                SingleController: N
 VirtualTargetID: 0x000                                    ScanTriggered: N
     ObjectCount: 0x000                                       AVTEnabled: N
             WWN: 600a0b80006e1bbe000000004d8093d6             RestoreCfg: N
    ModuleHandle: 0xFFFFFA80074EE270                        Page2CSubPage: Y
 FirmwareVersion: 7.77.18.0                                FailoverMethod: C
   ScanTaskState: 0x00000000
        LBPolicy: LeastQueueDepth

Controller 'A' Status:
----------------------
ControllerHandle: 0xFFFFFA80074E50A0                    ControllerPresent: Y
    UTMLunExists: N                                               Failed: N
   NumberOfPaths: 2                                      FailoverInProg: N
                                                            ServiceMode: N

                                                  Path #1
    ---------
 DirectoryVertex: 0xFFFFFA80076F6068                              Present: Y
       PathState: OPTIMAL
          PathId: 0x77070001 (P07P00I01)
   InterfaceType: iSCSI
```

```
   ReqInProgress: 0

    Path #2
    ---------
 DirectoryVertex: 0xFFFFFA80076F60F8                              Present: Y
       PathState: OPTIMAL
          PathId: 0x77040000 (P04P00I00)
   InterfaceType: FC
   ReqInProgress: 0


Controller 'B' Status:
----------------------
ControllerHandle: 0xFFFFFA800555A620                    ControllerPresent: Y
    UTMLunExists: N                                                Failed: N
   NumberOfPaths: 1                                       FailoverInProg: N
                                                             ServiceMode: N

    Path #1
    ---------
 DirectoryVertex: 0xFFFFFA80076F6300                              Present: Y
       PathState: OPTIMAL
          PathId: 0x77020000 (P02P00I00)
   InterfaceType: FC
   ReqInProgress: 0



Lun Information
---------------
    Lun #0 - WWN: 600a0b80006e32020000fcfd4e9eb175
    ----------------
       LunObject: 0x0                                   CurrentOwningPath: B
  RemoveEligible: N                                        BootOwningPath: B
   NotConfigured: N                                         PreferredPath: B
        DevState: N/A                                NeedsReservationCheck: N
        LBPolicy: LeastQueueDepth                                TASBitSet: Y
                                                                  NotReady: N
                                                                      Busy: N
                                                                 Quiescent: N

    Controller 'A' Path
    --------------------
   NumLunObjects: 1                                      RoundRobinIndex: 0
         Path #2: LunPathDevice: 0xFFFFFA80053D6010
                      DevState: OPTIMAL
                    PathWeight: 0
                  RemoveState: 0x0  StartState: 0x0  PowerState: 0x0

    Controller 'B' Path
    --------------------
   NumLunObjects: 1                                      RoundRobinIndex: 1
         Path #1: LunPathDevice: 0xFFFFFA80046F53B0
                      DevState: OPTIMAL
                    PathWeight: 0
                  RemoveState: 0x0  StartState: 0x0  PowerState: 0x0


    Lun #1 - WWN: 600a0b80006e32020000fc1d4e9c288c
    ----------------
```

```
     LunObject: 0x0                                    CurrentOwningPath: A
RemoveEligible: N                                        BootOwningPath: A
 NotConfigured: N                                         PreferredPath: A
      DevState: N/A                                NeedsReservationCheck: N
      LBPolicy: LeastQueueDepth                             TASBitSet: Y
                                                             NotReady: N
                                                                 Busy: N
                                                            Quiescent: N

  Controller 'A' Path
  -------------------
 NumLunObjects: 1                                      RoundRobinIndex: 0
       Path #1: LunPathDevice: 0xFFFFFA8004156D50
                    DevState: OPTIMAL
                  PathWeight: 0
                 RemoveState: 0x0  StartState: 0x0  PowerState: 0x0

  Controller 'B' Path
  -------------------
 NumLunObjects: 0                                      RoundRobinIndex: 1
```

4.3.8 iSCSI Software Initiator implementation

The DS Storage System has the option to attach your hosts using iSCSI interfaces. In this section, we show how to configure your Microsoft Windows Server to use a regular Ethernet network interface card (NIC) and software for iSCSI Software Initiator to connect to a DS5300 system with iSCSI host interface cards.

To configure a regular network adapter to access your DS Storage System using iSCSI Software Initiator, you need to install and configure an iSCSI software to provide the added protocol to the adapter. Windows Server 2008 R2 has iSCSI Initiator installed natively. On prior versions of Windows such as Windows Server 2003 and 2008, the software must be downloaded and installed separately.

Our implementation example is using Windows 2008 R2, with two Ethernet cards connected to a private network where the DS5300 iSCSI controllers resides. There is also a 10 GB disk configured as an iSCSI target disk using CHAP mutual authentication.

The DS Storage System iSCSI ports are defined as follows:

- IP Address - 192.168.130.101 - iSCSI Controller A
- IP Address - 192.168.130.102 - iSCSI Controller B

For more information about iSCSI configuration from the DS Storage side, see 3.1.6, “iSCSI configuration and management” on page 123.

Configuring iSCSI Software Initiator

As mentioned previously, iSCSI software is natively installed on the Windows 2008 R2 Server. To launch the iSCSI Initiator GUI, follow these steps:

1. Click **Start** → **Control Panel** → **Classic View** → **iSCSI Initiator.**

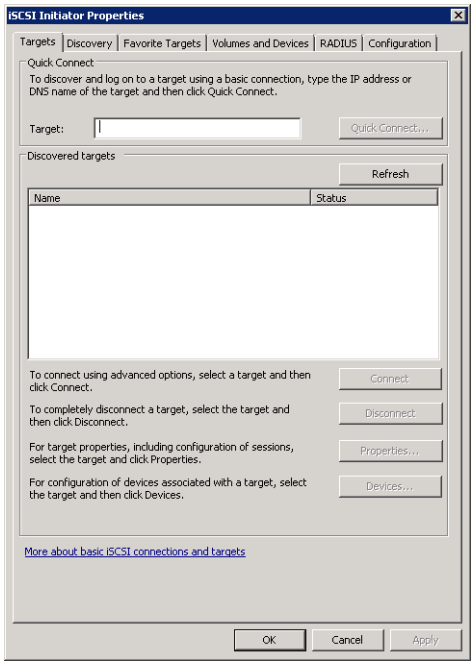


Figure 4-36 Launch iSCSI initiator GUI

2. To check our iSCSI Qualified Name (IQN), click the **Configuration** tab as shown in Figure 4-37.

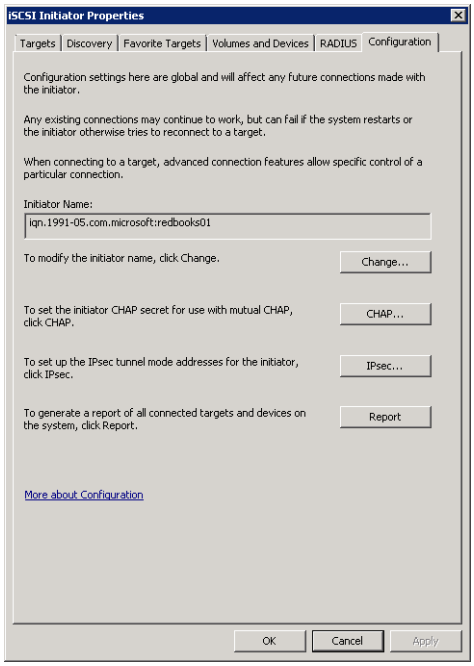


Figure 4-37 Checking iSCSI qualified name

3. To discover the iSCSI target manually, adding the DS iSCSI controller IP addresses on the “Discovery” tab, click **Discovery** as shown in Figure 4-38.

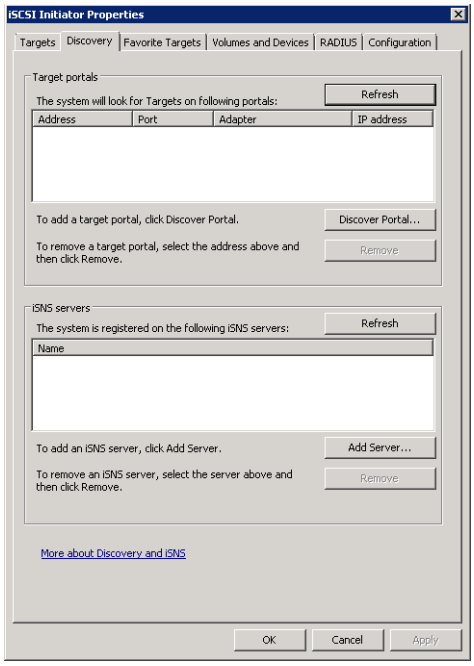


Figure 4-38 Adding iSCSI targets manually

4. Click the **Discovery Ports** button to add new portals. Enter the DS iSCSI controllerA IP address and click **Ok**. Repeat the same process to add the iSCSI controllerB, as shown in Figure 4-39.

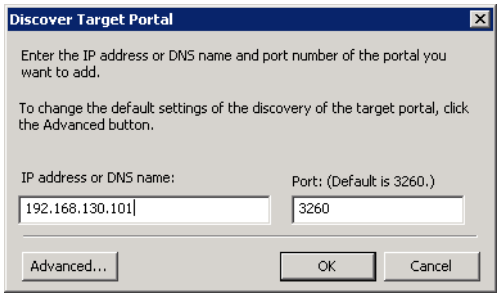


Figure 4-39 Adding new target portals

5. Check the **Targets** tab as shown in Figure 4-40. A new target has been discovered, after which we can add the iSCSI target controller IP addresses manually.

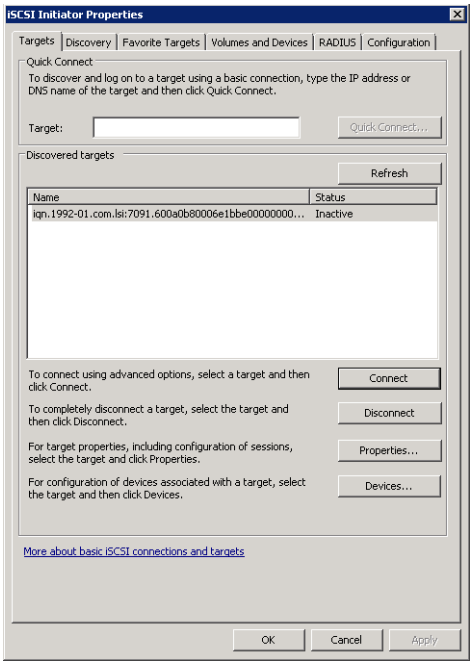


Figure 4-40 Checking discovered targets

6. After we have the target added, we need to proceed to connect them to the iSCSI client. Click the **Connect** button and follow the steps proposed as shown in Figure 4-41.

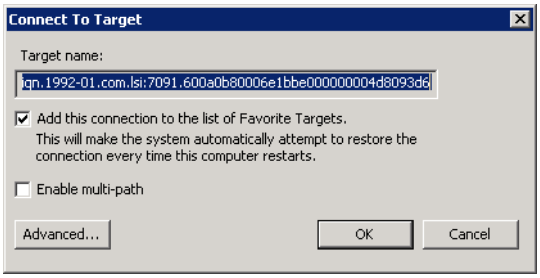


Figure 4-41 Connecting targets

Tip: To enable high availability and to boost performance, choose the Enable Multi-path check box. Make sure that multipathing (MPIO) requires multiple network adapters dedicated to the iSCSI tasks. MPIO Driver must be installed and configured beforehand.

7. If you are using CHAP authentication to secure your target, click the **Advanced** option to set the target CHAP secret key. as shown in Figure 4-42.

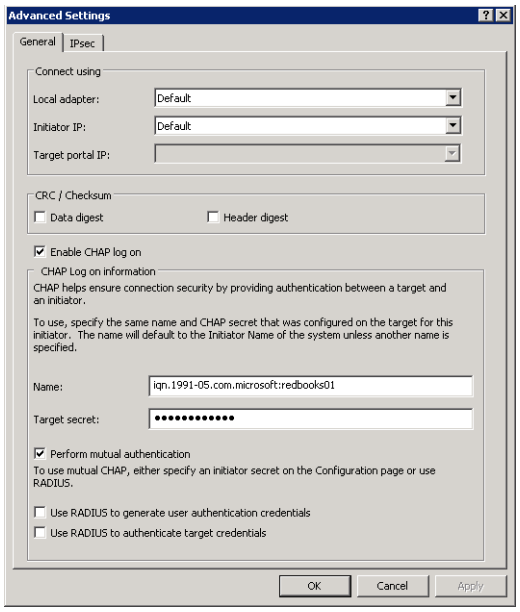


Figure 4-42 Configuring CHAP security

8. Because we already have a 10 GB iSCSI target disk configured and presented to our server, we will be able to refresh the discovered device by clicking the **Auto Configure** button on the **Mounted Devices** tab, as shown in Figure 4-43.

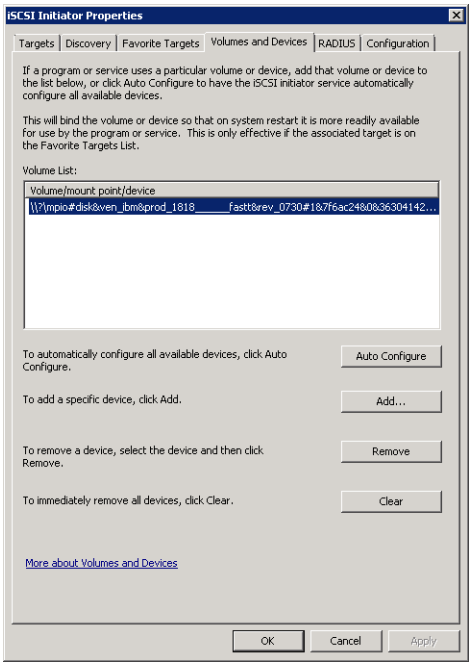


Figure 4-43 Auto Configuring target disk devices

9. When we have the target disk devices *auto configured* in the iSCSI Initiator, we will be able to see them in the Disk Administrator management console, as shown in Figure 4-44. Right-click **Disk 3** → **Select Initialize**. A window will be opened; select the disk and click **OK** to continue.

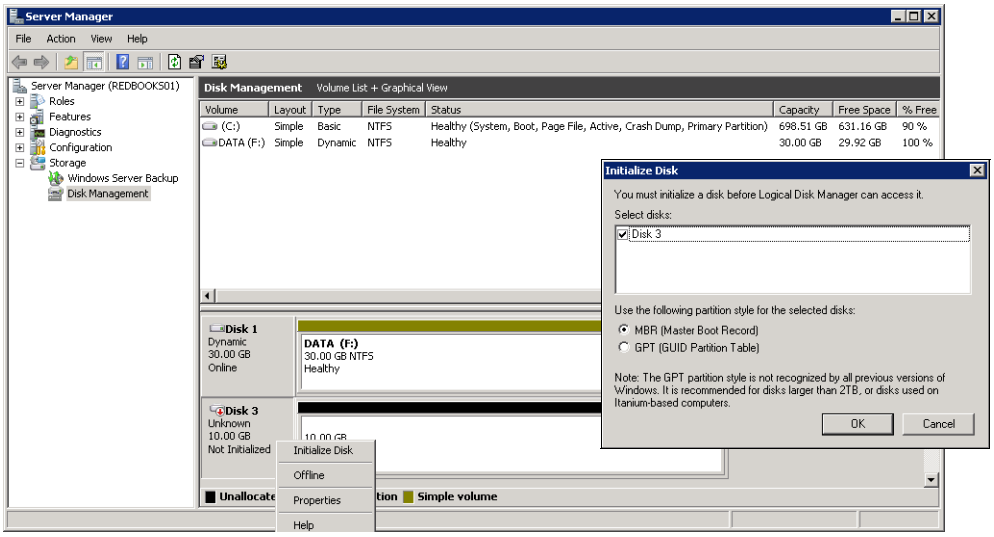


Figure 4-44 Identifying and Initializing new disks

The next configuration steps are basically the same that we already performed in 4.3.6, “Using Windows Disk Manager” on page 175.

iSCSI security suggestions and best practices

Here we list various security considerations:

- Configure additional Paths for High Availability; use either Microsoft MPIO or MCS (multiple connections per session) with additional NICs in the server to create additional connections to the iSCSI storage array through redundant Ethernet switch fabrics.
- Use Gigabit Ethernet connections for high speed access to storage.
- Use Server class NICs. It is suggested to use NICs which are designed for enterprise networking and storage applications.
- Use CAT6 rated cables for Gigabit Network Infrastructures and Cat-6a or Cat-7 for 10Gigabit implementations.
- Segregate SAN and LAN traffic. iSCSI SAN interfaces need to be separated from other corporate network traffic (LAN). Servers should use dedicated NICs for SAN traffic. Deploying iSCSI disks on a separate network helps to minimize network congestion and latency. Additionally, iSCSI volumes are more secure when… Segregate SAN & LAN traffic can be separated using port based VLANs or physically separate networks.
- Use non blocking switches and set the negotiated speed on the switches.
- Use Jumbo Frames if supported in your network infrastructure.
- Enable and configure CHAP Security to use either target or mutual authentication.

4.3.9 Collecting information

In addition to the information mentioned in *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023, you might be asked to send the following information from your Windows host system to perform problem determination over certain failures:

- System event log: Right-click the my computer icon on your desktop and select **Manage** to open the System Management window. From the system tools option under Computer Management, select **Event viewer** → **System**. Now click **Action** from the pull-down menu, and select **Save Log File as** to save the event view log file.
- Application log: Proceed as described for the system event log, but this time, select the application log.
- Dynamic system analysis, DSA log: IBM Dynamic System Analysis (DSA) collects and analyzes system information to aid in diagnosing system problems. This tool was developed initially for System x servers and not only collects the system event and application logs, but also information from your adapters, drivers, and configuration, which allows you to easily provide all the information that your support representative needs in one package. To access this tool, go to:

  `http://www-03.ibm.com/systems/management/dsa.html`
- SMdevices: Use this command and capture the output to show all the devices detected by the DSM driver.

4.4 AIX configuration

In the following section, we present the steps that you need to perform on your AIX server version 7.1 (similar to the procedure as on AIX 6.1) to install and manage your DS Storage System. In our example, we use the most recent version of DS Storage Manager (DSM) version 10.77.

Apart of the installation of DSM, we briefly cover the considerations that needs to be taken for the AIX SDDPCM and MPIO device driver, and specific steps to establish both FC and iSCSI connection to your target DS5000 storage subsystem. Remember that the DS5000 Storage subsystem with iSCSI ports does not support direct-attached connections from the host systems to the storage subsystem iSCSI ports. Always use the switched network instead.

For more details about the advanced AIX configuration, see Chapter 12, “DS5000 with AIX, PowerVM, and PowerHA” on page 541.

4.4.1 Installing DS Storage Manager software on an AIX host

This section covers the installation of the IBM System Storage DS Storage Manager software components attached to a host server running the AIX operating system. When referring to an AIX host system, it can be a stand-alone IBM System p machine, an LPAR running AIX, or even a Virtual I/O (VIO) server. Although it is possible, we do *not* suggest that you install Storage Manager on a VIO server.

The 10.77 version of the IBM DS Storage Manager host software for AIX is required for managing all DS3000, DS4000, and DS5000 storage models with controller firmware version 07.77.xx.xx. In addition, it is also suggested for managing models with controller firmware version 6.xx.xx.xx or higher installed. The similar logic applies to future versions of DS5000 firmware and Storage Manager levels.

Important: Do not install Storage Manager software on a VIO server.

Before installing SMclient, make sure that you have met the following conditions:

- The System p server hardware to which you connect the DS5000 Storage Server meets the required specifications for your storage model.
- The AIX host on which you install SMruntime meets the minimum software requirements.
- You have prepared the correct file sets for the installation on an AIX system.

Important: Review your system requirements and additional support information in the latest Storage Manager `readme` file for your specific DS5000 storage system and code level.

SMruntime provides the Java runtime environment required to run the SMclient. The AIX FCP disk array driver (also known as MPIO for AIX), provides redundancy in the I/O paths, and is part of the AIX operating system, and therefore is not included on the DS5000 Storage Manager software CD shipped with the storage server.

Depending on your configuration and environment, you have the option to install the software packages using two possible methods:

- From a graphic user interface using the installation wizard.
- From console mode of the installation wizard.
- Using the standard command line method or **`smitty`** panels.

Tip: The installation package requires sufficient free space in several file systems. It identifies and informs you about the space needed during the installation.

4.4.2 Instructions for each installation method

We describe both installation methods in the following sections. You might need to adjust the instructions for the specifics of your environment. The AIX reboot is not required during the installation process. The installation package consists of the following components that you can choose for installation:

SMruntime	Storage Manager Java compiler
SMesm	Storage Manager ESM firmware delivery package
SMclient	Storage Manager client package
SMagent	Storage Manager agent package
SMutil	Storage Manager utility package
Profiler	Storage Manager Support Monitor (10.60.x5.17 and higher) tool that installs Apache web server and MySQL database software

Installation through a graphical desktop

To install the Storage Manager software using the installation wizard, use the following steps:

1. Locate the installation code file, either in the appropriate CD-ROM directory or on the IBM support website, and transfer the code to a directory on your AIX system.

2. Make sure that the file has permissions to execute as shown in Example 4-5.

Example 4-5 Making the installation package executable

```
# ls -la SM*
-rw-r----- 1 root system    262690421 Oct 23 08:49 SMIA-AIX-10.77.G5.16.bin
# chmod +x SMIA-AIX-10.77.G5.16.bin
# ls -la SM*
-rwxr-x--x 1 root system    262690421 Oct 23 08:49 SMIA-AIX-10.77.G5.16.bin
```

3. Execute the file to begin installation. If you have insufficient free space in /tmp, you get the warning as shown in Figure 4-45.

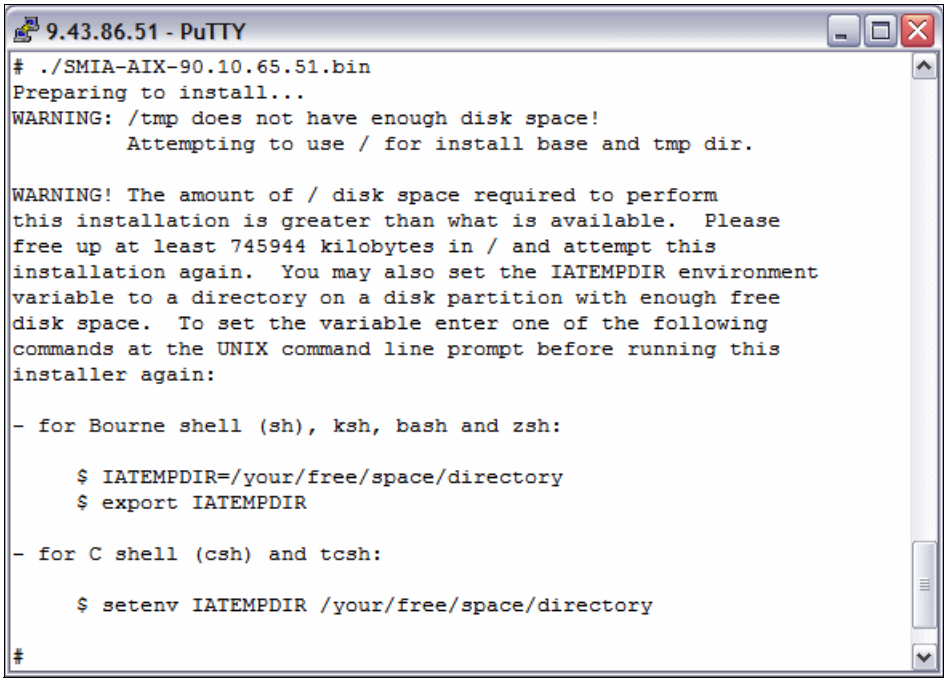


Figure 4-45 Insufficient space in target folder

4. To remedy this warning, either increase the /tmp file system to have sufficient free space, according to the warning given, or set the environment variable IATEMPDIR to a directory with sufficient free space.
5. After the installer is launched, proceed by selecting the language, and go through the introduction window, copyright statement, and license agreement windows. Acceptance of the license agreement is required to continue.
6. Now select the installation type, as shown in Figure 4-46:
   - The installation type that you select defines the components that are installed. For example, if you select Management Station, the agent component is not installed, because it is not required on the management computer. In most cases, you can select either the Management Station or Host installation type.
   - Because having only these two options might be a limitation, two additional choices are offered: full and custom installation. As the name indicates, a full installation installs all components, whereas a custom installation lets you choose from the following elements:
     - IBM Storage Manager 10 Client
     - IBM Storage Manager 10 Utilities
     - IBM Storage Manager 10 Agent

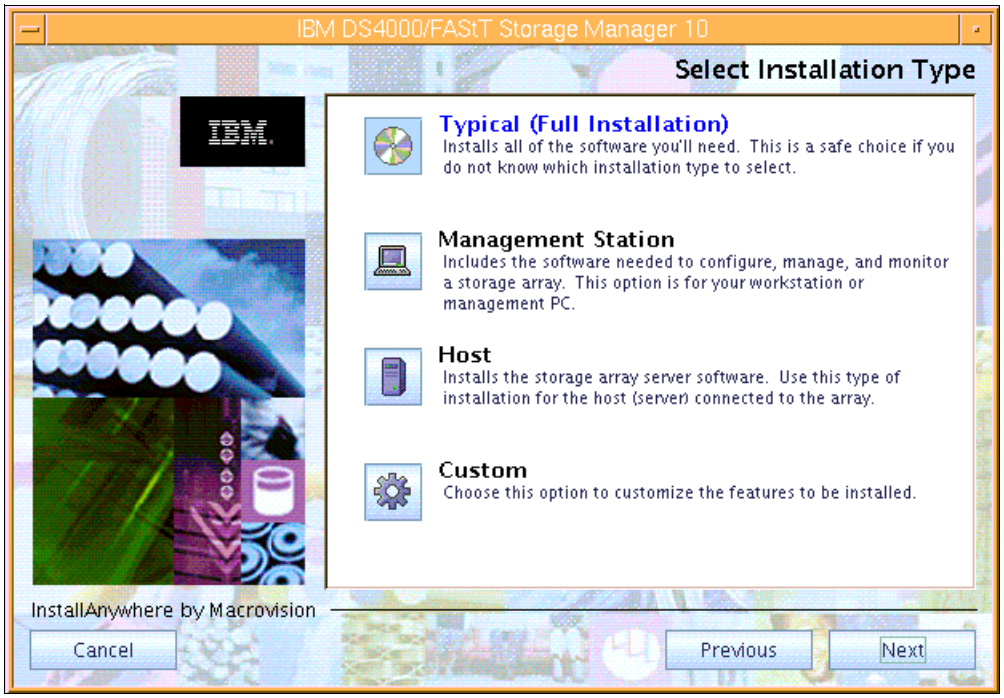


Figure 4-46 Installing Storage Manager software in AIX

7. After selecting an installation type, you are presented with the Pre-Installation Summary window. Verify that you have enough space in the designated file systems and click the **Install** button to proceed with the installation.

Installation in console mode

Follow these steps to install the Storage Manager in the console mode:

1. If you are trying to install Storage Manager from graphical user interface as described in previous section “Installation through a graphical desktop” on page 191, but your environment does not support this option, you will be automatically redirected to the console mode of your terminal session. The typical warning is shown, as in Example 4-6.

Example 4-6 Entering the console mode

```
# ./SMIA-AIX-10.77.65.16.bin
Preparing to install...
Extracting the JRE from the installer archive...
Unpacking the JRE...
Extracting the installation resources from the installer archive...
Configuring the installer for this system's environment...

Launching installer...

Graphical installers are not supported by the VM. The console mode will be used
instead...

Preparing CONSOLE Mode Installation...
===============================================================================
IBM System Storage DS Storage Manager 10          (created with InstallAnywhere)
```

In this case, select the appropriate language (English is marked as default) and go through the license terms and conditions by multiple pressing of Enter (approximately 180 times, it depends on your console window size). Respond to each prompt to proceed to the next step in the installation. If you want to change something on a previous step, type **back**. You can cancel the installation at any time by typing **quit**.

2. When you accept the license terms and conditions, the following screen appears as shown in Example 4-7.

Example 4-7 Select the type of installation

```
Select Installation Type
------------------------
Please choose the Install Set to be installed by this installer.

  ->1- Typical (Full Installation)
    2- Management Station
    3- Host
    4- Customize...

ENTER THE NUMBER FOR THE INSTALL SET, OR PRESS <ENTER> TO ACCEPT THE DEFAULT
```

3. Make sure you have sufficient space in the designated folder and review the pre-installation summary as in our Example 4-8.

Example 4-8 installation preview

```
Pre-Installation Summary
------------------------
Please Review the Following Before Continuing:

Install Folder:
    /opt/IBM_DS

Required Disk Space
    1,466 MB

Available Disk Space
    1,847 MB
```

As you can see, the typical installation automatically chosen the folder /opt/IBM_DS as a target filesystem, where we do not have optimal free space. To reject the default settings, type `back` in the command prompt and select customized process (Option 4) as shown in Example 4-7.

4. Select the features you want to install on your systems by entering a comma separated list of functions, as outlined in Example 4-9. We had only the requirements for Storage Manager Client, Utilities, and Agent; therefore we put `1,3,4` and pressed Enter.

Example 4-9 Customized installation

```
Choose Product Features
-----------------------
ENTER A COMMA_SEPARATED LIST OF NUMBERS REPRESENTING THE FEATURES YOU WOULD
LIKE TO SELECT, OR DESELECT. TO VIEW A FEATURE'S DESCRIPTION, ENTER
'?<NUMBER>'. PRESS <RETURN> WHEN YOU ARE DONE:

    1- [X] DS Storage Manager 10 Client
```

```
    2- [ ] Support Monitor
    3- [X] DS Storage Manager 10 Utilities
    4- [X] DS Storage Manager 10 Agent
```

5. Again, you can review the pre-installation summary and start the installation by pressing Enter. The process bar shows you the current progress. Depending on your system and selected features, the installation takes 5-15 minutes, when done the confirmation message appears with the specification of the target folder.

When you are sure that all your features are properly installed (Example 4-10) in the system and the Storage Manager functional, we suggest to remove the extracted installation packages to save some space in `/tmp` folder (in our case, type **`rm -R /tmp/DSM/AIX10p77`**).

Example 4-10 Verification of the installation

```
# lslpp -L | grep DS
  SMruntime.aix.rte     10.77.6501.0    C     F    IBM DS Storage Manager 10
#
# ps -ef | grep /opt/IBM_DS | more
# <watch for Java processes of /opt/IBM_DS/IBMStorageManagerProfiler_Server>
```

Installation from AIX SMIT interface

If you prefer to install the packages individually, you can use the AIX smit interface:

1. Copy all the individual packages to a temporary directory. For our example, we used `/tmp`.
2. Issue the following command to create the installable list:

   ```
   # inutoc /tmp
   ```

3. Invoke the AIX management interface calling the install program:

   ```
   # smitty installp
   ```

4. Select the “Install Software” option.
5. For the input directory, type in the chosen directory name (in our case `/tmp`), and press Enter.
6. Select all packages or any individual package to install.

Tip: If you install the packages individually, be sure to install them in the sequence shown in “Installation from the command line” on page 196.

Installation from the command line

In this section, we describe an alternative method for individual package installation done from the AIX command line. This method can be used when you only have access to a regular text terminal.

To successfully install the client software through the AIX command line, you must install the various packages in the following sequence:

1. SMruntime
2. SMesm
3. SMclient (optional)
4. SMagent (optional)
5. SMutil (optional)

Installing SMruntime

To install, follow these steps:

1. Install SMruntime by typing the following command:

   ```
   # installp -a -d /tmp/DSM/SMruntime-AIX-10.77.65.16.bff SMruntime.aix.rte
   ```

2. Verify that the installation was successful by typing the following command:

   ```
   # lslpp -ah SMruntime.aix.rte
   ```

   The verification process returns a table that describes the software installation, including the install package file name, version number, action, and action status:

   ```
   Fileset         Level     Action       Status       Date         Time
     ----------------------------------------------------------------------------
   Path: /usr/lib/objrepos
     SMruntime.aix.rte
                10.77.6516.0   COMMIT       COMPLETE     10/25/11 11:04:46
                10.77.6516.0   APPLY        COMPLETE     10/25/11 11:04:46
   ```

Tip: Check the Status column for the complete installation. The levels need to match your specific version.

Installing SMesm

This package provides firmware images for the disk expansion enclosure controller cards and can be used in case of compatibility problems. SMesm is required if installing SMclient. Follow these steps:

1. Install SMesm by typing the following command:

   ```
   # installp -a -d /tmp/DSM/SMesm-AIX-10.77.G5.16.bff SMesm.aix.rte
   ```

2. Verify it by typing the following command:

   ```
   # lslpp -ah SMesm.aix.rte
   ```

   This generates output showing the status of the package installed.

Installing SMclient

SMclient is required only if you plan to run the management client from the AIX host. Follow these steps:

1. Install SMclient by typing the following command:

   ```
   # installp -a -d /tmp/DSM/SMclient-AIX-10.77.G5.16.bff SMclient.aix.rte
   ```

2. Verify that the installation was successful by typing the following command:

   ```
   # lslpp -ah SMclient.aix.rte
   ```

The verification process returns a table that describes the software installation, including the install package file name, version number, action, and action status:

```
Fileset          Level     Action       Status       Date          Time
  ----------------------------------------------------------------------------
Path: /usr/lib/objrepos
  SMclient.aix.rte
              10.77.6516.0   COMMIT       COMPLETE     10/26/11 11:33:23
              10.77.6516.0   APPLY        COMPLETE     10/26/11 11:33:23
```

3. To start the client, issue the following command:

```
# /usr/SMclient/SMclient
```

Installing SMagent

SMagent is required only for in-band management (that is, through Fibre Channel) Follow these steps:

1. Install the SMagent by typing the following command:

```
# installp -a -d /tmp/DSM/SMagent-AIX-10.00.65.16.bff SMagent.aix.rte
```

2. Verify the installation by typing the following command:

```
# lslpp -ah SMagent.aix.rte
```

The verification process returns a table that describes the software installation, including the install package file name, version number, action, and action status:

```
Fileset          Level     Action       Status       Date          Time
  ----------------------------------------------------------------------------
Path: /usr/lib/objrepos
  SMagent.aix.rte
              10.0.6516.0   COMMIT       COMPLETE     10/26/11 11:20:12
              10.0.6516.0   APPLY        COMPLETE     10/26/11 11:20:12
```

Tip: If the Fibre Channel connections to the controllers are lost, the SMclient software cannot be accessed for problem determination. Having an additional IP network connectivity enables you to continue to manage and troubleshoot the storage subsystem even if there are problems with the Fibre Channel link.

Installing SMutil

The SMutil package provides additional commands for listing and adding DS5000 Storage Server devices. This utility is also available for other operating systems. Follow these steps:

1. Install the SMutil by typing in the following command:

```
# installp -a -d /tmp/DSM/SMutil-AIX-10.00.65.16.bff SMutil.aix.rte
```

2. Verify the installation:

```
# lslpp -ah SMutil.aix.rte
```

The SMutil utility provides the following commands:

- SMdevices: Lists the configured DS5000 Storage Servers found.
- hot_add: Scans and configures DS5000 devices.

In-band: Starting and stopping the SMagent

If you plan to use in-band connectivity to your DS5000 Storage Server, make sure that the (SMagent) process is active. Use the following command to manually start and stop the SMagent:

```
# SMagent stop
# SMagent start
```

This command is issued from the workstation with an FC connection to the DS5000 Storage Server and on which you have installed the SMagent package.

The AIX system running the agent uses the Fibre Channel path to recognize the storage. It also requires an access LUN to be configured and mapped. By default, the storage has an access LUN mapped to the default host group in order to allow an initial management connection.

4.4.3 Performing the initial configuration on AIX hosts

Complete the installation by defining logical drives. For instructions on how to do this task, see 3.1.2, “Creating arrays and logical drives” on page 99. Logical drives, partitioning, and all other related tasks can be done also from the AIX Storage Manager client. The interface is similar on all operating systems currently supported (see Figure 4-47).

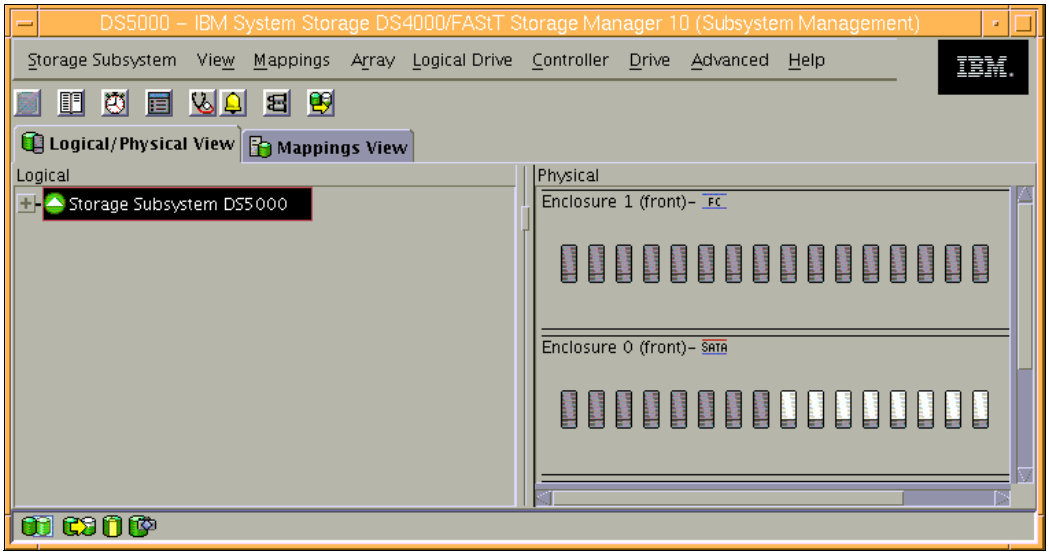


Figure 4-47 Storage Manager Client 10 for AIX

Follow the instructions in the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023 to create a host group, host, and host ports for the AIX server. After you set up storage partitioning, perform the following steps to verify that the host ports match the AIX host:

1. Type the following command:

   ```
   # lscfg -l fcs*
   ```

   A list is displayed containing all the HBAs in the system.

2. Use the second column of this list to determine the adapter slot location placement in your server when you need to cable the fibers:

   ```
   # lscfg -l fcs*
     fcs1             U0.1-P2-I5/Q1  FC Adapter
     fcs0             U0.1-P2-I4/Q1  FC Adapter
   ```

3. If you have multiple adapters, identify the fcs number of the HBAs to use with your DS5000 Storage Server.

   Reminder: Do not share the same HBAs for tape and disk traffic.

4. Check the adapter microcode by typing the following command for each of the HBAs used by the DS5000 Storage Server, as shown in Example 4-11.

   Example 4-11 Check HBA microcode level

   ```
   # lsmcode -d fcs0

   DISPLAY MICROCODE LEVEL
   802111
   fcs0    FC Adapter

   The current microcode level for fcs0 is 391101.
   ```

   Use the level seen in the output, together with the type or feature code of your HBA, to make sure that you have at least the minimum level listed at the IBM support website:

   http://www.ibm.com/systems/support/storage/config/ssic/

5. Type the following command to determine the adapter's WWN:

   ```
   # lscfg -vl fcsn |grep Network
   ```

   Where `fcsn` is the fcs number of the HBA that is connected to the DS5000 Storage Server.

   You will need this information later when mapping the logical drives, or configuring zoning if your DS5000 Storage Server is connected through a SAN. The network address number of the HBA is displayed, as shown in the following example:

   ```
   # lscfg -vl fcs0 |grep Network
           Network Address.............10000000C932A75D
   ```

6. Verify that the network address number matches the host port number that displays on the host partition table of SMclient.

7. Repeat this procedure to verify the other host ports.

8. Depending on the HBA, you need specific file sets, as shown in Table 4-4.

Table 4-4 FC HBA and AIX drivers

Feature code	Description	AIX driver
6228	2 Gb Fibre Channel Adapter PCI	devices.pci.df1000f9
6239	2 Gb Fibre Channel PCI-X	devices.pci.df1080f9
5716	2 Gb Fibre Channel PCI-X	devices.pci.df1000fa
5758	4 Gb Single-Port Fibre Channel PCI-X	devices.pci.df1000fd
5759	4 Gb Dual-Port Fibre Channel PCI-X	devices.pci.df1000fd
5773	4 Gb PCI Express Single Port	devices.pciex.df1000fe
5774	4 Gb PCI Express Dual Port	devices.pciex.df1000fe
5735	8 Gb PCI Express Dual Port	devices.pciex.df1000f114108a03

9. Always check for at least the minimum required levels for your adapter driver, as well as adapter microcode. You will find that information at the IBM support website.
10. Get your AIX installation CDs and install the foregoing file sets from the media.

Installing the MPIO driver

Important: RDAC under AIX is no longer supported with the DS5000 Storage Server and SDDPCM is being phased out. We strictly advise the use of MPIO.

The MPIO driver is part of the AIX operating system. Table 4-5 lists the current AIX release levels needed to support MPIO and the DS5000 Storage Server correctly. Although they are the current levels at the time of writing of this publication, when possible, always upgrade to the latest version available. Some of the previous versions of MPIO device driver might not fully support the enhancements of newer codes of DS5000. A fix pack is either a Service Pack or a Technology Level package. Use the **`oslevel -s`** command to determine the current level of your AIX operating system.

Table 4-5 Required AIX levels

AIX release	APAR number
53TL11 SP7	IZ99563
53TL12 SP4	IZ94521
61TL3 SP9	IZ93803
61TL4 SP9	IZ93742
61TL5 SP5	IZ93510
61TL6 SP5	IZ89402
71TL0 SP3	IZ89871

Ensure that the file set `devices.common.IBM.mpio.rte` is installed by issuing the command:

```
lslpp -ah devices.common.IBM.mpio.rte
```

Verification tasks

After the DS5000 storage system has been set up, the logical drives assigned to the host, and the MPIO driver installed, you must verify that all of your DS5000 Storage Server device names and paths are correct and that AIX recognizes your LUNs.

You must do it before you mount file systems and install applications. Type the following command to probe for the new devices:

```
# cfgmgr -v
```

Tip: You must run **cfgmgr** after the DS5000 Storage Server is cabled to allow the Storage Manager client to discover the AIX HBA WWN. Otherwise, you are not able to define the host port under the partitioning section.

Use the **lsdev -Cc disk** command to see whether the FCP array disk software recognizes each DS5000 Storage Server logical drive correctly. The following Example 4-12 illustrates the results of the command for a set of DS5300 storage server LUNs.

Example 4-12 lsdev command

```
# lsdev -Cc disk
hdisk2 Available 07-00-02 MPIO DS5100/5300 Disk
hdisk3 Available 07-00-02 MPIO DS5100/5300 Disk
```

Use the AIX **mpio_get_config** command to perform the following verification tasks:

- Correlate AIX `hdisk` numbers to the logical drive name displayed in the Storage Manager Client. The logical drive name from the Storage Manager Client is displayed under the User Label heading.
- Verify that MPIO driver correctly indicates the type of your storage subsystems (in our case DS5300). If "`MPIO Other FC SCSI Disk Drive`" is being displayed, consider the upgrade of your MPIO device driver to fully support used storage system.
- Make sure that the logical drives are on the preferred DS5000 Storage Server controller.
- Make sure that the correct number of storage controllers are discovered.
  - Controller count: 1 (indicates a single-controller configuration)
  - Controller count: 2 (indicates a dual-controller configuration)
- Make sure that the Partition count matches the number of storage partitions configured in the Storage Manager Client.

Example 4-13 shows an output of the **mpio_get_config -Av** command with a dual-controller DS5000 storage subsystem.

Example 4-13 mpio_get_config command

```
# mpio_get_config -Av
Frame id 0:
Storage Subsystem worldwide name: 60ab8006e1bbe00004d8093d6
    Controller count: 2
    Partition count: 1
    Partition 0:
    Storage Subsystem Name = 'DS5300'
        hdisk#           LUN #   Ownership          User Label
        hdisk2               0   B (preferred)      redbooks01_LUN3
```

There are several ways to correlate a system's configuration and monitor the state of DS5000 storage systems.

Example 4-14 shows that the two disks that we configured on the DS5000 Storage Server are in the available state. The third column shows the location code. In this example, it is 07-00-02. Each AIX system has its own set of location codes that describe the internal path of that device, including bus and host adapter locations. See the service manual for your system type to identify device locations.

Example 4-14 Show MPIO disks

```
# lsdev -C | grep -i mpio
hdisk2      Available 07-00-02    MPIO DS5100/5300 Disk
hdisk3      Available 07-00-02    MPIO DS5100/5300 Disk
```

Example 4-15 uses the **lsdev** command to show the status and location codes of two DS5300 storage server hdisks mapped and the local configured hdisk0. Notice that the location codes of the DS5300 storage server and the local disks are not the same.

Example 4-15 Disks status and location codes

```
# lsdev -Cc disk
hdisk0 Available 00-08-00       SAS Disk Drive
hdisk1 Available 00-08-00       SAS Disk Drive
hdisk2 Available 07-00-02       MPIO Other FC SCSI Disk Drive
hdisk3 Available 07-00-02       MPIO Other FC SCSI Disk Drive
```

Use the **lspath** command to verify that the correct number of paths are detected and that they are enabled. Example 4-16 shows the output for a dual-controller DS5300 storage server with a single-controller host configuration with two LUNs attached.

Example 4-16 Path detection

```
# lspath | sort
Enabled hdisk0 scsi0
Enabled hdisk2 fscsi0
Enabled hdisk2 fscsi0
Enabled hdisk3 fscsi0
Enabled hdisk3 fscsi0
```

Example 4-16 shows that all paths are enabled and each LUN on the DS5300 storage server has two paths.

The **lsattr** command provides detailed information about a disk drive, including information that allows you to map the system device name to the logical drive on the DS5000 storage system.

In Example 4-17, we run the **lsattr** command on the LUN named hdisk1. The command output provides the size information and LUN ID (0).

Example 4-17 Disk drive information

```
# lsattr -El hdisk1
PCM             PCM/friend/otherapdisk           Path Control Module              False
PR_key_value    none                             Persistant Reserve Key Value     True
algorithm       fail_over                        Algorithm                        True
clr_q           no                               Device CLEARS its Queue on error True
cntl_delay_time 0                                Controller Delay Time            True
cntl_hcheck_int 0                                Controller Health Check Interval True
dist_err_pcnt   0                                Distributed Error Percentage     True
dist_tw_width   50                               Distributed Error Sample Time    True
hcheck_cmd      inquiry                          Health Check Command             True
hcheck_interval 60                               Health Check Interval            True
hcheck_mode     nonactive                        Health Check Mode                True
location                                         Location Label                   True
lun_id          0x0                              Logical Unit Number ID           False
max_transfer    0x40000                          Maximum TRANSFER Size            True
node_name       0x200600a0b829eb78               FC Node Name                     False
pvid            0004362a0a2c1cd30000000000000000 Physical volume identifier       False
q_err           yes                              Use QERR bit                     True
q_type          simple                           Queuing TYPE                     True
queue_depth     10                               Queue DEPTH                      True
reassign_to     120                              REASSIGN time out value          True
reserve_policy  single_path                      Reserve Policy                   True
rw_timeout      30                               READ/WRITE time out value        True
scsi_id         0x6f1e00                         SCSI ID                          False
start_timeout   60                               START unit time out value        True
ww_name         0x204600a0b829eb78               FC World Wide Name               False
```

Changing ODM attribute settings in AIX

After a logical drive has been created and mapped to the host, devices can be configured and the Object Data Manager (ODM) is updated with the default parameters. In most cases and for most configurations, the default parameters are satisfactory. However, there are various parameters that can be modified for maximum performance and availability. See the *Installation and Support Guide for AIX, HP-UX, Solaris, and Linux*, which is provided with the Storage Manager software.

You can actually change the ODM attributes for the disk driver and DS5000 Storage Server; here we explain the basic functions of the **chdev** command.

As an example, we focus here on the `queue_depth` parameter. Setting the `queue_depth` attribute to the appropriate value is important for system performance. For large DS5000 Storage Server configurations with many logical drives and hosts attached, it is a critical setting for high availability. Reduce this number if the array is returning a `BUSY` status on a consistent basis. Possible values are from 1 to 64.

Use the following formula to determine the maximum queue depth for your system:

Maximum `queue depth` = DS5000 queue depth / (number-of-hosts * LUNs-per-host)

Where “DS5000 queue depth” = 4096.

For example, a DS5300 storage server with four hosts, each with 32 LUNs, has a maximum queue depth of 32:

Maximum queue depth = 4096 / (4 * 32) = 32

In this case, you can set the queue_depth attribute for `hdisk`*n* as follows:

```
#chdev -l hdiskn -a queue_depth=16 -P
```

We do not suggest to use the maximum calculated value. Rather, stay on the half value, unless advised by support. To make the attribute changes permanent, use the -P option to update the AIX ODM attribute.

Queue depth:

- Use the maximum queue depth as a guideline, and adjust the setting as necessary for your specific configuration.
- In systems with one or more SATA devices attached, you might need to set the queue depth attribute to a lower value than the maximum queue depth.

4.4.4 iSCSI configuration

In contrast with the Fibre Channel attached DS5000 storage, the iSCSI connection through the lossless 1 GbE Ethernet gives you an opportunity to utilize cheaper and easy-to-manage existing IP-based network. If your storage performance requirements are sufficient to use the capabilities of 1 GbE using iSCSI connection, you do not need to utilize more expensive, dedicated Fibre Channel components, such a HBAs, optical cables, expensive dual ports on SAN switches in separate fabrics, and so on.

The IBM DS5000 series now gives you a unique opportunity to attach your storage system to the 10 GbE Ethernet network using 10 Gbps Host Interface Cards. Unfortunately, at the time of writing of this publication, there are no 10 Gbps iSCSI adapter with TCP Offload Engine (TOE) available on the market for IBM Power systems. Using the software iSCSI initiator on 10 GbE without TOE brings the significant risk of saturated processor to manage TCP traffic related to iSCSI storage data transfer. Therefore, we always suggest to use TOE devices.

In our practical example, we provide readers the basic configuration steps to enable iSCSI storage attachment on your AIX 7.1 system using software initiator shipped with the basic AIX operating system, without hardware TCP Offload Engine. Differences to AIX version 6.1 are outlined, however the process is similar, with the same layout of smitty panels or syntax of the AIX shell commands.

Reminder: Do not connect the same storage systems using both protocols, iSCSI and FC.

Setup prerequisites

There are a few suggestions and mandatory steps that you need to complete to successfully establish the iSCSI connection to from your host systems (iSCSI initiators) to the DS5000 Storage subsystem (iSCSI target, in our case DS5300):

1. Make sure that you are implementing an iSCSI connection on switched Ethernet and DS5000 does not support direct-attached iSCSI Host Interface Cards to any host system.
2. Verify the correct configuration of you iSCSI targets on the DS5000 storage subsystem or ask your storage administrator to do so. The procedure for how to set up DS5000 iSCSI targets is explained in 3.1.6, “iSCSI configuration and management” on page 123.

3. Confirm the appropriate version of iSCSI device driver (software initiator) in your AIX system. Use the **lslpp** command to obtain the fileset **devices.iscsi_sw.rte** version. For reference, see Example 4-18, where we show the results on AIX 7.1 and 6.1.

Example 4-18 Determine the version of iSCSI driver

```
# lslpp -L |grep iscsi
  devices.common.IBM.iscsi.rte     7.1.0.15    C     F    Common iSCSI Files
  devices.iscsi.disk.rte           7.1.0.15    C     F    iSCSI Disk Software
  devices.iscsi.tape.rte           7.1.0.0     C     F    iSCSI Tape Software
  devices.iscsi_sw.rte             7.1.0.15    C     F    iSCSI Software Device Driver

# lslpp -L | grep iscsi
  devices.common.IBM.iscsi.rte     6.1.6.15    C     F    Common iSCSI Files
  devices.iscsi.disk.rte           6.1.6.15    C     F    iSCSI Disk Software
  devices.iscsi.tape.rte           6.1.0.0     C     F    iSCSI Tape Software
  devices.iscsi_sw.rte             6.1.6.15    C     F    iSCSI Software Device Driver
```

4. When using the hardware TCP Offload Engine (TOE), make sure it is correctly configured in your system and adapter firmware is up-to-date. Without TOE, make sure you have sufficient processor resources assigned from VIOS or physically available in a system.
5. Use two independent network connections to access both DS5000 controllers, utilizing two different network adapters (NIC) in your AIX. Refrain to use single NIC with dual ports, as it represents the Single Point of Failure if your dual-port NIC fails.
6. Enable jumbo frames on your network adapters and if necessary, do the same on LAN switch port. Consult this option with your network administrator and make sure that jumbo frames are enabled on DS5000 HICs as well. See Figure 4-48 as an example of setup of jumbo frames using HMC console of the AIX system. To do so, navigate to: **System Management** → **Hardware Information** → **Adapters** → **Host Ethernet**.

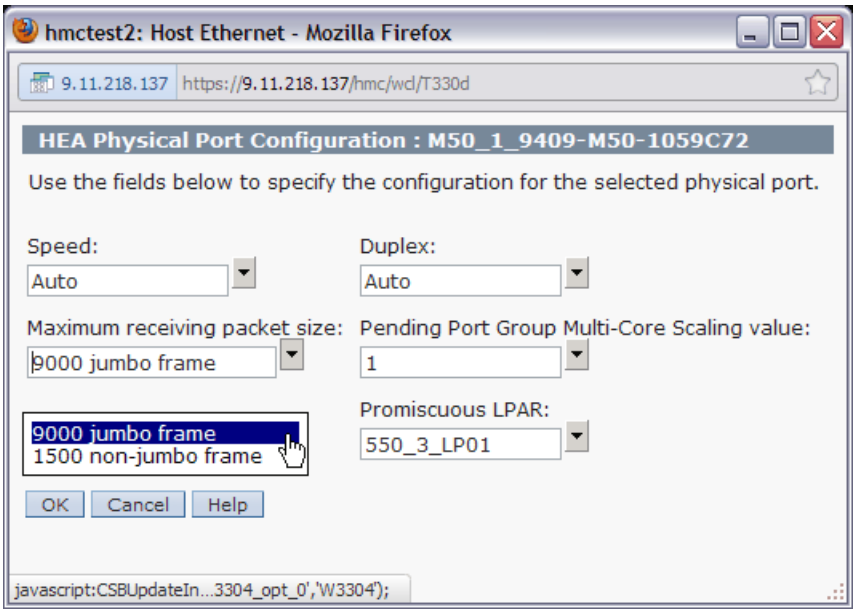


Figure 4-48 Configuration of jumbo frames using HMC console

Make the same settings on adapter in AIX of the associated LPAR using the command:

```
chdev -a mtu=9000 -l enx
```

Where *x* stands for the number of the adapter that you use for iSCSI connection.

7. Use server class, supported NICs. Strictly refrain from buying low-cost adapters from the nearest IT shop; they will probably not work with your AIX server, and if they do work, they might not perform to your expectations. Segregate the network traffic and SAN iSCSI traffic. Use dedicated adapters and LAN switches or Virtual LAN (VLAN).
8. Discover and record all your environment variables that you need to establish iSCSI connection to the DS5000. They include DS5000 controller IP addresses, target aliases, specific Ethernet adapters in your AIX and their IP addresses, optional CHAP authentication password given from DS5000 and network details (subnet mask, gateway). For our test scenario, we used the following variables:
   - 192.168.130.101 - iSCSI DS5300 Controller A
   - 192.168.130.102 - iSCSI DS5300 Controller B
   - 192.168.130.148 - iSCSI AIX Ethernet adapter en4
   - 192.168.130.149 - iSCSI AIX Ethernet adapter en5
   - 255.255.255.0 - iSCSI network subnet mask
   - 192.168.130.1 - iSCSI network gateway (not required, as we have internal iSCSI LAN)
   - iqn.1992-01.com.lsi - DS5300 target IQN
   - iqn.us.ibm.com.aixitso - AIX LPAR1 host IQN
   - DS5300_redbooks - iSCSI DS5300 target alias
   - AIX_LPAR1 - Host identifier as defined on DS5300 for storage allocation
   - iSCSI_AIX_LPAR1 - Host port identifier as defined on DS5300

Configuration procedure

After we have met all the prerequisites, the setup procedures includes the steps described in this section. In our lab, we used redundant connections over 1 GbE Ethernet, but using single LAN switch. This enablement shows the single point of failure, in the event of switch failure, we will loose both paths to the storage disks. In production we highly advise to use fully redundant, dual paths, including two LAN switches.

The configuration itself is simple and straight forward procedure, which includes the following steps:

1. Tune the network for large jumbo packets on adapters en4 and en5 as shown in Example 4-19.

Example 4-19 Enable jumbo packets

```
# ifconfig en4 down
# ifconfig en4 detach

# chdev –a ent4 –a jumbo_frames=yes
# chdev –l en4 –a tcp_recvspace=262144
# chdev –l en4 –a tcp_sendspace=262144
# chdev –l en4 –a rfc1323=1

# ifconfig en4 up
```

2. To configure the iscsi4 and iscsi5 interfaces as a software initiator, a unique network interface protocol for iSCSI needs to be created. Using SMIT, enter **`smitty chgiscsisw`** and specify “iSCSI Initiator Name”, for example: `iqn.us.ibm.com.aixitso`. Keep the discovery policy “File”. The “Maximum Targets Allowed” field corresponds to the maximum number of iSCSI targets that can be configured. If you reduce this number, you also reduce the amount of network memory pre-allocated for the iSCSI protocol driver during configuration. We changed this number from default 16 to 2 as we have only one target device with two storage controllers.

3. There are required entries into the /etc/iscsi/targets file. In the /etc/iscsi directory, there is a targets file that requires entries for each of the iSCSI targets ports, as outlined in your DS5000 Storage Manager. The entries need to be created in the format shown in Example 4-20. There are also examples in the targets file for levels of access and authentication. Each uncommented line in the file represents an iSCSI target.

Example 4-20 iSCSI targets configuration file in AIX host

```
########################################################################
# EXAMPLE 1: iSCSI Target without CHAP(MD5) authentication
#     The DS5300 target is at address 192.168.103.101 and 192.168.103.102,
#     the valid port is 3260,(default for DS5300)
#     the target name is iqn.1992-01.com.lsi
# The target line would look like:
# 192.168.130.101 3260 iqn.1992-01.com.lsi
   192.168.130.101 3260 iqn.1992-01.com.lsi
   192.168.130.102 3260 iqn.1992-01.com.lsi
#
# EXAMPLE 2: iSCSI Target with CHAP(MD5) authentication
#     The DS5300 target is at address 192.168.103.101 and 192.168.103.102,
#     the valid port is 3260
#     the name of the target is iqn.1992-01.com.lsi
#     the CHAP secret is "This is my password."
# The target line would look like:
# 192.168.103.101 3260 iqn.1992-01.com.lsi "This is my password."
# 192.168.103.102 3260 iqn.1992-01.com.lsi "This is my password."
#
# EXAMPLE 3: iSCSI Target with CHAP(MD5) authentication and line continuation
#     The DS5300 target is at address 192.168.103.101 and 192.168.103.102,
#     the valid port is 3260
#     the name of the target is iqn.1992-01.com.lsi
#     the CHAP secret is "123ismysecretpassword.fc1b.for.itso"
# The target line would look like:
# 192.168.103.101 3260 iqn.1992-01.com.lsi \
#               "123ismysecretpassword.fc1b.for.itso"
# 192.168.103.102 3260 iqn.1992-01.com.lsi \
#               "123ismysecretpassword.fc1b.for.itso"
########################################################################
```

4. After the changes are done in the targets file, run **cfgmgr** to make them effective:

   ```
   cfgmgr -vl iscsi0
   ```

   This command reconfigures the software initiator driver to attempt to communicate with the targets listed in the /etc/iscsi/targets file, and to define a new hdisk for each LUN on the targets that are found. Since then, identify newly established iSCSI session on your DS5300 Storage Manager under menu **Storage Manager** → **iSCSI** → **End Sessions**.

5. Confirm that newly discovered volumes are available in your AIX system using the command **lsdev** as shown in Example 4-21.

Example 4-21 Verification of available iSCSI attached disks

```
# lsdev -Cc disk
hdisk1 Available Other iSCSI Disk Drive
hdisk2 Available Other iSCSI Disk Drive
```

6. Finally, show the characteristics of the disk with the command **lsattr**; see Example 4-22.

Example 4-22 Attributes of the iSCSI disk

```
lsattr -El hdisk1
clr_q         no                                  Device CLEARS its Queue on error True
host_addr     192.168.130.148                     Hostname or IP Address           False
location                                          Location Label                   True
lun_id        0x0                                 Logical Unit Number ID           False
max_transfer  0x40000                             Maximum TRANSFER Size            True
port_num      0xcbc                               PORT Number                      False
pvid          00f602736c1a2ad80000000000000000    Physical volume identifier       False
q_err         yes                                 Use QERR bit                     True
q_type        simple                              Queuing TYPE                     True
queue_depth   1                                   Queue DEPTH                      True
reassign_to   120                                 REASSIGN time out value          True
rw_timeout    30                                  READ/WRITE time out value        True
start_timeout 60                                  START unit time out value        True
target_name   iqn.1992-01.com.lsi                 Target NAME                      False
```

4.4.5 AIX restrictions

When connecting the DS5000 Storage Server to AIX systems, there are various restrictions and guidelines that you need to take into account. This statement does not mean that the other configurations will not work, but you might end up with unstable or unpredictable results that are hard to manage and troubleshoot. All the references to an AIX host can be used for a stand-alone system, an LPAR, or a VIOS.

SAN and connectivity restrictions

The following restrictions apply:

- AIX hosts or LPARS can support multiple host bus adapters (HBAs) and DS5000 Storage Server devices. However, you can only connect up to two HBAs to any DS5000 Storage Server partition, and up to two partitions. Additional HBAs can be added for additional DS5000 Storage Servers and other SAN devices, up to the limits of your specific server platform.
- Storage Manager V10.30 and higher versions allow the creation of logical drives greater than 2 TB. However, when selecting your boot disk, remember that the AIX boot logical drive must reside within the first 2 TB.
- Single-switch configurations are allowed, but each HBA and DS5000 Storage Server controller combination must be in a separate SAN zone.
- Other storage devices, such as tape devices or other disk storage, must be connected through separate HBAs and SAN zones.
- Even if 10 Gbps Host Interface Cards (HIC) are available for DS5000, at the time of writing only 1 Gbps TCP Offload Engine iSCSI card is on the market for Power systems. Avoid using software iSCSI initiator on 10 GbE without TOE, as it might significantly saturate your processor resources.
- Single HBA configurations are allowed, but each single HBA configuration requires that both controllers in the DS5000 Storage Server be connected to the HBA through a switch with both controllers zoned to the HBA, as shown in Figure 4-49.

Important: Having a single HBA configuration can lead to a loss of data in the event of a path failure.

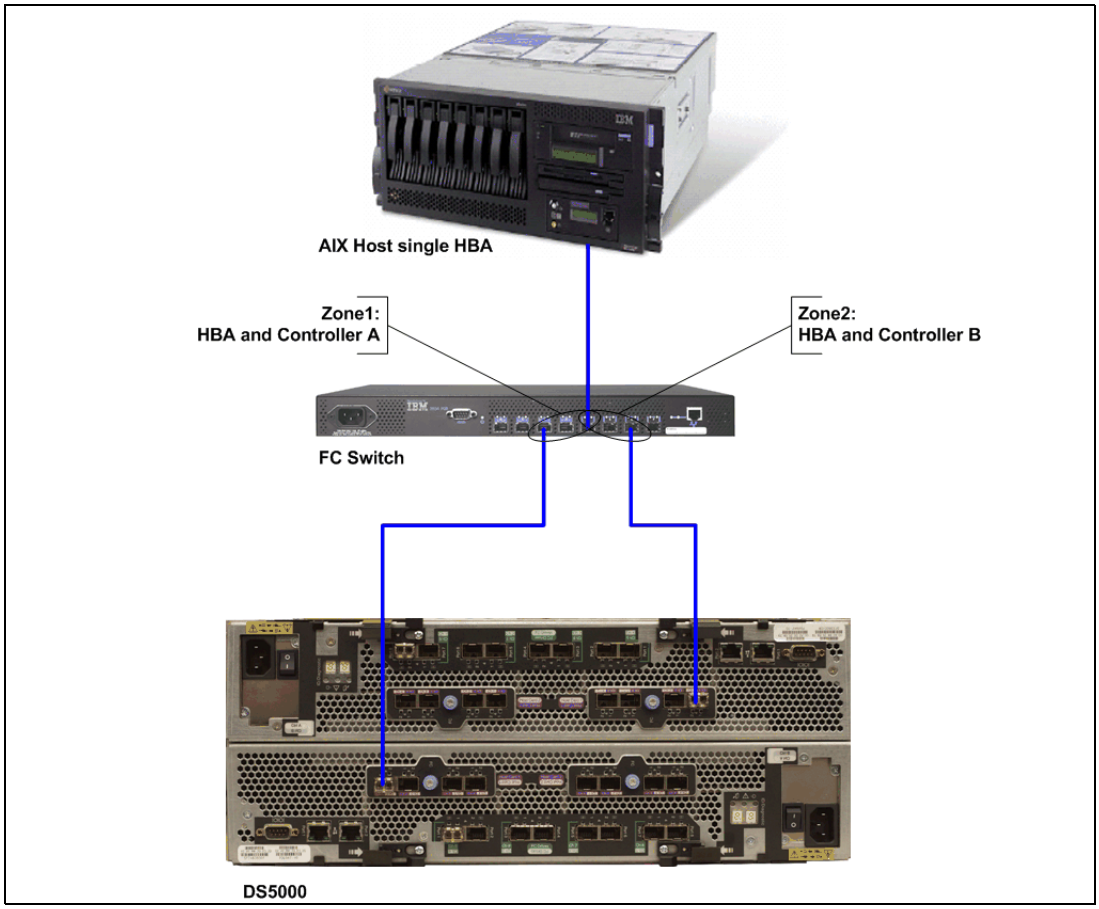


Figure 4-49 Single HBA configuration in an AIX environment

Limitations when booting your system from SAN

The following restrictions apply:

- If you create more than 32 LUNs on a partition, you cannot use the AIX 5L V5.1 release CD to install AIX on a DS5000 Storage Server device on that partition.
- When you boot your system from a DS5000 Storage Server device (SAN boot), both paths to the controller must be up and running.
- The system cannot use path failover during the AIX boot process. After the AIX host has started, failover operates normally.
- If you have FC and SATA drives in your DS5000 Storage Server, do not use SATA drives as boot devices.
- An iSCSI boot is supported on Power Blade systems using either the iSCSI Software Initiator or the iSCSI TOE daughter card using IPv4. IPv6 protocol is not yet supported.

Partitioning restrictions

The following restrictions apply:

- The maximum number of partitions per AIX host, per DS5000 Storage Server, is two.
- All logical drives that are configured for AIX must be mapped to an AIX host group. If using a default host group, change your default host type to AIX.
- On each controller, you must configure at least one LUN with an ID between 0 and 31 that is not an utm or access logical drive.

Interoperability restrictions

The following restrictions apply:

- Concurrent download is not supported with certain controller firmware versions.
- See the latest Storage Manager `readme` file for AIX to learn which firmware versions support concurrent download.

Dynamic volume expansion (DVE) is supported in AIX 5L Version 5.2 and later.

For further information regarding AIX restrictions on partitioning and SAN connectivity, see Chapter 12, “DS5000 with AIX, PowerVM, and PowerHA” on page 541.

4.5 Linux

In the following section, we present the steps that you need to perform in your Linux host to install and manage your DS Storage System from the Linux operating system.

We assume that you have already configured the DS Storage Manager, with the logical volumes already mapped to the Linux host. In this example, we cover Linux Red Hat 6, kernel version 2.6.32-71.el6. We also cover specific details for iSCSi implementation.

You will perform the following basic steps in this sequence:

1. Install the host bus adapter hardware and drivers to attach your DS Storage System.
2. Install the IBM DS Storage System Linux RDAC driver by using the instructions provided in the `readme` file located in the LinuxRDAC directory on the installation CD or downloaded with the device driver source package from the IBM support website.
3. If there is a previous version 7.x, 8.x or 9.xx of the IBM DS Storage Manager host software (that is, SMRuntime, SMclient, RDAC, SMUtil, and SMAgent packages) installed in the server, uninstall the previous version first before installing the new version of the Storage Manager host software.
4. Install the new Storage Manager host software version from the CD-ROM directory or from the host software package that you downloaded from the IBM Support website.

4.5.1 Installing DS Storage Manager software

In this section, we describe the installation of the host software on a System x server running a Linux distribution with kernel v2.6. There are versions of the components available for 32-bit and 64-bit architectures. In our example, we used Red Hat Enterprise Linux 6 32-bit - Kernel 2.6.32-71.el6.i686.

The DS Storage Manager software needed to install and manage your system from a Linux host is available for download at the following website:

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

In the website, select your DS Storage System product, and then click the **Download** option.

Select the Storage Manager software for your Linux distribution and also server type, because there are four separate IBM DS Storage Manager host software version 10.77 packages for Linux operating system environments:

- Linux x86 Only.
- Linux x86_64 Only.
- Intel Itanium - IA64.
- Linux on Power (LoP).

Make sure to check the readme file, because earlier or later versions of Linux kernels might not have been tested with this DS Storage Manager version, so you need to upgrade your system to a specific kernel version according to the `readme` file contained in the Storage MAnager package. Updated kernel packages are available for download for the various distributions.

For the Linux Red Hat updates, see the following websites:

`http://www.redhat.com`
`http://updates.redhat.com`

Also check the `readme` file for firmware prerequisites that might exist for your Storage System.

The DS Storage Manager software for Linux operating systems is available as a single package that you can install as individual packages based on your requirements.

Installing DS Storage Manager packages

Complete the following steps to install the IBM DS Storage Manager software. Adjust the steps as necessary for your specific installation:

1. Ensure that you have root privileges, which are required to install the software.
2. Get the DS Storage Manager software installation package file from the DS Storage Manager CD, or download it from the IBM Support website to an SM directory on your system.

   Tip: Always verify the latest version of the DS Storage Manager host software packages available for download on the IBM Support website.

3. Change into the directory where the SM software was transferred, and uncompress the installation package as shown in the following example:

   ```
   # tar -xvzf SM10.77_Linux_32bit_x86_single-10.77.x5.16.tgz
   ```

   In our case, this generates the following directory and file:

   ```
   /Linux_32bit_x86_10p77_single/Linux
   ```

4. Start the Storage Manager installation wizard, changing to the directory where the file was uncompressed, and execute the commands as shown in Example 4-24 on page 212.

Tip: SMruntime and SMesm needs to be installed prior SMclient package.

Important: Before installing the client software, install the runtime software. A system reboot is only required when installing the RDAC driver package. The Event Monitor software is installed automatically during the client software installation.

Installing the Storage Manager Runtime: SMruntime

Installation of the SMruntime package is necessary for both hosts and storage management stations. To install SMruntime, follow these steps:

1. Type the following command at the command prompt and press Enter:

   ```
   rpm -ivh SMruntime-LINUX<version number>.rpm
   ```

2. To verify the installation of the SMruntime package, type the following command at the command prompt and press Enter:

   ```
   rpm -qa SMruntime
   ```

3. If the installation was successful, and no failure was reported, continue with the next section. If the installation was unsuccessful, repeat steps 1 and 2. If the failure persists, see the Storage Manager Release Notes or contact your IBM technical support representative. See Example 4-23.

Example 4-23 Installing SMruntime .rpm package

```
[root@redbooks03 Linux]# rpm -ivh SMruntime-LINUX-10.77.A5.03-1.i586.rpm
Preparing...                ########################################### [100%]
   1:SMruntime ########################################### [100%]
```

Installing Storage Manager ESM: SMesm

Installation of the SMesm is required for the possibility of Environmental Service Module (ESM) firmware upgrades. To install the SMesm package, follow these steps:

1. Type the following command at the system prompt and press Enter.

   ```
   rpm -ivh SMesm-LINUX<version number>.rpm
   ```

2. To verify installation of the SMagent package, type the following command at the system prompt and press Enter:

   ```
   rpm -qa SMesm
   ```

3. If the installation was unsuccessful, repeat steps 1 and 2. If the failure persists, see the Storage Manager Release Notes or contact your IBM technical support representative.

The result is that the DS4000 Storage Manager software packages are installed in the `/opt/IBM_DS4000` directory.

The various launch scripts are located in the `/opt/IBM_DS4000` directory and respective sub-directories.

Installing the Storage Manager Client: SMclient

Installation of the SMclient package is necessary only if the system is a storage management station or if the system is a host acting as a storage management station. To install the SMclient (see Example 4-24), follow these steps:

1. Type the following command at the command prompt and press Enter:

   ```
   rpm -ivh SMclient-LINUX<version number>.rpm
   ```

2. To verify the installation of the SMclient package, type the following command at the command prompt and press Enter:

   ```
   rpm -qa SMclient
   ```

3. If the installation was unsuccessful, repeat steps 1 and 2. If the failure persists, see the Storage Manager Release Notes or contact your IBM technical support representative.

Example 4-24 Installing SMclient .rpm package

```
[root@redbooks03 Linux]# rpm -ivh SMclient-LINUX-10.77.A5.03-1.i586.rpm
Preparing...                ########################################### [100%]
   1:SMclient ########################################### [100%]
```

4. Launch the DS Storage Manager program running the following command as “root” :

   ```
   SMclient
   ```

The Storage Manager installation wizard’s introduction window opens. See Figure 4-50. Select the appropriate options and proceed with the installation.

Figure 4-50 DS Storage Manager load screen

Disabling and enabling the Event Monitor

The Event Monitor, which comes with the client software, installs automatically during the client software installation. The Event Monitor handles storage subsystem error notification through email or SNMP traps when the storage management software is not actively running on the management station or host computer.

You can disable and enable the Event Monitor while the Event Monitor is running, or you can disable and enable the boot-time reloading of the Event Monitor. If you disable the Event Monitor while it is running, it starts automatically at the next reboot.

Important: The Event Monitor must remain enabled if you intend to use automatic Environmental Service Modules (ESM) firmware synchronization.

If you installed the client software and configured alert notifications on multiple hosts, you might receive duplicate error messages from the same storage subsystem. To avoid receiving duplicate error messages, disable the Event Monitor on all but one system. You must run the Event Monitor on one host (server) that will run continually.

To disable the Event Monitor while the software is running, at the command prompt, type the following command and press Enter:

```
SMmonitor stop
```

When the program shutdown is complete, the system displays the following message, where *xxxx* represents the process ID number:

```
Stopping SMmonitor process <xxxx>
```

To enable the Event Monitor while the software is running, at the command prompt, type the following command and press Enter:

```
SMmonitor start
```

When the program startup begins, the system displays the following message:

```
SMmonitor started.
```

To disable boot-time loading of the Event Monitor, at the command prompt, type the following command and press Enter:

```
mv /etc/rc2.d/S99SMmonitor /etc/rc2.d/disabledS99SMmonitor
```

You are returned to the command prompt.

To enable boot-time loading of the Event Monitor, at the command prompt, type the following command and press Enter:

```
mv /etc/rc2.d/S99SMmonitor /etc/rc2.d/S99SMmonitor
```

You are returned to the command prompt.

Installing the Storage Manager Agent: SMagent

Installation of the Storage Manager Agent software is necessary only if the system is a host and the storage array is to be managed using the in-band storage array management method. In addition, RDAC must be installed. See Example 4-25. Follow these steps:

1. To install the SMagent package, type the following command at the system prompt and press Enter:

   ```
   rpm -ivh SMagent-LINUX<version number>.rpm
   ```

2. To verify installation of the SMagent package, type the following command at the system prompt and press Enter:

   ```
   rpm -qa SMagent
   ```

Example 4-25 Installing SMagent .rpm package

```
[root@redbooks03 Linux]# rpm -ivh SMagent-LINUX-10.77.G5.03-1.noarch.rpm
Preparing...                ########################################### [100%]
   1:SMagent ########################################### [100%]
```

If the installation was unsuccessful, repeat steps 1 and 2. If the failure persists, see the Storage Manager Release Notes or contact your IBM technical support representative.

Installing the Storage Manager Utility: SMutil

Installing the SMutil package is necessary only if the system is a host. See Example 4-26. Follow these steps:

1. To install the SMutil, type the following command at the system prompt and press Enter:

   ```
   rpm -ivh SMutil-LINUX<version number>.rpm
   ```

2. To verify installation of the SMutil package, type the following command at the system prompt and press Enter:

   ```
   rpm -qa SMutil
   ```

Example 4-26 Installing SMutil .rpm package

```
[root@redbooks03 Linux]# rpm -ivh SMutil-LINUX-10.77.G5.14-1.noarch.rpm
Preparing...                ########################################### [100%]
   1:SMutil ########################################### [100%]
SMmonitor started.
```

If the installation was successful, and no failure was reported, continue with the following section. If the installation was unsuccessful, repeat steps 1 on page 214 and 2 on page 214. If the failure persists, see the Storage Manager Release Notes or contact your IBM technical support representative.

Updating the host software

To update the host software in a Linux environment, follow these steps:

1. Uninstall the Storage Manager components in the following order:
   a. DS4000/5000 Runtime environment
   b. SMutil
   c. RDAC
   d. SMclient

   Verify that IBM host adapter device driver versions are current. If they are not current, see the `readme` file located with the device driver and then upgrade the device drivers.
2. Install the Storage Manager components in the following order:
   a. SMagent
   b. DS Runtime environment\
   c. RDAC
   d. SMutil
   e. SMclient

4.5.2 Installing the host bus adapter drivers

The device driver enables the operating system to communicate with the host bus adapter.

In order for your Linux distribution to support the FC or iSCSI adapter, you must ensure that you have the correct drivers installed. To download the Linux drivers for both FC and iSCSI host bus adapters, go to the following website:

`Http://www-1.ibm.com/servers/storage/support/disk`

In the website, select your DS Storage System product. Then click the **Download** option to find all the available packages for download for the various HBAs supported for connecting your DS Storage System. Select the package for your specific model. If you are not sure what your adapter model is, you can use your adapter Management software to obtain the information, similar to the SANsurfer shown in “Installing the QLogic SANsurfer” on page 216. You can also query the events in your Linux host for the adapter type as shown in Example 4-27.

Example 4-27 Determining HBA type installed

```
[root@redbooks03 ~]# dmesg | grep -i emulex
[root@redbooks03 ~]# dmesg | grep -i qlogic
QLogic Fibre Channel HBA Driver: 8.03.01.05.06.0-k8
QLogic Fibre Channel HBA Driver: 8.03.01.05.06.0-k8
QLogic QLA2340 - 133MHz PCI-X to 2Gb FC, Single Channel
QLogic Fibre Channel HBA Driver: 8.03.01.05.06.0-k8
QLogic QLA2340 - 133MHz PCI-X to 2Gb FC, Single Channel
```

In Example 4-27, there are two QLogic QLA2340 adapters installed. The device drive level is also displayed.

Depending on your adapter model, the driver might be already packaged with your Linux distribution. The `readme` file mentions that as an “in-distro” driver, because they are packed together with the Linux distribution. If the driver is indicated as “standard” in the `readme` file for your adapter, then you need to download and install the appropriate driver separately.

In this example, for FC attachment we used QLogic FC cards with Red Hat 6, which already has included the adapter driver of the QLogic Fibre Channel Host Bus Adapters, so there is no need to install it. If you are using iSCSI QLogic HBAs, install the specific package for your adapter and Linux distribution from the previous website.

Tip: For redundancy, use multiple HBAs to access your DS Storage System. Only like-type adapters can be configured as a failover pair.

Before installing the driver, make sure that your system meets the requirements stated in the `readme` file of the driver package regarding kernel versions, host adapter settings, and so on. You might need to install the kernel headers and kernel source for the supported version of kernel before you install the drivers. You can download them from the following website:

`http://updates.redhat.com`

Set your adapter settings according to the `readme` file specifications. For our example, we set all adapter settings, except for the following ones, which maintain the IBM defaults:

- Host Adapter settings:

  Loop reset delay - 8

- Advanced Adapter Settings:

  LUNs per target - 0
  Enable Target Reset - Yes
  Port down retry count - 12

To change your adapter settings and update the firmware or BIOS of your adapter, you can use the management software provided by your HBA manufacturer. In our case we used QLogic SANsurfer.

Installing the QLogic SANsurfer

The QLogic SANsurfer component is the graphical interface to manage your QLogic adapters. You can use this application to determine your adapter type; check its status; determine the firmware, BIOS, and driver levels; set specific parameters; and display the LUNs available. Here we cover the installation on a Linux host.

Important: When installing SANsurfer, do not enable FC HBA adapter failover, because we will be defining our own preferred multipath driver, like RDAC or DM-Multipath.

The QLogic SANsurfer contains these components:

- QLogic SANsurfer GUI: This software is a Java-based GUI application for the management of HBAs on host servers.
- Linux QLogic SANsurfer Agent (or qlremote agent): This software is required on servers where HBAs reside.

You can install either from the GUI or from the command line, as explained in the following two sections. If you do not have a graphical console in your Linux system, then you can still install just the agent, and run the GUI from a separate Management station.

Tip: Starting with Version 2.0.30b62, IBM no longer releases a specialized IBM version of the QLogic SANsurfer HBA Manager or SANsurfer PRO program. You can use the QLogic version 5.0.X buildX that supports Fibre Channel (FC) HBAs and iSCSI HBAs.

GUI installation

The name of the installation file indicates the SANsurfer version. At the time of writing this book, the file package name is `standalone_sansurfer5.0.1b34_linux_install.tgz` for the FC version. When working with iSCSi, select the appropriate package during the installation stage.

Follow these steps for the installation:

1. Change to the directory where you downloaded the file, and uncompress it using the following command:

   ```
   tar -xvzf standalone_sansurfer5.0.1b34_linux_install.bin.tgz
   ```

2. Change the file attributes to be an executable

   ```
   chmod a+x standalone_sansurfer5.0.1b34_linux_install.bin
   ```

3. Execute a change to the directory where the installation file is located and run the command:

   ```
   sh ./standalone_sansurfer5.0.1b34_linux_install.bin
   ```

   This initiates the installation process.

4. Confirm the first windows by clicking **Next**, and select the components to install. Continue the installation process. *Do not* select to Enable QLogic Failover Configuration, because we are planning to install RDAC as the multipath driver. See Figure 4-51.

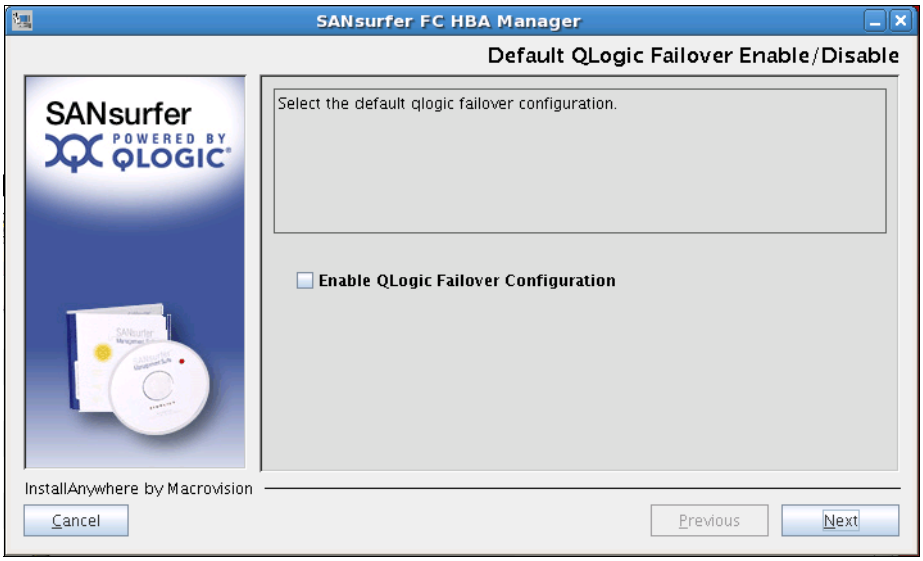


Figure 4-51 Linux SANsurfer install

Command line installation

In the following example, we show how to install the QLogic combined version from a terminal command line. Follow these steps:

1. Open a shell and change to the directory that contains the “sansurfer” installation file.
2. If you want to install Linux GUI Agent, type the following command:

   ```
   sh ./standalone_sansurfer5.0.1b34_linux_install.bin -i silent
   -DSILENT_INSTALL_SET="QMSJ_G_LA"
   ```

3. If you want to install the Linux GUI, use the following command:

```
sh ./standalone_sansurfer5.0.1b34_linux_install.bin -i silent
-DSILENT_INSTALL_SET="QMSJ_G"
```

4. If you want to install the Linux Agent, use the following command:

```
sh ./standalone_sansurfer5.0.1b34_linux_install.bin -i silent
-DSILENT_INSTALL_SET="QMSJ_LA"
```

For more details on QLogic SANsurfer and other HBA Management tools, see the “Advanced Maintenance” section of the *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

4.5.3 Installing the Linux multipath driver

The MPP Linux RDAC and DM-Multipath are the current multipath supported drivers for attaching your DS Storage System to your Linux host.

The Linux RDAC provides redundant failover/failback support for the logical drives in the DS5000 storage subsystems that are mapped to the Linux host server. The Linux host server must have host connections to the ports of both controllers A and B of the DS Storage subsystem. It is provided as a redundant path failover/failback driver alternative to the Linux FC host bus adapter failover device driver.

In addition, RDAC provides functions and utilities to get much better information about the DS Storage System from the operating system. After the RDAC driver is installed, you have access to several commands, useful for problem determination and fixing. We cover these utilities later.

Tip: In Linux, multipath device drivers like RDAC or DM-Multipath are not installed together with the Storage Manager software. You must download and install it separately.

The RDAC driver for Linux is not part of the Storage Manager software. You can download the RDAC driver from either of the following websites:

`http://www.ibm.com/servers/storage/support/disk`

Device Mapper Multipathing (DM-Multipath) is the native Linux multipath I/O support that has been added to the Linux 2.6 kernel tree with the release of 2.6.13. We cover this driver later in this chapter.

LSI RDAC Multipath Driver

At the time of writing, there are three versions of the Linux RDAC driver:

- Version 09.03.0C05.0504. For use in 2.6 Kernel, SLES 10-SP3, SLES10-SP4,SLES 11.1, RHEL5-u5, RHEL5-u6, RHEL 6, SLES10-SP4 operating system environments.
- Version 09.03.0B05.0439. For use in Linux 2.6 kernel RHEL4-u8 operating system environment.
- Version 09.00.A5.22. For the use is Linux 2.4 kernel, Red Hat EL 3-u8 (x86, IA64 and x86_64 versions and SuSe SLES 8: 2.4.21-295 (x86 and IA64 versions only)

Because in our example, we cover Red Hat 6, we used the first version. Just as with the HBA drivers, make sure that your system meets the requirements stated in the `readme` file of the driver package regarding kernel versions, HBA settings, and so on. Also pay attention to the installation sequence

Tip: Before you install RDAC, make sure that the HBA driver is installed, the partitions and LUNs are configured and mapped to a Linux host type, and script disable AVT was executed.

At the time of writing, the following Linux RDAC restrictions apply. You can find more information about them in the readme file:

- The RDAC driver cannot co-exist with an HBA-level failover driver such as the Emulex, QLogic or LSI Logic HBA failover drivers.
- Auto Logical Drive Transfer (ADT/AVT) mode is not supported.
- The Linux SCSI layer does not support *sparse/skipped* LUNs, which means that mapped LUNs must be sequentially numbered for the Linux host to see them.
- When using multiple FC HBAs on the host server and if each HBA port sees both controllers (by an un-zoned switch), the Linux RDAC driver might return I/O errors during controller failover.

  Tip: The guideline here is for each HBA to have a path to both the controllers in the DS Storage Server, either using multiple SAN Switches or using one SAN Switch with appropriate zoning. The intent here is to not have multiple initiators in the same zone.

- The Linux RDAC reports I/O failures immediately if it detects failure of all paths to the storage server. This behavior is unlike that of the HBA failover driver, which delayed for a certain period of time.
- The DS Storage Server must be connected to the HBAs in order for the virtual HBA driver to be loaded.

  Tip: Save a copy of the existing initial ramdisk before installing the RDAC driver. This backup copy, together with a separate stanza in /etc/lilo.conf or /boot/grub/menu.lst, is needed in case the system does not boot after installing the RDAC device driver.

Next we cover how to install the RDAC driver. If you are planning to use your Linux system to configure your DS Storage System, then skip to the next section to install the DS Storage Manager first, and then configure and map the logical drives planned to the Linux host, in order to meet the previous guidelines.

To install the Linux RDAC driver, proceed as follows (for the latest details, always check the readme file that comes with the driver package):

1. Install the kernel-source RPM files for the supported kernel. These files are available from your Linux vendor or your OS installation CD set. Also, be sure that the ncurses packages and the gcc packages have been installed:

```
# rpm -iv kernel-source*.rpm (SLES9/10/11)
# rpm -iv kernel-syms*.rpm (SLES9/10/11)
# rpm -iv kernel-devel*.rpm (RH)
# rpm -iv kernel-utils*.rpm
# rpm -iv ncurses*.rpm
# rpm -iv ncurses-devel*.rpm
# rpm -iv gcc*.rpm
# rpm -iv libgcc*.rpm
```

Query the installed packages with the following commands:

```
# rpm -qa | grep kernel
# rpm -qa | grep ncurses
# rpm -qa | grep gcc
```

2. This Linux RDAC release does not support auto-volume transfer/auto-disk transfer (AVT/ADT) mode. AVT/ADT is automatically enabled in the Linux storage partitioning host type. Disable it by using the script that is bundled in the IBM Linux RDAC Web package or in the \Scripts directory of the DSStorage Manager Version Linux CD. The name of the script file is DisableAVT_Linux.scr.

   Use the following steps to disable the AVT/ADT mode in your Linux host type partition. To change it:

   a. Start the SMclient on your management station.

   b. From the Enterprise Management window, highlight the storage system that is used with the Linux host. Select **Tools** → **Execute Script**.

   c. In the script editing window, select **File** → **Load Script**. Select the full path name for the script (<CDROM>/scripts/DisableAVT_Linux.scr).

   d. Select **Tools** → **Verify and Execute**.

      This script resets the controllers individually.

      We assume that the Linux host type is selected for the storage partition in which the Linux host server HBA port is defined. This Linux host type has AVT/ADT enabled as the default.

      Be sure that all connected hosts have redundant FC paths to the DS5000 subsystem, as each controller is rebooted, one at a time, and the FC paths swap temporarily.

3. The host server must have the non-failover Fibre Channel HBA device driver properly built and installed before the Linux RDAC driver installation.

4. Change to the directory when the RDAC package was downloaded, and type the following command to uncompress the file:

   ```
   # tar -xvzf rdac-LINUX-xx.xx.xx.xx-source.tar.gz
   ```

   Here, *xx. xx.xx.xx* is the release version of the RDAC driver.

   In our case, we used the command:

   ```
   # tar -xvzf rdac-LINUX-09.03.0C05.0214-source.tar.gz
   ```

5. Change to the directory where the files are extracted.

   In our case, we used the command:

   ```
   # cd /RDAC/linuxrdac-09.03.0c05.0214
   ```

6. Remove the old driver modules in case of previous RDAC installations. Type the following command and press Enter:

   ```
   # make clean
   ```

7. To compile all driver modules and utilities in a multiple-CPU server (SMP kernel), type the following command and press Enter:

   ```
   # make
   ```

8. To install the driver, type the following command:

   ```
   # make install
   ```

   This command:

   – Copies the driver modules to the kernel module tree

– Builds the new RAMdisk image (mpp-`uname -r`.img), which includes the RDAC driver modules and all driver modules that are needed at boot

9. A new line will be automatically added to the /boot/grub/menu.lst file with the new MPP drivers as in Example 4-28. Just for better understanding, edit the file, adding “with MPP Support” at the end of the new line title.

Example 4-28 Edit /boot/grub/menu.lst

```
title Red Hat Enterprise Linux (2.6.32-71.el6.i686)
        root (hd0,0)
        kernel /vmlinuz-2.6.32-71.el6.i686 ro root=/dev/mapper/vg_redbooks03-lv_root
rd_LVM_LV=vg_redbooks03/lv_root rd_LVM_LV=vg_redbooks03/lv_swap rd_NO_LUKS rd_NO_MD
rd_NO_DM LANG=en_US.UTF-8 SYSFONT=latarcyrheb-sun16 KEYBOARDTYPE=pc KEYTABLE=us
crashkernel=auto rhgb quiet
        initrd /initramfs-2.6.32-71.el6.i686.img
title Red Hat Enterprise Linux (2.6.32-71.el6.i686) with MPP Support
        root (hd0,0)
        kernel /vmlinuz-2.6.32-71.el6.i686 ro root=/dev/mapper/vg_redbooks03-lv_root
rd_LVM_LV=vg_redbooks03/lv_root rd_LVM_LV=vg_redbooks03/lv_swap rd_NO_LUKS rd_NO_MD
rd_NO_DM LANG=en_US.UTF-8 SYSFONT=latarcyrheb-sun16 KEYBOARDTYPE=pc KEYTABLE=us
crashkernel=auto rhgb quiet
        initrd /mpp-2.6.32-71.el6.i686.img
```

10. Reboot the system.

11. Verify that the following modules are loaded by running the **/sbin/lsmod** command, as shown in Example 4-29:

– sd_mod
– sg
– mpp_Upper
– qla2xxx
– mpp_Vhba

Example 4-29 Validating loaded modules - lsmod output

```
mbcache                 5918  1 ext4
jbd2                   73876  1 ext4
mppVhba               121729  0
sr_mod                 14187  0
cdrom                  34035  1 sr_mod
qla2xxx               246038  0
scsi_transport_fc      40098  1 qla2xxx
scsi_tgt               10035  1 scsi_transport_fc
pata_acpi               2487  0
ata_generic             2555  0
ata_piix               19016  2
i915                  290623  2
drm_kms_helper         29029  1 i915
drm                   162327  3 i915,drm_kms_helper
i2c_algo_bit            4600  1 i915
i2c_core               25799  5 i2c_i801,i915,drm_kms_helper,drm,i2c_algo_bit
video                  16662  1 i915
output                  1779  1 video
mppUpper              137238  1 mppVhba
sg                     24778  0
sd_mod                 33344  3
crc_t10dif              1191  1 sd_mod
dm_mod                 63859  11 dm_mirror,dm_log
```

12. Use the command mppUtil to display the RDAC driver version just installed:

```
[root@redbooks03 /]# mppUtil -V
Linux MPP Driver Version: 09.03.0C05.0504
```

13. Scan the mapped LUNs without rebooting the server. You can use the command **hot_add** or **mppBusRescan**, as shown in Example 4-30.

Example 4-30 Re scanning HBAs - mppBusRescan output

```
[root@redbooks03 ~]# mppBusRescan
scan qla2 HBA host /sys/class/scsi_host/host4...
        found 4:0:0:1
scan qla2 HBA host /sys/class/scsi_host/host5...
        found 5:0:0:1
run /usr/sbin/mppUtil -s busscan...
scan mpp virtual host /sys/class/scsi_host/host6...
        found 6:0:0:1->/dev/sdc
```

14. Verify that the RDAC driver discovered the LUNs by running the **ls -lR /proc/mpp** command and checking for virtual LUN devices, as shown in Example 4-31.

Example 4-31 LUN Verification

```
[root@redbooks03 ~]# ls -lR /proc/mpp
/proc/mpp:
total 0
dr-xr-xr-x. 4 root root 0 Oct  7 16:24 DS5300

/proc/mpp/DS5300:
total 0
dr-xr-xr-x. 3 root root 0 Oct  7 16:24 controllerA
dr-xr-xr-x. 3 root root 0 Oct  7 16:24 controllerB
-rw-r--r--. 1 root root 0 Oct  7 16:24 virtualLun0
-rw-r--r--. 1 root root 0 Oct  7 16:24 virtualLun1

/proc/mpp/DS5300/controllerA:
total 0
dr-xr-xr-x. 2 root root 0 Oct  7 16:24 qla2xxx_h5c0t0

/proc/mpp/DS5300/controllerA/qla2xxx_h5c0t0:
total 0
-rw-r--r--. 1 root root 0 Oct  7 16:24 LUN0
-rw-r--r--. 1 root root 0 Oct  7 16:24 LUN1
-rw-r--r--. 1 root root 0 Oct  7 16:24 UTM_LUN31

/proc/mpp/DS5300/controllerB:
total 0
dr-xr-xr-x. 2 root root 0 Oct  7 16:24 qla2xxx_h4c0t0

/proc/mpp/DS5300/controllerB/qla2xxx_h4c0t0:
total 0
-rw-r--r--. 1 root root 0 Oct  7 16:24 LUN0
-rw-r--r--. 1 root root 0 Oct  7 16:24 LUN1
-rw-r--r--. 1 root root 0 Oct  7 16:24
UTM_LUN31
```

In our example, we have a DS5300 Storage System with two LUNs, each mapped to this Linux host. The host has two HBAs, each mapped to both controllers on each Storage System.

If any changes are made to the MPP configuration file (/etc/mpp.conf) or persistent binding file (/var/mpp/devicemapping), then the **mppUpdate** executable can be used to rebuild the RAMdisk.

The RDAC multipath driver consists of two parts. One part (mppUpper) is loaded before any HBA is loaded and prevents the HBA from advertising the LUNs to the system, whereas the other part (mppVhba) is a virtual HBA on top of the real HBA, which is responsible for bundling the various paths to the same LUN to a single (multipath) device.

To dynamically re-load the driver stack (scsi_mod, sd_mod, sg, mpp_Upper, <physical HBA driver>, mpp_Vhba) without rebooting the system, follow these steps:

1. Comment out all scsi_hostadapter entries in /etc/module.conf.
2. Issue the command **modprobe -r mpp_Upper** to unload the driver stack.
3. Then issue the command **modprobe mpp_Upper** to reload the driver stack.

Reboot the system whenever you need to unload the driver stack.

Linux Device mapper multipathing Driver (DMM / DM-Multipath)

Device mapper multipathing (DM-Multipath) allows you to configure multiple I/O paths between server nodes and storage arrays into a single device. These I/O paths are physical SAN connections that can include separate cables, switches, and controllers. Multipathing aggregates the I/O paths, creating a new device that consists of the aggregated paths, providing redundant failover/failback support for the logical drives in the DS5000 storage subsystems that are mapped to the Linux host server.

DMM Driver can be used for the following purposes:

Redundancy DM-Multipath can provide failover in an active/passive configuration. In an active/passive configuration, only half the paths are used at any time for I/O. If any element of an I/O path (the cable, switch, or controller) fails, DM-Multipath switches to an alternate path.

Improved Performance DM-Multipath can be configured in active/active mode, where I/O is spread over the paths in a round-robin fashion. In some configurations, DM-Multipath can detect loading on the I/O paths and dynamically re-balance the load.

In this section, we are covering the DM-Multipath configuration for *redundancy*.

Tip: DM-Multipath is only supported on DS Systems running firmware v7.77 and using Red Hat Enterprise Linux 6 and Novell SUSE Linux Enterprise Server 11 SP1 as the operating system.

Red Hat Enterprise Linux 6 and Novell SUSE Linux Enterprise Server 11 SP1 include the DM-Multipath driver (DMM Driver) by default. In the following exercise, we use Red Hat 6.

The DMM Driver and libraries are included in the installation media in the package directory as an RPM file. To check if you already have the DMM Driver and library dependencies installed, use the commands shown in Example 4-32.

Example 4-32 Checking DM-Multipath Driver packages

```
[root@redbooks03 ~]# rpm -qa | grep device-mapper
device-mapper-libs-1.02.53-8.el6.i686
device-mapper-1.02.53-8.el6.i686
device-mapper-multipath-0.4.9-31.el6.i686
```

```
device-mapper-event-libs-1.02.53-8.el6.i686
device-mapper-event-1.02.53-8.el6.i686
device-mapper-multipath-libs-0.4.9-31.el6.i686
```

If DM-Multipath driver is not installed on your systems, run the .rpm installation steps as shown in Example 4-33.

Example 4-33 Basic DM-Multipath .rpm installation

```
[root@redbooks03 ~]# mount /dev/cdrom /media/RHEL_6.0\ i386\ Disc\ 1/
[root@redbooks03 ~]# cd /media/RHEL_6.0\ i386\ Disc\ 1/
[root@redbooks03 RHEL_6.0 i386 Disc 1]# cd Packages/
[root@redbooks03 Packages]# pwd
/media/RHEL_6.0 i386 Disc 1/Packages
[root@redbooks03 Packages]# rpm -ivh device-mapper-XXX.rpm
```

Important: Before starting to work with the DM-Multipath driver configuration, make sure that no other multipath driver is loaded on the system, such as LSI RDAC, or any vendor HBA multipath driver.

Basic DM-Multipath configuration (basic failover/redundancy)

DM-Multipath daemon requires a basic configuration file, this configuration is stored on `/dev/multipath.conf`. We can set up multipath with the **`mpathconf`** utility, which creates the multipath configuration file `/etc/multipath.conf`.

You can enable the multipath support and create the basic configuration rules with the command line shown in Example 4-34.

Example 4-34 Enabling multipath on your system

```
[root@redbooks03 ~]# mpathconf --enable --with_multipathd y
[root@redbooks03 ~]# ll /etc/multipath.conf
-rw-------. 1 root root 2968 Oct 10 13:27 /etc/multipath.conf
```

After running this command line, you will have the daemon loaded on your system and added to the init sequence.

The default settings for DM-Multipath are compiled in to the system and do not need to be explicitly set in this file. To check the DM-Multipath Status, enter this command as shown in Example 4-35.

Example 4-35 Checking DM-Multipath Status

```
[root@redbooks03 ~]# mpathconf
multipath is enabled
find_multipaths is enabled
user_friendly_names is enabled
dm_multipath module is loaded
multipathd is chkconfiged on
```

To discover the targeted devices, run the command as shown in Example 4-36.

Example 4-36 Running multipath queries with multipath Command

```
[root@redbooks03 ~]# multipath -ll
mpathc (3600a0b80006e32020000fbb04e8f3a23) dm-4 IBM,1818      FAStT
size=10G features='1 queue_if_no_path' hwhandler='0' wp=rw
```

```
|-+- policy='round-robin 0' prio=6 status=active
| `- 5:0:0:1 sde 8:64 active ready running
`-+- policy='round-robin 0' prio=1 status=enabled
  `- 4:0:0:1 sdc 8:32 active ghost running
mpathb (3600a0b80006e32a000001d0b4e8f2c35) dm-3 IBM,1818      FAStT
size=10G features='1 queue_if_no_path' hwhandler='0' wp=rw
|-+- policy='round-robin 0' prio=6 status=active
| `- 5:0:0:0 sdd 8:48 active ready running
`-+- policy='round-robin 0' prio=1 status=enabled
  `- 4:0:0:0 sdb 8:16 active ghost running
```

For troubleshooting purposes, DM-Multipath has an interactive interface console to the multipathd daemon. Entering the **# multipathd -k** command brings up an interactive multipath console. After entering this command, you can enter **help** to get a list of available commands, you can enter a interactive command, or you can enter CTRL-D to quit, as shown in Example 4-37.

Example 4-37 DM-Multipath interactive interface console

```
[root@redbooks03 ~]# multipathd -k
multipathd> help
multipath-tools v0.4.9 (04/04, 2009)
CLI commands reference:
 list|show paths
 list|show paths format $format
 list|show status
 list|show maps|multipaths
 list|show maps|multipaths status
 list|show maps|multipaths stats
 list|show maps|multipaths format $format
 list|show maps|multipaths topology
 list|show topology
 list|show map|multipath $map topology
 list|show config
 list|show blacklist
 list|show devices
 list|show wildcards
 add path $path
 remove|del path $path
 add map|multipath $map
 remove|del map|multipath $map
 switch|switchgroup map|multipath $map group $group
 reconfigure
 suspend map|multipath $map
 resume map|multipath $map
 resize map|multipath $map
 disablequeueing map|multipath $map
 restorequeueing map|multipath $map
 disablequeueing maps|multipaths
 restorequeueing maps|multipaths
 reinstate path $path
 fail path $path
 paths count
 quit|exit
multipathd> show paths
hcil    dev dev_t pri dm_st  chk_st dev_st  next_check
0:0:0:0 sda 8:0   1   undef  ready  running orphan
4:0:0:0 sdb 8:16  1   active ghost  running XX........ 11/40
4:0:0:1 sdc 8:32  1   active ghost  running XX........ 11/40
5:0:0:1 sde 8:64  6   active ready  running XX........ 11/40
```

```
5:0:0:0 sdd 8:48  6   active ready  running XX........ 11/40
multipathd> show status
path checker states:
up                  3
ghost               2
```

For more information about the DM-Multipath driver, configuration, and troubleshooting, check the Red Hat documentation support page:

http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/index.html

4.5.4 Managing the Disk Space with LVM

In this section, we cover to how create a new Volume Group and how to expand it using native OS LVM Support.

First, we need to check which space we have assigned on the DS Storage system. Run **`ls -lR /proc/mpp`** on the server console. See Example 4-31 on page 222 for more details.

For this test scenario, we have assigned two disks with 10 Gb each, configured with multipath support provided by the LSI RDAC driver. Device name /dev/sda is the local disk supporting Red Hat 6 OS. Also, /dev/sdb and /dev/sdc are the DS Storage SAN Disks Space (IBM 1818) attached to the server.

To create a new LVM Volume, run the commands shown in Example 4-38.

Example 4-38 Creating a new LVM Volume

```
[root@redbooks03 ~]# cat /proc/partitions <----------------------- Check active partitions
major minor  #blocks  name

   8        0  245117376 sda
   8        1     512000 sda1
   8        2  244603904 sda2
   8       16   10485760 sdb
   8       32   10485760 sdc
 253        0   52428800 dm-0
 253        1    6045696 dm-1
 253        2  186126336 dm-2
253        3   10477568 dm-3

[root@redbooks03 ~]# pvcreate /dev/sdb <- Create a new physical volume, Initialize the disk
  Physical volume "/dev/sdb" successfully created

[root@redbooks03 ~]# pvdisplay /dev/sdb <---------------------------- Check physical volume
  "/dev/sdb" is a new physical volume of "10.00 GiB"
  --- NEW Physical volume ---
  PV Name               /dev/sdb
  VG Name
  PV Size               10.00 GiB
  Allocatable           NO
  PE Size               0
  Total PE              0
  Free PE               0
  Allocated PE          0
  PV UUID               qVtEwv-foKI-moZb-wNSN-O2tp-s8Wk-ganEj9

[root@redbooks03 ~]# vgcreate vg_ITSO2011 /dev/sdb <------------- Create a new volume group
  Volume group "vg_ITSO2011" successfully created
```

```
[root@redbooks03 ~]# vgdisplay vg_ITSO2011 <---------------------------- Check volume group
  --- Volume group ---
  VG Name               vg_ITSO2011
  System ID
  Format                lvm2
  Metadata Areas        1
  Metadata Sequence No  1
  VG Access             read/write
  VG Status             resizable
  MAX LV                0
  Cur LV                0
  Open LV               0
  Max PV                0
  Cur PV                1
  Act PV                1
  VG Size               10.00 GiB
  PE Size               4.00 MiB
  Total PE              2559
  Alloc PE / Size       0 / 0
  Free  PE / Size       2559 / 10.00 GiB
  VG UUID               XFHEmF-xOiY-sTOF-8Gtp-1SyR-c6W2-IUnsru

[root@redbooks03 ~]# lvcreate -L 9.99Gib vg_ITSO2011 <------------- Create a logical volume
  Rounding up size to full physical extent 9.99 GiB

[root@redbooks03 ~]# lvdisplay vg_ITSO2011/lvol0 <---------------- Check the logical volume
  --- Logical volume ---
  LV Name                /dev/vg_ITSO2011/lvol0
  VG Name                vg_ITSO2011
  LV UUID                jDOe1c-t371-Sa3q-UZBH-p5YL-r3rH-W7A2xr
  LV Write Access        read/write
  LV Status              available
  # open                 0
  LV Size                9.99 GiB
  Current LE             2558
  Segments               1
  Allocation             inherit
  Read ahead sectors     auto
  - currently set to     256
  Block device           253:3
```

When we have the SAN Space managed by LVM, we will be able to format and mount the filesystem to server. Run the commands shown in Example 4-39.

Example 4-39 Formatting and mounting disk partitions

```
[root@redbooks03 ~]# mkfs.ext4 /dev/vg_ITSO2011/lvol0 <---------- Format the logical volume
mke2fs 1.41.12 (17-May-2010)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
655360 inodes, 2619392 blocks
130969 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=2684354560
80 block groups
32768 blocks per group, 32768 fragments per group
```

```
8192 inodes per group
Superblock backups stored on blocks:
        32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632

Writing inode tables: done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 27 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.

[root@redbooks03 ~]# mkdir /opt/ITSO2011_Space <---------------------- Create a mount point

[root@redbooks03 ~]# mount /dev/vg_ITSO2011/lvol0 /opt/ITSO2011_Space/ <-- Mount filesystem

[root@redbooks03 ~]# df -h <------------------------- Check filesystem mounted and capacity
Filesystem            Size  Used Avail Use% Mounted on
/dev/mapper/vg_redbooks03-lv_root
                       50G  3.7G   44G   8% /
tmpfs                 1.9G   88K  1.9G   1% /dev/shm
/dev/sda1             485M   60M  400M  14% /boot
/dev/mapper/vg_redbooks03-lv_home
                      175G  754M  166G   1% /home
/dev/mapper/vg_ITSO2011-lvol0
                      9.9G   23M  9.4G   1% /opt/ITSO2011_Space
```

Example 4-40 shows the necessary commands to add a new physical disk (Storage Space) to an existing Volume Group and show to the resize the space dynamically on a existing filesystem.

Example 4-40 Extending LVM & Filesystem Capacity

```
[root@redbooks03 ~]# pvcreate /dev/sdc <- Create a new physical volume, Initialize the disk
  Physical volume "/dev/sdc" successfully created

[root@redbooks03 ~]# vgextend vg_ITSO2011 /dev/sdc <--------------- Extend the volume group
Volume group "vg_ITSO2011" successfully extended

[root@redbooks03 ~]# vgdisplay vg_ITSO2011 <---------------------------- Check volume group
  --- Volume group ---
  VG Name               vg_ITSO2011
  System ID
  Format                lvm2
  Metadata Areas        2
  Metadata Sequence No  5
  VG Access             read/write
  VG Status             resizable
  MAX LV                0
  Cur LV                1
  Open LV               0
  Max PV                0
  Cur PV                2
  Act PV                2
  VG Size               19.99 GiB
  PE Size               4.00 MiB
  Total PE              5118
  Alloc PE / Size       2558 / 9.99 GiB
  Free  PE / Size       2560 / 10.00 GiB
  VG UUID               I2yIrY-Iqak-W1qO-Nosw-VMvt-ZcD9-3F971X

[root@redbooks03 ~]# lvextend -L +9.99Gib vg_ITSO2011/lvol0 <-------- Extend logical volume
```

```
Rounding up size to full physical extent 9.99 GiB
  Extending logical volume lvol0 to 19.98 GiB
  Logical volume lvol0 successfully resized

[root@redbooks03 ~]# lvdisplay vg_ITSO2011/lvol0 <------------------ - Check logical volume

  --- Logical volume ---
  LV Name                /dev/vg_ITSO2011/lvol0
  VG Name                vg_ITSO2011
  LV UUID                dO6SMD-n6P2-t89T-ORwY-VziS-MTcW-e5VJlb
  LV Write Access        read/write
  LV Status              available
  # open                 1
  LV Size                19.98 GiB
  Current LE             5116
  Segments               2
  Allocation             inherit
  Read ahead sectors     auto
  - currently set to     256
  Block device           253:3

[root@redbooks03 ~]# resize2fs -p /dev/vg_ITSO2011/lvol0 <----------- Resize the filesystem

resize2fs 1.41.12 (17-May-2010)
Filesystem at /dev/vg_ITSO2011/lvol0 is mounted on /opt/ITSO2011_Space; on-line resizing
required
old desc_blocks = 1, new_desc_blocks = 2
Performing an on-line resize of /dev/vg_ITSO2011/lvol0 to 5238784 (4k) blocks.
The filesystem on /dev/vg_ITSO2011/lvol0 is now 5238784 blocks long.

[root@redbooks03 ~]# df -h <----------------------------------- Check filesystem capacity

Filesystem            Size  Used Avail Use% Mounted on
/dev/mapper/vg_redbooks03-lv_root
                       50G  3.7G   44G   8% /
tmpfs                 1.9G   88K  1.9G   1% /dev/shm
/dev/sda1             485M   60M  400M  14% /boot
/dev/mapper/vg_redbooks03-lv_home
                      175G  754M  166G   1% /home
/dev/mapper/vg_ITSO2011-lvol0
                       20G  156M   19G   1% /opt/ITSO2011_Space
```

Because the goal of the prior example is to show the basic configuration steps attaching SAN Space and managing the space dynamically using LVM, some OS suggestions have been excluded on the procedure, such as the following ones:

- Label the Logical Volume using `e2label` command.
- Mount the Filesystem permanently adding the mount point on the `/etc/fstab` file.
- Use the **`blkid`** command to get disk additional information such as UUID. Note that UUID is best practice for mounting the partition in /etc/fstab (for stability).

For more information using LVM and Linux commands, visit the vendor support documentation portal and man pages.

4.5.5 Configuring Linux for iSCSI attachment

The DS Storage Systems has the option to attach your hosts using iSCSI interfaces. In this section, we show how to configure your Linux Server to use a regular Ethernet network interface card (NIC), using software iSCSI Initiator to connect to a DS5300 system with iSCSI host interface cards.

Our implementation example is using Red Hat Linux Enterprise 6, two Ethernet network cards connected to a different Ethernet switches, the traffic is isolated on a dedicated private network where the DS5300 iSCSI controllers resides.

The DS Storage System iSCSI ports are defined as follows:

- 192.168.130.101 - iSCSI Controller A
- 192.168.130.102 - iSCSI Controller B

Installing and configuring the iSCSI Software Initiator

Red Hat Enterprise Linux includes an iSCSI Software Initiator. The software initiator is included in the installation media in the package directory as an RPM file.

To check if you already have the iSCSI initiator software installed, use the commands shown in Example 4-41.

Example 4-41 Checking initiator software

```
[root@redbooks03 ~]# rpm -qa | grep iscsi-initiator-utils
iscsi-initiator-utils-6.2.0.872-10.el6.i686
[root@redbooks03 ~]# rpm -qi iscsi-initiator-utils-6.2.0.872-10.el6.i686
Name        : iscsi-initiator-utils        Relocations: (not relocatable)
Version     : 6.2.0.872                         Vendor: Red Hat, Inc.
Release     : 10.el6                         Build Date: Wed 18 Aug 2010 02:24:31 AM MST
Install Date: Thu 06 Oct 2011 12:07:37 PM MST      Build Host: hs20-bc2-4.build.redhat.com
Group       : System Environment/Daemons    Source RPM:
iscsi-initiator-utils-6.2.0.872-10.el6.src.rpm
Size        : 1707344                          License: GPLv2+
Signature   : RSA/8, Wed 18 Aug 2010 10:02:18 AM MST, Key ID 199e2f91fd431d51
Packager    : Red Hat, Inc. <http://bugzilla.redhat.com/bugzilla>
URL         : http://www.open-iscsi.org
Summary     : iSCSI daemon and utility programs
Description :
The iscsi package provides the server daemon for the iSCSI protocol,
as well as the utility programs used to manage it. iSCSI is a protocol
for distributed disk access using SCSI commands sent over Internet
Protocol networks.
```

If you do not see a similar output, install the software initiator. Mount the installation media and install the iSCSI Software Initiator RPM file from the Packages directory, as in Example 4-42.

Example 4-42 Installation of the iSCSI Software Initiator RPM file

```
[root@redbooks03 ~]# mount /dev/cdrom /media/RHEL_6.0\ i386\ Disc\ 1/
[root@redbooks03 ~]# cd /media/RHEL_6.0\ i386\ Disc\ 1/
[root@redbooks03 RHEL_6.0 i386 Disc 1]# cd Packages/
[root@redbooks03 Packages]# pwd
/media/RHEL_6.0 i386 Disc 1/Packages
[root@redbooks03 Packages]# rpm -ivh iscsi-initiator-utils-6.2.0.872-10.el6.i686.rpm
warning: iscsi-initiator-utils-6.2.0.872-10.el6.i686.rpm: Header V3 RSA/SHA256 Signature,
key ID fd431d51: NOKEY
```

```
Preparing...                ########################################### [100%]
   1:iscsi-initiator-utils  ########################################### [100%]
[root@redbooks03~Packages]#
```

After the iSCSI Software Initiator is installed, we need to check the configuration files stored in the /etc/iscsi directory.

Determining the initiator name

The initiator name assigned by default is in file */etc/iscsi/initiatorname.iscsi*. After the iSCSI daemon is started for the first time, this file will be populated with the iSCSI qualified name of the host.

Example 4-43 shows the file with an automatically generated iSCSI Qualified Name (IQN). This IQN is regenerated to the same value when the software initiator gets reinstalled.

Example 4-43 /etc/iscsi/nitiatorname.iscsi after first iSCSI daemon start

```
[root@redbooks03 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1994-05.com.redhat:aa75593df85a
```

You can generate a new IQN to assign to your host using the **iscsi-iname** command, which will generate a unique IQN with the correct format and output it to the console. Example 4-44 shows how to use the **iscsi-iname** command to define the IQN of a given host. The output can be redirected into /etc/initiatorname.iscsi to change the actual host initiator name.

Example 4-44 iscsi-iname

```
[root@redbooks03 /]# iscsi-iname
iqn.1994-05.com.redhat:5c2b7a56315
```

Starting the iSCSI service

With the iscsi tools configured, then we start the iscsi daemon, and we configure it to start automatically after each reboot. Use the commands, as shown in Example 4-45.

Example 4-45 Starting iSCSI daemon

```
[root@redbooks03 /]# service iscsid start
Turning off network shutdown. Starting iSCSI daemon:       [  OK  ]
                                                           [  OK  ]
[root@redbooks03 /]#
```

To load the iSCSI driver during reboot, it is necessary to modify the service database. Run the command **chkconfig iscsi on** to enable the load of the iSCSI driver during system start, as shown in Example 4-46.

Example 4-46 Enable iSCSI during system start

```
[root@redbooks03 ~]# chkconfig iscsi on
[root@redbooks03 ~]# chkconfig --list | grep iscsi
iscsi           0:off   1:off   2:on    3:on    4:on    5:on    6:off
iscsid          0:off   1:off   2:off   3:on    4:on    5:on    6:off
```

Configuring the connection to the target

With the iSCSI service running, we can start using the utility **iscsiadm**. This command is used for managing and configuring iSCSi in general, so we will continue to use it later.

To obtain a list of available targets from a given host, we use the **iscsiadm** command to discovery any available targets, providing the IP address of any of your DS Storage System iSCSI host ports. In our example, the DS5300 has its iSCSI host port addresses set to 192.168.130.101 and 192.168.130.101.

Example 4-47 Discovering targets

```
[root@redbooks03 ~]# iscsiadm -m discovery -t sendtargets -p 192.168.130.101
192.168.130.101:3260,1 iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
[fe80:0000:0000:0000:02a0:b8ff:fe5e:115c]:3260,1
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
192.168.131.101:3260,1 iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
192.168.130.102:3260,2 iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
192.168.131.102:3260,2 iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
```

Alternatively, you can also define all the iSCSI parameters in the file `/etc/iscsi/iscsi.conf`. This information is well documented in the configuration file itself. After installation, the file contains only comments, and it needs to be modified at least with a discovery address. Without the discovery address, the daemon will fail to start. If you plan to use CHAP authentication, then you also need to configure here. We cover security configuration for iSCSI later in “Enhancing connection security” on page 236.

Use the Storage Manager to confirm your DS Storage System iscsi target name. Select the storage subsystem from the Logical view, and go down in the right frame to find the iSCSI target name, as shown in Figure 4-52.

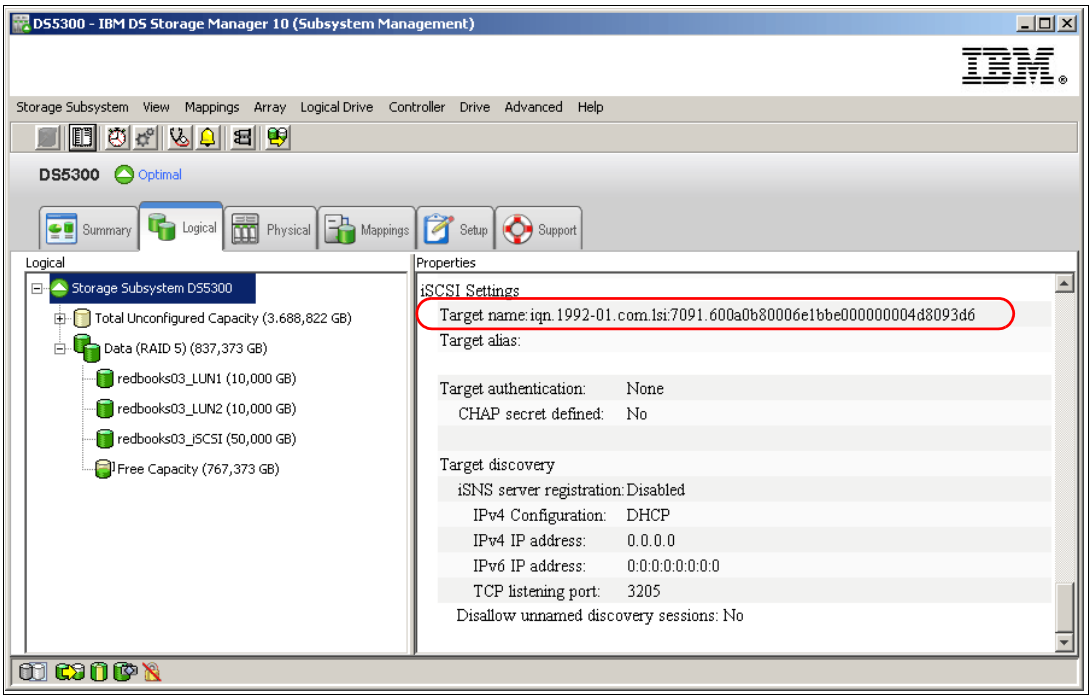


Figure 4-52 Identifying target iSCSI Qualified Name (IQN)

Now use the command in Example 4-48 on page 233 to log in to the available target, using the name from the output of the previous command, confirmed with the Storage Manager.

Repeat the command in order to set two sessions, one to each controller, which also enables the target to be accessed upon reboots.

Example 4-48 Login to Desired DS target

```
[root@redbooks03 ~]# iscsiadm -m node -T
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6 -p 192.168.130.101 -l
Logging in to [iface: default, target:
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6, portal: 192.168.130.101,3260]
Login to [iface: default, target:
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6, portal: 192.168.130.101,3260]
successful.

[root@redbooks03 ~]# iscsiadm -m session
tcp: [1] 192.168.130.101:3260,1 iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6

[root@redbooks03 ~]# iscsiadm -m node -T
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6 -p 192.168.130.101 -l
Logging in to [iface: default, target:
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6, portal: 192.168.130.101,3260]
Login to [iface: default, target:
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6, portal: 192.168.130.101,3260]:
successful
```

Defining the host in the DS Storage System

After successfully logging in to the desired DS Storage System, go to the Storage Manager software and create a host, adding the newly found host port identifier corresponding to the Linux server initiator name.

From the Subsystem Management window of the Storage Manager, select **Mappings** → **Define** → **Host.** After assigning a name, you need to select the connection interface type, and host port identifier. Because the server is already logged to the DS Storage System, you can pick the host port identifier from the known unassociated list, which has to match with the initiator name of your server. Display the file /etc/iscsi/initiatorname from your server to make sure to assign the correct identifier for the host, as shown in Figure 4-53.

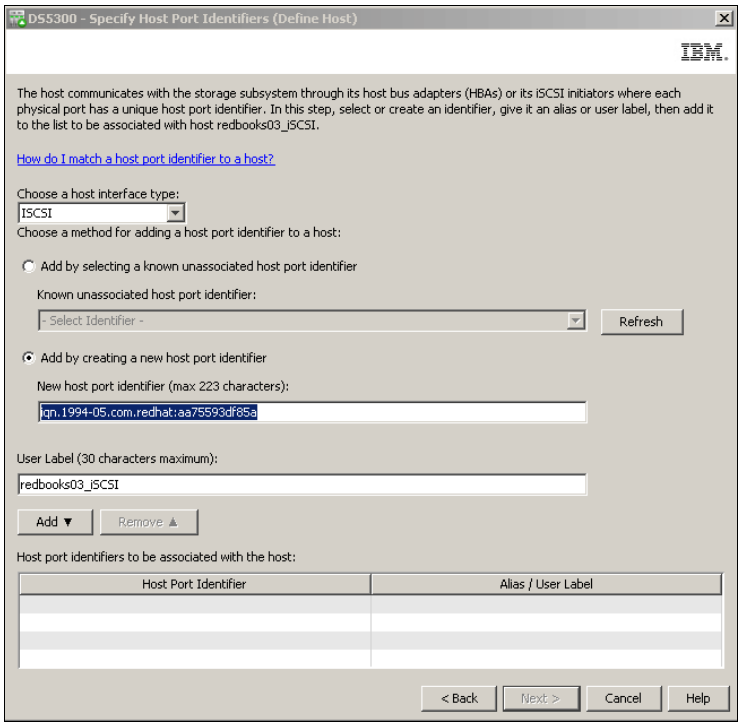


Figure 4-53 Defining Host with discovered Initiator name

Now, with the Host defined, proceed to map the logical volumes planned to attach to the Linux host. After it is mapped, proceed to install the RDAC driver or DM-Multipath in your Linux operating system, as detailed in 4.5.3, “Installing the Linux multipath driver” on page 218.

Displaying iSCSI sessions

An iSCSI session is a communication link established between an initiator and a target. In our installation example, we have a single NIC with connections established to both controllers during the login, executed as described in “Configuring the connection to the target” on page 231.

The DS Storage System implementation does not allow for multiple connections per target, so each session will have only one connection established against a single controller port. If you have a single Ethernet card in your host attaching your storage, and only one ISCSI port defined per controller, then you must have two sessions established, one to each controller, as in this implementation example.

Use the **`iscsi adm`** command to display the sessions in your host. To display the session information from the DS Storage Manager, select from your Storage subsystem Management window, **Storage Subsystem** → **iSCSI** → **End Sessions,** as shown in Figure 4-54 and Example 4-49.

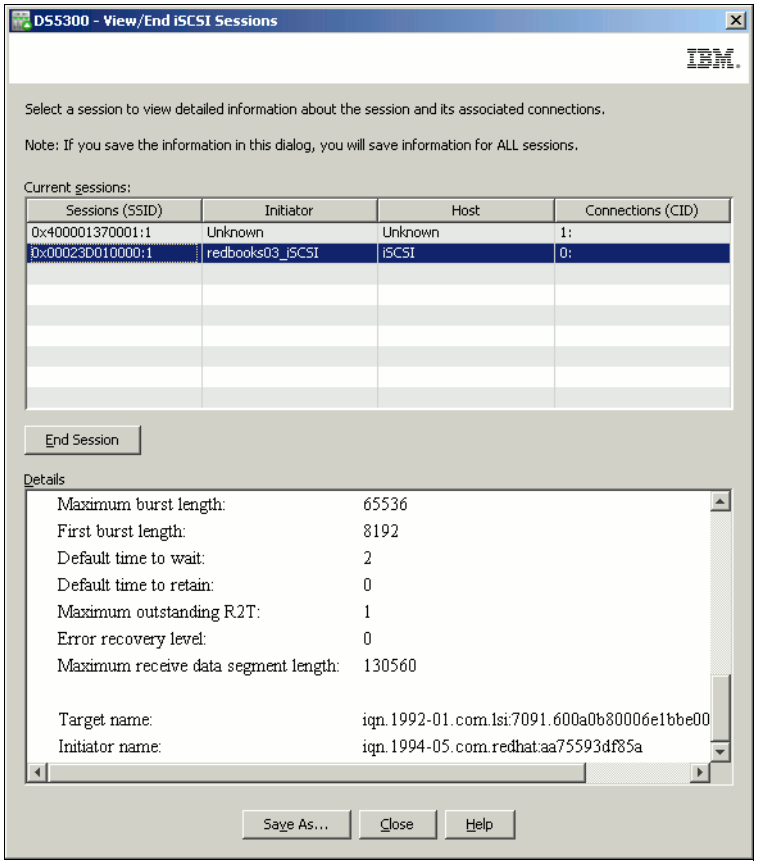


Figure 4-54 Displaying sessions established

Example 4-49 Displaying sessions established

```
[root@redbooks03 ~]# iscsiadm -m session
tcp: [1] 192.168.130.101:3260,1
iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
```

Always check that you have established connections to each controller, as in the example. In the details section of the View/End iSCSI Sessions window, you can confirm the IP address participating in each session.

To display the details of each section from your Linux host, use the **iscsiadm** command, with the parameters shown in Example 4-50.

Example 4-50 Displaying sessions details

```
[root@redbooks03 ~]# iscsiadm -m session
tcp: [1] 192.168.130.101:3260,1 iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
Session ID 1, connecting to A controller IP

[root@redbooks03 ~]# iscsiadm -m session -P 3 -r 1 <------- where 1 is the session ID
iSCSI Transport Class version 2.0-870
version 2.0-872
Target: iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
        Current Portal: 192.168.130.101:3260,1
        Persistent Portal: 192.168.130.101:3260,1
                **********
                Interface:
                **********
                Iface Name: default
                Iface Transport: tcp
                Iface Initiatorname: iqn.1994-05.com.redhat:aa75593df85a
                Iface IPaddress: 192.168.130.20
                Iface HWaddress: <empty>
                Iface Netdev: <empty>
                SID: 1
                iSCSI Connection State: LOGGED IN
                iSCSI Session State: LOGGED_IN
                Internal iscsid Session State: NO CHANGE
                ************************
                Negotiated iSCSI params:
                ************************
                HeaderDigest: None
                DataDigest: None
                MaxRecvDataSegmentLength: 262144
                MaxXmitDataSegmentLength: 65536
                FirstBurstLength: 8192
                MaxBurstLength: 65536
                ImmediateData: Yes
                InitialR2T: Yes
                MaxOutstandingR2T: 1
                ************************
                Attached SCSI devices:
                ************************
                Host Number: 6  State: running
                scsi6 Channel 00 Id 0 Lun: 0
                        Attached scsi disk sdd          State: running
                scsi6 Channel 00 Id 0 Lun: 31
```

Use these details to ensure that all hardware paths established by your network configuration cabling are correctly recognized, especially when troubleshooting connection problems.

Discovering the mapped volumes

With the logical drives mapped, the RDAC or the DM-Multipath installed, the sessions established, and the host communicating with the target storage through both controllers, proceed to detect them using the hot_add utility.

In Example 4-51, we show the output of the hot_add utility executed to dynamically discover the volumes.

Example 4-51 Discovering mapped volumes

```
[root@redbooks03 /]# hot_add
scan iSCSI software initiator host /sys/class/scsi_host/host2...
        found 5:0:0:1
        found 4:0:0:1
scan iSCSI software initiator host /sys/class/scsi_host/host1...
        found 5:0:0:0
        found 4:0:0:0
run /usr/sbin/mppUtil -s busscan...
ls: /sys/bus/scsi/drivers/sd/*/block:*: No such file or directory
ls: /sys/bus/scsi/drivers/sd/*/block:*: No such file or directory
[root@TC-2008 /]
```

Now the iSCSI driver and MPP are loaded, recognizing the multiple paths per disk volume. We verify the volume and its connections later in 4.5.6, “Collecting information” on page 238.

Enhancing connection security

After everything is working and tested, you must implement security for the iSCSI connection; essentially, this means configuring initiator and target authentication. Initiator authentication means that an initiator must prove its identity with a password that is known by the target when the initiator attempts access. Target authentication is the opposite; the target authenticates itself to the initiator with a password.

Because an iSCSI qualified name can be modified within Storage Manager, this does not protect against spoofing of the qualified name, and hence security can be compromised. We describe the steps required to set up initiator and target authentication.

Just like a hardware based initiator, the software initiator requires initiator authentication as a prerequisite for target authentication. Follow these steps:

1. Use the **iscsiadm** command or directly edit the configuration file /etc/iscsi/iscsid.conf of the iSCSI Software Initiator. Define the incoming and outgoing user names and passwords in this file. Incoming means that the target has to authenticate itself against the initiator, and is also called target authentication. Outgoing represents the initiator authentication. The initiator has to authenticate against a target.

   Incoming and outgoing user names are limited to valid IQNs by the DS5300 defined as host ports.

   Example 4-52 shows the /etc/iscsi/iscsid.conf file with the incoming and outgoing account details. The incoming account was configured as a local option for the target and not a global option. Other subsystems might use another password.

Example 4-52 Configuration of the iSCSI Software Initiator

```
HeaderDigest=always
DataDigest=always
OutgoingUsername=iqn.1994-05.com.redhat:aa75593df85a
OutgoingPassword=ITSO2011_Secured

Targetname=iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
        Enabled=yes
        IncomingUsername=iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6
        IncomingPassword=ITSO2011_Secured
        ConnFailTimeout=15
```

2. Shut down the server until the DS5300 is also configured.
3. Use the Storage Manager CLI commands (**`set iscsiInitiator`**) shown in Example 4-53 to set up the CHAP secret (Challenge Handshake Authentication Protocol) for the already defined host ports of host “redbooks03”.

Example 4-53 SMcli commands to set CHAP secrets for initiators

```
# SMcli -n DS5300 -c "set iscsiInitiator [\"redbooks03\"] host=\"redbooks03\"
chapSecret=\"ITSO2011_Secured\"; " -S
#
```

Tip: CHAP (RFC1944) is the most basic level of iSCSI security available.

Security suggestions: Use strong passwords for all accounts. Use CHAP authentication because it ensures that each host has its own password. Mutual CHAP authentication is also suggested. See 3.1.6, “iSCSI configuration and management” on page 123.

4. Clarify if there are any initiators without target authentication configured that access the DS5300. In that case, use the command in Example 4-54; otherwise, use the command shown in Example 4-55.

Example 4-54 SMcli - Set target authentication - CHAP only

```
# SMcli -n DS5300 -c "set iscsiTarget
<\"iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6\">
authenticationMethod=chap chapSecret=\"ITSO2011_Secured\";" -S
#
```

Example 4-55 SMcli - Set target authentication - CHAP and no CHAP

```
# SMcli -n DS5300 -c "set iscsiTarget
<\"iqn.1992-01.com.lsi:7091.600a0b80006e1bbe000000004d8093d6\">
authenticationMethod=none authenticationMethod=chap chapSecret=\"ITSO2011_Secured\";" -S
#
```

After modifying the configuration on the DS5300, reboot the server to get access to the logical drives by using initiator and target authentication.

After you have your iSCSI Software Initiator installed and configured, you will be able to start managing your disk space. We already covered this topic before, for additional information refers to section 4.5.4, “Managing the Disk Space with LVM” on page 226

iSCSI security suggestions and best practices

Consider the following suggestions:

- Configure additional Paths for High Availability; use either LSI RDAC driver or DM-Multipath with additional NICs in the server to create additional connections to the iSCSI storage array through redundant Ethernet switch fabrics.
- Use Gigabit Ethernet connections for high speed access to storage.
- Use Server class NICs. It is suggested to use NICs which are designed for enterprise networking and storage applications.
- Use CAT6 rated cables for Gigabit Network Infrastructures and Cat-6a or Cat-7 for 10Gigabit implementations.

- Segregate SAN and LAN traffic. iSCSI SAN interfaces need to be separated from other corporate network traffic (LAN). Servers should use dedicated NICs for SAN traffic. Deploying iSCSI disks on a separate network helps to minimize network congestion and latency. Additionally, iSCSI volumes are more secure when… Segregate SAN & LAN traffic can be separated using port based VLANs or physically separate networks.
- Use non blocking switches and set the negotiated speed on the switches.
- Use Jumbo Frames if supported in your network infrastructure.
- Enable and configure CHAP Security to use either target or mutual authentication.

4.5.6 Collecting information

You can use the following commands to collect additional information about your Linux host server.

- **`SMdevices`**: Lists all DS Storage volumes recognized by your server (if SMutil is installed).
- **`rpm -qa`**: Lists the installed software.
- **`uname -r:`** Displays the kernel version.
- **`mppUtil -V`**: Lists the installed disk driver version.
- **`mppUtil -a`** *`DSname:`* Lists all the details for the Storage name provided.
- **`ls -lR /proc/mpp`**: Lists devices recognized by the RDAC driver.
- **`cat /proc/scsi/scsi`**: Displays LUNs recognized by the HBAs, with all the paths. Because it is not under the RDAC driver usage, then you see all LUNs from all the paths.
- **`/opt/mpp/lsvdev`**: Provides a relation between the logical volumes and the Linux disk assigned names.
- **`/opt/mpp/mppSupport`**: Script provided by RDAC to collect information. Generates a compressed file in the /tmp folder that you can use yourself or share with your support representative for debugging problems.
- IBM Dynamic System Analysis (DSA) log: DSA collects and analyzes system information to aid in diagnosing system problems. This tool was developed initially for System x servers, available for Linux and Windows, and when executed in your server, can provide all the information that your support representative will need in one package. To find the DSA tools, see the following website:

  `http://www-03.ibm.com/systems/management/dsa.html`

4.6 i5/OS

The DS Storage System is supported now connected to i5/OS, both in client partitions of VIOS, and as a direct attach to the newest DS5000 Systems.

For more details about the Midrange Storage attachment to i5/OS, see the *IBM i and Midrange External Storage*, SG24-7668.

4.7 VMware

The DS Storage Systems are supported with VMware vSphere products. There is a separate IBM Redpapers™ publication covering all the information about this topic, and much more.

See *VMware Implementation with IBM System Storage DS5000*, REDP-4609. This document is a compilation of best practices for planning, designing, implementing, and maintaining IBM Midrange storage solutions and, more specifically, configurations for VMware vSphere ESXi. Setting up an IBM Midrange Storage Subsystem can be a challenging task, and the principal objective of this book is to give users a sufficient overview to effectively implement IBM Midrange Storage and VMWare vSphere products.

4.8 Hyper-V

Microsoft Hyper-V is the hypervisor-based virtualization product that allows customers to consolidate workloads onto a single physical server. It provides additional functions such as workload administration, dynamic memory, live migration, cluster shared volume support and expanded processor and memory support for host systems.

At the time that we wrote this publication, Microsoft offers two forms to access to the Hypervisor:

- The Hyper-V server role, which you can install on systems running Windows Server 2008 R2 Standard, Enterprise, or Datacenter edition. (host-based)
- Microsoft Hyper-V Server 2008 R2, a standalone, hypervisor-based server virtualization product that lets you virtualize workloads onto a single physical server. (bare-metal)

Both products are based on the same hypervisor technology. Microsoft Hyper-V Server, however, has the following characteristics:

- It is totally free.
- It has no graphical user interface (think Server Core; you need to manage it remotely).
- It has no guest virtualization rights, which means you will need licenses for any Windows operating systems that you run on it.

For more information about Microsoft Hyper-V virtualization products, see this website:

`http://www.microsoft.com/en-us/server-cloud/hyper-v-server/default.aspx`

5

SAN boot with the IBM System Storage DS5000 storage subsystem

SAN boot is a technique that allows servers to utilize an OS image installed on external SAN-based storage to boot up, rather than booting off their own local disk or direct attached storage (internal server disk drives).

Now that iSCSI has become more popular, we need to use the terms *remote boot* or *network boot* rather than *SAN boot* because iSCSI is typically Ethernet communication protocol, not typical SAN storage term. However, because the iSCSI and Fibre Channel worlds are merging together, especially with the introduction of 10 Gbps iSCSI, we use the term *SAN boot* in this chapter to explain booting utilizing both the Fibre Channel and iSCSI techniques.

Here, we introduce only the concept of SAN boot using various operating systems on several platforms and mandatory prerequisites. We do not go to the deep technical details and do not show any practical examples. An IBM Redbooks publication is available for implementation scenarios, *SAN Boot Implementation and Best Practices Guide for IBM System Storage*, SG24-7958. That publication also covers the best practices for booting various platforms from IBM Midrange Storage Systems, IBM DS5300, respectively:

- AIX FC SAN boot for IBM Power Systems
- Windows 2008 SAN boot with Fibre Channel and iSCSI
- FC SAN boot for RHEL5.5 on IBM system x servers
- iSCSI SAN boot for RHEL5.4 on IBM system x servers
- FC SAN boot for SLES 11 on IBM system x servers
- iSCSI SAN boot for SLES 11on IBM system x servers
- FC SAN Boot implementation for VMware ESXi 4
- Implementing Windows Server 2008 Failover Clustering with SAN boot

5.1 Introduction to SAN boot

Boot from SAN refers to the server configuration where the server OS is installed on a logical drive (LUN) that does not reside inside the server chassis. SAN boot utilizes drives located in a disk storage subsystem like an IBM System Storage DS5000, that is connected by a Fibre Channel or iSCSI host bus adapter located in the server chassis. Using SAN boot has the following distinct advantages:

- Interchangeable servers:

  By allowing boot images to be stored on the SAN, servers are no longer physically bound to their startup configurations. Therefore, if a server were to fail, it becomes very easy to replace it with another generic server and resume operations with the exact same boot image from the SAN (only minor reconfiguration is required on the storage subsystem). This quick interchange will help reduce downtime and increase host application availability.

- Provisioning for peak usage:

  Because the boot image is available on the SAN, it becomes easy to deploy additional servers to temporarily cope with high workloads.

- Centralized administration:

  SAN boot enables simpler management of the startup configurations of servers. Rather than needing to manage boot images at the distributed level at each individual server, SAN boot empowers administrators to manage and maintain the images at a central location in the SAN. This feature enhances storage personnel productivity and helps to streamline administration.

- Utilizing high-availability features of SAN storage:

  SANs and SAN-based storage are typically designed with high availability in mind. SANs can utilize redundant features in the storage network fabric and RAID controllers to ensure that users do not incur any downtime. Most boot images that are located on local disk or direct attached storage do not share the same protection. Using SAN boot allows boot images to take advantage of the inherent availability built-in to most SANs, which helps to increase availability and reliability of the boot image and reduce downtime.

- Efficient disaster recovery process:

  Assuming that data (boot image and application data) is mirrored over the SAN between a primary site and a recovery site, servers can take over at the secondary site in case a disaster destroys servers at the primary site.

- Reduced overall cost of servers:

  Placing server boot images on external SAN storage eliminates the need for local disk in the server, which helps lower costs and allows SAN boot users to purchase servers at a reduced cost but still maintain the same functionality. In addition, SAN boot minimizes the IT costs through consolidation, what is realized by electricity, floor space cost savings, and by the benefits of centralized management.

Boot from SAN provides more resilient and robust architecture and offers better protection from hardware failure. This chapter is intended to provide readers practical guidebook how to set up and bring to the production servers booting from SAN either utilizing FC protocol or iSCSI on various platforms.

5.1.1 SAN boot implementation

Setting up and executing SAN boot requires performing certain steps on both the server and the storage sides of the SAN. We first provide a brief overview of how a user can implement SAN boot and how SAN boot works.

SAN boot relies on configuring servers with a virtual boot device, which enables the server to access the actual boot information stored on a specific LUN in the storage subsystem. The virtual boot device is configured on the HBA in the server and thus assumes that the HBA BIOS supports booting from SAN attached storage.

Multiple LUNs can be configured with various boot images if there is a need for separate operating system images. Each LUN that contains a boot image for a specific operating system is referred to as a boot partition. Boot partitions are created, configured, and mapped just like normal storage partitions on the DS5000 storage subsystem.

Figure 5-1 shows a 4+P RAID 5 array that is carved up into boot partition LUNs. Each of these LUNs corresponds to a separate server. The LUNs can be connected to various homogeneous hosts, or they can be connected to heterogeneous hosts.

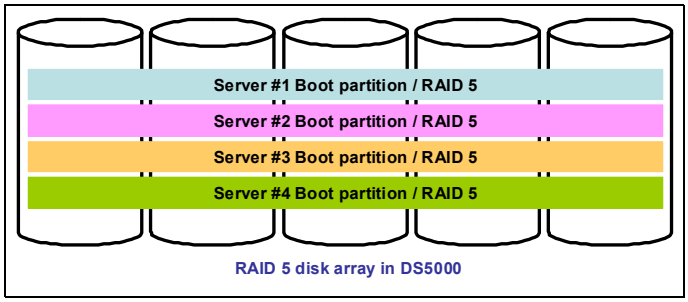


Figure 5-1 Boot partition LUNs on a RAID 5 array

Figure 5-2 shows five heterogeneous servers that use SAN boot and store their boot images on the same DS5000 storage subsystem. However, this diagram shows the concept similar to Fibre Channel or iSCSI implementation.

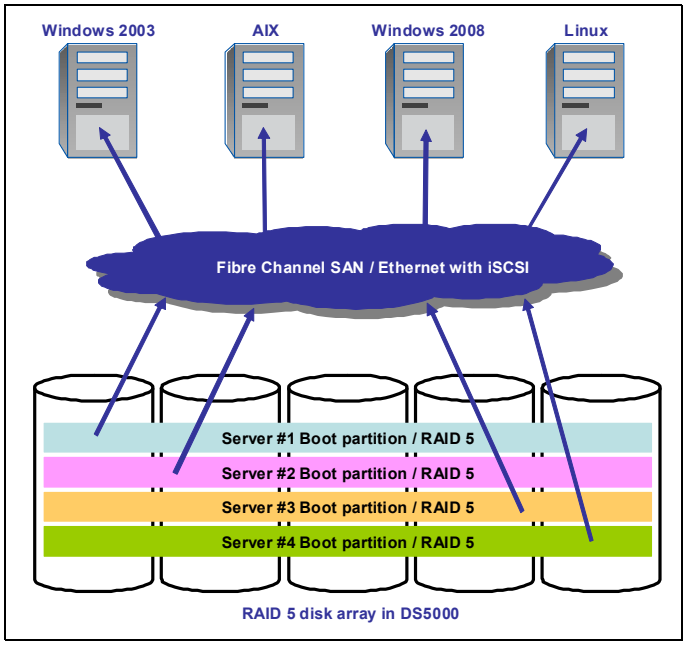


Figure 5-2 Logical diagram of SAN boot partitions

In order to implement SAN boot, you must first ensure that all your hosts are properly zoned to the storage subsystems in the SAN fabric. Zoning emulates a direct connection from the storage to each server. Storage partitioning, which needs to be done on the DS5000, only ensures that authorized hosts can access their dedicated storage units through the SAN.

The Host Bus Adapters (HBAs) and SAN-based storage subsystem both require setup. The specific syntax and steps differ from vendor to vendor for HBAs, so we take more of a general approach when describing the setup with the DS5000 storage subsystem.

The general steps are as follows:

1. Establish the physical connectivity between the HBA, switch, and DS5000 storage subsystem.
2. Provision LUNs on the DS5000 storage subsystem to handle host images. Create one LUN per boot image.
3. Select the proper port (HBA WWPN) and LUN from which the host must launch its boot.
4. Ensure that storage partitioning properly maps the appropriate LUN with the appropriate host server. It is also required that no other server can view that LUN.
5. Configure the HBAs of the hosts that are to utilize SAN boot to point toward the external storage unit. The HBA BIOS tells the hosts that its boot image is now located on the SAN.
6. Install the boot image onto the appropriate LUN.

Here we list the requirements and guidelines for SAN boot:

- SAN configuration, zoning of boot devices, and multipath configurations
- Active path to boot LUN
- Only one path to the boot LUN being enabled, prior to installing and enabling a multipath driver
- HBA BIOS selectable boot, or boot BIOS, must be enabled

These general steps can serve as a guide for how to configure SAN boot. It is important that you analyze additional prerequisites and specific directions according to the particular host, HBA, fabric, and storage subsystem you might have.

Important:

- When installing the operating system, you can only have one path to the storage device (LUN). Because servers are generally equipped with two HBAs, and because most DS5000 storage subsystems have two RAID controllers, you must isolate (disconnect) the second HBA or do an appropriate zoning.
- For Windows and Linux based operating systems, the boot LUN must be assigned as LUN 0 when doing storage partitioning.
- The boot LUN must be accessible only by the host that is utilizing it to boot, which can be achieved through the storage partitioning feature of the DS5000 storage subsystem.
- For iSCSI configuration, follow the vendor instructions for hardware initiators, make sure your firmware version is at required level; for software initiators study the online operating system documentation for mandatory prerequisites.

5.1.2 Installing local hard disk for high-load environments

It is always best to install the local hard disk for high-load environments. The operating system or applications can experience issues and become unstable due to latency with accessing the page file. Thus, having a page file on a local disk ensures reliable access to the page file. It becomes crucial especially in Windows server environments where we can identify highly swapping paging cache. Specifically, we elect to use the local disk for swap/RAS.

See the following Microsoft Knowledge Base (KB) articles on Boot from SAN issues:

- Microsoft SAN support (Technet):

  http://technet.microsoft.com/en-us/library/cc786214(WS.10).aspx

- Support for booting from a SAN:

  http://support.microsoft.com/default.aspx?scid=kb;en-us;305547

- Support for multiple clusters attached to the same SAN device:

  http://support.microsoft.com/kb/309395

5.1.3 Comparison: iSCSI and FCoE versus Fibre Channel

The iSCSI protocol is a transport layer for SCSI over TCP/IP. Until recently, the standard IP protocol infrastructure (100 MBps Ethernet) was not able to provide the enough bandwidth and less latency to accommodate storage traffic. Dedicated infrastructure with respective communication protocols such as Fibre Channel Protocol (SCSI over Fibre Channel), was developed to achieve high volume data interchange over Storage Area Networks. With the recent advances in Ethernet technology, it is now practical (from a performance and cost perspective) to access storage devices through an IP network. 10 Gigabit Ethernet is now widely available in many datacenters and the results are competitive to 4 and 8 Gbps Fibre Channel.

Similar to FC protocol, Fibre Channel over Ethernet (FCoE) and iSCSI allows storage to be accessed over a Storage Area Network, allowing shared access to the devices. Opposed to a dedicated TCP and FC networks, the investment into Converged Network utilizing Converged Network Adapters (CNA) and convergence-capable LAN switches or SAN directors, clients significantly reduce the management cost by operating single network, thus saving on power, cooling, and floor space in the expensive datacenters.The key advantage of iSCSI over FCP is that iSCSI can utilize standard, off-the-shelf Ethernet network components. In addition, the network, that incorporates iSCSI SAN only, exploits a single kind of network infrastructure only (1 Gbps or 10 Gbps Ethernet) for both data and storage traffic, whereas use of FCP requires a separate type of infrastructure (Fibre Channel) and administration for the storage. Furthermore, FCoE and iSCSI based SANs can expand over arbitrary distances, and are not subject to distance restrictions that currently limit FCP. This concept helps clients consolidate their strategic datacenters with remote branch offices or departments into the single, centrally managed infrastructure.

Because an iSCSI and FCoE are designed to run on an IP network, it takes the advantage of existing features and tools that were already in place for IP networks. Today's Ethernet network standards guarantee delivery of data and congestion control. Lossless Ethernet is one of the key requirements for implementation of storage networking on 10 Gbps IP-based networks. IPSec can be utilized to provide security for an iSCSI SAN, whereas a new security mechanism might need to be developed for the Fibre Channel. Service Location Protocol (SLP) can be used by iSCSI to discover iSCSI entities in the network. Thus, in addition to iSCSI running on standard, cheaper, off-the-shelf hardware, iSCSI also benefits from using existing, standard IP-based tools and services.

Tip: Fibre Channel security standards are generally available and defined by the Technical Committee T11 as a part of InterNational Committee for Information Technology Standards (INCITS):

http://www.t11.org/t11/hps.nsf/guesthome766kx?

Figure 5-3 shows an example of typical datacenter networking utilizing FC and Ethernet components separately, as opposed to the storage networking solutions that benefit from iSCSI or FCoE technology. As mentioned in the beginning of the chapter, we focus on FC and iSCSI configurations because FCoE is not natively supported by the DS5000 family at the time of writing. There are no FCoE HICs available yet.

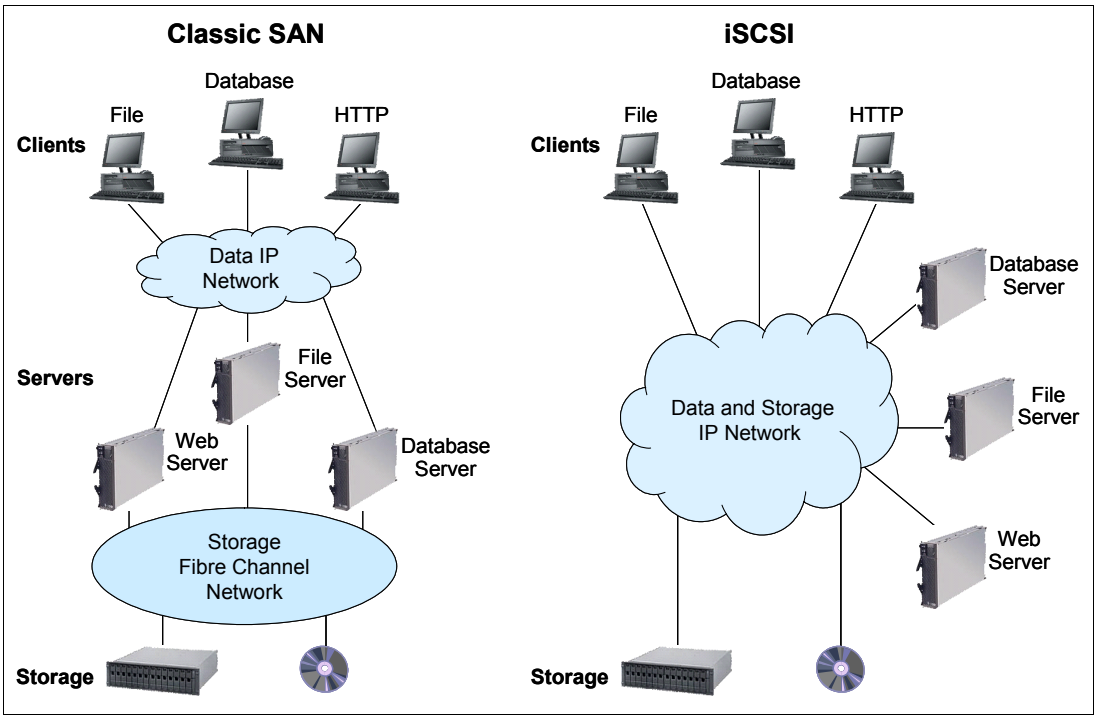


Figure 5-3 Typical FC SAN and an iSCSI SAN topology

Tip: Where the Boot from iSCSI or Fibre Channel SAN methods can serve and satisfy the requirements of several client environments, there are certain engineering and scientific applications that require the use of the local disk for better performance. Such applications might not be feasible for a Boot from iSCSI or Fibre Channel SAN solution:

- Nastran (linear finite element analysis codes for engineering)
- Dytran (non-linear finite element analysis)
- Marc (another non-linear code)
- Fluent (fluid dynamics)
- Gaussian (computation chemistry)
- Amber (computational chemistry)
- GCG (computational chemistry)

Therefore, you need to consider the performance requirements for each application installed on your systems and analyze if they are feasible for SAN boot.

5.1.4 iSCSI initiators

An iSCSI initiator can be either an iSCSI HBA installed in a host server, or you can configure a software iSCSI initiator by using an iSCSI device driver and an Ethernet network adapter. With the available 10 Gbps Ethernet and 10 Gbps iSCSI adapter, make sure that used cables satisfy the requirements for lossless IP-based storage networking.

Software initiators

A configuration that uses software initiators includes the following components:

- Microsoft iSCSI software initiator or equivalent: The Microsoft Server 2008 already has the iSCSI software initiator built in.
- One or two Ethernet cards: There is no iSCSI card required on the server to implement a connection to an iSCSI SAN environment.

For more information about iSCSI on Microsoft, visit the Microsoft support pages:

http://technet.microsoft.com/en-us/library/ee338474(WS.10).aspx

A similar concept applies for UNIX platforms:

- Majority of UNIX systems (AIX, Linux, HP-UX, and so on) have already built-in SW packages to support iSCSI, usually certain level of OS is required to best utilize all features. Consult OS product support for details.
- Minimum one 10 Gbps Ethernet card for 10 Gbps iSCSI emulation and 10 Gbps capable LAN switches. In the storage environments with required high data throughput, we do not suggest to utilize only a 1 Gbps network.

The software requirements for the most common UNIX platforms on IBM systems can be found at these websites:

- AIX:

  http://www.ibm.com/developerworks/aix/library/au-iscsi.html

- Linux Red Hat:

  http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/6/html/6.1_Technical_Notes/index.html#iscsi-initiator-utils

- Linux SuSE:

  http://www.suse.com/documentation/sles10/book_sle_reference/?page=/documentation/sles10/book_sle_reference/data/sec_inst_system_iscsi_initiator.html

Hardware initiators

A configuration that uses hardware initiators includes the following components:

- One or two iSCSI cards for each server, which access the storage array and associated drivers.

  For BladeCenters, you use the QLogic iSCSI Expansion Card for IBM BladeCenter. This iSCSI Expansion Card option is a hardware initiator that provides iSCSI communication from the blade server to an iSCSI storage device. It delivers full TCP/IP Offload Engine (TOE) functionality to reduce CPU processing. For more information, see the Web link:

  http://www.ibm.com/common/ssi/rep_ca/4/897/ENUS105-194/ENUS105-194.PDF

  System x servers currently support these iSCSI TOE adapters:

  - QLogic QLE8142-SR-CK Dual port 10 GBps iSCSI HBA for System x
  - Emulex OCe10102-IX-D Dual port 10 Gbps iSCSI HBA for System x

- QLogic QLE4060C 1 Gbps iSCSI Single Port PCIe HBA for IBM System x
- QLogic QLE4062C 1 Gbps iSCSI Dual Port PCIe HBA for IBM System x
- IBM iSCSI 1 Gbps Server TX Adapter (30R5201)
- IBM iSCSI 1 Gbps Server SX Adapter (30R5501)
- IBM iSCSI 1 Gbps Server Adapter (73P3601)

For more information, visit this website:

http://www.ibm.com/support/entry/portal/docdisplay?brand=5000008&lndocid=MIGR-57073

- One or two Ethernet switches, preferably using two Ethernet switches for redundancy.

TCP Offload Engine benefits

We know that the processing of TCP packets from an Ethernet connection consumes many processor resources, and iSCSI protocol only adds another layer of processing. With the number of packets and their corresponding interrupts required for iSCSI, the software iSCSI packet processing can burden the host system with 50–65% CPU usage. Depending upon signaling options used, high CPU usage might even render certain host applications unusable.

Therefore, it is important that efficient iSCSI systems depend on a hardware TCP Offload Engine (TOE) to handle the transportation protocols of iSCSI. A TOE network interface card (NIC) is a dedicated interface card specifically designed for interfacing a server to the IP-SAN including iSCSI offloading and additionally TCP/IP encapsulation from the server processors. A hardware TOE implements the entire standard TCP and iSCSI protocol stacks on the hardware layer. This approach completely offloads the iSCSI protocol from the primary processors, leveraging storage communications efficiently and enabling applications to run faster and more reliable. By using the TCP Offload Engine, a single system can run multiple initiators for improved throughput.

Choosing between hardware and software initiators

Using hardware initiators offers the following key benefits:

- Their performance is noticeably faster.
- They do not interfere with data networking traffic if the network topology is designed to segregate storage traffic such as by using a separate set of switches to connect servers to iSCSI storage.
- The traffic that passes through them will not load the server's CPU to the same extent that might be the case if the storage traffic passed through the standard IP stack.
- It is possible to implement *iSCSI boot from SAN* with hardware initiators.

Using software initiators offers the following key benefits:

- The cost of the iSCSI hardware is avoided.
- It is possible to use one set of switches for both data and storage networking, avoiding the cost of additional switches but possibly impacting performance.
- It is possible to access other network storage devices such as NAS, NFS, or other file servers using the same network interfaces as are used for iSCSI.

Attention: The iSCSI software initiator support is limited to Ethernet adapters. Software initiators with physical Converged Network Adapters (CNAs) are not supported.

5.2 SAN boot of AIX on IBM POWER systems

This section contains a step-by-step illustration of SAN boot implementation for the IBM POWER® (formerly IBM eServer™ pSeries) in an AIX 6.1 and AIX 7.1 environment managed from Korn Shell interpreter and its utilities. We do not include procedures valid for AIX GUI as its deployment is seen as very rare and only very minor number of system administrators operates it. We describe practical experience with setup of SAN boot using Fibre Channel protocol and iSCSI connection.

5.2.1 Implementation options

Implementation of SAN boot with AIX include the following options, common for both types of SAN connections:

- To implement SAN boot on a system with already installed AIX operating system, you have these options:
  - Use the alt_disk_install system utility.
  - Mirror an existing SAN Installation to several other LUNs using Volume Copy.
- To implement SAN boot on a new system, you have these options:
  - Start the AIX installation from a bootable AIX CD install package.
  - Use the Network Installation Manager (NIM).

The methods known as *alt_disk_install* or *mirroring* are simpler to implement than the more complete and more sophisticated method using the Network Installation Manager (NIM).

5.2.2 General prerequisites and considerations

Before implementing SAN boot, consider the following requirements and stipulations:

- Booting from a DS5000 storage subsystem utilizing SATA drives for the boot image is supported but is not preferred due to performance reasons.
- Boot images can be located on a volume in DS5000 storage subsystem partitions that have greater than 32 LUNs per partition if utilizing AIX release CDs 5.2 TL4 (5.2H) and 5.3.0 TL0 or higher. These CDs contain DS5000 storage subsystem drivers that support more than 32 LUNs per partition.
- When booting from a DS5000 storage subsystem, both storage controller fiber connections to the boot device must be up and operational. Single HBA configurations are supported, but must be zoned, enabling connection to both controllers.

  Single HBA configurations are allowed, but each single HBA configuration requires that both controllers in the DS5000 storage subsystem be connected to the host. In a switched environment, both controllers must be connected to the switch within the same SAN zone as the HBA.
- Path failover is not supported during the AIX boot process. After the AIX host has booted, failover operates normally.

5.2.3 AIX boot with iSCSI considerations

The IBM AIX supports booting from an iSCSI attached disk when running on POWER6 or later processors and using the iSCSI software initiator and IPv4. An iSCSI boot does not support IPv6. An iSCSI boot is supported on Power Blade systems using either the iSCSI software initiator or the iSCSI TOE daughter card using IPv4.

The iSCSI boot is supported with AIX version 5.3 or higher (6.1 and 7.1). Even if IPv6 support has been introduced in AIX 6.1, IPv6 still does not support booting from SAN disks using iSCSI protocol.

Beginning with AIX 5L with 5200-04, the iSCSI protocol driver (software initiator) is included as part of AIX base operating system. The iSCSI protocol allows the access of storage devices over gigabit Ethernet TCP/IP networks. The iSCSI support is in the filesets `devices.iscsi_sw.rte`, `devices.iscsi.disk.rte`, and `devices.common.IBM.iscsi.rte`. These filesets supersede the `iscsi_sw.rte` fileset that was previously included in the AIX Bonus Pack.

When you are booting the iSCSI software initiator, ensure that the Ethernet network is configured so that the link comes up without delay. After the Ethernet link is enabled, the AIX iSCSI software initiator will attempt to contact the iSCSI target for approximately 30 seconds before declaring that the boot disk cannot be found and indicating **“**`554 Unknown Boot Disk`**.”**. Some Ethernet protocols, such as spanning tree protocols, might prevent the link from coming up in 30 seconds and will cause boot failures. Such protocols must be disabled or overridden on the Ethernet switch if they prevent the Ethernet link from coming up in less than 30 seconds.

The iSCSI protocol driver follows the iSCSI standards specified in RFC3720[1], with these limitations:

- During installation, the iSCSI driver creates a default initiator name. However, this generated iSCSI name might not comply with the format specified by the iSCSI String Profile document. You can use the iSCSI SMIT panels (under **`smit iscsi`**) to change the initiator name to comply with the standard or to match local iSCSI name conventions.
- The iSCSI protocol driver can connect to a maximum of 16 unique targets at one time. If fewer targets are in use, you can change the Maximum Targets Allowed field in the SMIT panel to reduce memory usage by the iSCSI driver.
- This implementation of iSCSI supports only one TCP/IP connection per iSCSI session.
- This implementation of iSCSI supports login redirection to numeric IP addresses only. Any received login redirection that specifies a host name instead of a numeric IP address is considered a login failure. The login redirection cannot occur between IPv4 and IPv6 networks. If the original login was on an IPv4 network address, the redirection must be to an IPv4 network address as well.

The list of releases for the iSCSI driver fileset is available at IBM AIX support web pages. Use the following link to look at the history of releases and to determine the latest fileset versions available for each release of AIX:

`https://www.ibm.com/support/docview.wss?uid=isg1fileset-2118302142`

[1] The Internet Engineering Task Force (IETF) - `http://www.ietf.org/rfc/rfc3720.txt`

5.3 Windows 2008 SAN boot with Fibre Channel and iSCSI

Booting from SAN creates a number of possibilities that are not available when booting from local disks. It means that the operating systems and configuration of SAN based computers can be centrally stored and managed. It can provide advantages with regards to deploying servers, backup, and disaster recovery procedures. For x86/x64 based systems, both scenarios are available, Fibre Channel SAN boot and iSCSI SAN boot.

5.3.1 Configuration overview of FC SAN and iSCSI boot

To boot from SAN, you need to go into the HBA configuration mode, set the HBA BIOS to be Enabled, select at least one DS5000 target port, and select a LUN to boot from. In practice you will typically configure 2 DS5000 ports as targets and you might need to enable the BIOS on two HBAs, but this will depend on the HBA, driver, and operating system. Study the documentation that comes with your HBA and operating systems.

5.3.2 Example of FC SAN and iSCSI boot environment

In Figure 5-4 we show an example of the environment's configuration when booting from SAN is enabled using iSCSI Host Bus Adapters (HBA). The concrete procedures for setting up your server and HBA to boot from SAN can vary. They are mostly dependent on whether your server has an Emulex or QLogic HBA (or the OEM equivalent). The configuration panels of HBA BIOS mode might differ, however the principal process remains the same.

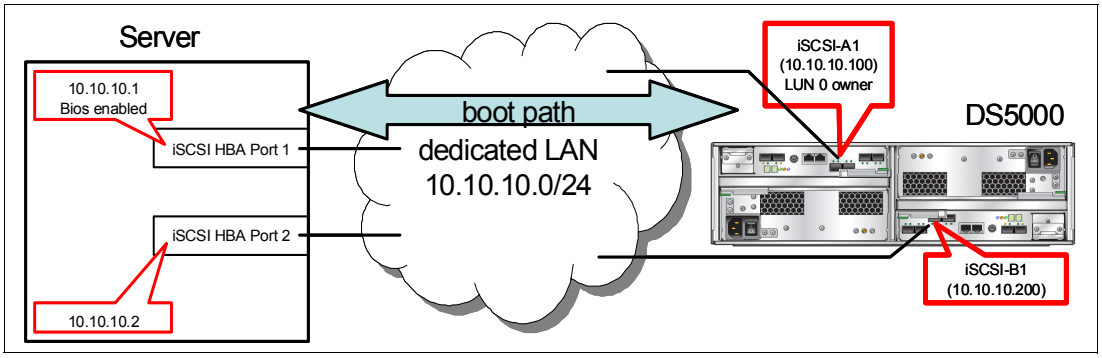


Figure 5-4 Boot from SAN - iSCSI topology

Make sure that your topology supports controller failover and also has information about the controller owning the boot LUN.

5.4 Linux SAN boot on IBM system x servers

To understand the configuration steps required to boot a Linux system from SAN-attached DS5000 volumes, you need a basic understanding of the Linux boot process. Therefore we briefly summarize the steps a Linux system goes through until it presents the login prompt.

1. The operating system loader:

   The system firmware provides functions for rudimentary input-output operations (for example, the BIOS of x86 servers). When a system is turned on, it first performs the Power On Self Test (POST) to check which hardware is available and whether everything is working. Then it runs the operating system loader (OS loader) that uses those basic I/O tines to read a specific location on the defined system disk and starts executing the code it contains. This code either is part of the boot loader of the operating system or it branches to the location where the boot loader resides.

If we want to boot from a SAN attached disk, we must make sure that the OS loader can access this disk. FC HBAs provide an extension to the system firmware for this purpose. In many cases it must be explicitly activated.

On x86 systems, this location is called the *Master Boot Record* (MBR).

2. The boot loader:

   The boot loader's purpose is to start the operating system kernel. To do it, the boot loader must know the physical location of the kernel image on system disk, read it, unpack if compressed, and start it. All of it is still done using the basic I/O routines provided by the firmware. The boot loader also can pass configuration options and the location of the `InitRAMFS` to the kernel. These Linux boot loaders are the most common:

   - `GRUB` (Grand Unified Boot Loader) on x86 systems
   - `yaboot` for Power systems
   - `zipl` for IBM System z®

3. The kernel and `InitRAMFS`:

   After the kernel is unpacked and running, it takes control over the system hardware. It starts and sets up the memory management, interrupt handling, and the built in hardware drivers for the hardware that is common on all systems (MMU, clock, and so on).

   It reads and unpacks the `InitRAMFS` image, again using the same basic I/O routines. The `InitRAMFS` contains additional drivers and programs that are needed to set up the Linux file system tree (root file system). To be able to boot from a SAN attached disk, the standard `InitRAMFS` must be extended with the FC or iSCSI HBA driver and the multipathing software. In modern Linux distributions, it is done automatically by the tools that create the `InitRAMFS` image.

   After the root file system is accessible, the kernel starts the `init()` process.

4. The `init()` process:

   The `init()` process brings up the operating system itself: networking, services, user interfaces, and so on. At this point the hardware is already completely abstracted. Therefore `init()` is not platform dependent, nor are there any SAN-boot specifics.

A detailed description of the Linux boot process for x86 based systems can be found on IBM Developerworks at this website:

`http://www.ibm.com/developerworks/linux/library/l-linuxboot/`

The practical instructions on how to set up Linux SAN boot are described in detail steps in *SAN Boot Implementation and Best Practices Guide for IBM System Storage*, SG24-7958.

5.4.1 Linux SAN boot considerations

Certain steps are similar to those presented in the Windows 2008 SAN boot section, but there are several key differences.

Tip: Always check the hardware and software, including firmware and operating system compatibility, before you implement SAN boot. Use the System Storage Interoperation Center (SSIC) website:

`http://www.ibm.com/systems/support/storage/config/ssic/`

The DS5000 still supports Linux with the RDAC driver only. There are particular versions of drivers for the various Linux kernel versions available. See each RDAC readme file for a list of compatible Linux kernels.

The following Linux kernels are supported with the current RDAC version 09.03.0C05.0214:

- SLES 11: 2.6.27.19-5
- SLES 10-SP2: 2.6.16.60-0.21
- RHEL5-u2: 2.6.18-92
- RHEL5-u3: 2.6.18-128

Consider the following precautions and guidelines when using RDAC with Linux:

- Always consult the RDAC `readme` file for a full list of limitations.
- Auto Logical Drive Transfer (ADT/AVT) mode is automatically enabled in the Linux storage partitioning host type. This mode causes contention when an RDAC is installed.

  If using the "Linux" host type, *it must be disabled* using the script that is bundled in this Linux RDAC Web package or in the `\Scripts` directory of this DS Storage Manager version 10 support for Linux CD. The name of the script file is `DisableAVT_Linux.scr`.

  If using the "LNXCLVMWARE" host type, you do not need to disable AVT, because this host type has AVT disabled by default. This host type is preferred for Linux with RDAC.
- The Linux SCSI layer does support skipped (sparse) LUNs with 2.6 and higher kernels. However, it is good practise that all mapped LUNs are contiguous.
- The guideline is to use multiple unconnected FC switches to zone the FC switch into multiple zones so that each HBA port sees only one controller in a DS5000 storage subsystem.
- The RDAC driver reports I/O failures immediately after all paths are failed. When the RDAC driver detects that all paths to a DS4000/DS5000 storage subsystem are failed, it will report I/O failure immediately. This behavior varies from the IBM Fibre Channel (FC) HBA failover device driver. The FC HBA failover device driver will wait for a certain time out/retry period before reporting an I/O failure to the host application. There is no work-around.

5.4.2 Linux SAN boot: Configuration overview

Here we present a summary of the Linux SAN boot configuration.

1. Server HBA configuration:
   a. Power on the server.
   b. Disable the local such as IDE, SAS, and SATA drives in the server BIOS.
   c. Record the WWPN and enable the BIOS of the first FC HBA using the FastT!Util menu.
2. SAN switches configuration:

   Ensure that the Fibre Channel SAN switches are correctly set, according to the following guidelines:

   a. Complete the necessary fiber cable connections with the DS5000 storage subsystem controllers, SAN switches, and server's HBAs, and ensure that all the devices are properly connected.
   b. Verify that all the switches in the fabric are configured with separate domain ID and IP addresses.
   c. Verify and confirm that all of the switches are running the latest and compatible firmware version.
   d. Define and activate zoning. Include only one HBA and one storage controller port in one zone.

3. DS5000 storage subsystem configuration for the primary path:
   a. Create a logical drive to be used as the operating system disk for the server.
   b. Complete the host group, host, and first host port definitions.
   c. Define the storage partition and map the logical drive with LUN ID 0 to the first host port on the host group.
4. Fibre Channel host configuration for primary path:

   Configure the boot device (FC BIOS parameter) for the primary path from the server to the DS5000 storage subsystem controller A by identifying and selecting the boot device.

   Tip: Verify that the displayed WWPN matches the WWPN of the DS5000 storage controller A zoned with the first FC HBA and the LUN ID=0.

5. Operating system installation:
   a. Install the Linux operating system, update the QLogic HBA driver, and install the RDAC driver and the optional storage management software.
   b. Verify that the server successfully boots from the logical drive on the primary path with a power off/on or by restarting the server.
6. Fibre Channel switched fabric configuration for the secondary path:
   a. Add another zone to include the second FC HBA of the server and the DS5000 storage subsystem controller B as members of that zone.
   b. Modify the active zone config to add the new zone.
   c. Enable the zone config.
   d. Verify that the active zone config lists the correct configuration.
7. DS5000 storage subsystem configuration for the secondary path:
   a. Define the host port for the second port of the HBA under the same host group used for the primary path configuration. The host type must be the same as for the first host port.
   b. Add the access logical drive with LUN ID = 31 (optional).
8. Verify and test access to the secondary path from the host:
   a. Check whether both controllers are visible from the host.
   b. Test path redundancy.

5.5 OS support for SAN boot

For a list of supported operating systems for SAN boot and their limitations, see the IBM System Storage Interoperation Center website:

http://www.ibm.com/systems/support/storage/config/ssic/

Table 5-1 shows a matrix for SAN boot supported OS and their supported HBAs.

Table 5-1 SAN boot Support matrix

Operating system	HBA
Windows 2003 -SP2 Windows 2008 -SP1 (including Vista and Hyper-V) Solaris 9 (SPARC only) Solaris 10 (SPARC and x86) Red Hat 4.7, 4.8 Red Hat 5.3 SLES 10.2 SLES 11 HPUX 10, 11	Supported with all HBAs except in IA64 systems
AIX 5.3 -TL10 AIX 6.1 -TL3 VMware 3.5-U3, 3.5-U4 VMware 4	Supported with all HBAs

6

DS5000 performance tuning

In this chapter, we consider performance topics as they apply to the IBM Midrange Storage Subsystems. Although the focus here is the IBM Midrange Storage Subsystem performance, we also describe various workload types generally observed in storage, as well as how their impact on performance can be addressed by the configuration and parameters.

With all Midrange Storage Subsystems, good planning and data layout can make the difference between having excellent workload and application performance, and having poor workload, with high response times resulting in poor application performance. It is therefore not surprising that first-time clients ask for advice on how to optimally layout their storage subsystems.

This chapter covers the following topics:

- Understanding workload
- Solution wide considerations
- Host considerations
- Application considerations
- Midrange storage subsystem considerations
- Performance data analysis
- Midrange storage subsystem tuning

6.1 Workload types

In general, you can expect to see two types of data workload (processing):

- Transaction based
- Throughput based

These two workloads are vary widely in their nature, and you need to plan for them quite differently. Knowing and understanding how your host servers and applications handle their workload is important to your success in your storage configuration efforts, and the resulting performance of the Midrange Storage Subsystem.

To best understand what is meant by transaction based and throughput based, we must first define a workload. The workload is the total amount of work that is performed at the storage subsystem, and is measured through the following formula:

```
Workload = [transactions (number of host IO's sent)] * [ throughput (amount of
data sent in all the IO's)] for the measured timeframe
```

Knowing that a storage subsystem can sustain a given maximum workload, we can see with the previous formula that if the number of host transactions increases, then the throughput must decrease. Conversely, if the host is sending large volumes of data with each I/O, the number of transactions must decrease.

A workload characterized by a high number of transactions (IOPS) is called a *transaction based* workload. A workload characterized by large I/Os is called a *throughput based* workload. These two workload types are conflicting in nature, and consequently work best with a wide range of configuration settings across all the pieces of the storage solution.

But first, let us describe each type of processing in greater detail, and explain what you can expect to encounter in each case.

6.1.1 Transaction based processes (IOPS)

High performance in transaction based environments cannot be created with a low cost model (with a small quantity of physical drives) of a storage subsystem. Transaction intense applications frequently use a small *random* data block pattern to transfer data. Read cache is far less effective, and the misses need to be retrieved from disk.

This results in these applications being heavily dependent on the number of back-end drives that are available for parallel processing of the host's workload. With too few drives, high queuing with longer response times will be encountered.

Determining at what point the performance can be accepted is important in deciding how large a subsystem is needed. Willingness to have higher response times can help in reducing the cost, but might not be acceptable by the user community. Cost and this factor frequently result in decisions needing to be made on just how many drives will be enough.

In many cases, slow transaction performance problems can be traced directly to "hot" files that cause a bottleneck on a particular critical component (such as a single physical disk). This situation can occur even when the overall storage subsystem is seeing a fairly light workload. When bottlenecks occur, they can present a very difficult and frustrating task to resolve.

Because workload content can be continually changing throughout the course of the day, these bottlenecks can be very mysterious in nature and appear and disappear, or move from one location to another over time.

Generally, I/O and therefore application performance are best when the I/O activity is evenly spread across most, if not all, of the I/O subsystem.

6.1.2 Throughput based processes (MBps)

Throughput based workloads are seen with applications or processes that require massive amounts of data sent, and frequently use large *sequential* blocks to reduce disk latency. Throughput rates are heavily dependent on the storage subsystem's internal bandwidth. Generally, a much lower number of the drives are needed to reach maximum throughput rates with the storage subsystem. This number can vary between the various members of the midrange family and their respective bandwidth handling capabilities. Newer storage subsystems with broader bandwidths are able to reach higher numbers and bring higher rates to bear. In this environment, read operations make use of the cache to stage greater chunks of data at a time, to improve the overall performance.

6.1.3 Optimizing both workload types

With the Midrange Storage Subsystem, these two workload types have various parameter settings that are used to optimize their specific workload environments. These settings are not limited strictly to the storage subsystem, but span the entire solution being used. With care and consideration, it is possible to create an environment of very good performance with both workload types and share the same Midrange Storage Subsystem. However, it must be understood that portions of the storage subsystem configuration will be tuned to better serve one workload over the other.

For maximum performance of both workloads, consider using two separate smaller storage subsystems each tuned for their specific workload, rather than one large server being shared. When this model is not financially feasible, a DS5100 or DS5300 can be used with planning and careful configuring of the solution to gain very good performance for both of the mixed workloads with a single subsystem solution.

6.2 Solution-wide considerations for performance

Considering the various pieces of the solution that can impact performance, we first look at the host and the operating systems settings, as well as how volume managers come into play. Then we look at the applications; what their workload is like, as well as how many particular types of data patterns they might need to use and that we must plan for.

Of course, we also look at the Midrange Storage Subsystem and the configuration and many parameter settings that must be considered in accordance to the environment where the storage subsystem is deployed. And finally, we look at specific SAN settings that can affect the storage environment as well.

When looking at performance, we must first consider the location of the data in three areas:

- With regard to the path that the host uses to access it: Here we consider the host volume manager, HBA, SAN fabric, and the storage subsystem controller used in accessing the logical drive. Many performance issues stem from mis-configured settings in these areas.
- Within the storage subsystem on its configured array and logical drive: We check that the array has been laid out to give best performance, and the RAID type is the most optimal for the type of workload. Also, if multiple logical drives reside in the same array, we check for interference from other logical drive members when doing their workload.

- On the back-end drives that make up the array: We consider how the data is carved up (segmented and striped) across all the members of the array, including the number of drives being used, as well as their size, and speed. This area can have a great deal of impact that is very application specific, and in many cases can be addressed by tuning to get the best results.

In the following sections, we describe each of these areas.

6.3 Host considerations

In regard to performance, we need to consider far more than just the performance of the I/O workload itself. Many settings within the host frequently affect the overall performance of the system and its applications. We must check all areas to ensure that we are not focusing on a result rather than the cause. However, in this book we are concerned with the I/O subsystem part of the performance puzzle; so we describe items that affect its operation.

Certain settings and parameters described in this section are defined and must match both for the host operating system and for the HBAs that are being used as well. Many operating systems have built-in definitions that can be changed to enable the HBAs to be set to the new values. In this section, we present two examples using AIX and Windows operating systems as illustrations of these concepts.

6.3.1 Host based settings

Certain host operating systems can set values for the Midrange Storage Subsystem logical drives assigned to them. For instance, hosts can change the *write cache* and the *cache read-ahead* values through attribute settings. These settings can affect both the transaction and throughput workloads. Settings that affect cache usage can have a great impact in most environments.

High transaction environments

Other host device attributes that can affect high transaction environments are those affecting the *blocksize* and *queue depth* capabilities of the logical drives:

- The blocksize value used by the host I/O helps to determined the best *segment size* choice. Set the segment size at least to twice the size of the I/O blocksize being used by the host for the high transaction workload types.
- The queue depth value cannot exceed the storage subsystem maximum of 4096 for Midrange Storage Subsystems running firmware 7.1x and later; and a maximum of 2048 for firmware 6.60.x. All logical drives on the storage subsystem must share these queue limits. Certain hosts define the queue depth only at the HBA level, whereas others might also define this limit at the device level, which ties to the logical drive. You can use the following formulas to determine a good starting point for your queue depth value on a per logical drive basis:
  - For firmware level 7.1x and higher: 4096 / *(number-of-hosts * logical drives-per-host)*
  - For firmware level 6.60.xx: 2048 / *(number-of-hosts * logical drives-per-host)*

  As an example, a storage subsystem with 4 hosts with 12, 14, 16, and 64 logical drives attached respectively can be calculated as follows:

  4096 / 4 * 64 *(largest number of logical drives per host)* = 16
  2048 / 4 * 64 *(largest number of logical drives per host)* = 8

  If configuring only at the HBA level, for the respective firmware levels, use the formula:

  4096 / *(total number-of-HBAs)*, and 2048 / *(total number-of-HBAs)*

Important: Setting queue depth too high can result in high IO queuing, slow response times, busy being sent to the host, possible loss of data, and possible file corruption; therefore, being conservative with these settings is better than pushing the limits.

In the high throughput environments, we are interested in settings that affect the large I/O blocksize. Try to use a host I/O blocksize that is equal to, or an even multiple of, the stripe width of the logical drive being used.

Also check for settings mentioned earlier that might force a cache read-ahead value from the host. Ensure that the cache read-ahead value is being enabled. We describe this function in detail later in this chapter; but certain operating system environments might have a variable value for changing it through device settings. Allowing the Midrange Storage Subsystem to manage this setting is the best method for handling.

Finally, there are settings that might also impact performance with various host servers and HBA types that enhance FC tape support. This setting must not be used with FC disks attached to the HBA.

Best practice: As supported in theory, it is always best to keep the *Fibre Channel tape* and *Fibre Channel disks* on separate *HBAs*. These devices have two very unique data patterns when operating in their optimal mode, and the switching between them can cause undesired overhead and performance slowdown for the applications.

Host data layout

When possible, ensure the host operating system aligns its device data partitions or slices, with those of the logical drive. Misalignment can result in multiple back-end drive IOs, referred to as *boundary crossings,* which are responsible for unnecessary multiple drive I/Os. Certain operating systems do it automatically, and you just need to know the alignment boundary they use. Others, however, might require manual intervention to set their start point to a value which can align them.

Understand how your host based volume manager (if used) defines and makes use of the logical drives after they are presented as an important part of the data layout. As an example, the AIX Logical Volume Manager (LVM) is described in 2.6.1, “Planning for systems with LVM: AIX example” on page 69).

Volume managers are generally set up to place logical drives into groups for their use. The volume manager then creates volumes by carving up the logical drives into *partitions* (sometimes referred to as a *slice*); and then building a volume from them by either striping, or concatenating them to form the volume size desired. How the partitions are selected for use and laid out can vary from system to system. In all cases, you need to ensure that spreading of the partitions is done in a manner to achieve maximum I/Os available to the logical drives in the group.

Generally, large volumes are built across a number of separate logical drives to bring more resources to bear. The selection of logical drives when doing it must be made carefully so as not to use logical drives that will compete for resources, and degrade performance. Important rules to follow when designing and building these volumes are as follows:

- Ensure that the logical drives are selected from *separate array groups*, and the *preferred paths are spread across various controllers*, thus the logical drives will not be in conflict with each other when the volume is used and both slices are accessed. See 6.5.6, “Arrays and logical drives” on page 275 for further details.

Best practice: For best performance, when building (host) volumes from a single storage subsystem using any RAID type, use logical drives from various groups of arrays with their preferred paths evenly spread between the two controllers of the Midrange Storage Subsystem.

- If striping is used, ensure that the stripe size chosen is a value that is complementing the size of the underlying *stripe width* defined for the logical drives (see “Logical drive segments” on page 284). The value used here will be dependent on the application and host I/O workload that will be using the volume. If the stripe width can be configured to sizes that complement the logical drives stripe; then benefits can be seen with using it. In most cases this model requires larger stripe values and careful planning to properly implement.

  The exception is for single threaded application processes with sequential I/O streams that have a high throughput requirement. In this case a small LVM stripe size of 64 K or 128 K can allow for the throughput to be spread across multiple logical drives on multiple arrays and controllers, spreading that single I/O thread out, and potentially giving you better performance. This model is generally not preferred for most workloads as the small stripe size of LVM can impair the storage subsystem’s ability to detect and optimize for high sequential I/O workloads.
- With file systems, you again will want to revisit the need to ensure that they are aligned with the volume manager or the underlying RAID. Frequently, an offset is needed to ensure that the alignment is proper. Focus here on avoiding involving multiple drives for a small I/O request due to poor data layout. Additionally, with certain operating systems, you can encounter interspersing of file index (also known as an inode) data with the user data, which can have a negative impact if you have implemented a *full stripe write* model. To avoid this issue, you might want to use raw devices (volumes) for full stripe write implementations.

6.3.2 Host setting examples

The following example settings can be used to start off your configuration in the specific workload environment. These settings are guidelines, and they are not guaranteed to be the answer to all configurations. Always try to set up a test of your data with your configuration to see if there is further tuning that might be of more help (see Chapter 8, “Storage Manager Performance Monitor” on page 343 for more information). Again, knowledge of your specific data I/O pattern is extremely helpful.

AIX operating system settings

The following section outlines the settings that can affect performance on an AIX host. We look at these in relation to how they impact the two particular workload types.

All attribute values that are changeable can be changed using the **chdev** command for AIX. See the AIX man pages for details on the usage of **chdev**.

Transaction settings

Early AIX driver releases allowed for the changing of the *cache read-ahead* through the logical drive attribute settings, but has been discontinued with current releases. Cache read-ahead is now set by the storage subsystem value, and can only report the value from the operating system.

One of the most common settings that has been found to need adjusting is the queue_depth. This value can be increased to help improve high transaction based applications to keep the storage subsystem always busy.

However, if set too high, the result can be very high response time values, IO retries, and application slowdowns. Changing this setting to higher values must be done in steps of smaller changes to see where the peak of the desired performance is. Monitoring the performance over time can also indicate that further changes (up or down) might be necessary. To make any change to this value for the logical drive, the change is performed on the hdisk in AIX, and the setting is the attribute queue_depth:

```
# chdev -l hdiskX -a queue_depth=Y -P
```

In the previous example, *"X"* is the hdisk number, and *"Y"* is the value for the queue_depth that you are setting it to.

Another setting of interest is for an HBA setting of the attribute num_cmd_elem for the fcs device being used. This value must not exceed 512:

```
chdev -l fcsX -a num_cmd_elem=256 -P
```

Best practice: For high transactions on AIX, set num_cmd_elem to 256 for the fcs devices being used.

Throughput based settings

In the throughput based environment, you might want to decrease the queue depth setting to a smaller value such as 16. In a mixed application environment, you do not want to lower the "num_cmd_elem" setting, because other logical drives might need this higher value to perform. In a pure high throughput workload, this value will have no effect.

AIX settings which can directly affect throughput performance with large I/O blocksize are the lg_term_dma, and max_xfer_size parameters for the fcs device.

Best practice: The best start values for high throughput sequential I/O environments are lg_term_dma = 0x800000 and max_xfr_size = 0x200000**.**

Note that setting the max_xfer_size affects the size of a memory area used for data transfer by the adapter. With the default value of max_xfer_size=0x100000, the area is 16 MB in size, and for other allowable values of max_xfer_size, the memory area is 128 MB in size.

Also see 12.6, "HBA and device settings" on page 557.

AIX LVM impact

AIX uses Logical Volume Manager (LVM) to manage the logical drives and physical partitions. By default, with standard and big VGs, LVM reserves the first 512 bytes of the volume for the *Logical Volume Control Block*. Therefore, the first data block will start at an offset of 512 bytes into the volume. Care must be taken when laying out the segment size of the logical drive to enable the best alignment. You can eliminate the Logical Volume Control Block on the LV by using a scalable VG, or by using the -T 0 option for big VGs.

Additionally, in AIX, file systems are aligned on a 16 K boundary. Remembering these two items helps when planning for AIX to fit well with the DS5000 segment size. JFS and JFS2 file systems intersperse inode data with the actual user data, and can potentially disrupt the *full stripe write* activity. To avoid this issue, you can place files with heavy sequential writes on raw logical volumes. See also the guidelines defined in "Logical drive segments" on page 284.

With AIX LVM, it is generally best to spread high transaction logical volumes across the multiple logical drives that you have chosen, using the maximum interpolicy setting (also known as maximum range of physical volumes) with a random ordering of PVs for each LV.

Ensure that your logical drive selection is done as explained previously, and is appropriate for the RAID type selected.

In environments with very high rate, sequentially accessed structures and a large I/O size, try to make the segment size times the (N-1 for RAID 5, or N/2 for RAID 10) to be equal to the application I/O size. And keep the number of sequential I/O streams per array to be less than the number of disks in the array.

Windows operating system settings

In this section, we describe settings for performance with the Windows operating system and the DS5000 storage subsystem, including the following topics:

- Fabric settings
- Windows disk types
- Disk alignment
- Allocation unit size

Fabric settings

With Windows operating systems, the queue depth settings are the responsibility of the host adapters, and are configured through the BIOS setting, varying from vendor to vendor. See your manufacturer's instructions on how to configure your specific cards.

For IBM FAStT FC2-133 (and QLogic based HBAs), the queue depth is known as *execution throttle*, which can be set with either the QLogic SANsurfer tool, or in the BIOS of the QLogic based HBA, by pressing CTL+Q during the boot process.

Disk types: Basic disks and dynamic disks

With Windows 2000 and later versions, there are two types of disks, basic disks and dynamic disks. By default, when a Windows system is installed, the basic disk is used. Disks can be changed from basic to dynamic at any time without impact on system or data.

Basic disks use partitions. These partitions are created at the size defined. With Windows 2003 and later, a primary partition on a basic disk can be extended using the **`extend`** command in the diskpart.exe utility.

In Windows 2000 and later, dynamic disks allowed for expansion, spanning, striping, software mirroring, and software RAID 5 to be implemented.

With the Midrange Storage Subsystem, you can use either basic or dynamic disks. The appropriate type depends on your individual circumstances:

- In large installations where you might have the requirement to span or stripe logical drives and controllers to balance the workload, then dynamic disks might be your only choice.
- For smaller to mid-size installations, you might be able to simplify and just use basic disks.

When using the Midrange Storage Subsystem, the use of software mirroring and software RAID 5 is not required. Instead, configure the storage subsystem for the RAID redundancy level required for protection needs.

Basic disks and basic volumes are the storage types most often used with Windows operating systems. Basic disk is the default disk type during initial installation. A basic disk refers to a disk that contains basic volumes, such as primary partitions and logical drives. A basic volume refers to a partition on a basic disk. Basic disks are used in both x86-based and Itanium-based computers.

Basic disks support clustered disks. Basic disks do not support spanning, striping, mirroring and software level RAID 5. To use these functions, you must convert the basic disk to a dynamic disk. If you want to add more space to existing primary partitions and logical drives, you can extend the volume using the **extend** command in the diskpart utility.

With dynamic disks, you use the Disk Management GUI utility to expand logical drives.

Dynamic disks offer greater flexibility for volume management because they use a database to track information about dynamic volumes on the disk, they also store information about other dynamic disks in the computer. Because each dynamic disk in a computer stores a replica of the dynamic disk database, you can repair a corrupted database on one dynamic disk by using the database from another dynamic disk in the computer.

The location of the database is determined by the partition style of the disk:

- On MBR disks, the database is contained in the last 1 megabyte of the disk.
- On GPT disks, the database is contained in a 1 MB reserved (hidden) partition known as the Logical Disk Manager (LDM) Metadata partition.

All online dynamic disks in a computer must be members of the same disk group, which is a collection of dynamic disks. A computer can have only one dynamic disk group, called the primary disk group. Each disk in a disk group stores a replica of the dynamic disk database. A disk group usually has a name consisting of the computer name plus a suffix of `Dg0`.

To determine the best choice of disk type for your environment, see the following website:

`http://support.microsoft.com/kb/816307/`

Disk alignment

With the midrange logical drives, as with physical disks that maintain 64 sectors per track, the Windows 2003 and earlier operating systems always create the partition starting with the sixty-fourth sector. With Windows 2008, it has changed somewhat with new alignment practices being introduced that use an alignment default of 1024 KB (2048 sectors), thus allowing all current segment size offerings to align with the partition offset being used. However, with the earlier Windows releases, it is important to ensure that the alignment of the partition and the logical drive are the same; failure to align can result in added processing being required for handling the host IO. To ensure that they are both aligned, you can use the **diskpar.exe** or **diskpart.exe** command, to define the start location of the partition.

The **diskpar.exe** command is part of the Microsoft Windows Server 2000 Resource Kit; it is for Windows 2000 and Windows 2003. The **diskpar.exe** functionality was put into **diskpart.exe** with Windows Server 2003 Service Pack 1 and is included with Windows Server 2008.

Using this tool, you can set the starting offset in the Master Boot Record (MBR) by selecting the number of 512 byte sectors desired to be offset. By setting this value to 128, you will skip the first 64 K before the start of first partition. If you set this value to 256, you will skip the first full 128 K (where the MBR resides) before the start of the partition. The setting that you define should depend on the segment size defined for the logical drive being used.

Doing so ensures track alignment and improves the performance. Other values can be defined, but these two offer the best chance to start out with the best alignment values. As a best practice, you must ensure that you aligned to a minimum of one segment size boundary for complete alignment. Failure to do so can cause a single I/O operation to require the storage subsystem to perform multiple I/O operations on its internal processing, causing extra work for a small host I/O, and resulting in performance degradation.

Important: The use of diskpart is a data destructive process. The diskpart utility is used to create partitions with the proper alignment. When used against a disk that contains data, all the data and the partitions on that disk must be wiped out, before the partition can be recreated with the storage track boundary alignment. Therefore, if the disk on which you will run diskpart contains data, you must back up the disk before performing the following procedure.

For more information, see the following website:

http://www.microsoft.com/downloads/details.aspx?FamilyID=5B343389-F7C9-43D0-9892-DDCF55890529&displaylang=en&displaylang=en

In Example 6-1, we align disk 3 to the 256th sector.

Example 6-1 Using diskpart in Windows for aligning the disks

```
Microsoft Windows [Version 6.0.6002]
Copyright (c) 2006 Microsoft Corporation. All rights reserved

C:\Users\Administrator>diskpart

Microsoft DiskPart version 6.0.6002
Copyright (C) 1999-2007 Microsoft Corporation.
On computer: TC-W2008

DISKPART> select disk 3

Disk 3 is now the selected disk.

DISKPART> create partition primary align 256

DiskPart succeeded in creating the specified partition.

DISKPART>
```

Allocation unit size

An allocation unit (or cluster) is the smallest amount of disk space that can be allocated to hold a file. All file systems used by Windows 2000, and Windows 2003 organize hard disks based on an allocation unit size, which is determined by the number of sectors that the cluster contains. For example, on a disk that uses 512-byte sectors, a 512-byte cluster contains one sector, whereas a 4 KB cluster contains eight sectors. See Table 6-1.

Table 6-1 Default cluster sizes for volumes

Volume size	Default NTFS allocation unit size
7 MB to 512 MB	512 bytes
513 MB to 1024 MB	1 KB
1025 MB to 2 GB	2 KB
2 GB to 2 TB	4 KB

In the Disk Management snap-in, you can specify an allocation unit size of up to 64 KB when you format a volume. If you use the format command to format a volume, but do not specify a an allocation unit size by using the /a:size parameter, the default values shown in Table 6-1 are used. If you want to change the cluster size after the volume is formatted, you must reformat the volume and select the new allocation size from the drop-down box selection list, which must be done with care, as all data is lost when a volume is formatted. The available allocation unit sizes when formatting are 512 bytes, 1 K, 2 K, 4 K, 8 K, 16 K, 32 K, and 64 K.

Best practice: When creating a disk partition for use with Microsoft SQL Server, format the partition with an allocation unit size of 64 K.

Important: The allocation unit size is set during a format of a volume. This procedure is data destructive, so if the volume on which you will run a format contains data, you must back up the volume before performing the format procedure.

Restriction: In Windows 2000 and Windows 2003, setting an allocation unit size larger than 4 K will disable file or folder compression on that partition.

In Windows 2000, the disk defragmenter ceased to function with an allocation size greater than 4 K. In Windows 2003 and 2008, the disk defragmenter functions correctly.

Always check the documentation for your environment and test to ensure that all the functionality remains after any changes.

For more information about formatting an NTFS disk, see the Microsoft Windows documentation for your specific environment's release.

6.4 Application considerations

When gathering data for planning from the application side, again it is important to first consider the workload type for the application.

If multiple applications or workload types will be sharing the system, you need to know the type of workloads each have; and if mixed (transaction and throughput based) which will be the most critical. Many environments have a mix of transaction and throughput workloads; with generally the transaction performance being considered the most critical.

However, in dedicated environments (for example, a TSM backup server with a dedicated Midrange Storage Subsystem attached), the streaming high throughput workload of the backup itself is the critical part of the operation; and the backup database, though a transaction centered workload, is the less critical piece.

6.4.1 Transaction environments

Applications that use high transaction workloads are OnLine Transaction Processing (OLTP), mostly databases, mail servers, Web servers, and file servers.

If you have a database, you can tune the server type parameters as well as the database's logical drives to meet the needs of the database application. If the host server has a secondary role of performing nightly backups for the business, you need another set of logical drives that are tuned for high throughput for the best backup performance you can get within the limitations of the mixed storage subsystem's parameters.

So, what are the traits of a transaction based application? In the following sections, we explain this concept in more detail.

As mentioned earlier, you can expect to see a high number of transactions and a fairly small block size. Various databases use particular I/O sizes for their logs (see the following examples), these vary from vendor to vendor. In all cases the logs are generally high write workloads. For table spaces, most databases use between a 4 KB and 32 KB blocksize. In certain applications, larger chunks (for example, 64 KB) will be moved to host application cache memory for processing. Understanding how your application is going to handle its I/O is critical to laying out the data properly on the storage subsystem.

In many cases the table space is generally a large file made up of small blocks of data records. The records are normally accessed using small I/Os of a *random* nature, which can result in about a 50% cache miss ratio.

For this reason, and to not waste space with unused data, plan for the storage subsystem to read and write data into cache in small chunks. Avoid also doing any cache read-ahead with the logical drives, due to the random nature of the I/Os (Web servers and file servers frequently use 8 KB as well, and generally follow these rules as well).

Another point to consider is whether the typical I/O is read, or write. In most OLTP environments, it is generally seen to be a mix of about 70% reads and 30% writes. However, the transaction logs of a database application have a much higher write ratio, and as such, perform better in another RAID array. This reason also adds to the need to place the logs on a separate logical drive, which for best performance, must be located on a separate array that is defined to better support the heavy write need. Mail servers also frequently have a higher write ratio than read. Use the RAID array configuration for your specific usage model, which is covered in detail in “RAID array types” on page 276.

Best practice: Database table spaces, journals, and logs must never be co-located on the same logical drive or RAID array. See 6.5, “Midrange storage subsystem considerations” on page 269 for more information about RAID types to use.

6.4.2 Throughput environments

With throughput workloads, you have fewer transactions, but much larger I/O sizes, normally 128 K or greater; and these I/Os are generally of a sequential nature. Applications that typify this type of workload are imaging, video servers, seismic processing, high performance computing (HPC), and backup servers.

With large size I/O, it is better to use large cache blocksizes to be able to write larger chunks into cache with each operation. Ensure that the storage subsystem is configured for this type of I/O load, covered in detail in “Cache blocksize selection” on page 273.

Such environments work best when defining the I/O layout to be equal to or an even multiple of the storage subsystems *stripe width*. There are advantages with writes in the RAID 5 configurations that make setting the I/O size to equal that of the *full stripe* performant. See Figure 6-1 on page 277 for the amount of read operations that will be required with writes to fewer drives. Generally, the intent here is to make the sequential I/Os take as few back-end operations as possible, and to get maximum throughput from them. So, care must be taken when deciding how the logical drive will be defined. We describe these choices in greater detail in “Logical drive segments” on page 284.

Another application consideration is when storing sequentially accessed files separately to use larger segment size on the LUNs to help keep the disk heads in position for the sequential I/Os, which reduces the amount of time required to seek the data. In this type environment, using a 256 KB or larger segment size can be beneficial.

6.4.3 Application examples

For general guidelines and tips to consider when implementing certain applications with the Midrange Storage Subsystems, see Chapter 7, “IBM Midrange Storage Subsystem tuning with typical applications” on page 309.

6.5 Midrange storage subsystem considerations

In this section, we look at the specific details surrounding the Midrange Storage Subsystem itself when considering performance planning and configuring. We cover the following topics:

- Which model fits best
- Storage subsystem processes
- Storage subsystem modification functions
- Storage subsystem parameters
- Disk drive types
- Arrays and logical drives
- Additional NVSRAM parameters of concern

6.5.1 Which model fits best

When planning for a Midrange Storage Subsystem, the first thing to consider is the choice of an appropriate model for your environment and the type of workload that it will handle.

If your workload is going to require a high number of disks, you want to be sure that the system chosen will support them, and the I/O workload that they will be processing. If the workload you have requires a high storage subsystem bandwidth, then you want your storage subsystem to have a data bus bandwidth in line with the throughput and workload needs.

Table 1-1 on page 15 shows details of the IBM Midrange Storage Subsystem family comparison.

In addition to doing your primary processing work, you might also have various storage subsystem background processes you want to run. All work being performed requires resources, and you need to understand how they all will impact your storage subsystem.

6.5.2 Storage subsystem processes

When planning for the system, remember to take into consideration any background processes, additional premium features, and background management utilities that you are planning to implement with the storage solution.

Midrange copy services

With the Midrange Storage Subsystem, a complete suite of copy services features is available. All of these features run internally on the storage subsystem, and therefore use processing resources. It is important that you understand the amount of overhead that these features require, and how they can impact your primary I/O processing performance.

Enhanced Remote Mirroring

Enhanced Remote Mirroring (ERM) is a critical back-end process to consider, as in most cases, it is expected to be constantly running with all applications while it is mirroring the primary logical drive to the secondary location. After the initial synchronization is complete, we have continuous mirroring updates that will run. These updates can be performed by metro mirror (synchronous), global mirror (asynchronous with write order consistency (WOC)) or global copy (asynchronous without WOC) methods of transfer. For further details on these methods and their specific impacts, see *IBM Midrange System Storage Copy Services Guide*, SG24-7822.

When planning to use ERM with any application, you must carefully review the data workload model of the production environment. With ERM feature implemented there is about an initial 25% overhead that can be expected on the production subsystem. Various additional processes required to establish and re-synchronized mirrored pairs can add further cost while they are running as well. There is also a requirement to dedicate the last host side Fibre Channel port from each controller to the mirroring network alone.

Another major factor that can be of impact here is the networking path followed by the remote data. Bandwidth is a critical factor in whether or not the ERM solution will be successful. Making sure that the path is not a bottleneck to the passing of IO traffic between the two mirrors becomes crucial to the success of both metro mirroring and global mirroring methods. A restricted network can make any ERM solution behave poorly and, in certain cases, can result in production side slowdowns of performance.

The DS5300 is the best choice when ERM is a strategic part of your storage plan.

Consider also the impact of the initial synchronization overhead. When a storage subsystem logical drive is a primary logical drive and a full synchronization is necessary, the controller owner performs the full synchronization in the background while processing local I/O writes to the primary logical drive and associated remote writes to the secondary logical drive. Because the full synchronization diverts controller processing resources from I/O activity, it will impact performance on the host application. The *synchronization priority* allows you to define how much processing time is allocated for synchronization activities relative to other system work so that you can maintain accepted performance.

The synchronization priority rates are lowest, low, medium, high, and highest.

Tip: The lowest priority rate favors system performance, but the full synchronization takes longer. The highest priority rate favors full synchronization, but system performance can be compromised.

Understanding the timing of these various priorities can be critical to the planning of the disaster recovery model being used for the environment. It might be better to plan for synchronizing of the mirrored pair during a low usage period when highest priority can be used to ensure the recovery point and time objectives can be met. See *IBM Midrange System Storage Copy Services Guide*, SG24-7822 for a description and guidelines on the using the various priority settings.

The synchronization progress bar at the bottom of the Mirroring tab of the logical drive Properties dialog box displays the progress of a full synchronization.

VolumeCopy function

With VolumeCopy, several factors contribute to system performance, including I/O activity, logical drive RAID level, logical drive configuration (number of drives in the array or cache parameters), and logical drive type.

When copying from a FlashCopy, logical drives might take more time to copy than standard logical drives. A major point to consider is whether you want to be able to perform the function while the host server applications are functioning, or during an outage. If the outage is not desired, using the FlashCopy volume as the source will allow you to perform your VolumeCopy in the background while normal processing continues.

Like the ERM functions, VolumeCopy does have a background process penalty which you need to decide how much you want it to affect your front-end host. With the FlashCopy image you can use lower priority and leave it run for more extended time, which will make your VolumeCopy creation take longer but decrease the performance hit. Because this value can be adjusted dynamically, you can increase it when host processing is slower.

You can select the copy priority when you are creating a new logical drive copy, or you can change it later using the Copy Manager. The copy priority rates are lowest, low, medium, high, and highest.

Tip: The lowest priority rate supports I/O activity, but the logical drive copy takes longer. The highest priority rate supports the logical drive copy, but I/O activity can be affected.

FlashCopy function

If you no longer need a FlashCopy logical drive, you need to disable it. As long as a FlashCopy logical drive is enabled, your storage subsystem performance is impacted by the copy-on-write activity to the associated FlashCopy repository logical drive. When you disable a FlashCopy logical drive, the copy-on-write activity stops.

If you disable the FlashCopy logical drive instead of deleting it, you can retain it and its associated repository. Then, when you need to create another FlashCopy of the same base logical drive, you can use the re-create option to reuse a disabled FlashCopy, which takes less time.

6.5.3 Storage subsystem modification functions

The Midrange Storage Subsystems have many modification functions that can be used to change, tune, clean, or redefine the storage dynamically. Various functions help improve the performance as well. However, all of these will have an impact on the performance of the storage subsystem and its host I/O processing while the inspection or conversion is being performed. All of these functions use the *modification priority* rates to determine their process priority. Values to choose from are lowest, low, medium, high, and highest.

Tip: The lowest priority rate favors system performance, but the modification operation takes longer. The highest priority rate favors the modification operation, but system performance can be compromised.

In the following sections, we describe the use of each of these functions and their impact on performance.

Media scan

Media scan is a background check performed on all logical drives in the Midrange Storage Subsystem when selected to ensure that the blocks of data are good, which is accomplished by reading the logical drives one data stripe at a time into cache, and if successful it moves on to the next stripe. If a bad block is encountered, it will retry three times to read the block, and then go into its recovery process to rebuild the data block.

Media scan is configured to run on selected logical drives; and has a parameter for defining the maximum amount of time allowed to complete its run through all the logical drives selected. If the media scan process sees it is reaching its maximum run time and calculates that it is not going to complete in the time remaining, it will increase its priority and can impact host processing. Generally, it has been found that media scan scheduled with a “30 day” completion schedule, is able to complete if controller utilization does not exceed 95%. Shorter schedules require lower utilization rates to avoid impact.

Best practice: Setting media scan to 30 days has been found to be a good general all around value to aid in keeping media clear and server background process load at an acceptable level.

Defragmenting an array

A fragmented array can result from logical drive deletion resulting in free space node or not using all available free capacity in a free capacity node during a logical drive creation.

Because creation of new logical drives cannot spread across several free space nodes, the logical drive size is limited to the greatest amount of a free space node available, even if there is more free space in the array. The array needs to be defragmented first to consolidate all free space nodes to one free capacity node for the array. Then, a new logical drive can use the whole available free space.

Use the defragment option to consolidate all free capacity on a selected array. The defragmentation runs concurrently with normal I/O; it impacts performance, because the data of the logical drives must be moved within the array. Depending on the array configuration, this process continues to run for a long period of time.

Important: After this procedure is started, it cannot be stopped; and no configuration changes can be performed on the array while it is running.

The defragmentation done on the Midrange Storage Subsystem only applies to the free space nodes on the array. It is not connected to a defragmentation of the file system used by the host operating systems in any way.

Copyback

Copyback refers to the copying of data from a hot spare drive (used as a standby in case of possible drive failure) to a replacement drive. When you physically replace the failed drive, a copyback operation automatically occurs from the hot spare drive to the replacement drive.

With the new 7.x code releases, whether or not to perform the copyback function has become an option that must be selected if desired. In a high performance environment where data layout and array and LUN configuration matter, using the copyback feature is necessary to ensure continued balance of the workload spread.

Initialization

Initialization is the deletion of all data on a drive, logical drive, or array. In previous versions of the storage management software, it was called *format*.

Dynamic Segment Sizing

Dynamic Segment Sizing (DSS) is a modification operation where the segment size for a select logical drive is changed to increase or decrease the number of data blocks that the segment size contains. A segment is the amount of data that the controller writes on a single drive in a logical drive before writing data on the next drive.

Dynamic Reconstruction Rate (DRR)

Dynamic Reconstruction Rate (DRR) is a modification operation where data and parity within an array are used to regenerate the data to a replacement drive or a hot spare drive. Only data on a RAID 1, RAID 3, or RAID 5 logical drive can be reconstructed.

Dynamic RAID Level Migration

Dynamic RAID Level Migration (DRM) is a modification operation used to change the RAID level on a selected array. The RAID level selected determines the level of performance and parity of an array.

Dynamic Capacity Expansion

Dynamic Capacity Expansion (DCE) is a modification operation used to increase the available free capacity on an array. The increase in capacity is achieved by selecting unassigned drives to be added to the array. After the capacity expansion is completed, additional free capacity is available on the array for the creation of other logical drives. The additional free capacity can then be used to perform a Dynamic logical drive Expansion (DVE) on a standard or FlashCopy repository logical drive.

Dynamic Logical Drive Expansion

Dynamic Logical Drive Expansion (DVE) is a modification operation used to increase the capacity of a standard logical drive or a FlashCopy repository logical drive. The increase in capacity is achieved by using the free capacity available on the array of the standard or FlashCopy repository logical drive.

6.5.4 Storage subsystem parameters

Settings on the DS5000 storage subsystem are divided into two groups:

- Storage subsystem wide parameters affecting all workloads that reside on the storage.
- Settings that are specific to the array or logical drive where the data resides

Cache blocksize selection

On the Midrange Storage Subsystem, the cache blocksize is a variable value that can be set to 4 K, 8 K, 16 K or 32K. The general default setting is 8 K. The main goals with setting this value are to minimize the number of cache IO's needed, and at the same time, not waste space. This value is a storage subsystem wide parameter, and when set, it is the value to be used by all cache operations.

For example, if the I/O of greatest interest is taken from your database operations during the day rather than from your weekly backups, you want to tune this value to handle the high transactions best. Knowing that the higher transactions will have smaller I/O sizes, using the 4 K settings is generally best for transaction intense environments.

Best practice: Set the cache blocksize to 4 K or 8 K for the Midrange Storage Subsystem, normally for transaction intense environments with smaller I/O blocksizes.

In a throughput intense environment, as we described earlier, you want to get as much data into cache as possible. In this environment, it is generally best to use the 16 K blocksize for the cache.

Best practice: Set the cache blocksize to 16 K or 32 K for the DS5000 storage subsystem, normally for throughput intense environments with large host I/O blocksizes.

In mixed workload environments, you must decide which workload type is most critical and set the system wide settings to best handle your business needs.

Best practice: Set the cache blocksize to 8 K for the Midrange Storage Subsystem, normally for mixed workload environments.

Tip: Throughput operations, though impacted by smaller cache blocksize, can still perform reasonably if all other efforts have been accounted for. Transaction based operations are normally the higher concern, and therefore must be the focus for setting the server wide values if applicable.

Cache flush control settings

In addition to the cache blocksize, the Midrange Storage Subsystem also has a cache control, which determines the amount of data that can be held in write cache. With the *cache flush* settings, you can determine what level of write cache usage can be reached before the server will start to flush the data to disk, and at what level the flushing will stop.

By default, these parameters are set to the value of “80” for each, which means that the server will wait until 80% of the write cache is used before it will flush the data to disk. In a fairly active write environment, this value might be far too high. You can adjust these settings up and down until you find a particular value that best suits your environment. If the values are not the same, then back-end drive inactive time increases, and you have surging with peaks and valleys occurring instead of a steady usage of back-end disks.

You can also vary the maximum amount of time the write data can remain in cache prior to being forced out, and written to disks. This value by default is set to ten seconds but can be changed by using the Storage Manager (SM) command line interface command:

```
‘set logical Drive [LUN] cacheflushModifier=[new_value];’
```

Best practice: Begin with Start/Stop flush settings of 50/50, and adjust from there. Always keep them equal to each other.

6.5.5 Disk drive types

With the Midrange Storage Subsystem, many types of disk drives are available for you to choose from. The following list shows available drive types and sizes at the time of writing.

- FC drives without encryption:
  - 300 GB/15K 4 Gbps FC E-DDM
  - 450 GB/15K 4 Gbps FC E-DDM
  - 600 GB/15K 4 Gbps FC E-DDM
- FC disk with encryption:
  - 300 GB/15K 4 Gbps FC encryption-capable E-DDM
  - 450 GB/15K 4 Gbps FC encryption-capable E-DDM
  - 600 GB/15k 4 Gbps FC encryption-capable E-DDM

- SATA disks:
  - 1000 GB/7.2K SATA E-DDM
  - 2000 GB/7.2K SATA E-DDM
- SAS drives (with FC-SAS interposer) without encryption:
  - 300 GB/10k FC-SAS E-DDM
  - 600 GB/10k FC-SAS E-DDM
  - 900 GB/10k FC-SAS E-DDM
- SAS drives (with FC-SAS interposer) with encryption
  - 300 GB/10k FC-SAS encryption-capable E-DDM
  - 600 GB/10k FC-SAS encryption-capable E-DDM
  - 900 GB/10k FC-SAS encryption-capable E-DDM
- SAS SSD (with FC-SAS interposer)
  - 200 GB 4Gbps FC-SAS SSD E-DDM
  - 400 GB 4Gbps FC-SAS SSD E-DDM

With the new SSD technology, solid state drives are the highest performing drives available for use. However, with the max limit of 20 400 GB SSD's per subsystem, this limits the capacity that can be defined. So for the larger capacity environments, 15K RPM drives provide the best performance. The 10K RPM drives are a close third; whereas the 7200 RPM drives are the slowest. SATA drives can be used in lower transaction intense environments where maximum performance needs are less important, and high storage capacity, or price are main concerns. SATA drives do provide good throughput performance and can be a very good choice for these environments.

If large drives are used to store a heavy transaction environment on fewer drives, then the performance will be impacted. If using large size drives and high numbers of them, then how the drive will be used becomes another variable. In certain cases, where you prefer RAID 1 to RAID 5, a larger drive might be a reasonable cost compromise; but only testing with your real data and environment can show for sure.

Best practice: For transaction intense workloads SSD's or 15K RPM drives provide the best performance.

The current Midrange Storage Subsystems support a drive side I/O queue depth of "16" for the Fibre Channel disks. The SATA drives support a queue depth of "4" I/Os. See 2.2.2, "Drive types" on page 22 additional information.

6.5.6 Arrays and logical drives

When setting up the Midrange Storage Subsystem, the configuration of the arrays and logical drives is most likely the single most critical piece in your planning. Understanding how your workload will use the storage is crucial to the success of your performance, and your planning of the arrays and logical drives to be used.

Next, we further describe various considerations for the Midrange Storage Subsystems.

RAID array types

When configuring a Midrange Storage Subsystem for the transaction intense environment, you need to consider also whether it will be a read or write intensive workload. As mentioned earlier, in a database environment, we actually have two separate environments with the table space, and the journals and logs. Because the table space is normally high reads and low writes, and the journals and logs are high writes with low reads, this environment is best served by two RAID types.

RAID 0, which is striping, without mirroring or parity protection, is generally the best choice for almost all environments for maximum performance; however, there is no protection built into RAID 0 at all, and a drive failure requires a complete restore. For protection it is necessary to look toward either a software mirroring solution of one of the other RAID types.

On the Midrange Storage Subsystem, RAID 1 is disk mirroring for the first pair of disks, and for larger arrays of four or more, disks mirroring and striping (RAID10 model) is used. This RAID type is a very good performer for high random read and write environments. It outperforms RAID 5 due to additional reads that RAID 5 requires in performing its parity check when doing the write operations. With RAID1, there are two writes performed per operation; where as with RAID 5, there are two reads and two writes required for the same operation, totaling four I/Os.

A common use for RAID 1 is for the mail server environment, where random writes can frequently out-weigh the reads.

With database journals and logs being generally sequential write intensive I/Os, they are also a good fit for a RAID 1 as well, which is generally due to the small IO size that they use for their updates. However, it is something that must be reviewed for your specific environment. If these processes are found to be of an IO size that is better served by a design with RAID 5 and a full stripe write, then better performance can be gained from that layout.

Random reads can be served well by both RAID 1 and RAID 5. RAID 1 when laid out with a set of disks that are designed to handle a specific capacity is compared to a RAID 5 design of few disks to handle the same capacity can outperform the RAID 5 due to the extra spindles it is designed with. In many cases this advantage has been found to be too small to justify the added cost of the extra drives required.

When the RAID 5 array is created with an equal number of spindles to handle the same workload as the RAID 1 array, it will outperform the RAID 1 array, due to the positioning of the LUN on the RAID 5 array and its being striped across the outside tracks, whereas the RAID 1 array needs to use the full disks for its LUN layout. For this reason, it is best to use RAID 5 for the OLTP table space even for large database environments.

i

Tip: There are no guaranteed choices as to which type of RAID to use, because it is very much dependent on the workload read and write activity. A good general guide might be to consider using RAID 1 if random writes exceed about 25%, with a peak sustained I/O rate that exceeds 50% of the storage subsystem's capacity.

In the sequential high throughput environment, RAID 5 performance is excellent, as it can be configured to perform just one additional parity write when using "full stripe writes" (also known as "full stride writes") to perform a large write I/O, as compared to the two writes per data drive (self, and its mirror) that are needed by RAID 1. As shown in Figure 6-1, this model is a definite advantage with RAID 5 array groups.

Optimized RAID-5 (7+P) write activity

HDDs Written	Read Data	Read Parity	Write Data	Writes Parity	RAID 5	RAID 1
1	1	1	1	1	4	2
2	2	1	2	1	6	4
3	3	1	3	1	8	6
4	3	0	4	1	8	8
5	2	0	5	1	8	10
6	1	0	6	1	8	12
7	0	0	7	1	8	14

Figure 6-1 RAID 5 write penalty chart

The decrease in the overhead read operations with the *full stripe write* operation is the advantage you are looking for. You must be very careful when implementing this type of layout to ensure that your data pattern does not change, and decrease its effectiveness. However, this layout might work well for you in a large sequential write environment. Due to the small size of segments, reads might suffer, so mixed I/O environments might not fare well, which might be worth testing if your writes are high.

When performing high sequent IO workloads, it is often overlooked that the seek time for the small blocksize on the same drive becomes “zero”. This result in itself is a great benefit to be had when using a larger segment size with mixed read and write sequential IO workloads which use a small host blocksize, which is also why storing sequentially accessed files separately keeps the disk heads in position for sequential I/O, which reduces the amount of time required to locate data.

Applications where it has shown high success are with backups, imaging, video and high performance computing (HPC).

The differences in RAID 1 and RAID 5 make them both better suited for certain workload types. Take care not to try to force a fit for an unsuitable workload.

Best practice: The differences described so far outline the major reasons to keep the journals and logs on arrays other than the table spaces for database applications.

With the amount of variance that can exist with each customer environment, it is a good idea to test your specific case for best performance; and decide which layout to use. With the write cache capability, in many cases RAID 5 write penalties are not noticed, as long as the back-end disk configuration is capable of keeping up with the front-end I/O load so processing is not being slowed, which again points to ensuring that proper spreading is done for best performance.

Number of disks per array

In the transaction intense environment, it is more important to ensure that there are enough disk drives configured to perform the I/Os demanded by the host application, than to focus on the amount of possible storage space on the storage subsystem.

With the Midrange Storage Subsystems, you can purchase the new 73 GB SSD disk drives; or the 36 GB, 73 GB, 146 GB, 300 GB, 450 GB, or 600 GB 15k RPM Fibre Channel disk drive types. Previous 10 K RPM Fibre Channel disk drives are also supported in the Midrange Storage Subsystems. Obviously, with the larger drives, you can store more data using fewer drives. Also available are the 500 GB and 1 TB SATA 2 7500 RPM disk drives. Obviously, with the larger drives you can store more data using fewer drives; however, this might not serve your purpose as well as you might want.

Transaction intensive workload

In a transaction intense environment, you want to have higher drive numbers involved. It can be done by creating larger arrays with more disks. The storage subsystems can have a maximum of 30 drives per RAID 5 array/logical drive and 448 drives (with the DS5100 and DS5300) for a full subsystem RAID 10 array.

With a RAID 10 array, the logical drive size can be configured to encompass the entire array, although operating system limitations on maximum logical drive size might restrict the usefulness of this capability. Though there are circumstances where this model can work well, it is best that smaller arrays be used to better balance the drive usage and access patterns across the controllers so to avoid contusion from intermixed IO.

Configuring multiple LUNs striped across a single large array can be used to make use of all the capacity. However, with this layout, consideration must be given to the workload types for which these LUNs are used, so as not to mix throughput and transaction based IO on the same array.

Another factor to consider is congestion when accessing the drives on the back-end loops. This situation can be avoided by using multiple arrays.

Generally, an array of 8 to 16 disks provides the best performance for RAID 5 workloads that are OLTP based.

Best practice: For high transaction environments requiring highest redundancy and protection, logical drives must be built on arrays with 8 to 16 disks when using RAID 5. With RAID 10, use the highest number of available drives / 4 to build two equally sized arrays/LUNs to be preferred across the two controllers.

For large size databases, consider using the host volume management software to spread the workload evenly across multiple arrays/LUNs to evenly balance the workload on all. Build the volume across sets of logical drives laid out per the RAID type in the previous description. In using multiple arrays, you will also be able to increase the controllers which are involved in handling the load, therefore getting full use of the storage subsystems resources.

For example: If needing to build a database that is 2 TB in size, you can use five 600 GB drives in a 4 + 1parity RAID 5 single array with one logical drive; or you can create a RAID 5 arrays of 8 + 1parity using 300 GB drives, and carve either one large logical drive or create two 1 TB logical drives across them on which to build the 1 TB database. In most cases using the second method of greater drive count for large databases will work best, as it brings double the number of disks into play for handling the host side high transaction workload. In many cases building even more arrays of this size and carving logical drives out of them will enhance the performance even more. With many host supporting methods of software volume management, it can be even a greater performance value.

Large throughput workload

In the large throughput environment, it typically does not take high numbers of disks to reach the maximum sustained throughput. Considering that this type of workload is usually made of sequential I/O, which reduces disk latency, in most cases about 20 to 28 drives are enough to reach the maximum throughput.

This does, however, require that the drives be spread evenly across the Midrange Storage Subsystem to best utilize the server bandwidth. The storage subsystem is optimized in its firmware to give increased throughput when the load is spread across all parts. Here, bringing all the Midrange Storage Subsystem resources into play is extremely important. Keeping the drives, channels, and bus busy with high data throughput is the winning answer, which is also a perfect model for using the high capacity drives, as we are looking to push a large volume of data and it will likely be large blocks of sequential reads and writes.

Consider building smaller arrays with single logical drives for higher combined throughput.

Best practice: For high throughput, logical drives must be built on arrays with 4+1, or 8+1 drives in them when using RAID 5. The total number of data drives multiplied by the *segment size* must equal host I/O blocksize for full stripe write. Use multiple logical drives on separate arrays for maximum throughput.

An example configuration for this environment is to have a single logical drive /array with 16+1 parity 300 GB disks doing all the transfers through one single path and controller; An alternative consists of two 8 + 1 parity defined to the two controllers using separate paths, doing two separate streams of heavy throughput in parallel and filling all the channels and resources at the same time, which keeps the whole server busy with a cost of one additional drive.

Further improvements can be gained by splitting the two 8 + 1 parity into four 4 + 1 parity arrays giving four streams, but addition of three drives is needed. A main consideration here is to plan for the array data drive count to be a number such that the host I/O blocksize can be evenly spread using one of the storage subsystem's segment size selections, which will enable the full stripe write capability described in the next section.

DS5000 special considerations

When setting up the DS5100 or DS5300, you need to consider how the storage subsystem handles its back-end loops and how best to spread the workload for maximum performance and balance of the workload. We describe various features of the subsystem's design to help with your understanding.

In Figure 6-2, we see that the subsystem's drive ports are outlined with their associated channels. The numbers shown in the center are the preferred sequence of connecting the first eight expansions to gain the best use of all the subsystem's resources for smaller configurations. As the subsystem grows and expands beyond the eight channels, daisychaining second expansions to each of the channels in the same order will help to maintain a continued balanced environment.

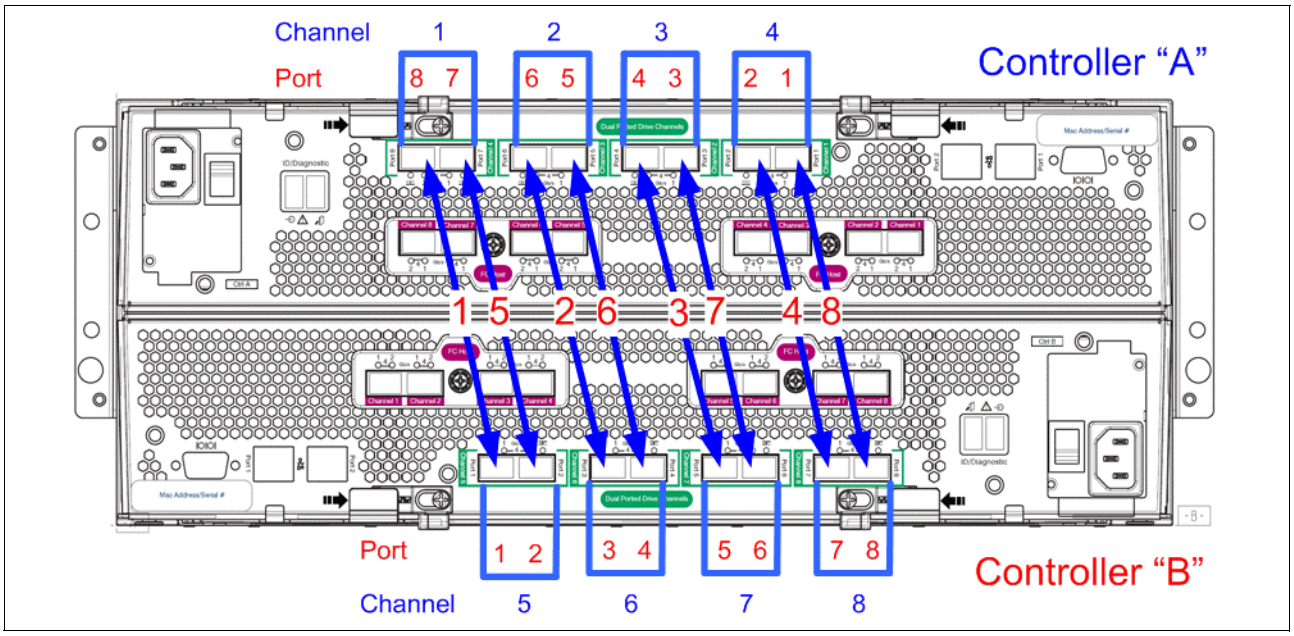


Figure 6-2 DS5100 and DS5300 back-end drive ports and channels defined

With the DS5100 and DS5300 storage subsystems, the guidelines for layout also extend to the channels and ports the expansion enclosures are attached to. With previous Midrange Storage Subsystems, fewer channels and loop pairs were used to support the back-end drives. With the new DS5100 and DS5300 we can see there are now eight channels with which loop pairs created.

With these eight unique loops, you can split them equally between the two controllers for best balance and spread of the workload; as well as to avoid contention with bus access. As examples Figure 6-3 and Figure 6-4 show, we use layouts of four expansions on four channels and sixteen expansions on all eight channels.

In these examples, the expansions which are attached to the odd numbered stacks are used to create the arrays and Logical drives which are assigned to Controller A, and the odd stack arrays and logical drives are assigned to Controller B, which will ensure that drive IO to the disks do not need to deal with channel conjestion between the two controllers.

These guidelines also greatly help to improve the internal operations of the controllers as well as the channel loop operations.

To help understand this situation better, let us look at RAID 5 4+P array laid out for use with controller A as its preferred owner. This array group is laid out to include enclosure loss protection as well.

There are many variables that can come into play with the design of the layout that will be needed. They can include LUN size, desired RAID protection level, number of enclosures, number of drives, as the more common ones. The main point to remember is” First, and foremost to make use of as many resources of the subsystem as possible.” then align to include the other best practices. Smaller subsystem designs have greater difficulty with covering all the best practices as they frequently do not have the number of drive enclosures to spread out over all the back-end channels. To help with possible guidance the following sample configurations are offered for review. These configurations can be tuned to best fit the need of the environment.

		EXP Slots															
Tray ID	Loop Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	1	A1	A3	A1	A3	A1	A3	A7	A5	A7	A5	A9	A11	A9	A11	A9	HS
25	2	B2	B4	B2	B4	B2	B4	B8	B6	B8	B6	B10	B12	B10	B12	HS	B10
31	3	A3	A1	A3	A1	A5	A7	A5	A7	A5	A7	A11	A9	A11	A9	HS	A11
45	4	B4	B2	B4	B2	B6	B8	B6	B8	B6	B8	B12	B10	B12	B10	B12	HS

Figure 6-3 Example layout of a four expansion 4 + P array configuration

In the layout used in Figure 6-4, we are giving up enclosure loss protection to have a greater amount of user capacity with RAID 5 versus RAID 10, and only use four expansions. In Figure 6-4, we show a configuration of sixteen expansions being used to layout 7 + P, RAID 5 array groups with enclosure loss protection. This configuration is frequently used for database application layouts with high random transaction processing.

		EXP Slots															
Tray ID	Loop Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	1	A1	A3	A5	A7	A9	A11	A13	A15	A17	A19	A21	A23	A25	A27	A29	A31
12	1	A3	A1	A7	A5	A11	A9	A15	A13	A19	A17	A23	A21	A27	A25	H	A29
25	2	B4	B2	B6	B8	B10	B12	B14	B16	B18	B20	B22	B24	B26	B28	B30	H
26	2	B2	B4	B8	B6	B12	B10	B16	B14	B20	B18	B24	B22	B28	B26	B32	B30
31	3	A1	A3	A5	A7	A9	A11	A13	A15	A17	A19	A21	A23	A25	A27	A29	H
32	3	A3	A1	A7	A5	A11	A9	A15	A13	A19	A17	A23	A21	A27	A25	A31	A29
45	4	B4	B2	B6	B8	B10	B12	B14	B16	B18	B20	B22	B24	B26	B28	B30	B32
46	4	B2	B4	B8	B6	B12	B10	B16	B14	B20	B18	B24	B22	B28	B26	H	B30
51	5	A1	A3	A5	A7	A9	A11	A13	A15	A17	A19	A21	A23	A25	A27	A29	A31
52	5	A3	A1	A7	A5	A11	A9	A15	A13	A19	A17	A23	A21	A27	A25	H	A29
65	6	B4	B2	B6	B8	B10	B12	B14	B16	B18	B20	B22	B24	B26	B28	B30	H
66	6	B2	B4	B8	B6	B12	B10	B16	B14	B20	B18	B24	B22	B28	B26	B32	B30
71	7	A1	A3	A5	A7	A9	A11	A13	A15	A17	A19	A21	A23	A25	A27	A29	H
72	7	A3	A1	A7	A5	A11	A9	A15	A13	A19	A17	A23	A21	A27	A25	A31	A29
81	8	B4	B2	B6	B8	B10	B12	B14	B16	B18	B20	B22	B24	B26	B28	B30	B32
82	8	B2	B4	B8	B6	B12	B10	B16	B14	B20	B18	B24	B22	B28	B26	H	B30

Figure 6-4 Example layout of a sixteen expansion 7+ P array configuration

If you need a full RAID 10 configuration, you can use the arrangement in Figure 6-5 as a sample to tune to meet the need.

		EXP Slots															
Tray ID	Loop Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	1	A1	A1	A1	A3	A3	A3	A5	A5	A5	A7	A7	A7	xx	xx	xx	HS
25	2	B2	B2	B2	B4	B4	B4	B6	B6	B6	B8	B8	B8	xx	xx	xx	HS
31	3	xx	A1	A1	A1	A3	A3	A3	A5	A5	A5	A7	A7	A7	xx	xx	HS
45	4	xx	B2	B2	B2	B4	B4	B4	B6	B6	B6	B8	B8	B8	xx	xx	HS

Figure 6-5 Four expansion 3 + 3 RAID 10 example

Array and logical drive creation

An array is the grouping of drives together in a specific RAID type format on which the logical drive will be built for presentation to the hosts. As described in 3.1.2, “Creating arrays and logical drives” on page 99, there are a number of ways to create an array on the Midrange Storage Subsystem using the Storage Manager Client.

For best layout and optimum performance, you can manually select the drives when defining arrays.

Build the array groups by selecting drives equally from odd and even drive slots in the expansion enclosures. This task can be performed using either a “barber pole” pattern as shown in Figure 6-6 on page 283. or a “zig zag” (aka “zipper”) pattern as shown used in Figure 6-3 on page 281 and Figure 6-4 on page 281.

This type of balancing can also be extended to RAID10 arrays in a simplified manner as shown in Figure 6-5. These layouts help to spread the IO across both of the paths in the channel so that greater bandwidth and resources are being used.

To help understand this situation better, let us look at how the RAID 5 4+P array laid out for use with controller A as its preferred owner is defined in Figure 6-6. With this array group, the layout includes enclosure loss protection as well. For a balanced subsystem design, we have a second set of five enclosures on channels 2, 4, 6, and 8 as well.

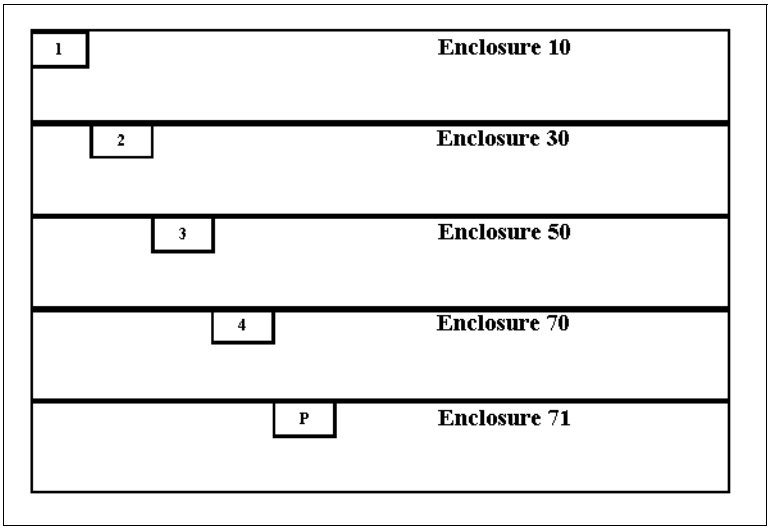


Figure 6-6 Barber pole example

However you plan them, try to focus on keeping the load spread evenly. The main goal in all cases must be to bring as many resources into play as possible. With a small configuration, try to bring as many channels into play as you can.

With the DS storage subsystem, it is a good idea to build the arrays and logical drives across expansions that are on a redundant channel drive loop pair as described in “Enclosure layout and loss protection planning” on page 34 and 3.1.2, “Creating arrays and logical drives” on page 99.

A logical drive is the portion of the array that is presented to the host by the storage subsystem. A logical drive can be equal to the entire size of the array, or just a portion of the array. A logical drive will be striped across all the data disks in the array. Generally, try to keep the number of logical drives created on a array to a low number. However, in certain cases it is not possible, and the planning of how the logical drives are to be used becomes very important. You must remember that each logical drive will have its own host I/Os that will be queuing up against the drives that make up the array. Multiple heavy transaction applications using the same array of disks can result in that array having poor performance for all its logical drives.

Best practice: Create a single logical drive for each array whenever possible.

When configuring multiple logical drives on an array, try to spread out their usage evenly to have a balanced workload on the disks making up the arrays. It is also best if the workload types, transaction based and throughput based, are not mixed on the same array. IOs with a large blocksize can result in small IO’s being starved for bandwidth, and being queued for a greater period of time. A major point to remember here is to try to *keep all the disks busy.* Also, you will want to tune the logical drive separately for their specific workload environments.

Drive loop communications

The drive loop channels can be used to transfer IOs to disks, for transfer of controller to controller internal communications and cache mirroring operations. These operations are performed through the use of the controller's "exchange control blocks" (XCB). XCB's are used to track the handling of all back-end and control processes that require the use of back-end Fibre Channel transfers to be performed, which includes operations between the controllers or for disk operations.

Each controller is allocated 512 XCBs per channel for its own operations. They are not shared XCBs, but dedicated to that controller. When a controller to controller operations is performed an XCB is used from both controllers to handle it from both sides. If a controller's XCB's are depleted a "queue full" condition is reported to the Destination Driver Event (DDE) and further operations from the controller to that specific channel are prevented until an XCB is freed up and write cache becomes internally disabled.

This makes the spreading of the workload across the subsystem resources very important to avoid running out of XCBs. As a recovery, when this event occurs, the controller will try to automatically send the request down a second channel to see if it can succeed through it. However, as expected, any recovery steps needing to be run will have a performance impact that might be better avoided when possible.

With the design of the DS5100 and DS5300, there are separate cache mirroring busses that are used for the write mirroring processes that allows for the XCB's to be used only for the inter-controller communications and Fibre Channel disks operations only, which improves the amount of XCB resources available to the production disk IO operations. However, it is still important for performance to properly balance the IO workload.

Logical drive segments

The segment size is the maximum amount of data that is written or read from a disk per operation before the next disk in the array is used. For small host I/Os, as mentioned earlier, set the segment size larger than the host I/O size. Doing it makes it unnecessary to access a second drive for a single small host I/O. For certain storage subsystems, having the segment size equal to the host I/O size is preferred, which is *not* the case with the Midrange Storage Subsystems.

There is no advantage in using a smaller segment size with RAID 1; only in a few instances does this help with RAID 5 (which we describe later). Because only the data that is to be written to the disk is written to cache for an I/O, there is no cache penalty encountered either. As mentioned earlier in the host sections, aligning data on segment boundaries is very important for performance. With larger segment sizes, there are less occasions to have misaligned boundaries impacting your performance, as more small I/O boundaries reside within a single segment decreasing the chance of a host I/O spanning multiple drives. This technique can be used to help mask the effect of poor layout of the host data on the disks due to boundary differences.

Best practice: For most high transaction workloads with the Midrange Storage Subsystems, the segment size of 64 KB to 128 KB (default) works best.

With high throughput workload, the focus is on moving larger but fewer I/Os. This workload is generally sequential in nature.

Best practice: In the high throughput environment, you want the *stripe size* to be equal to, or an even multiple of, the host I/O size.

The total of all the segments for one pass of all the back-end data disks is a *stripe*. So, large segment sizes that can equal the I/O size might be desired to accomplish the higher throughput you are looking for. For high read throughput, you want to have large segments (128 K or higher) to get the most from each stripe. For example if the host I/O is 512 KB, and the write is to a RAID 10 array, you want to use a segment size of 512 KB to limit the disk operations as much as possible.

When the workload is high writes, and we are using RAID 5, we can use a method known as *full stripe (stride) write*, which can work well to improve your performance. With RAID 5 the parity is based on the value calculated for a stripe. As described earlier in “RAID array types” on page 276, when the I/O being written is spread across the entire stripe width, no reads are required to calculate the parity; and the I/O completes with fewer back-end I/Os being required.

This design can use a smaller segment size to align the host I/O size with the size of the stripe width. For instance with our previous example, if the array is a 4+P RAID 5 you want to use a 128 KB segment size to achieve a full stripe write. This type of management requires that very few host I/Os not equal a full stripe width.

Logical drive cache settings

Now, to help enhance the use of the storage subsystem data cache, the Midrange Storage Subsystem’s have very useful tuning parameters which help the specific logical drive in its defined environment. One such parameter is the *cache read-ahead multiplier* (see also 2.2.11, “Cache parameters” on page 42). This parameter is used to increase the number of segments that are read into cache to increase the amount of data that is readily available to present to the host for sequential I/O requests. To avoid excess read I/Os in the random small transaction intense environments you must disable the cache read-ahead multiplier for the logical drive by setting it to 0.

In the throughput environment where you want more throughput faster, you must generally enable this parameter to deliver more than one segment to the cache at a time, which can be done by simply setting the value to 1 (or any non-zero value).

Best practice: For *high throughput with sequential I/O*, enable the cache read-ahead multiplier. For high transactions with random I/O, disable this feature.

For write I/O in a transaction based environment, you can *enable write cache*, and *write cache mirroring* for cache protection. Doing it allows the write I/Os to be acknowledged even before they are written to disks as the data is in cache, and backed up by the second mirror in the other controller’s cache. improving write performance dramatically.

This sequence is actually a set of two operations doing both write caching, and mirroring to the second cache. If you are performing a process that can handle loss of data, and can be restarted, you can choose to disable the mirroring, for further improved performance.

With the DS5100 and DS5300, the amount of improvement might not be worth the gain as the new subsystem uses a dedicated private high speed bus for internal data cache operations and the impact is far less than that seen with the DS5020 systems. With these subsystems the design of the mirror path was to use the back-end Fibre Channel loops which shared this function with other controller communications and back-end disk IO operations.

Be aware that, in most transaction intense environments where the data is a live OLTP update environment, it is not an option that can be chosen; however, in cases such as table updates, where you are loading in new values, and can restart the load over, it can be a way of reducing the load time; and therefore shortening your downtime outage window. For the earlier subsystem design, the improvement can be as much as three times.

For write I/O in the throughput sequential environment, these two parameters again come into play and can give you the same basic values. It must be noted that many sequential applications are more likely to be able to withstand the possible interrupt of data loss with the no cache mirroring selection, and therefore are better candidates for having the mirroring disabled. However, knowing the ability to recover is critical before using this feature for improved performance.

Tip: The setting of these values is dynamic and can be varied as needed online. Starting and stopping of the mirroring can be implemented as a part of an update process when used.

In addition to usage of write cache to improve the host I/O response, you also have a control setting that can be varied on a per logical drive basis that defines the amount of time the write can remain in cache. This value by default is set to ten seconds, which generally has been found to be a very acceptable time; in cases where less time is needed, it can be changed by using the following Storage Manager command line interface command:

```
set logical Drive [LUN] cacheflushModifier=[new_value];
```

This value can also be adjusted to higher values when application needs are better served by holding data in cache for longer periods. Acceptable values to choose from are: *immediate, 250ms, 500ms, 750ms,1sec, 1500ms, 2sec, 5sec,10sec, 20sec, 60sec, 120sec, 300sec, 1200sec, 3600sec and infinite.*

Additional NVSRAM parameters of concern

There are a number of Midrange Storage Subsystem parameters that are defined specifically for the host type that is planned to be connected to the storage. These parameters are stored in the NVSRAM values that are defined for each host type. Two of these parameters can impact performance if not properly set for the environment. NVSRAM settings requiring change must be made using the Storage Manager Enterprise Management window and selecting the script execution tool.

The *Forced Unit Access* setting is used to instruct the storage subsystem to not use cache for I/O, but rather go direct to the disks. This parameter must be configured to ignore.

The *Synchronize Cache* setting is used to instruct the storage subsystem to honor the SCSI *cache flush to permanent storage* command when received from the host servers. This parameter must be configured to ignore.

The AVT/ADT enable setting can also have an effect on how the cache is used by the logical drives. With AVT/ADT enabled on any of the host types the cache usage is limited to 16 MB per logical drive. With AVT/ADT disabled for *all* host types this limitation is lifted. However, it is only an viable option when all host types being used do not make use of the AVT/ADT option for path management and failover. This setting is desired to be used when supporting VMWare host on their own private storage subsystem.

Tip: When defining a storage subsystem for usage with VMWare host servers alone, it is desired that AVT/ADT be disabled for all host type regions.

6.5.7 Special considerations for use of the EXP5060

With the EXP5060, several design factors influence the configuration of your arrays and LUNs for maximum performance. Consider these major areas:

- Number of disks per array group
- Trunking
- Selection of the members of the array group
- Number of LUNs per array group
- Spreading and balancing the workload evenly across all resources

We cover each of these areas in detail.

Number of disks per array group

With the EXP5060, the number of drives per array depends on the environment with which the subsystem is used. With the SATA drives, small arrays of only a few drives can result in serious performance bottlenecks. In certain cases, small arrays of only a few drives cause the processes to back up. However, it is also important not to make the array group so large that the rebuild times are long. For these reasons, we suggest using array groups that consist of five to nine drives for RAID5. And, we suggest using array groups that consist of six to 10 drives for RAID6. In certain cases, you can use larger arrays with the understanding that they have longer rebuild times. But in most cases, smaller arrays encounter conflicts and excessive I/Os per drive and are unable to align block size for matching the “full stripe write” pattern.

In testing, we used four to eight data drives to gain the optimal throughput performance with the SATA drives. These configurations attained the maximum throughputs for these array layouts.

Trunking

With the EXP5060, trunking is a new feature that enables you to achieve high throughput with a minimum number of storage expansions and disks. For this capability, we use the four ports of a channel pair to drive the arrays and LUNs in a single EXP5060. Therefore, we effectively give the 60 drives a full 1.6 GB of bandwidth for their use. This capability can enhance the throughput of a single EXP5060 up to twice that of a standard non-trunked configuration, as outlined in the configuration section of *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679. However, you must follow certain rules and be aware of specific restrictions with this configuration. We describe both methods of attaching the EXP5060, so that you can determine the best method to implement for your specific environment.

Using trunking does not provide a transaction (IOPS) rate advantage. I/O capability is purely a result of the number of spindles. When the configurations are the same, there is no benefit with the increase in bandwidth.

With these facts understood, it is easy to see that the advantage of the EXP5060 lies with its improved throughput handling capabilities or, when needed, its large capacity for use as a low-cost archival system.

EXP5060 non-trunked configurations

In the non-trunked configuration, you can attain high capacities with mixed environments of EXP5000 and EXP5060 expansions sharing channels and subsystem resources. In these environments, we suggest that you follow the best practices that are used with the EXP5000 as the basis for the layout of the arrays and LUNs. This way, you can avoid contention with the drive channel switches and controllers. Also, spread the slot selection evenly between odd and even slots to balance the loop port utilization for all arrays and LUNs.

Figure 6-7 shows an example of a layout for a configuration of four EXP5060s that are used for a layout of 8 + P RAID5 or 7 + P + Q RAID6 array groups using two expansions to provide arrays and LUNs to controller A and two expansions to provide arrays and LUNs for controller B. The layout in Figure 6-7 aligns with the dedicated channel model, which is a best practice for both EXP5000 and EXP5060 in the non-trunked configuration.

Tip: You can use any RAID5 layout as a RAID6 layout for a configuration with one less user data drive capacity.

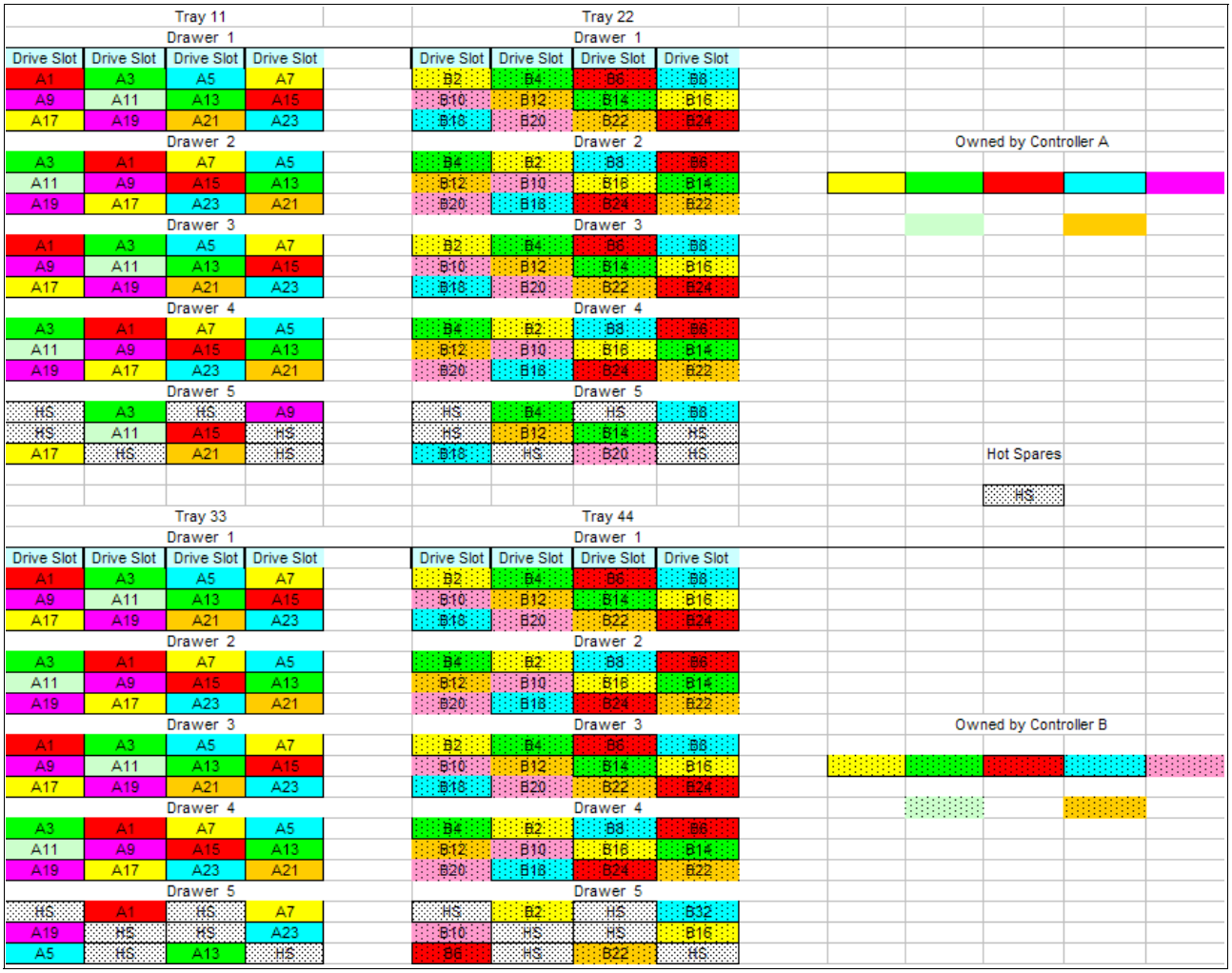


Figure 6-7 Example of four EXP5060s in a non-trunked 8 + P configuration

In dedicated EXP5060 environments, you can use the entire DS5000 storage subsystem to drive up to eight EXP5060 expansions. *When configuring the maximum configuration of eight EXP5060s, there must not be any EXP5000s in the configuration.*

Figure 6-8 shows an example of a layout for a configuration of eight EXP5060s that are used for a maximum configuration with a layout of 8 + P RAID5 or 7 + P + Q RAID6 array groups across all eight expansions. With fewer EXP5060s, you can create shared configurations with EXP5000s on certain channels or cascaded with the EXP5060, creating mixed environments for shared workload patterns.

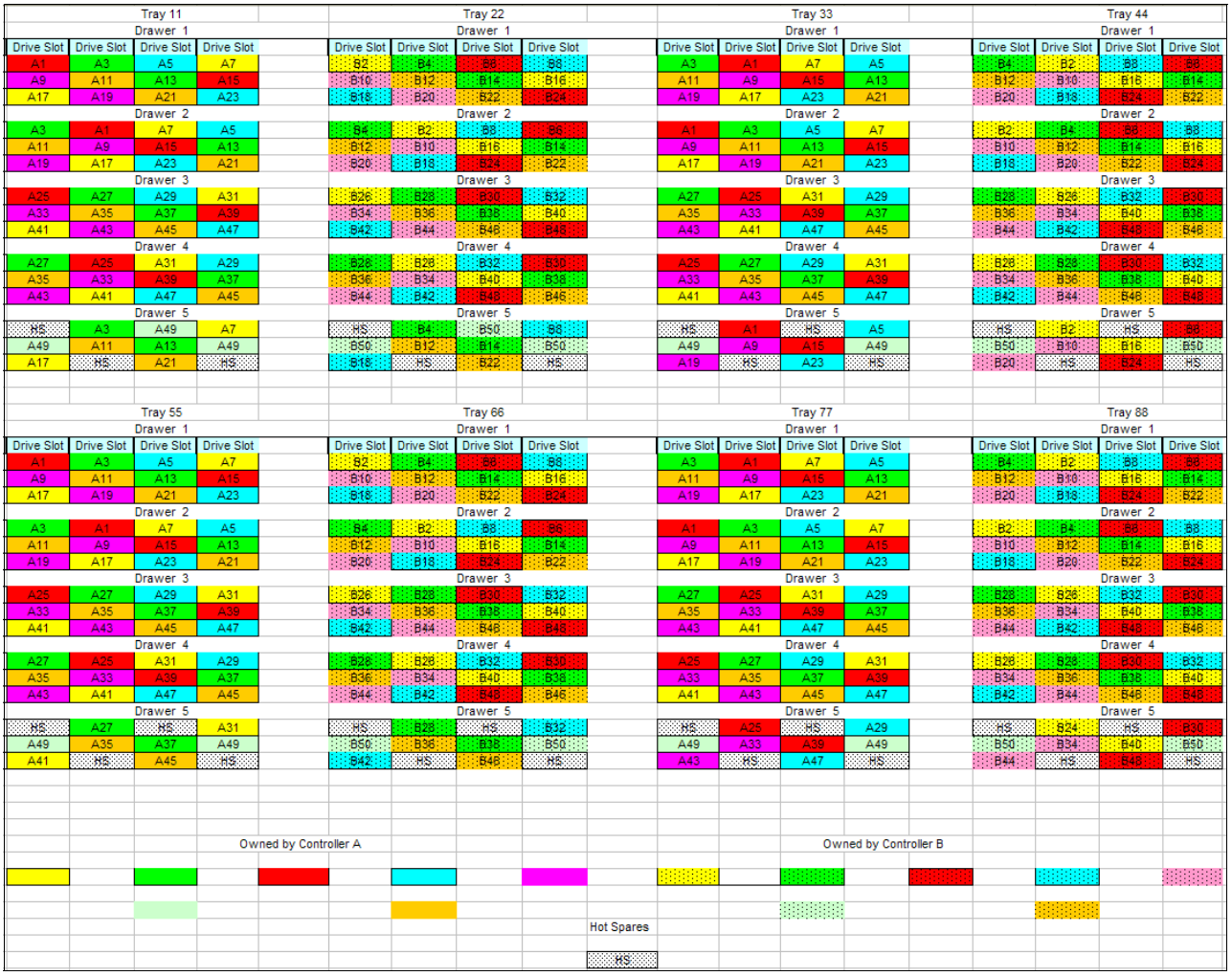


Figure 6-8 Example 8 + P arrays striped across eight non-trunked EXP5060s

With this configuration, you must build all of the disks in all of the drawers of the same expansion into arrays that are owned by the same controller for the best performance of the channel and the path. The EXP5000 expansions use this same best practice. You need to follow the same configuration suggestions for this configuration as you do for EXP5000 layouts:

1. It is critically important that you evenly spread odd and even slots across all channels to a controller.
2. Base the loop pair numbering on the port numbers that are used on Controller B. Use port numbers 1, 3, 5, and 7 to build the arrays and LUNs that are owned by controller A. Use port numbers 2, 4, 6, and 8 to build the arrays and LUNs that are owned by controller B.

The layout in Figure 6-8 follows suggestions that are critical to both load balancing and performance.

Trunked configuration of the EXP5060

With the new trunked design of the EXP5060 expansion, there is more to consider when planning the layout of arrays and LUNs. With the new expansion, we also have two additional switches that are built into the Environmental Service Module (ESM) that are not available in the earlier designs. With the EXP5060, the Fibre Channel drawer consists of a two drawer control monitor (DCM) with one drawer for each of the loop pairs. The DCM uses a micro controller to manage the environmental data and enclosure control over the I/O loop and System on a Chip (SOC) switch path controls.

When configuring the EXP5060 in a trunked configuration for maximum throughput, avoid sharing the drive drawers between controllers whenever possible. Create arrays and LUNs by using the disks from two drawers for arrays and LUNs to be owned by controller A, and two other drawers to be owned by controller B. Then, you can use the fifth drawer for hot spares. If all 60 drives are needed, limit the drives using the fifth drawer to only arrays that are built to be used by only one controller to prevent the effect of shared DCMs. Figure 6-9 shows a full four EXP5060 configuration that is built with 8 + P RAID5 or 7 + P + Q RAID6.

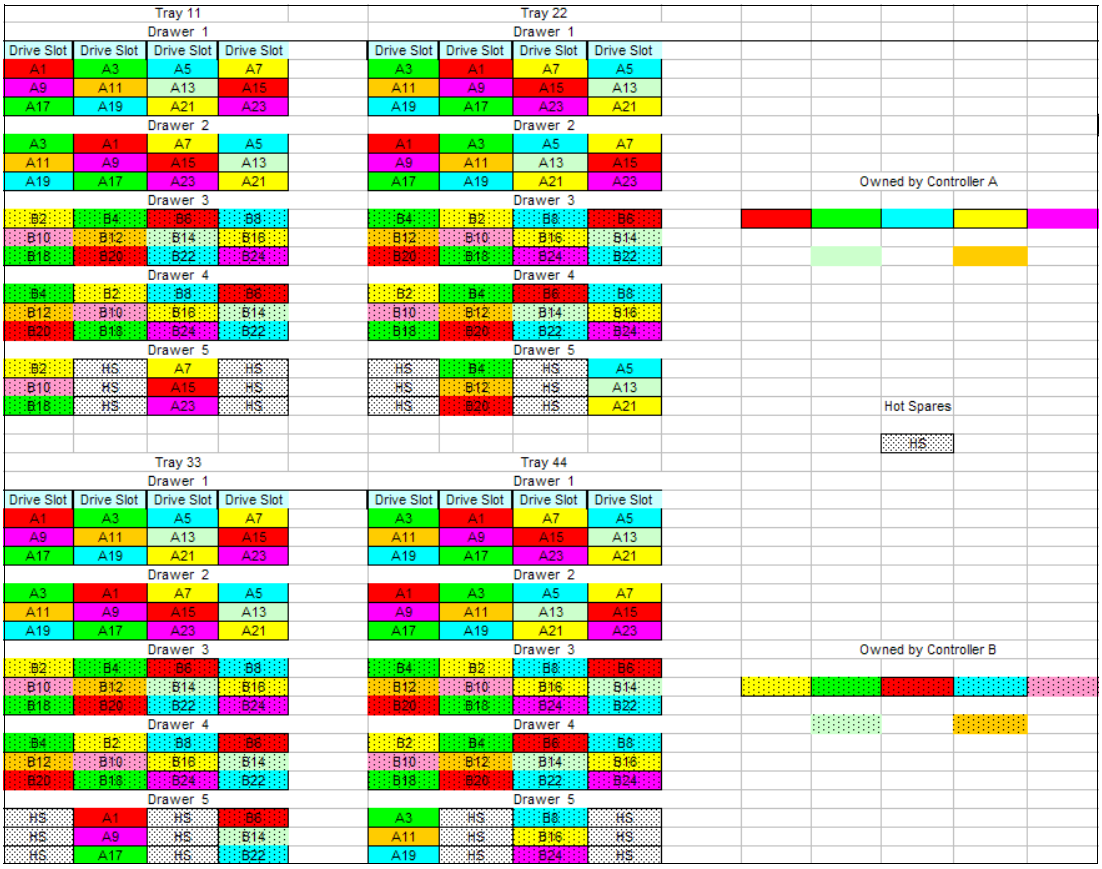


Figure 6-9 Trunked EXP5060 configured with 8 + P arrays with drawer protection

Follow these suggestions to help you plan for smaller configurations or when you want maximum performance:

1. Even when building a small configuration, make good use of all of the available resources for the best performance.
2. Dedicate two drawers, drawers 1 and 2 in the example in Figure 6-9, from all of the expansion trays to build arrays and LUNs that are owned by Controller A.
3. Dedicate two other drawers, drawers 3 and 4 in the example in Figure 6-9 on page 290, from all of the expansion trays to build arrays and LUNs that are owned by Controller B.
4. When using the fifth drawer, build separate array groups (A21, A23, B22, and B24 in the example in Figure 6-9 on page 290) that are split equally between Controller A and B. This design allows the shared DCM to have minimal effect and still spreads the workload as evenly as possible across all of the drive channels.

Tip: When less user capacity is needed and layout requirements allow, using only the four disks that are required in drawer 5 for hot spares is the best model for maximum performance. You can create this layout with a 7 + P configuration.

You can design many layouts with the trunked configuration, and if you follow the suggestions, you can achieve good throughput.

High availability features to use

Clients have always been able to build arrays with enclosure loss protection with the drive layout selection. With earlier expansions and code releases, this capability carried a high degree of importance. However, with the newer switched technology and more robust code design, this requirement has lesser importance. The new EXP5060 expansion design adds a new level of protection at the drawer level. With this protection, you can create an array that spans the expansion trays and is built across multiple drives in separate drawers to incorporate multiple paths for improved throughput performance. This capability helps to strengthen the robustness of your solution. This protection helps to alleviate concerns about enclosure (tray) loss protection when you build an environment that spans four EXP5060s in a trunked environment. To help in this area, you can build the array group by selecting a drive of each of the associated drawers that are preferred to the specific controller, allowing you to use an increased array group size. Additionally, with RAID6, you can extend the number of members in an enclosure to two and still have parity coverage.

You can create many configurations that can provide either optimal performance or maximum protection, but only a few configurations can provide the best levels for both performance and protection. Building these environments requires extensive planning and careful implementation to avoid conflicts and degraded performance. We provide the following sample configurations as additional references to help you with your planning efforts. The following examples are only a subset of all of the possible choices.

Figure 6-10 shows a RAID10 configuration that incorporates the drawer protection best practice.

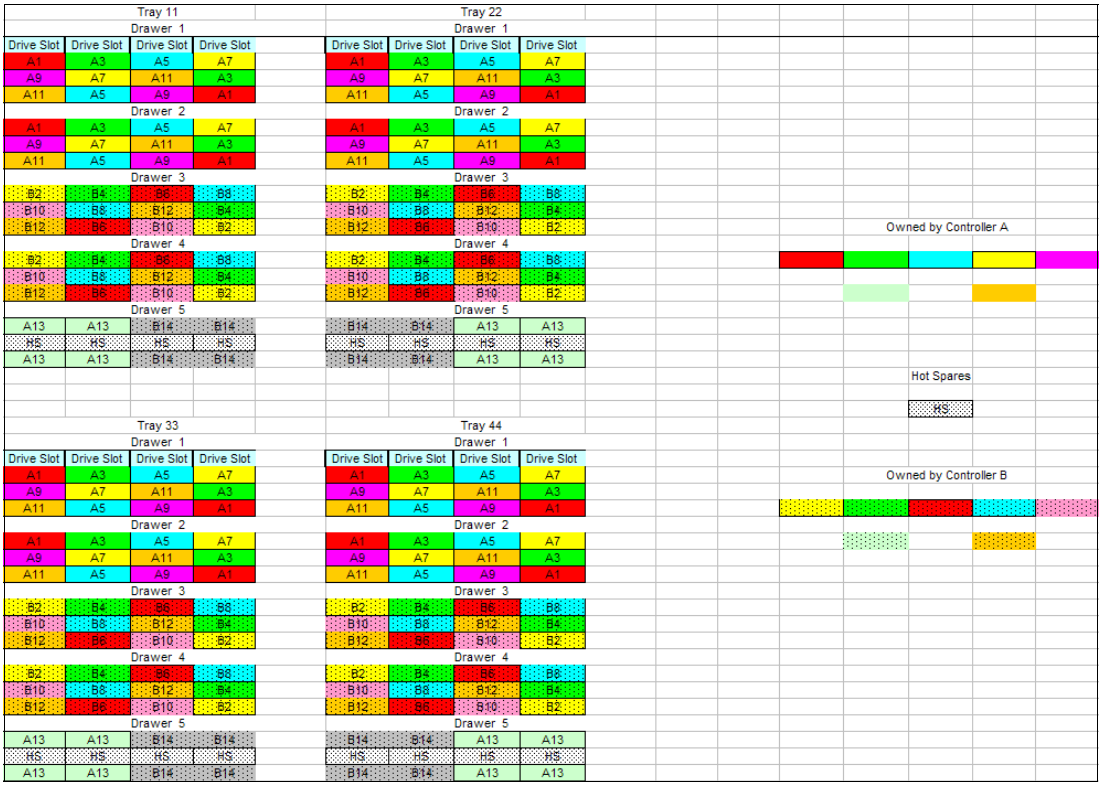


Figure 6-10 Trunked EXP5060 configured with 8 + P RAID10 arrays with drawer protection

For "*High Performance Computing*" environments, we are frequently forced to use small arrays of 4 + P RAID5 layouts. For these environments, we suggest using the layout in Figure 6-11. In this configuration, we build the arrays to include drawer loss protection. Without this requirement, you can improve performance; however with this configuration, you can attain reasonable levels of throughput. For details about performance differences, see the *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679.

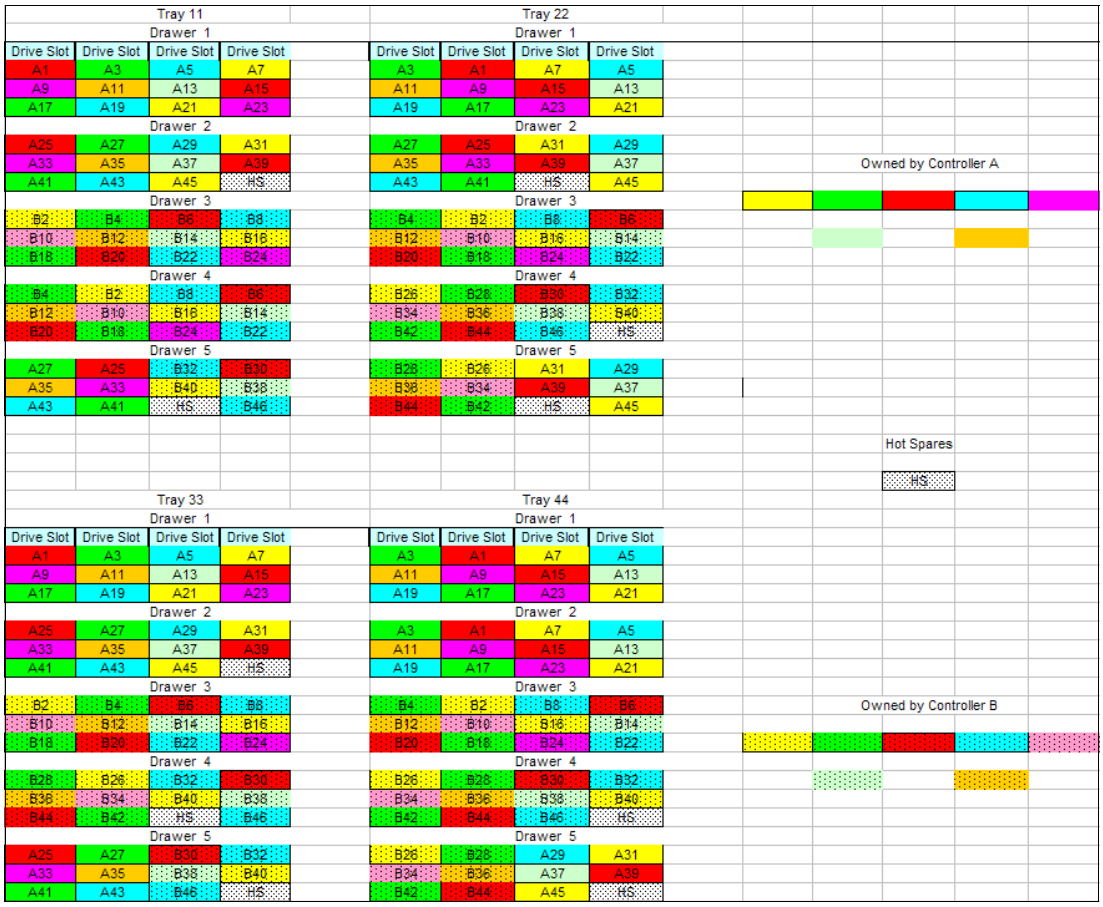


Figure 6-11 Trunked EXP5060 with 4 + P RAID5 arrays

Although you can use the 4 + P layout in Figure 6-11 for a 3 + P + Q RAID6 configuration, we do not advise it, due to having too few data drives in the arrays.

6.5.8 EXP5060 performance

This chapter provides the performance results that we collected during various test scenarios. We used the DS5300 and a configuration of four IBM System Storage EXP5060 High Density Storage Enclosures (EXP5060s) that were installed in a trunked configuration.

Important: We reached the performance numbers in this chapter by using test programs in a lab environment. Your results probably will vary from case to case.

Because the major focus of the EXP5060 expansion is the throughput environment, most of the tests that we performed were to show the maximum throughputs possible with these configurations. In many of the configurations that we used, we did not consider a need for a high availability solution. Include planning for high availability when you plan your solution.

We gathered part of these results with best practice configurations for performance, but not all. We wanted to show the differences that you can encounter. In certain cases, the configurations that show the best performance do not fit a high availability model. In these cases, you need to base your choices on how best to meet your business needs, as described in *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679. In this chapter, we explain options to help in the decision-making process.

Optimal performance layouts

To attain the best possible performance, you must follow these rules to avoid any chance of contention or workload imbalance in the configuration of the array and logical unit number (LUN) layout. We list these rules in the order of their effect on performance:

1. You must build array groups across a balance of odd and even drive slots.
2. You must share each expansion tray evenly between the two controllers by assigning two drawers to controller A and two drawers to controller B from each expansion tray.
3. There is no imbalance in the use of the channels.
4. There are no shared drawers being used.
5. Define hot spares in the front row of the drives in the fifth drawers of the trays to use the required disks that are installed in that drawer.

Sometimes, you might not be able to follow these rules. In certain cases, you might need to make exceptions to these rules to meet the configuration needs of your environment. In these cases, plan to implement the exception so that it has the least effect on the overall configuration. An example of an exception is when the fifth drawer of the EXP5060 is used for arrays and LUNs to be built. See *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679 for the best practices in this situation.

Availability: Depending on the array group size, it might be difficult to ensure enclosure and drawer loss protection. You need to understand the availability requirements.

We defined the following example test case layouts in 8 + P RAID5 array groups (Figure 6-12). Our testing demonstrated the maximum throughput performance numbers that are shown in Table 6-2.

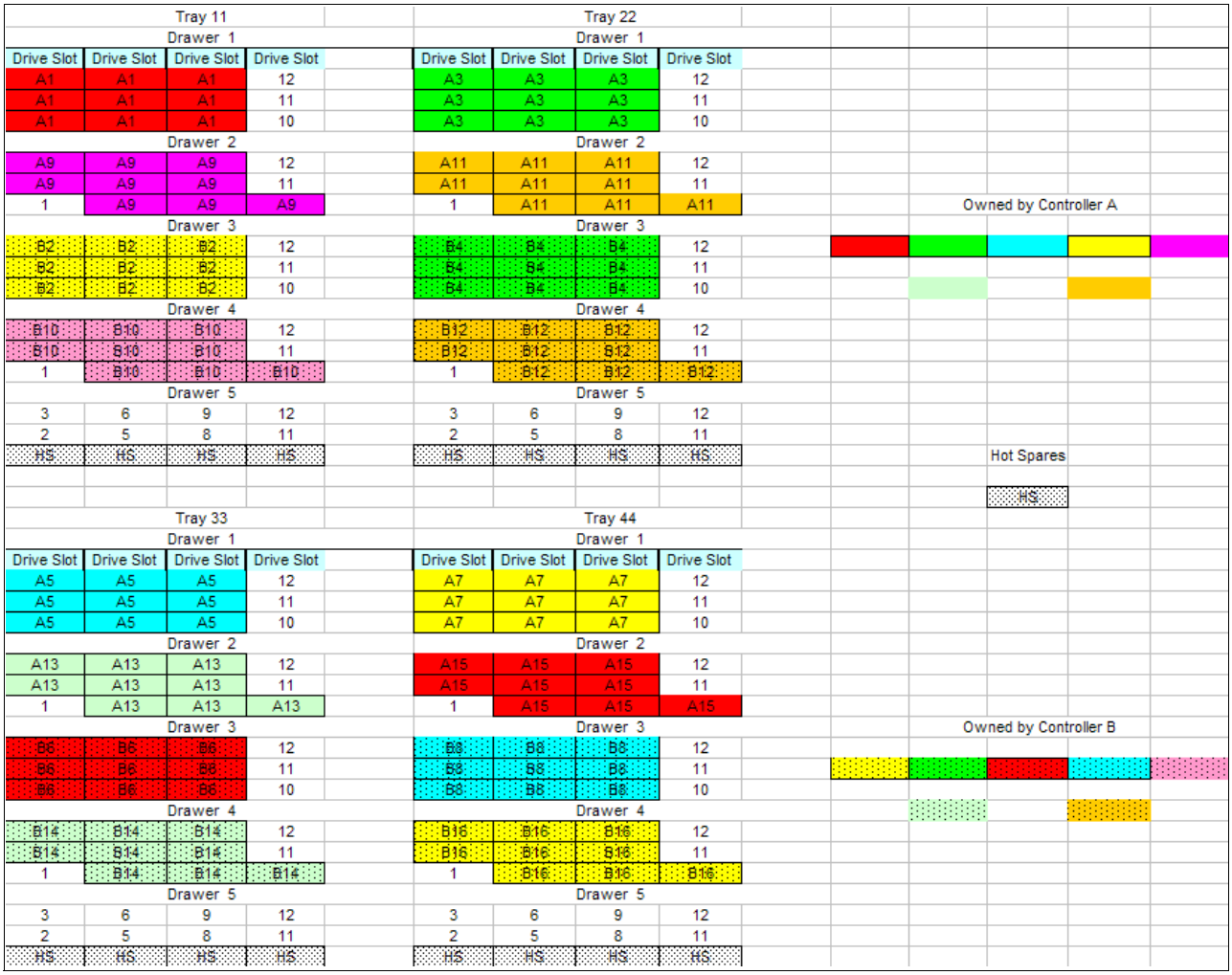


Figure 6-12 An 8 + P single enclosure and a single drawer array group layout

Table 6-2 Maximum performance test results as seen with the Figure 6-12 configuration

Measurement	Result
Throughput read performance	6339.48 MB/sec
Throughput write performance	5369.19 MB/sec
Throughput write with Cache Mirroring Enabled (CME) performance	3803.21 MB/sec

Figure 6-13 shows 8 + P shared enclosures with dedicated drawers for the array group layout. Table 6-3 shows the maximum performance test results for the Figure 6-13 configuration.

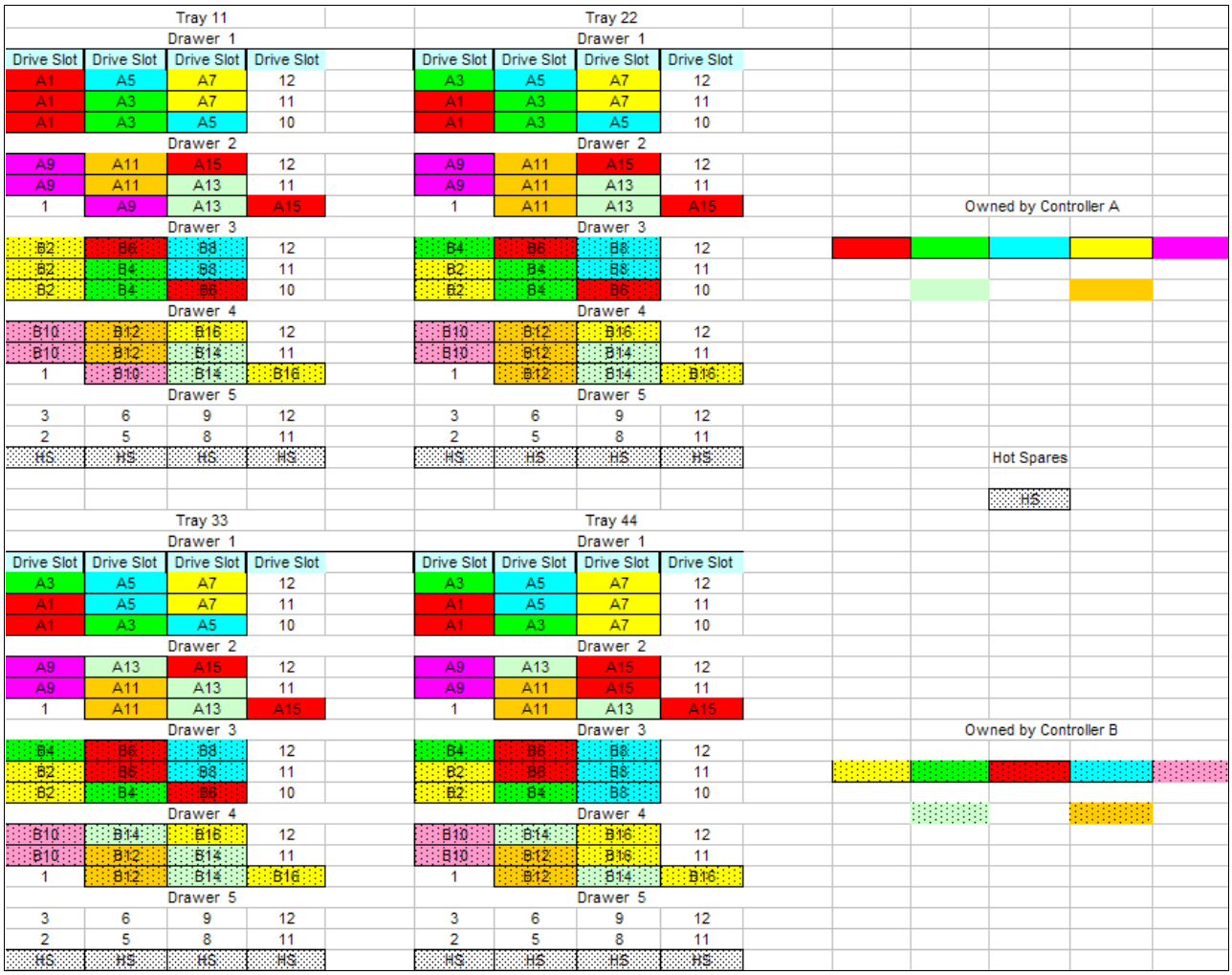


Figure 6-13 Test with an 8 + P shared enclosures and dedicated drawers for array group layout

Table 6-3 Maximum performance test results as seen with the Figure 6-13 configuration

Measurement	Result
Throughput read performance	6320.76 MB/sec
Throughput write performance	5373.92 MB/sec
Throughput write with CME performance	3801.89 MB/sec

As shown in Table 6-2 on page 295 and Table 6-3 on page 296, with Figure 6-12 on page 295 and Figure 6-13 on page 296, these tests adhered to the following best practice suggestions:

- Dedicate two drawers (in this example, drawers 1 and 2) to build arrays and LUNs that are owned by Controller A. Dedicate two drawers (in this example, drawers 3 and 4) to build arrays and LUNs that are owned by Controller B.
- Ensure that the array groups were created with as equal a balance of odd and even disks slots as possible for balanced channel or Environmental Service Module (ESM) workloads. Figure 6-13 on page 296 shows this practice. We have an odd number of drives in an array, but we have made sure to build them as equal arrays as much as possible by alternating the ninth drive of each array between the odd and even slots to keep the channels balanced.

Enclosure protection and drawer loss protection are not provided with these configurations. Make sure that you understand your availability requirements. For guidance in this area, see the *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679. In certain cases, you might only need a slight change in the configuration to meet the high availability requirement (at a small cost to the performance expectations).

Non-optimal performance layouts

The following configurations are examples of how poorly the EXP5060 environment can perform when you do not adhere to the best practice rules. We defined these example test case layouts in 8 + P RAID5 array groups so that you can easily compare them to the previous optimal test cases. These examples are worst-case scenarios. While you are in the planning stages, use these examples to help you estimate the effect on performance.

Figure 6-14 shows non-shared drawers and expansion trays. Table 6-4 shows the maximum performance test results as seen with the Figure 6-14 configuration.

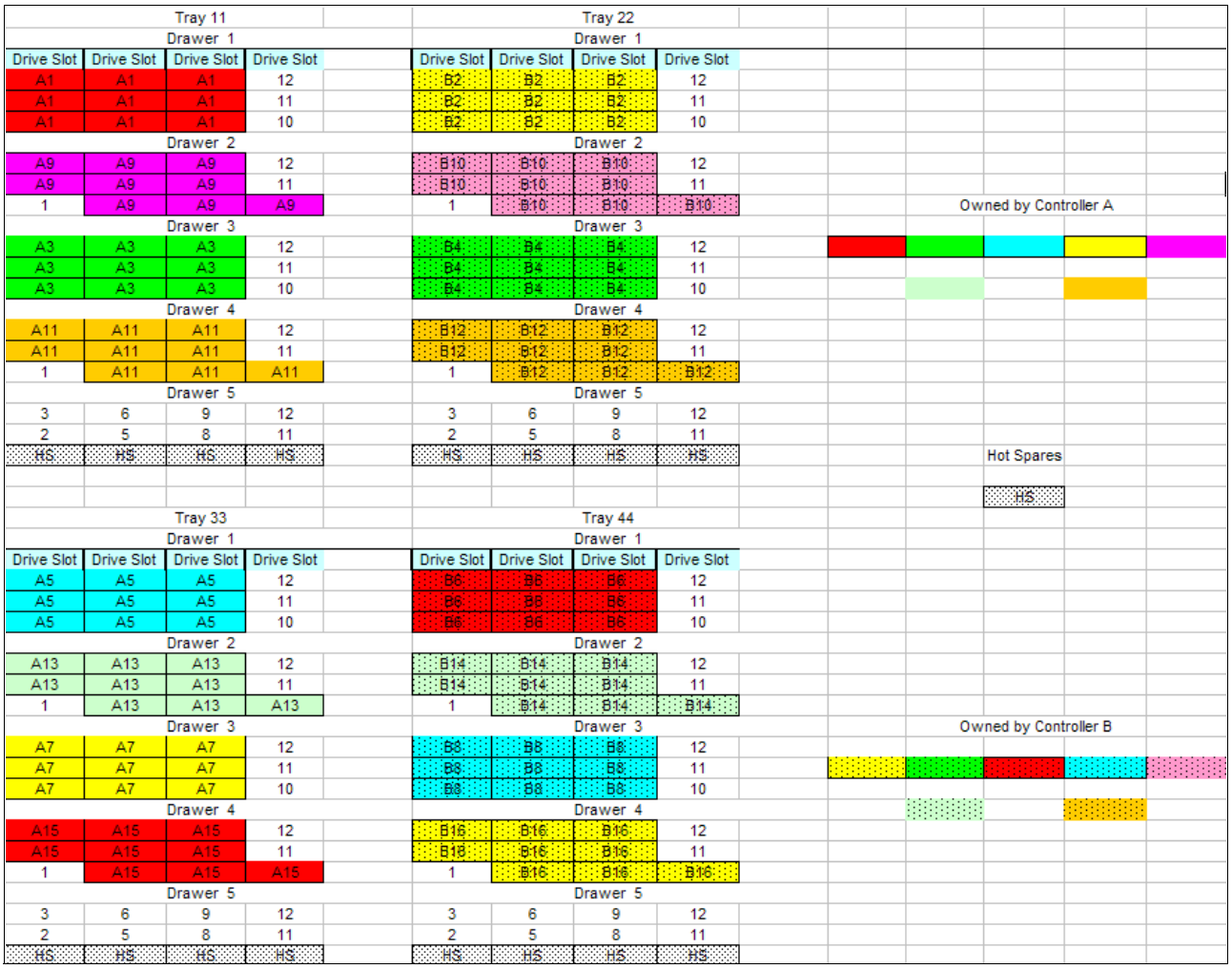


Figure 6-14 A balanced, non-shared drawers or expansions 8 + P layout

Table 6-4 Maximum performance test results as seen with the Figure 6-14 configuration

Measurement	Result
Throughput read performance	3172.75 MB/sec
Throughput write performance	2700.24 MB/sec
Throughput write with CME performance	2692.88 MB/sec

As shown in Table 6-4, the performance of the configuration that is shown in Figure 6-14 is about half of the performance of the two trunked connected channels. The same controller owns all of the array groups and LUNs in the expansion; therefore, it only uses half of the bandwidth.

Important: Dividing the drawers in an expansion tray between the controllers is the most critical of all guidelines when using the trunking method to gain increased performance. Failure to adhere to this rule nullifies the performance benefit.

Figure 6-15 shows non-balanced access across the drive channels by controllers in an 8 + P layout. Table 6-5 gives maximum performance test results from the Figure 6-15 configuration.

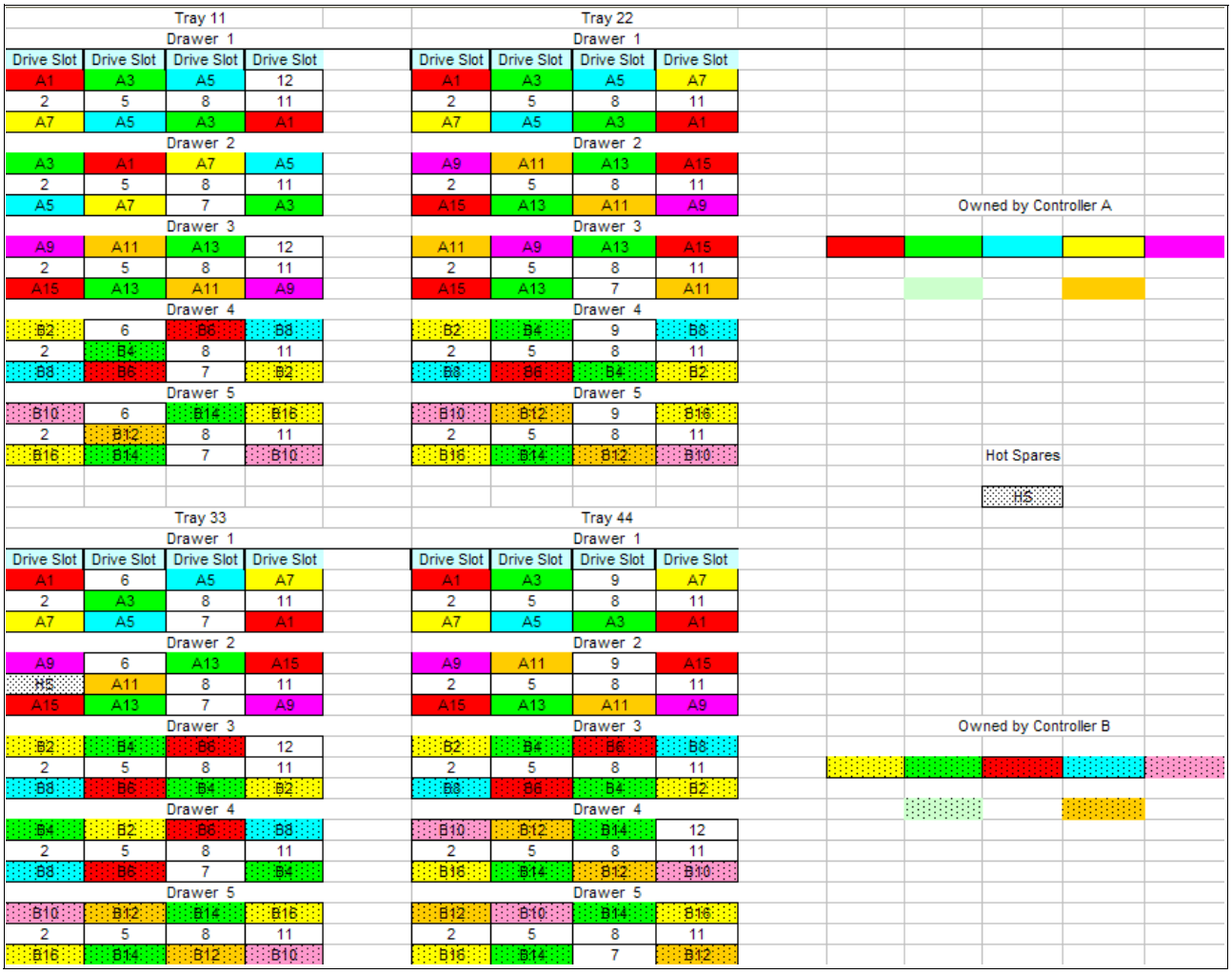


Figure 6-15 Non-balanced access across the drive channels by controllers in an 8 + P layout

Table 6-5 Maximum performance test results that were seen with the Figure 6-15 configuration

Measurement	Result
Throughput read performance	5183.80 MB/sec
Throughput write performance	4412.15 MB/sec
Throughput write with CME performance	3800.59 MB/sec

In Figure 6-15, we share the expansions between the two controllers, which helps to gain access to the full bandwidth of the trunked pair. However, with the addition of the fifth drawers being brought into the mix, we see an imbalance in the channel usage. Controller A has a higher drive count on expansion trays 11 and 12, and controller B has a higher number on expansion trays 33 and 44. As shown in Table 6-5, the performance is fair and might be acceptable, but it is not at an optimal level. When you need to use all five drawers, it is a better layout to split the fifth drawer's disks between the two controllers and to build dedicated arrays and LUNs with them to minimize the performance effect. For an example that helps to minimize this negative effect, see the configuration section of the *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679.

Figure 6-16 shows sharing the device control manager (DCM) with controllers in an 8 + P layout. Table 6-6 shows the maximum performance test results that were seen with the Figure 4-5 configuration.

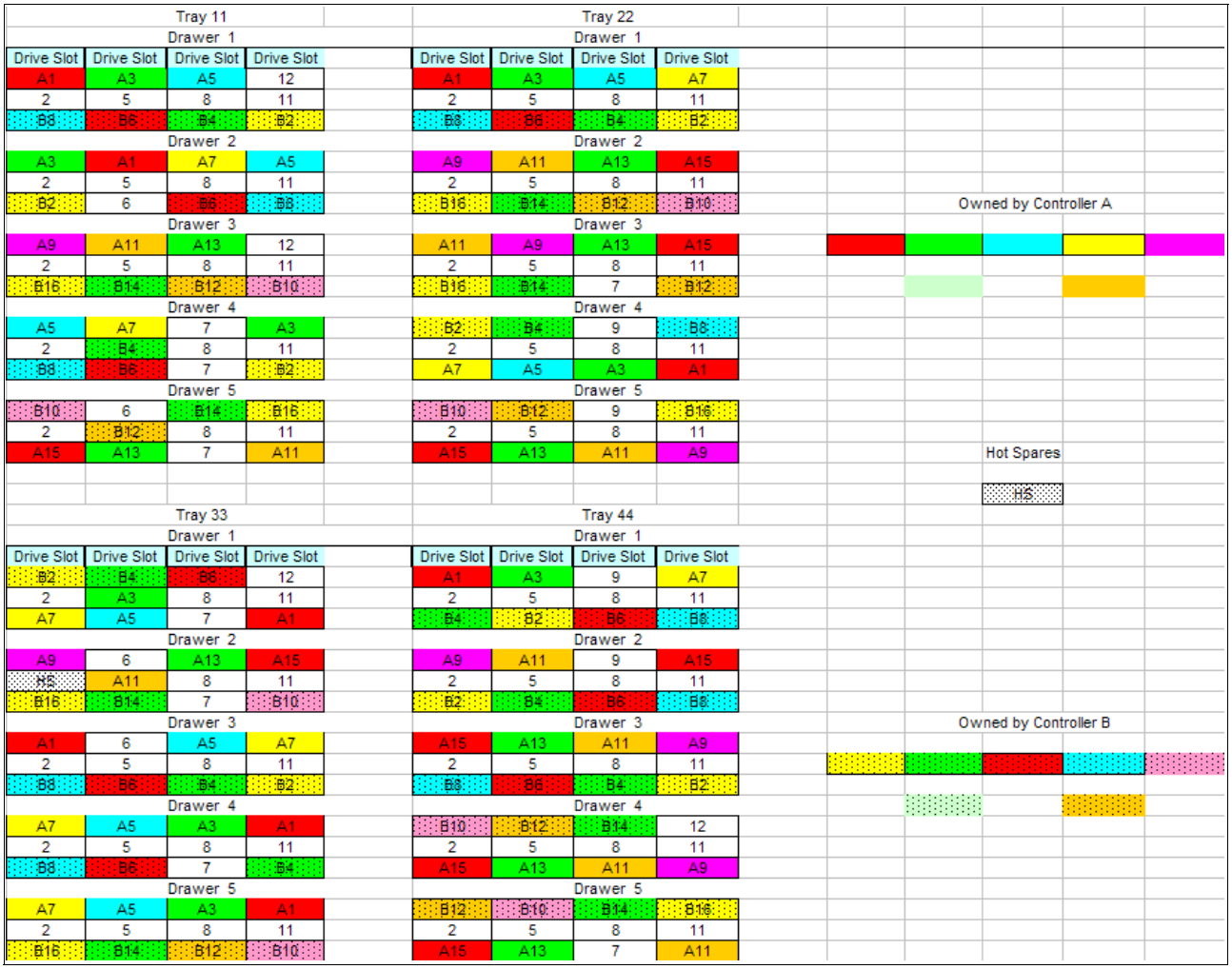


Figure 6-16 Sharing the DCM with controllers in an 8 + P layout

Table 6-6 Maximum performance test results as seen with the Figure 6-16 configuration

Measurement	Result
Throughput read performance	5451.83 MB/sec
Throughput write performance	4516.85 MB/sec
Throughput write with CME performance	3799.36 MB/sec

In Figure 6-16, we share the expansions between the two controllers, which helps to gain access to the full bandwidth of the trunked pair. However, with the addition of the fifth drawers being brought into the mix, we see an imbalance in the DCMs. As shown in Table 6-5 on page 299, the performance is fair and might be acceptable, but it is not at an optimal level.

For ways to minimize negative effects when you need to use all five drawers for storage capacity and drive usage, see the suggestions outlined in the configuration section of the *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679.

Figure 6-17 shows a non-balanced single tray in a single drawer 8 + P configuration. Table 6-7 shows the maximum performance test results that were seen with configuration in Figure 4-6.

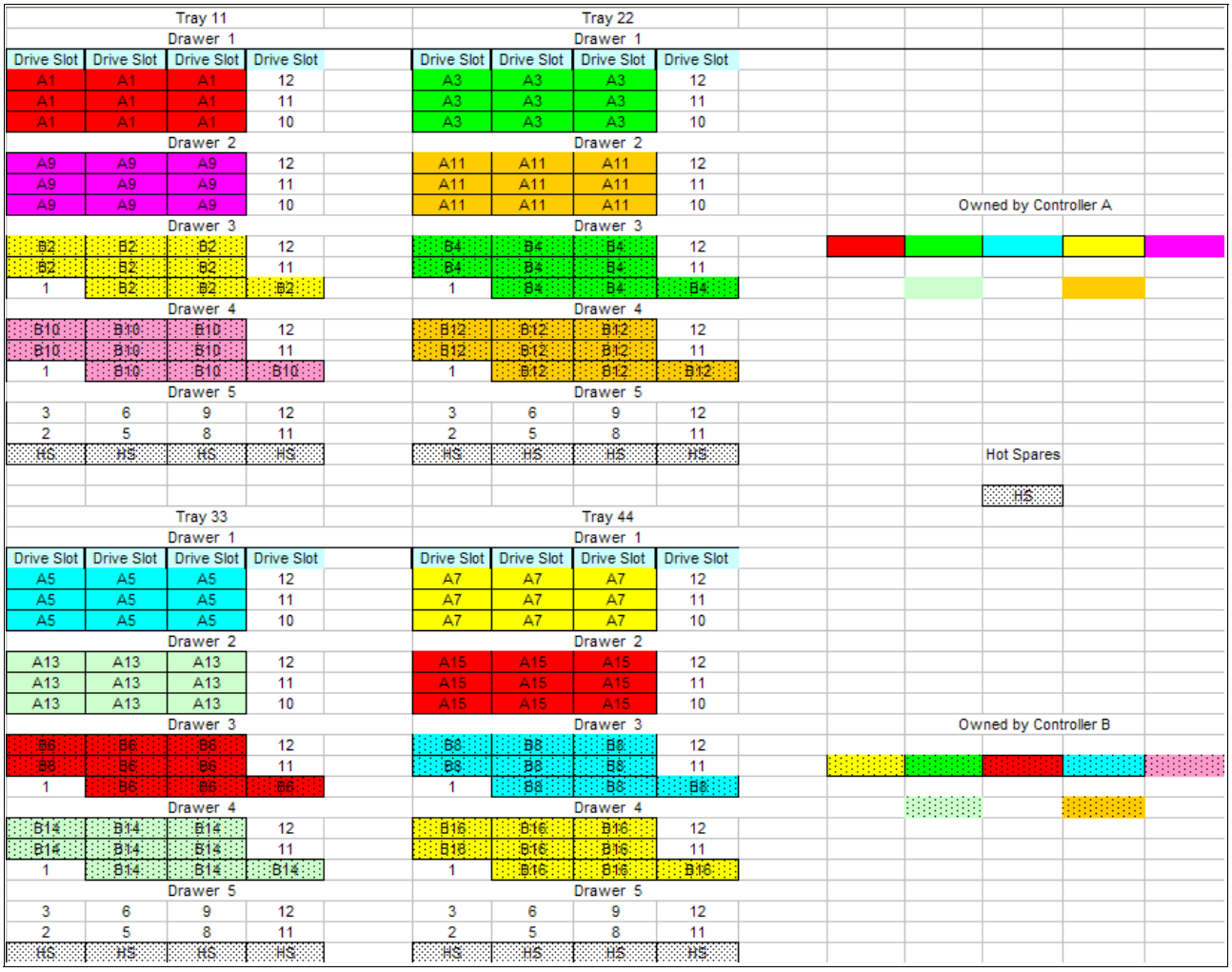


Figure 6-17 Non-balanced single tray in a single drawer 8 + P configuration

Table 6-7 Maximum performance test results as seen with the Figure 6-17 configuration

Measurement	Result
Throughput read performance	5714.90 MB/sec
Throughput write performance	4844.83 MB/sec
Throughput write with CME performance	3802.34 MB/sec

In the Figure 6-17 configuration, we see the results of an array layout that meets most of the rules for a best practice configuration, but the layout fails to evenly balance the arrays across the odd and even drives for a balanced use of the connected loops. Table 6-7 shows performance numbers that might be at an acceptable level, but they miss the optimal performance that can be achieved with a slight drive selection layout change (see Figure 6-12 on page 295, for example).

Mixed configuration test runs

The best usage environment for the EXP5060 expansion enclosure and the Serial Advanced Technology Attachment (SATA) drives is to support a high throughput-based transaction volume. However, clients can use this high-capacity expansion enclosure in many nearline environments. Often, the environment supports a mix of workloads. To help you in this planning area, we created a scenario with various arrays and LUNs and tested these arrays and LUNs with multiple host I/O workloads. We focused our testing on the best practices in these environments. We used the *iometer* data generator test tool.

Multiple arrays and LUNs in a variety of configurations

Figure 6-18 shows the initial configuration of various 4 + P arrays and LUNs that was used to test a variety of workload types. With the DS5300 and the EXP5060s configured in this manner, we built an array group that used many creation patterns of trays and drawers. Two LUNs on each tray used 128 KB and 512 KB segment sizes. We used various I/O block sizes to gather data points for a variety of sequential I/O workloads as well as a random pattern test run. We tested these areas:

- Maximum throughput capabilities with various sequential host block sizes
- Maximum I/O per second (IOPS) capabilities with various random host block sizes
- A test case of the effects on a non-optimal configuration

Figure 6-18 shows a test of a trunked EXP5060 with variety of 4 + P test array configurations.

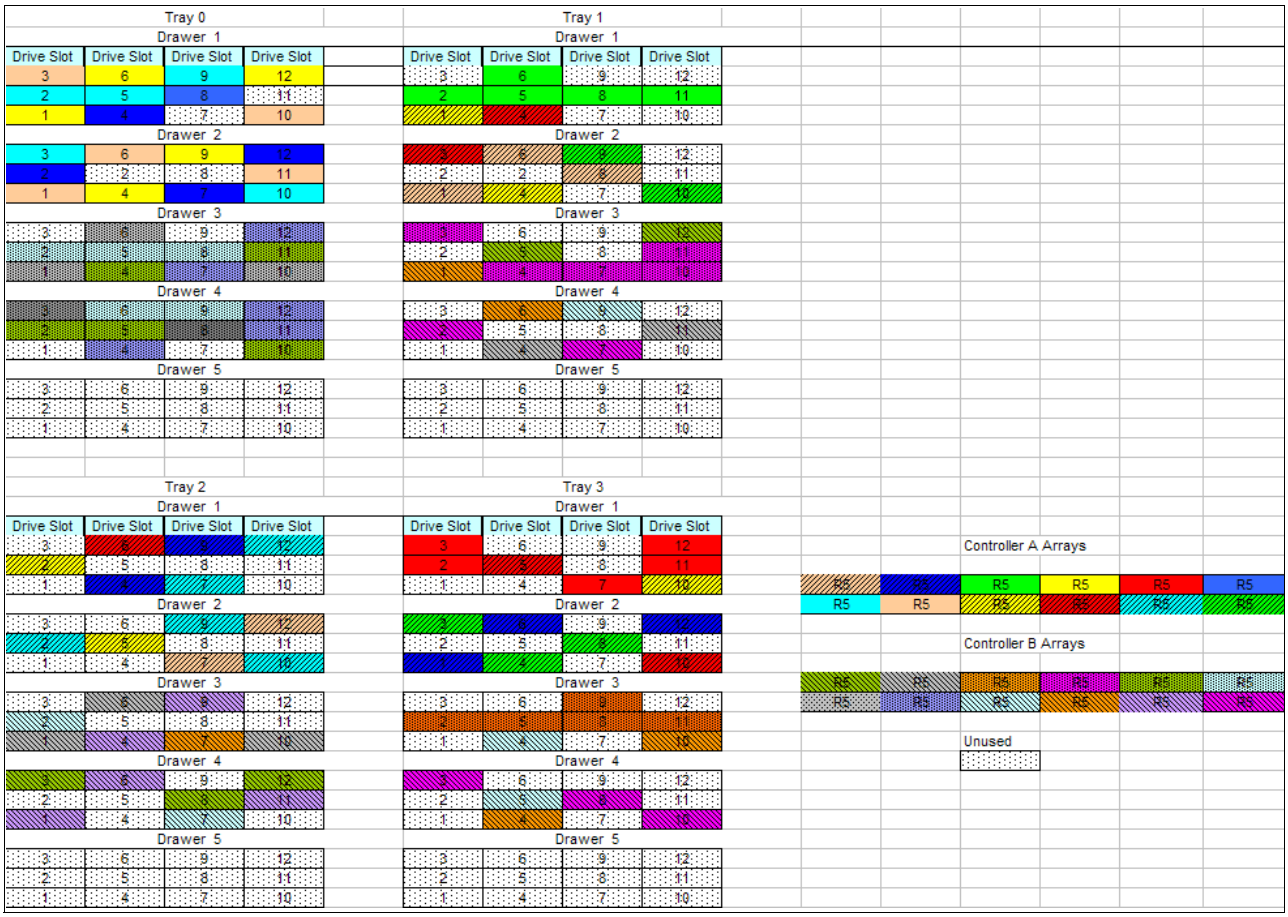


Figure 6-18 Trunked EXP5060 with variety of 4 + P test array configurations

We focused only on the best practices that affected the performance in this test configuration layout. We created the array groups in the following manner:

- Single EXP5060 with multiple drawers
- Single EXP5060 with a single drawer
- Multiple EXP5060s with multiple drawers

To avoid affecting the performance, we did not use Drawer 5.

Drawer 5: When using Drawer 5 for arrays and LUNs, use the drives in the drawer to build separate arrays and LUNs that are evenly spread across the two controllers. This layout helps to minimize the effect of the shared drawer on the overall performance.

We configured each array group with two LUNs of equal size using 128 KB segment size for one LUN and 256 KB segment size for the other LUN. We wanted to test the effects of multiple host I/O block sizes using separate segment sizes. Table 6-8 shows the various results for numerous sequential workload runs with 96 user sessions configured across four hosts. We used host block sizes of 256 KB, 512 KB, 1 MB, and 2 MB to simulate various sequential application types. We collected these results for read and write runs with a queue depth of 1 for reads and a queue depth of 121 for writes. We disabled all cache (both read and write) for the runs that are reported in Table 6-8 to obtain results that were as close to disk throughput results as possible.

Table 6-8 Maximum throughput with no cache enabled

Workload operation type	Host block size of I/O	Throughput MB/sec
Sequential read	256 KB	6.889 GB
Sequential read	512 KB	6.915 GB
Sequential read	1 MB	6.914 GB
Sequential read	2 MB	6.914 GB
Sequential write	256 KB	4.008 Gb
Sequential write	512 KB	4.428 GB
Sequential write	1 MB	4.644 GB
Sequential write	2 MB	4.632 GB

The following results show running the same workload test with all cache enabled. In this case, we used varying queue depths to improve the throughput levels. Table 6-9 shows the throughput results for queue depths of 1 and 121, as seen in our test runs.

Table 6-9 Maximum throughput with cache mirroring enabled

Workload operation type	Host block size of I/O	Throughput MB/sec
Sequential read	256 KB	4.702 GB/6.915 GB
Sequential read	512 KB	5.695 GB/6.917 GB
Sequential read	1 MB	6.509 GB/6.916 GB
Sequential read	2 MB	6.899 GB/5.857 GB
Sequential write	256 KB	3.561 GB/3.239 GB
Sequential write	512 KB	3.976 GB/3.265 GB

Workload operation type	Host block size of I/O	Throughput MB/sec
Sequential write	1 MB	3.330 GB/3.871 GB
Sequential write	2 MB	3.822 GB/4.191 GB

As shown in Table 6-9, the throughput is affected by enabling write cache mirroring. The effect on the throughput is lessened by the increases in the amount of full stripe performance that is reached (with the larger block size). The effect on the throughput is lessened due in part to the number of LUNs that are able to perform full stripe writes at this block size verses the smaller block sizes. In Table 6-10 on page 304, we show a snapshot of the LUN's performance rates.

In Table 6-10 on page 304, we show the results of transaction (IOPS)-based workloads when used with the multiple array/LUN (trunked or "T") configuration that is shown in Figure 6-18 on page 302. We set the queue depth for this test to 1 to minimize the effect on response time results. We ran the tests with cache enabled. We wanted to show that although this layout is not the best layout, it provides performance equal to the capability of the disk type that is being used.

Table 6-10 Random I/O read and write performance data with a multiple array/LUN configuration T

Workload type	Host block size of I/O	IOPS	Response times
Random read	4 KB	8940	10.73 ms
Random read	8 KB	8840	10.85 ms
Random read	16 KB	8722	11 ms
Random read	32 KB	8489	11.30 ms
Random read	64 KB	8037	11.94 ms
Random read	128 KB	7040	13.12 ms
Random read	256 KB	5536	17.33 ms
Random read	512 KB	3627	26.45 ms
Random write	4 KB	5416	17.72 ms
Random write	8 KB	5470	17.54 ms
Random write	16 KB	5940	16.15 ms
Random write	32 KB	6526	14.70 ms
Random write	64 KB	4431	21.65 ms
Random write	128 KB	3815	25.15 ms
Random write	256 KB	3054	31.40 ms
Random write	512 KB	2664	35.92 ms

In Table 6-10, understand that we set the queue depth per LUN to 1 to minimize the amount of time that was spent sitting in the queue waiting. As queue depth is increased, the response time value also increases.

Single array and multiple LUN configuration

In this configuration, we configured the subsystem with a single array, as shown in Figure 6-19 on page 305. We then divided this array into 48 LUNs, which were spread evenly between the two controllers and presented to the hosts, which were running an iometer test suite.

Figure 6-19 Single, large RAID10 array that was built across all expansion trays and drawers

We do *not* advise this configuration, because it does not honor any of the throughput best practice rules. However, because the DS5300 allows you to build a large array that uses a large number of disks, we tested this configuration to see how the I/O workload compared to the previously tested configuration. Table 6-11 and Table 6-12 show the performance that was seen with this configuration. We built this configuration as a RAID10 array with 180 disks; therefore, the test results are not a fair comparison to the test results that were seen in the previous test scenarios. Table 6-11 shows the throughput results for queue depths of 1 and 121 in our test runs.

Table 6-11 Single array throughput with cache but without cache mirroring enabled

Workload operation type	Host block size of I/O	Throughput MB/sec
Sequential read	256 KB	1.345 GB/3.076 GB
Sequential read	512 KB	2.140 GB/6.737 GB
Sequential read	1 MB	4.569 GB/6.762 GB
Sequential read	2 MB	6.454 GB/5.476 GB
Sequential write	256 KB	4.058 GB/1.892 GB
Sequential write	512 KB	4.211 GB/1.053 GB
Sequential write	1 MB	4.320 GB/1.422 GB
Sequential write	2 MB	4.3.96 GB/1.353 GB

Table 6-12 Random I/O read and write performance data with one large array with multiple LUNs T

Workload type	Host block size of I/O	IOPS	Response times
Random read	4 KB	6817	21.11 ms
Random read	8 KB	7110	20.23 ms
Random read	16 KB	7008	20.54 ms
Random read	32 KB	6908	20.83 ms
Random read	64 KB	6659	21.62 ms
Random read	128 KB	6185	23.27 ms
Random read	256 KB	5383	26.73 ms
Random read	512 KB	4280	33.62 ms
Random write	4 KB	6541	18.58 ms
Random write	8 KB	8632	16.67 ms
Random write	16 KB	7565	19.03 ms
Random write	32 KB	7803	18.44 ms
Random write	64 KB	5785	24.88 ms
Random write	128 KB	5108	28.17 ms
Random write	256 KB	4442	32.39 ms
Random write	512 KB	3361	42.80 ms

You can achieve greater performance in a RAID10 mixed workload configuration where nearline (SATA) storage is an acceptable solution if you build a large array that follows the suggestions that we have outlined. The largest possible array size is one-half of the total disk population, or 120 drives in a 60x60 RAID10 configuration. In this manner, you can build two of these arrays with one array that is owned by controller A and one array that is by controller B. The same controller, to which the array is defined, must own all of the LUNs that are created on these arrays.

Figure 6-20 shows a suggested layout for this type of configuration with two 56x56 RAID10 arrays leaving hot spares for protection.

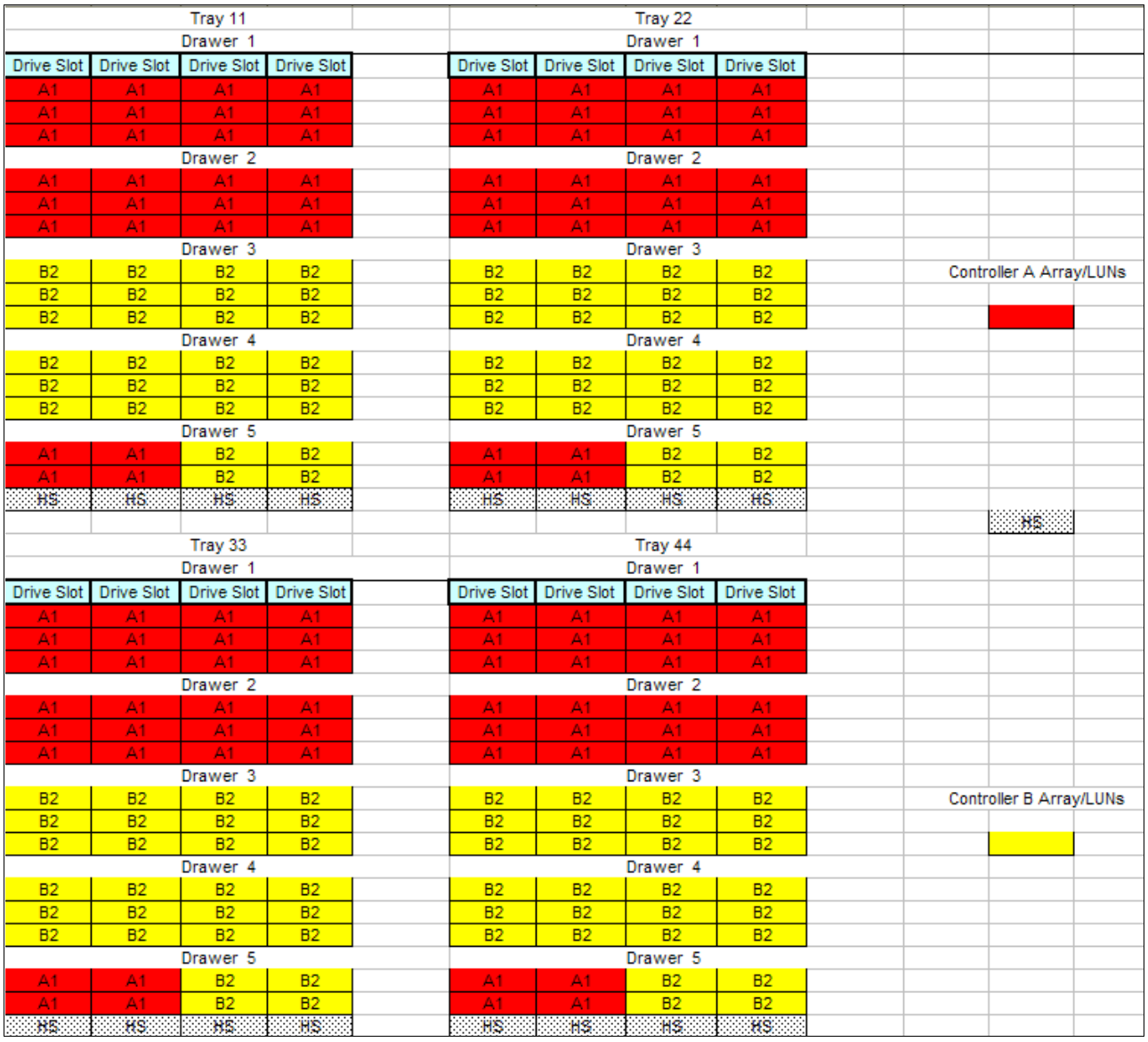


Figure 6-20 Suggested large RAID10 array layout

6.6 Fabric considerations

When connecting the Midrange Storage Subsystem to your SAN fabric, it is best to consider what are all the other devices and servers that will share the fabric network, which is related to how you configure your zoning. See 2.4, “Planning your host attachment method” on page 51, for guidelines and details on how to establish the SAN infrastructure for your environment. Remember that a noisy fabric is a slow fabric. Unnecessary traffic makes for poor performance.

Specific SAN switch settings that are of particular interest to the Midrange Storage Subsystem environment, and can impact performance, are those that help to ensure *in-order-delivery* (IOD) of frames to the endpoints. The Midrange Storage Subsystems cannot manage out of order frames, and retransmissions will be required for all frames of the transmitted packet. See your specific switch documentation for details on configuring parameters.

7

IBM Midrange Storage Subsystem tuning with typical applications

In this chapter, we provide general guidelines and tips to consider when implementing certain popular applications with the IBM Midrange Storage Subsystem.

Our intent is not to present a single way to set up your solution, but rather to show various areas of interest that you need to consider when implementing these applications. Each situation and implementation will have its own specific requirements.

We provide general guidance and tips for using the following software products:

- IBM DB2
- Oracle Database
- Microsoft SQLserver
- IBM Tivoli Storage Manager
- Microsoft Exchange

7.1 DB2 database

In this section, we describe the usage of the IBM Midrange Storage Subsystem with a DB2 database. We cover the following topics:

- Data location
- Database structure
- Database RAID type
- Redo logs RAID type
- Volume management

7.1.1 Data location

DB2 applications generally have two types of data:

- Data consisting of application programs, indexes, and tables, and stored in *table spaces*.
- Recovery data, made up of the database logs, archives, and backup management.

In an OLTP environment, it is a good idea to store these two data types separately, that is, on separate logical drives, on separate arrays. Under certain circumstances, it can be advantageous to have both logs and data co-located on the same logical drives, but they are special cases and require testing to ensure that the benefit will be there for you.

7.1.2 Database structure

Table spaces can be configured in three possible environments:

- Database Managed Storage (DMS) table space
- System Managed Storage (SMS) table space
- Automatic Storage (AS) table space — new with V8.2.2

In a DMS environment, all the DB2 objects (data, indexes, large object data (LOB) and long field (LF)) for the same table space are stored in the same files. DB2 also stores metadata with these files as well for object management.

In an SMS environment, all the DB2 objects (data, indexes, LOB, and LF) for the same table space are stored in separate files in the directory.

Restriction: When using concurrent or direct I/O (CIO or DIO) with earlier versions than AIX 5.2B, you must separate the LOB and LF files on separate table spaces due to I/O alignment issues.

In both DMS and SMS table space environments, you must define the container type to be used; either file system or raw device.

In the AS table space environment, there are no containers defined. This model has a single management method for all the table spaces on the server that manages where the data is located for them on the storage.

In all cases, striping of the data is done on an *extent* basis. An extent can only belong to one object.

DB2 performs data retrieval by using three type of I/O *prefetch*:

- RANGE: Sequential access either in the query plan or through sequential detection at run time. Range request can be affected most by poor configuration settings.
- LIST: Prefetches a list of pages that are not necessarily in sequential order.

  Tip: DB2 will convert a LIST request to RANGE if it detects that sequential ranges exist.

- LEAF: Prefetches an index leaf page and the data pages pointed to by the leaf.
  - LEAF page is done as a single I/O.
  - Data pages on a leaf are submitted as a LISTrequest.

Prefetch is defined by configuring the application for the following parameter settings:

- PREFETCHSIZE (PS): A block of contiguous pages requested. The block is broken up into prefetch I/Os and placed on the prefetch queue based on the level of I/O parallelism that can be performed:
  - PS must be equal to the size of all the storage subsystem's logical drives stripe sizes so that all drives that make up the container are accessed for the prefetch request.
  - For example, if a container resides across two logical drives that were created on two separate RAID arrays of 8+1p, then when the prefetch is done, all 16 data drives need to be accessed in parallel.
- Prefetch is done on one extent at a time; but with careful layout planning can be paralleled.
- EXTENTSIZE (ES): It is both the *unit of striping granularity*, and the *unit of prefetch I/O size*. Good performance of prefetch is dependent on a well configured ES:
  - Choose an extent size that is equal to or a multiple of the segment size used by the subsystem's logical drives.

    Best practice: The ES value must be a multiple of the segment size, and be evenly divisible into the stripe size.

  - In general, configure the extent size to be between 128 KB and 1 MB, but at least must be equal to 16 pages. DB2 supports page sizes equal to 4 KB, 8 KB, 16 KB, or 32 KB in size, which means that an ES must not be less than 64 KB (16 X 4 KB (DB2's smallest page size)).
- Prefetch I/O parallelism for midrange storage performance requires DB2_PARALLEL_IO to be enabled, which allows you to configure for all or one table space to be enabled for it.
- NUM_IOSERVERS: The number of parallel I/O requests that you will be able to perform on a single table space.
- Starting with V8.2 of DB2, a new feature AUTOMATIC_PREFETCHSIZE was introduced.

  A new database tablespace will have DFT_PREFETCH_SZ= AUTOMATIC.

  The AUTOMATIC setting assumes a RAID 5 array of 6+1p, and will not work properly with an 8+1p size array. See the DB2 documentation for details on proper settings to configure this new feature.

Figure 7-1 provides a diagram showing how all these pieces fit together.

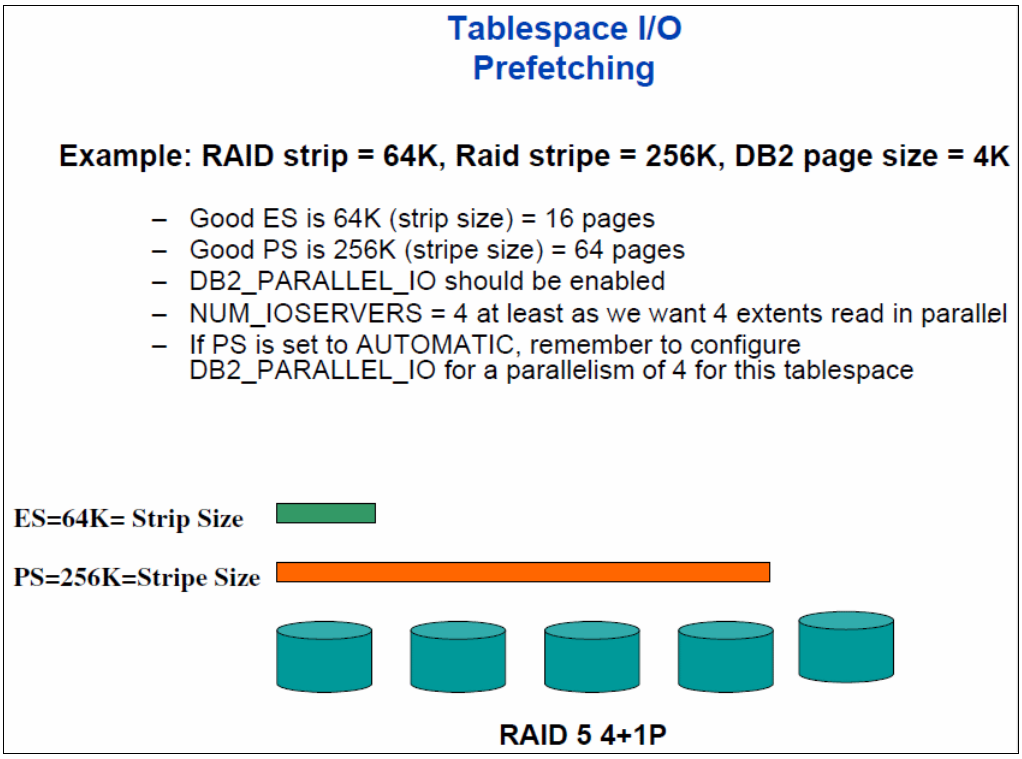


Figure 7-1 Diagram of tablespace I/O prefetch structure

7.1.3 Database RAID type

In many cases, OLTP environments contain a fairly high level of read workload. In this area, your application might vary, and behavior is very unpredictable, so try to test performance with your actual application and data.

In most cases, it has been found that laying out the datafile tables across a number of logical drives that were created across several RAID 5 arrays of 8+1 parity disks, and configured with a segment size of 64 KB or 128 KB, is a good starting point to begin testing with. This, coupled with host guidelines to help avoid offset and striping conflicts, seem to provide a good performance start point to build from. A point to remember is that high write percentages might result in a need to use RAID 10 arrays rather than the RAID 5, which is environment specific and will require testing to determine. A rule of thumb is, if there are greater than 25–30% writes, then you might want to look at RAID 10 over RAID 5.

Best practice: Spread the containers across as many drives as possible, and ensure that the logical drive spread is evenly shared across the Midrange Storage Subsystem's resources. Use multiple arrays where larger containers are needed.

DB2 uses a block based buffer pool to load the prefetched RANGE of I/O into. Though the RANGE is a sequential prefetch of data; in many cases the available blocks in the buffer might not be sequential, which can result in a performance impact. To assist with this management, certain operating system primitives can help. They are VECTOR or SCATTER/GATHER I/O primitives. For certain operating systems, you might need to enable DB2_SCATTERED_IO to accomplish this function. There are also page cleaning parameters that can be configured to help clear out the old or cold data from the memory buffers. See your DB2 documentation for details and guidelines.

7.1.4 DB2 logs and archives

The DB2 logs and archive files generally are high write workloads, and sequential in nature. RAID 10 logical drives are best suited for this type of workload.

Because they are critical files to protect in case of failures, it is a good idea to keep two full copies of these files on separate disk arrays in the storage server, thus protecting you from the extremely unlikely occurrence of a double disk failure, which can result in data loss.

Also, because they are generally smaller files and require less space, you can use two separate arrays of 1+1 or 2+2 RAID1 to hold the logs and the mirror pair separately.

Logs in DB2 use the operating system's default blocksize for I/O (generally 4 K) of sequential data at this time. Because the small write size has no greater penalty on the Midrange Storage Subsystem with higher segment size, you can configure the logical drive with a 64 KB or 128 KB segment size.

Also, it is best to place the redo logs on raw devices or volumes on the host system versus on the file system.

7.2 Oracle databases

In this section, we describe the usage of the IBM Midrange Storage Subsystem with an Oracle database application environment. We cover the following topics:

- Data types
- Data location
- Database RAID and disk type
- Redo logs RAID type
- Volume management

7.2.1 Data types

Oracle stores data at three levels:

- The first or lowest level consists of data blocks, also called logical blocks or pages.

  One data block corresponds to a specific number of bytes corresponding to physical database space on the disk. A data block is the smallest unit of data used by a database. In contrast, at the physical, operating system level, all data is stored in bytes. Each operating system has a block size. Oracle requests data in multiples of Oracle data blocks, not operating system blocks.

- The second level consist of extents.

  An extent is a specific number of contiguous data blocks, which are allocated for storing a specific type of information.

  A table space that manages its extents locally can have either uniform extent sizes or variable extent sizes that are determined automatically by the system:

  – For uniform extents, the default size of an extent is 1 MB.

  – For system-managed extents, Oracle determines the optimal size of additional extents, with a minimum extent size of 64 KB. If the table spaces are created with a *segment space management auto*, and if the database block size is 16 KB or higher, then Oracle manages segment size by creating extents with a minimum size of 1 MB, which is the default for permanent table spaces.

- The third level of database storage greater than an extent is called a segment.

  A segment is a set of extents, each of which has been allocated for a specific data structure and all of which are stored in the same table space. Oracle allocates space for segments in units of one extent, as extents are allocated as needed. The extents of a segment might or might not be contiguous on disk. A segment and all its extents are stored in one table space. Within a table space, a segment can include extents from more than one file. The segment can span datafiles. However, each extent can contain data from only one datafile.

7.2.2 Data location

With Oracle applications, there are generally two types of data:

- Primary data, consisting of application programs, indexes, and tables.
- Recovery data, consisting of database backups, archive logs, and redo logs.

For data recovery reasons, in the past there has always been a guideline that several categories of the RDBMS files be isolated from each other and placed in separate physical disk locations. Thus the redo logs had to be separated from your data, indexes separated from the tables, and rollback segments as well. Today, the guideline is to keep user datafiles separated from any files needed to recover from any datafile failure.

This strategy ensures that the failure of a disk that contains a datafile does not also cause the loss of the backups or the redo logs needed to recover the datafile.

Because indexes can be rebuilt from the table data, it is not critical that they be physically isolated from the recovery-related files.

Because the Oracle control files, online redo logs, and archived redo logs are crucial for most backup and recovery operations, it is a good idea to keep at least two copies of these files stored on separate RAID arrays, and that both sets of these files must be isolated from your base user data as well.

In most cases with Oracle, the user data application workload is transaction based with high random I/O activity. With an OLTP application environment, you might see that an 8 KB database block size has been used, whereas with Data Warehousing applications, a 16 KB database block size is typical. Knowing what these values are set to, and how the devices on which the datafiles reside were formatted, can help in prevention of added disk I/O due to layout conflicts. For additional information, see the previous description of host parameters in 2.5, “Host support and multipathing” on page 57.

7.2.3 Database RAID and disk types

In many cases, OLTP environments contain a fairly high level of read workload. In this area your application might vary, and behavior is very unpredictable. Try to test performance with your actual application and data.

With the default extent size of 1 MB, the segment size must be selected to allow a full stripe write across all disks. In most cases, it has been found that laying out the datafile tables across a number of logical drives that were created across several RAID 5 arrays of 8+1 disks, and configured with a segment size of 64 KB or 128 KB, is a good starting point. This, coupled with host guidelines to help avoid offset and striping conflicts, seems to provide a good performance start point to build from.

Keep in mind also that high write percentages might result in a need to use RAID 10 arrays rather than RAID 5, which is environment specific and will require testing to determine.

A rule of thumb is if there are greater than 25–30% writes, then you might want to look at RAID 10 over RAID 5.

Best practice: Spread the datafiles across as many drives as possible, and ensure that the logical drive spread is evenly shared across the Midrange Storage Subsystems resources.

Enclosure loss protection must always be employed, which will ensure that in the unlikely event of an enclosure failure, then all the arrays and databases will continue to operate, although in a degraded state. Without enclosure loss protection, any failure of an enclosure will have major impact on those LUNs residing within that enclosure.

With the introduction of the new SSD drives, if the table space is of a size where it will fit on these devices, you might want to place it on them. However, in many cases the table space can be much larger and then the use of 15 K rpm drives is best to ensure maximum throughput. The use of the 15 K rpm drives gives a 20–30% performance increase over 10 K rpm disks. Using more spindles to spread the workload across provides a definite advantage to building a large logical drive on high capacity disks to use fewer spindles, which gives definite transaction response time advantages, but does make use of more disks.

Ensure that enough spare drives are on hand to ensure that any disk failure can fail over onto a spare disk. Also ensure that for each type and speed of disk utilized that enough spares are employed to cover them. It is possibly good practice to use only high-speed versions of the drives used, because this will ensure that when a spare is in use that it does not degrade the array by having a slower speed drive introduced on a disk failure. A slower disk introduced to any array will cause the array to run at the slower speed.

7.2.4 Redo logs: RAID types

The redo logs and control files of Oracle generally are both high write workloads, and are sequential in nature. As they are critical files, it is a good idea to keep two full copies on separate disk arrays in the storage server. You do it to protect against the (extremely) unlikely occurrence of a double disk failure, which can result in data loss.

Place the redo logs on RAID 10 logical drives. Avoid RAID 5 for these logs; dedicate a set of disks for the redo logs.

Because they are generally smaller files and require less space, you can use two separate arrays of 1+1 or 2+2 RAID 1 to hold the logs and the mirror pair separately.

Redo logs in Oracle use a 512 byte I/O blocksize of sequential data at this time. Because the small write size has no greater penalty on the Midrange Storage Subsystem with higher segment size, configure the logical drive with a 64 KB or 128 KB segment size.

Also, place the redo logs on raw devices or volumes on the host system rather than the file system. Furthermore, keep the archive logs on separate disks, because disk performance can degrade when the redo logs are being archived.

7.2.5 TEMP table space

If the files with high I/O are datafiles that belong to the TEMP table space, then investigate whether to tune the SQL statements performing disk sorts to avoid this activity, or to tune the sorting process. After the application has been tuned and the sort process checked to avoid unnecessary I/O, if the layout is still not able to sustain the required throughput, then consider separating the TEMP table space onto its own disks.

7.2.6 Cache memory settings

The default cache memory on the Midrange Storage Subsystem is 8 KB (on firmware 7.xx and higher), because 8 KB is an optimal database block size. Setting the cache block size to 16 KB can allow for better cache write performance by lessening the operations, but can decrease the cache availability for usage by other operations and applications. Doing it might or might not be a good solution; a better choice might be to set the blocksize to the new 8 KB size that is available with the 7.xx firmware levels.

Setting the Start/Stop Flushing size to 50/50 is a good place to start. Monitor the cache settings for best performance and tune the settings to requirements. See Figure 7-2.

The Oracle Automatic Storage Management (ASM) can be utilized to provide various levels of redundancy to the environment. The methods available are external redundancy, normal redundancy, and high redundancy.

External redundancy means that we are counting on a fault-tolerant disk storage subsystem. So that is what we use with DS3K/4K/5K with RAID 5 or RAID 10.

Normal redundancy means that ASM does software mirroring with two separate devices either from separate or common storage subsystems.

High redundancy means that ASM does software mirroring with 3 copies of the data, which is generally used to create a disaster recovery solution for the environment by placing the mirrors on two separate servers locally, and a third that is remote.

These methods can be used enhance the midrange RAID level to provide either an extra measure of protection or for remote disaster recovery.

In all cases, it is always best that write cache mirroring on the storage subsystem be enabled, even though an additional copy is being made, because the write to each storage device must be protected to decrease the need to for recovery actions.

Though using write cache in this situation does have a small performance cost, unless it is a data base update or initial load, there is no good justification to risk the data on the mirror. Write cache mirroring being disabled must only be done during large loads, and then only when they can be restarted again if errors occur. Finally, as described earlier with the new DS5100 and DS5300 the design improvements in the new subsystems make the mirrored operations perform far faster than previous subsystems even for the large loads so the benefit becomes marginal.

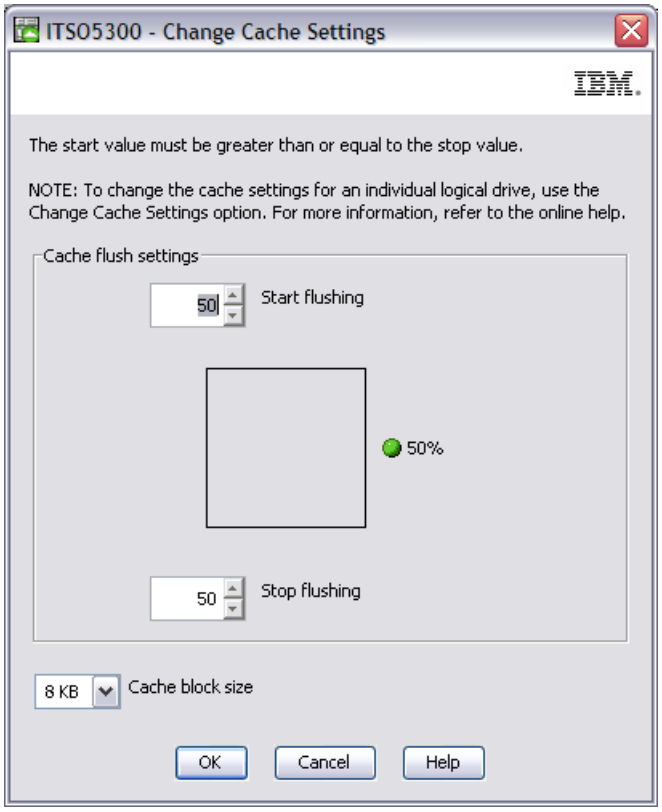


Figure 7-2 Start/stop flush settings and cache block size

7.2.7 Load balancing between controllers

Balance the load between the two controllers to ensure that the workload is evenly distributed. Load balancing is achieved by creating the arrays/LUNs evenly for both controllers as described in Chapter 4, “Host configuration guide” on page 151, thus ensuring balanced I/Os and bandwidth that can meet all requirements.

7.2.8 Volume management

Generally, the fewer volume groups, the better. However, if multiple volume groups are needed by the host volume manager due to the size of the databases, try to spread the logical drives from each array across all of the groups evenly. It is best to keep the two sets of recovery logs and files in two separate volume groups. Therefore, as a rule, you want to start with a minimum of three volume groups for your databases.

7.2.9 Performance monitoring

With Oracle it is important to optimize the system to keep it running smoothly. In order to know if the system is maintaining its performance level, a baseline is required so that you can compare the current state with a previously known state. Without the baseline, performance is based on perception rather than a real, documented measurement, which will also give you a level of workload that you can compare to for measuring actual growth.

A common pitfall is to mistake the symptoms of a problem for the actual problem itself. It is important to recognize that many performance statistics indicate the symptoms, and that identifying the symptom is not sufficient data to implement a remedy.

For example, consider the situation of a slow physical I/O, which is generally caused by poorly configured disks. However, it might also be caused by a significant amount of unnecessary physical I/O on those disks due to poorly tuned SQL statements or queries.

Ideally, baseline data gathered includes these statistics:

- Application statistics
- Database statistics
- Operating system statistics
- Midrange storage subsystem statistics

Application statistics consist of figures such as active session histories, transaction volumes, and response times.

Database statistics provide information about the type of load on the database, as well as the internal and external resources used by the database.

Operating system statistics are CPU, memory, disk, and network statistics. For Windows environments, Windows Performance Monitor can be used to gather performance data. For UNIX and Linux environments, a range of tools can be employed to gather the performance data. See Table 7-1.

Table 7-1 Linux tools commonly used to gather performance data

Component	Linux/UNIX tool to collect statistics
CPU	sar, vmstat, mpstat, or iostat
Memory	sar, vmstat
Disk	sar, iostat
Network	netstat

For the Midrange Storage Subsystem, the *IBM System Storage DS Storage Manager 10* offers a performance monitoring tool to gather data of what the Midrange Storage Subsystem sees its workload to be. A collection of the *Support Data* can also be helpful when trying to see what resources are being impacted.

CPU statistics

CPU utilization is the most important operating system statistic in the tuning process. Make sure to gather CPU utilization for the entire system and for each individual CPU in a multiprocessor environments. The utilization figure for each CPU can help detect single threading and scalability issues.

Most operating systems report CPU usage as time spent in user mode and time spent in kernel mode. These additional statistics allow better analysis of what is actually being executed on the CPU.

On an Oracle data server, where there is generally only one application running, the server runs database activity in user mode. Activities required to service database requests, such as scheduling, synchronization, and memory management, run in kernel mode.

Virtual memory statistics

Virtual memory statistics mainly ought to be used as a check to validate that there is very little paging on the system. System performance degrades rapidly and unpredictably when excessive paging activity occurs.

Individual process memory statistics can detect memory leaks due to a programming fault to deallocate memory. These statistics can be used to validate that memory usage does not rapidly increase after the system has reached a steady state after startup.

Disk statistics

Because the database resides on a set of disks, the performance of the I/O subsystem is very important to the performance of the database. Most operating systems provide extensive statistics about disk performance. The most important disk statistics are the throughput, current response times and the length of the disk queues. These statistics show if the disk is performing optimally or if the disk is being overworked.

Measure the normal performance of the I/O system. If the response times are much higher than the normal performance value, then it is performing badly or is overworked. If disk queues start to exceed two, then the disk subsystem is a potential bottleneck of the system.

Comparing the host reported IO statistics with those of the Midrange Storage Subsystem can help in isolating down the location of a problem.

Network statistics

Network statistics can be used in much the same way as disk statistics to determine whether a network or network interface is overloaded or not performing optimally. For today's range of networked applications, network latency can be a large portion of the actual user response time. For this reason, these statistics are a crucial debugging tool.

7.3 Microsoft SQL Server

In this section, we describe various considerations for Microsoft SQL server and the Midrange Storage Subsystem environment. Review the information in 4.3, "Microsoft Windows Server 2008 configuration" on page 156 for information specific to Windows OS. As with all guidelines, these settings must be checked to ensure that they suit your specific environment. Testing your own applications with your own data is the only true measurement.

We cover the following topics:

- Allocation unit size and SQL Server
- RAID levels
- Disk drives
- File locations
- Transaction logs
- Databases
- Maintenance plans

7.3.1 Allocation unit size

When running on Windows 2003, or Windows 2008 SQL Server must be installed on disks formatted using NTFS, because NTFS gives better performance and security to the file system. In all the Windows versions, setting the file system allocation unit size to 64 KB will improve performance. Allocation unit size is set when a disk is formatted.

Adjusting the allocation unit other than the default does affect features, for example, file compression. Use this setting first in a test environment to ensure that it gives the desired performance level and that the required features are enabled.

For more information about formatting an NTFS disk, see the appropriate Windows release documentation.

7.3.2 RAID levels

Redundancy and performance are required for the SQL environment.

- RAID 1 or RAID 10 must be used for the databases, tempdb, and transaction logs.
- RAID 1, RAID 5, or RAID 10 can be used for the maintenance plans.

7.3.3 File locations

As with all database applications, the database files and the transaction logs must be kept on separate logical drives and separate arrays, for best protection. Also, the tempdb and the backup area for any maintenance plans must be kept separate as well. Limit other uses for these arrays to minimize contention.

It is not a good idea to place any of the database, transaction logs, maintenance plans, or tempdb files in the same location as the operating system page file.

7.3.4 User database files

General guidelines for user database files are as follows:

- Create the databases on a physically separate RAID array. The databases are being constantly being read from and written to; therefore, using separate, dedicated arrays does not interfere with other operations such as the transaction logs, or maintenance plans. Depending upon the current size of the databases and expected growth, either a RAID 1 or RAID 10 array can give best performance and redundancy. RAID 5 can also be used, but with a slightly lower performance. Data redundancy is critical in the operation of the databases.
- Consider the speed of the disk, which will also affect performance. Use the 15K RPM disks rather than 10K RPM disks. Avoid using SATA drives for the databases.
- Spread the array over many drives. The more drives that the I/O operations are being sent to, the better the performance. Keep in mind that best performance is between 5 to 12 drives.

Midrange storage array settings: Use the following settings, as appropriate:

- Segment size 64 K or 128 K (dependent on I/O profile)
- Read cache on
- Write cache on
- Write cache with mirroring on
- Read ahead multiplier enabled (1)

7.3.5 Tempdb database files

Tempdb is a default database created by SQL Server. It is used as a shared working area for a variety of activities, including temporary tables, sorting, subqueries, and aggregates with GROUP BY or ORDER BY queries using DISTINCT (temporary worktables need to be created to remove duplicate rows), cursors, and hash joins.

It is good to enable tempdb I/O operations to occur in parallel to the I/O operations of related transactions. Because tempdb is a scratch area and very update intensive, use RAID 1 or RAID 10 to achieve optimal performance benefits. RAID 5 is not suited for this use. The tempdb is reconstructed with each server restart.

The ALTER DATABASE command can be used to change the physical file location of the SQL Server logical file name associated with tempdb; hence the actual tempdb database.

Here are general guidelines for the physical placement and database options set for the tempdb database:

- Allow the tempdb database to expand automatically as needed, which ensures that queries generating larger than expected intermediate result sets stored in the tempdb database are not terminated before execution is complete.
- Set the original size of the tempdb database files to a reasonable size to prevent the files from automatically expanding as more space is needed. If the tempdb database expands too frequently, performance can be affected.
- Set the file growth increment percentage to a reasonable size to avoid the tempdb database files from growing by too small a value. If the file growth is too small compared to the amount of data being written to the tempdb database, then tempdb might need to constantly expand, thereby affecting performance.
- If possible, place the tempdb database on its own separate logical drive to ensure good performance. Stripe the tempdb database across multiple disks for better performance.

Midrange storage array settings: Use the following settings, as appropriate:

- Segment size 64 K or 128 K (dependent on I/O profile)
- Read cache on
- Write cache on
- Write cache with mirroring on
- Read ahead multiplier disabled (0)

7.3.6 Transaction logs

General guidelines for creating transaction log files are as follows:

- Transaction logging is primarily sequential write I/O, favoring RAID 1 or RAID 10. Note that RAID 5 is not suited for this use. Given the criticality of the log files, RAID 0 is not suited either, despite its improved performance.

  There are considerable I/O performance benefits to be gained from separating transaction logging activity from other random disk I/O activity. Doing so allows the hard drives containing the log files to concentrate on sequential I/O. Note that there are times when the transaction log will need to be read as part of SQL Server operations such as replication, rollbacks, and deferred updates. SQL Servers that participate in replication must pay particular attention to making sure that all transaction log files have sufficient disk I/O processing power because of the read operations that frequently occur.
- The speed of the disk will also affect performance. Whenever possible, use the 15K rpm disks rather than 10K rpm disks. Avoid using SATA drives for the transaction logs.
- Set the original size of the transaction log file to a reasonable size to prevent the file from automatically expanding as more transaction log space is needed. As the transaction log expands, a new virtual log file is created, and write operations to the transaction log wait while the transaction log is expanded. If the transaction log expands too frequently, performance can be affected.
- Set the file growth increment percentage to a reasonable size to prevent the file from growing by too small a value. If the file growth is too small compared to the number of log records being written to the transaction log, then the transaction log might need to expand constantly, affecting performance.

- Manually shrink the transaction log files rather than allowing Microsoft SQL Server to shrink the files automatically. Shrinking the transaction log can affect performance on a busy system due to the movement and locking of data pages.

Midrange storage array settings: Use the following settings, as appropriate:

- Segment size 64 K or 128 K (dependent on I/O profile)
- Read cache on
- Write cache on
- Write cache with mirroring on
- Read ahead multiplier disabled (0)

7.3.7 Maintenance plans

Maintenance plans are used to perform backup operations with the database still running. For best performance, it is advisable to place the backup files in a location that is separate from the database files. Here are general guidelines for maintenance plans:

- Maintenance plans allow you to back up the database while it is still running, The location for the database backups must be in a dedicated array that is separate from both the databases and transaction logs. For the most part, they are large sequential files.
- This array needs to be much larger than the database array, as you will keep multiple copies of the database backups and transaction log backups. A RAID 5 array will give good performance and redundancy.
- The speed of the disk will also affect performance, but will not be as critical as the database or transaction log arrays. The preference is to use 15K disks for maximum performance, but 10K or even SATA drives can be used for the maintenance plans; this depends on your environment's performance needs.
- Spread the array over many drives. The more drives that the I/O operations are being sent to, the better the performance. Keep in mind that best performance is between 5 to 12 drives. For more details on array configurations see 2.2.6, "Array configuration" on page 33.
- Verify the integrity of the backup upon completion. Doing it performs an internal consistency check of the data and data pages within the database to ensure that a system or software problem has not damaged data.

Midrange storage array settings: Use the following settings, as appropriate:

- Segment size 128 K or higher (dependent on I/O profile)
- Read cache on
- Write cache on
- Write cache with mirroring off
- Read ahead multiplier enabled (1)

7.4 IBM Tivoli Storage Manager backup server

With an IBM Tivoli Storage Manager (ITSM) backup server environment, the major workload to be considered is the backup and restore functions which are generally throughput intensive environments. Therefore, try to ensure that the Midrange Storage Subsystem's server-wide settings are set for the high throughput settings specified for cache blocksize. For a Midrange Storage Subsystem dedicated to the ITSM environment, the cache blocksize must be set to 16 KB.

Best practice: For a Midrange Storage Subsystem dedicated to the ITSM environment, the cache blocksize must be set to 16 KB.

The ITSM server application has two separate sets of data storage needs. ITSM uses an instance database to manage its storage operations and disk storage pools for storing the backup data, either temporarily before migration to tapes or permanently using for example ITSM server deduplication function.

In our scenario, we use as an example a site with three ITSM host servers sharing a Midrange Storage Subsystem DS5000, and managing twelve ITSM instances across them. That means each physical TSM server machine manages 4 ITSM instances with dedicated disk storage pools on DS5000 system and with dedicated DB2 database instance. These servers manage the data stored in a 16 TB storage pool that is spread across them in fairly equal portions. It is estimated that the database needs for this will be about 1.7 TB in size, giving us about 100-150 GB per instance.

In our example, the customer has chosen to use 146 GB FC drives for the databases, and 250 GB SATA drives for the storage pools. The conceptual diagram of the environment is seen in Figure 7-3.

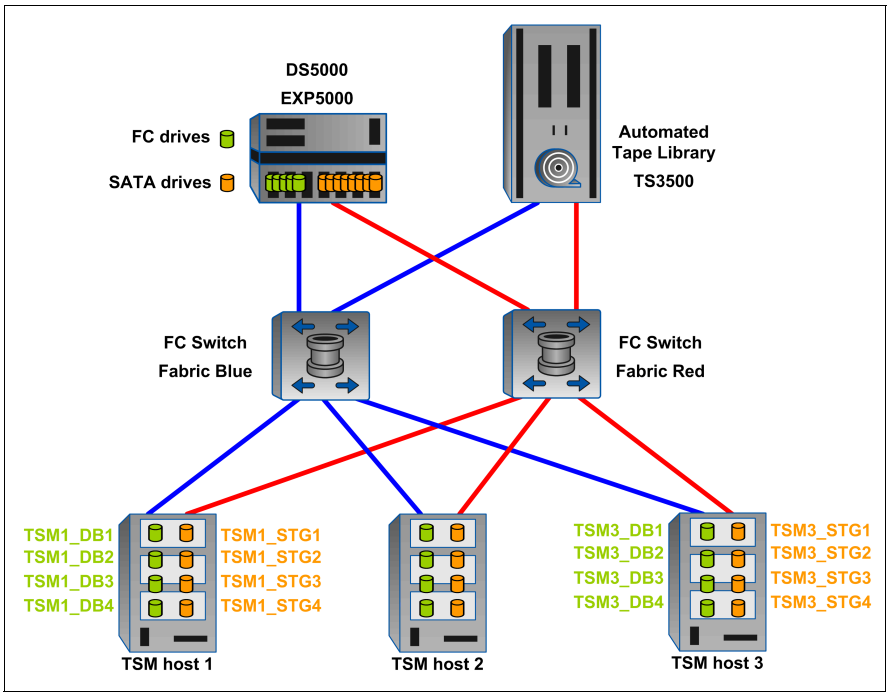


Figure 7-3 Deployment of DS5000 as a shared storage for TSM servers

Here are the general guidelines we followed for creating the ITSM databases:

- For a Midrange Storage Subsystem being used for the ITSM databases, you have a fairly high random write workload. We used the following general guideline guidelines for our TSM site with three ITSM host servers (TSM1, 2, and 3), and database needs to handle twelve ITSM instances per server:
  - Use a RAID 10 array, with four or more drives (remember, the higher the drive count, the better the performance with high transaction workloads). If you have a number of ITSM host servers, create a large RAID 10 array out of which you can create the logical drives that will handle the database needs for all the hosts applications. In our scenario, we created a single RAID 10 array of 13 x 13 drives.

- With ITSM databases of 1-100 and 11-150 GB size being requested, we created logical drives of 50 GB striped across the previous RAID 10 array; giving us 35 logical drives.
- For ITSM databases, we observed that the logical drives must have a large segment size defined. The guideline of 256 K has been found to work well.
- Ensure that cache read-ahead is disabled by setting it to “0”.

- Use partition mapping to assign each ITSM server the logical drives it will use for the databases it will have. Ensure that the correct host type setting is selected.
- Because the fabric connections will support both high transaction, and high throughput, you will need to set the host and any HBA settings available for high I/O support and high throughput both of these means:
  - HBA setting for high IOPS to be queued
  - Large blocksize support, in both memory and I/O transmission drivers
- Set the host device and any logical drive specific settings for high transaction support, Especially, ensure that you have a good queue depth level set for the logical drives being used. A queue depth of “32” per logical drive and 256 per host adapter are good values to start with.
- Using the host volume manager, we created a large volume group in which we placed all of the logical drives we have assigned to handle each of the instances managed by the ITSM server. As an example, suppose TSM1 has four instances to handle, each requiring 150 GB databases. In this case we have twelve logical drives for which we will build the volume group. Create the volume group using the following parameters:
  - Logical drives need to be divided into small partitions to be used for volume creation across them.

    Tip: Use a partition size that is one size larger than the minimum allowed.

  - For each ITSM instance, create a volume of 150 GB in size spread across three of the logical drives using *minimum interpolicy*.
  - Configure the ITSM database volume to have a file system on it.
  - Ensure that each ITSM instance is on its own separate file system built as defined in the previous steps.
  - In the ITSM application, create ten files for each instance, and define in a round-robin fashion to spread the workload out across them.

General guidelines for ITSM storage pools are as follows:

- Create as many RAID 5 arrays using a 4+1parity scheme as you can, using the drives that you have allocated for the storage pools. In the previous example we have enough drives to create sixteen arrays.
- Create a logical drive of equal size on each of the arrays for each of the ITSM host servers (in our example, three).
  - Make each of the logical drives of equal size. For example, a 4+1p RAID 5 of 250 GB SATA drives can give us about 330 GB if divided by three. A good plan is to have an even number of arrays to spread, if dividing arrays into an odd number of logical drives.

    Best practice: The objective is to spread the logical drives evenly across all resources. Therefore, if configuring an odd number of logical drives per array, it is a good practice to have an even number of arrays.

- Use a segment size of 512 KB for very large blocksize and high sequential data.
- Define cache read-ahead to “1” (enabled) to ensure we get best throughput.

- Define one logical drive from each array to each ITSM host server. Use partition mapping to assign each host its specific logical drives. Ensure the correct host type setting is selected.
- Using the host volume manager, create a large volume group containing one logical drive from each array defined for the specific storage pool’s use on the Midrange Storage Subsystem as outlined before:
  - These logical drives need to be divided into small partitions to be used for volume creation across them.
  - Create two raw volumes of even spread and size across all of the logical drives.
  - With high throughput workloads, set the logical drive queue depth settings to a lower value such as 16.
- The ITSM application has a storage pool parameter setting that can vary for use with tape drives:

  `txngroupmax=256 (change to 2048 for tape configuration support).`

  For further details and additional considerations related to the implementation and configuration of TSM server version 6.1 and higher, specifically with regard to the deployment of its embedded IBM DB2 database engine, see the *Tivoli Storage Manager V6.1 Technical Guide*, SG24-7718.

7.5 Microsoft Exchange 2003

In this section, we build on Microsoft best practices, providing guidelines based around storage design for deploying Microsoft Exchange 2003 messaging server on the family of Midrange Storage Subsystems.

The configurations described here are based on Exchange 2003 storage best practice guidelines and a series of lengthy performance and functionality tests. The guidelines can be found at this website:

`http://www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/fa839f7d-f876-42c4-a335-338a1eb04d89.mspx`

This section is primarily concerned with the storage configuration and does not go into the decisions behind the Exchange 2003 configurations referenced. For more information about Exchange design, see the following website:

`http://www.microsoft.com/technet/prodtechnol/exchange/2003/library/default.mspx`

We assume that:

- Exchange 2003 Enterprise Edition is running in a stand-alone configuration (non-clustered).
- Windows 2003 operating system, page file, and all application binaries are located on locally attached disks.
- All additional data, including Exchange logs, storage groups (SG), SMTP queues, and RSG (Recovery Storage Groups) are located on a Midrange Storage Subsystem.

7.5.1 Exchange configuration

All Exchange data is located in the Exchange store, consisting of three major components:

- Jet database (`.edb` file)
- Streaming database (`.stm` file)
- Transaction log files (`.log` files)

Each Exchange store component is written to separately. Performance will be greatly enhanced if the `.edb` files and corresponding `.stm` files are located on the same storage group, on one array, and the transaction log files are placed on a separate array.

The following list shows how the disk read/writes are performed for each Exchange store component:

- Jet database (`.edb file`):
  - Reads and writes are random
  - 4 KB page size
- Streaming database (`.stm file`):
  - Reads and writes are sequential
  - Variable page size that averages 8 KB in production
- Transaction log files (`.log files`):
  - 100% sequential writes during normal operations
  - 100% sequential reads during recovery operations
  - Writes vary in size from 512 bytes to the log buffer size, which is 5 MB

Additional activities that affect I/O

Here is a list of such activities:

- Zero out deleted database pages
- Content indexing
- SMTP mail transmission
- Paging
- MTA message handling
- Maintenance mode
- Virus scanning

User profiles

Table 7-2 lists mailbox profiles that can be used as a guideline for capacity planning of Exchange mailbox servers. These profiles represent mailbox access for the peak of an average user Outlook (or Messaging Application Programming Interface (MAPI) based) client within an organization.

Table 7-2 User profiles and corresponding usage patterns

User type	Database volume IOPS	Send/receive per day	Mailbox size
Light	0.18	10 sent/50 received	< 50 MB
Average	0.4	20 sent/100 received	50 MB
Heavy	0.75	30 sent/100 received	100 MB

7.5.2 Calculating theoretical Exchange I/O usage

To estimate the number of IOPS that an Exchange configuration might need to support, you can use the following formula:

```
Number of users (mailboxes) x I/O profile of user = required IOPS for database
drives
```

Consider the following example:

```
1500 (users/mailboxes) x 0.75 (heavy user) = 1125 IOPS
```

Using a ratio of two reads for every write, which is 66% read and 33% writes, you can plan for 742.5 IOPS for read and 371.5 IOPS for writes.

All writes are committed to the log drive first and then written to the database drive. Approximately 10% of the total IOPS seen on the database drive will be seen on the log drive. The reason for a difference between the log entries and the database is that the log entries are combined to provide for better streaming of data.

Therefore, 10% of the total 1125 IOPS seen on the database drive will be seen on the log drive:

```
1125/100 x 10 = 112.5
```

In this example, the drives need to support the following IOPS:

- Logs = 112.5 IOPS
- Database = 1125 IOPS
- Total = 1237.5 IOPS

Tip: It is assumed that Exchange is the only application running on the server. If other services are running, then the I/O profile needs to be amended to take into account the additional tasks running.

7.5.3 Calculating Exchange I/O usage from historical data

If an Exchange environment is already deployed, then historical performance data can be used to size the new environment.

This data can be captured with the Windows Performance Monitor using the following counters:

- Logical disk
- Physical disk
- Processor
- MS Exchange IS

To get the initial IOPS of the Exchange database, monitor the Logical Disk → Disk Transfers/sec → Instance=Drive letter that houses the Exchange Store database. (Add all drive letters that contain Exchange Database files). Monitor this situation over time to determine times of peak load.

Next we provide an example of how to calculate the I/O requirements for an Exchange deployment based on a DS4000 Storage System using RAID 10 arrays and having all Exchange transaction log and storage groups on their own individual arrays.

Assume the following values:

- Users/mailboxes = 1500

- Mailbox size = 50 MB
- Database IOPS = 925

To calculate the individual IOPS of a user, divide database IOPS by the number of users.

```
925/1500 = 0.6166
```

To calculate the I/O overhead of a given RAID level, you need to have an understanding of the various RAID types. RAID 0 has no RAID penalty, as it is just a simple stripe over a number of disks — so, 1 write will equal 1 I/O. RAID 1 or RAID 10 is a mirrored pair of drives or multiple mirrored drives striped together, because of the mirror, it means that for every write committed 2 I/Os will be generated. RAID 5 uses disk striping with parity so, for every write committed, this will often translate to 4 I/Os, due to the need to read the original data and parity before writing the new data and new parity. For further information about RAID levels, see 2.2.5, “DS5000 arrays and RAID levels” on page 26.

To calculate the RAID I/O penalty in these formulas, use the following substitutions:

- RAID 0 = 1
- RAID 1 or RAID 10 = 2
- RAID 5 = 4

Using the formula:

```
[(IOPS/mailbox × READ RATIO%)] + [(IOPS/mailbox × WRITE RATIO%) x RAID penalty]
```

Gives us:

```
[(925/1500 x 66/100) = 0.4069] + [(925/1500 x 33/100) = 0.2034 > (x 2) = 0.4069]
```

That is a total of 0.8139 IOPS per user, including the penalty for RAID 10.

Exchange 2003 EE supports up to four storage groups (SGs) with five databases within each storage group. Therefore, in this example, the Exchange server will have the 1500 users spread over three storage groups with 500 users in each storage group. The fourth storage group will be used as a recovery storage group (RSG). For more information about RSGs, see the following website:

http://www.microsoft.com/technet/prodtechnol/exchange/guides/UseE2k3RecStorGrps/d42ef860-170b-44fe-94c3-ec68e3b0e0ff.mspx

Here we describe our calculations:

- To calculate the IOPS required for a storage group to support 500 users, apply the following formula:

  ```
  Users per storage group x IOPS per user = required IOPS per storage group
  500 x 0.8139 = 406.95 IOPS per storage group
  ```

  A percentage must be added for non-user tasks such as Exchange Online Maintenance, anti-virus, mass deletions, and various additional I/O intensive tasks. Best practice is to add 20% to the per user I/O figure.

- To calculate the additional IOPS per storage group, use the following formula:

  ```
  Users per storage group x IOPS per user x 20% = overhead in IOPS per storage group
  500 x 0.8139/100 x 20 = 81.39 IOPS overhead per SG
  ```

- The final IOPS required per storage group is determined by adding the user IOPS per storage group with the overhead IOPS per storage group.

  ```
  User IOPS per SG + overhead in IOPS per SG = total IOPS per SG
  406.95 + 81.39 = 488.34
  ```

- The total required IOPS for the Exchange server in general is as follows:

  ```
  Total IOPS per SG x total number of SGs
  488.34 x 3 = 1465.02
  ```

- The new IOPS user profile is obtained by dividing total IOPS by total number of users:

  ```
  1465.02/1500 = 0.9766 IOPS per user
  ```

- Taking this last figure and rounding it up gives us a 1.0 IOPS per user. This figure allows for times of extraordinary peak load on the server.
- Multiplying by the 1500 users supported by the server gives a figure of 1500 IOPS across all three storage groups, divided by the three storage groups on the server, means that each storage group will need to be able to sustain 500 IOPS.
- Microsoft best practice is that log drives be designed to take loads equal to 10% of those being handled by the storage group logical drive, for example:

  ```
  500/100 x 10 = 50 IOPS
  ```

- Microsoft best practice is that log files be kept on separate spindles (physical disks) from each other and the storage groups.

After extensive testing, it has been determined that a RAID 1 (mirrored pair) provides the best performance for Exchange transaction logs on the DS4000 series, which is consistent with Microsoft best practices. In addition, Microsoft best practice is that the storage groups be placed on RAID 10. Again, it has proved to provide the best performance, however, RAID 5 will also provide the required IOPS performance in environments where the user I/O profile is less demanding.

Taking the same data used in the example for RAID 10, for the Exchange storage groups only and substituting the RAID 10 penalty, which is 2 I/Os for the RAID 5 penalty, which can be up to 4 I/Os, the new RAID 5 storage groups each need to deliver an additional 500 IOPS than for the RAID 10 configuration.

7.5.4 Path LUN assignment (MPIO)

With the recent release of 7.xx IBM DS Controller Firmware, the Midrange Storage Subsystems moved from using the previous proprietary RDAC/MPP (multi-path proxy) driver to fully supporting and using the Microsoft Windows MPIO driver for handling multipathing and failover functionality. See 4.3, “Microsoft Windows Server 2008 configuration” on page 156 for details on the use of MPIO for Windows.

7.5.5 Storage sizing for capacity and performance

Mailbox quota, database size, and the number of users are all factors that you need to consider during capacity planning. Considerations for additional capacity must include NTFS fragmentation, growth, and dynamic mailbox movement. Experience has determined that it is always a best practice to double capacity requirements wherever possible to allow for unplanned needs, for example:

- A 200 MB maximum mailbox quota
- About 25 GB maximum database size*
- Total mailbox capacity to be 1500 users
- Mailboxes to be located on one Exchange server

Maximum database size for Exchange 2003 SE (Standard Edition) is 16 GB, increasing to 75 GB with SP2, and the theoretical limit for Exchange 2003 EE (Enterprise Edition) is 16 TB.

Exchange SE supports four storage groups with one mailbox store and one public folder database store. Exchange EE supports four storage groups with five mailbox or public folder database stores, located within each storage group to a maximum of 20 mailbox or public store databases across the four storage groups. For more information, see the website:

http://support.microsoft.com/default.aspx?scid=kb;en-us;822440

Capacity planning calculations

Using the previous data, the capacity planning was done as follows:

```
Database size / maximum mailbox size = number of mailboxes per database
25GB / 200MB = 125 mailboxes per database
```

There is a maximum of five databases per storage group:

```
Maximum mailboxes per database x database instances per SG = maximum mailboxes
per SG
125 x 5 = 625 mailboxes per SG
```

There are three active storage groups on the Exchange server:

```
Storage groups per server x maximum mailboxes per SG = maximum mailboxes per
server
3 x 625 = 1875 mailboxes per server
```

In addition to the database storage requirements listed previously, logical drives and capacity must be provided for the following items:

- Log files
- Extra space added to each storage group for database maintenance and emergency database expansion
- A 50 GB logical drive for the SMTP and MTA working directories
- Additional logical drive capacity for one additional storage group, for spare capacity or as a recovery group to recover from a database corruption
- An additional logical drive for use with either the additional storage group or recovery storage group

Logical view of the storage design with capacities

Figure 7-4 shows a logical view of the storage design.

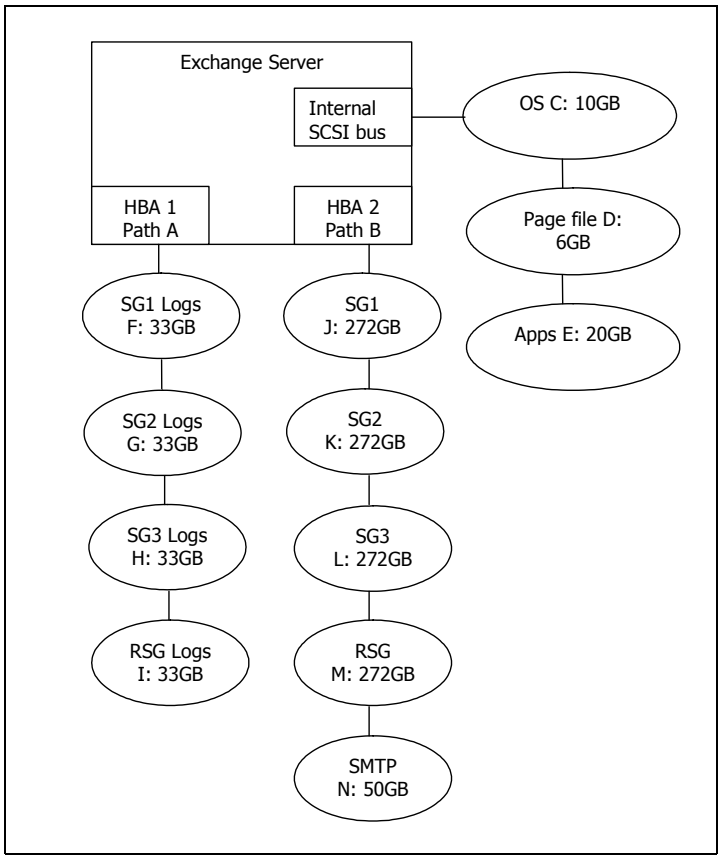


Figure 7-4 Logical view of the storage design with capacities

Table 7-3 details the drive letters, RAID levels, disks used, capacity, and role of the logical drives that will be presented to Windows.

Table 7-3 Logical drive characteristics

Drive Letter	Size (GB)	Role	Location	RAID level	Array	Disks used
C	10 GB	Operating system	Local	RAID1	N/A	N/A
D	6 GB	Windows page file	Local	RAID1	N/A	N/A
E	18 GB	Applications	Local	RAID1	N/A	N/A
F	33 GB	SG1 Logs	SAN	RAID1	1	2 x 36 GB 15k
G	33 GB	SG2 Logs	SAN	RAID1	2	2 x 36 GB 15k
H	33 GB	SG3 Logs	SAN	RAID1	3	2 x 36 GB 15k
I	33 GB	RSG Logs	SAN	RAID1	4	2 x 36 GB 15k
J	272 GB	SG1 + maintenance	SAN	RAID10	5	8 x 73 GB 15k

Drive Letter	Size (GB)	Role	Location	RAID level	Array	Disks used
K	272 GB	SG2 + maintenance	SAN	RAID10	6	8 x 73 GB 15k
L	272 GB	SG3 + maintenance	SAN	RAID10	7	8 x 73 GB 15k
M	272 GB	Recovery Storage Group + maintenance	SAN	RAID10	8	8 x 73 GB 15k
N	50 Gb	SMTP Queues & MTA data	SAN	RAID10	9	4 x 73 GB 15k

7.5.6 Storage system settings

Use the following settings:

- Global cache settings:
  - 4 k
  - Start flushing 50%
  - Stop flushing 50%
- Array settings:
  - Storage group type: RAID 10 or 5 depending on user I/O profile
  - Log drives:
    - Segment size 64 KB or 128 KB
    - Read ahead 0
    - Write cache on
    - Write cache with mirroring on
    - Read cache on

7.5.7 Aligning Exchange I/O with storage track boundaries

In this section, we describe various considerations regarding disk partitioning.

Basic precautions

With a physical disk that maintains 64 sectors per track, Windows always creates the partition starting at the sixty-fourth sector, therefore misaligning it with the underlying physical disk. To be certain of disk alignment, use diskpart.exe, a disk partition tool. The diskpart.exe utility is contained within Windows Server 2003 and later, or with Windows 2000 Server and can explicitly set the starting offset in the master boot record (MBR).

By setting the starting offset, you can track alignment and improve disk performance. Exchange Server 2003 writes data in multiples of 4 KB I/O operations (4 KB for the databases and up to 32 KB for streaming files). Therefore, make sure that the starting offset is a multiple of 32KB. Failure to do so can cause a single I/O operation to span two tracks, causing performance degradation.

Important: Using the diskpart utility to align storage track boundaries is data destructive. When used against a disk, all data on the disk will be wiped out during the storage track boundary alignment process. Therefore, if the disk on which you will run diskpart contains data, back up the disk before performing the operation.

Best practice: Set the offset to equal the segment size used for the logical drive, to ensure that the user data will start on a segment boundary.

For more information, see the website:

http://www.microsoft.com/technet/prodtechnol/exchange/guides/StoragePerformance/0e24eb22-fbd5-4536-9cb4-2bd8e98806e7.mspx

The diskpart utility supersedes the functionality previously found in diskpar.exe. Both diskpar and diskpart must only be used if the drive is translated as 64 sectors per track.

Additional considerations

It is important in planning an Exchange configuration to recognize that disks not only provide capacity but determine performance.

After extensive testing, it has been proven that the log drives perform best on a RAID 1 mirrored pair.

In line with Microsoft Exchange best practices, it has also been proven that RAID 10 for storage groups outperforms other RAID configurations. That said, depending on the user profile, RAID 5 will provide the required IOPS performance in certain instances.

Laying out the file system with diskpart provides additional performance improvements and also reduces the risk of performance degradation over time as the file system fills.

Dedicating HBAs for specific data transfer has a significant impact on performance. For example, HBA1 to controller A for log drives (sequential writes), and HBA 2 to controller B for storage groups (random read and writes). However, in a two HBA environment with the DS5100 and DS5300, it can restrict the back-end performance for the handling of the workload.

Disk latency can be relieved by adding additional HBAs to the server. Four HBAs with logs and storage groups assigned to pairs of HBAs can provide a significant improvement in limiting disk latency and also reduces performance problems in the event of a path failure.

The Windows Performance Monitor can be used effectively to monitor an active Exchange server. Here are the counters of interest:

- Average disk sec/write
- Average disk sec/read
- Current disk queue length
- Disk transfers/sec

The average for the average disk sec/write and average disk sec/read on all database and log drives must be less than 20 ms.

The maximum of the average disk sec/write and average disk sec/read on all database and log drives must be less than 40 ms.

7.6 Guidelines specific to Windows Exchange Server 2007

With Exchange Server 2007, we have found the following areas that you need to consider when creating your layout on the Midrange Storage Subsystems.

7.6.1 Storage layout across the storage subsystem

The Exchange Server is a disk-intensive application. Based on the tests we performed, here is a summary of best practices to improve the storage performance:

- Be sure to create arrays and logical drives (LUNs) by spreading them across as much of the storage subsystem's resources as possible. Following the transaction based model outlined in 6.1, "Workload types" on page 258 will help to ensure that you are spreading the workload out evenly and keep all the resources busy as well as avoiding conjestion and conflicts.
- With the Exchange Server 2007, we have seen a variety of IO sizes and workload patterns that were both random and sequential in nature. Due to these various operational patterns, a 64 K NTFS allocation unit size is best suited.
- Keep the Exchange 2007 LUNs isolated from other disk intensive applications. Sharing arrays with other applications can have a negative impact on the performance of your Exchange server.
- Isolate Exchange 2007 database and log disk I/O on separate array groups. Separate the transaction log files (sequential I/O) from databases (random I/O) to maximize I/O performance and increase fault tolerance. From a recoverability perspective, separating a storage group's transaction logs and databases ensure that a catastrophic failure of a particular set of physical disks will not cause a loss of both database and transaction logs.
- The cache block size is a global parameter for the storage subsystem. Set the cache block size closest to the typical I/O size, which is 16 K size for large block and sequential I/O.
- Configure the entire capacity of a array group in a single LUN. Multiple LUNs on one volume group typically increases the seek time penalty.
- Use all available host and drive side channels if at all possible. Switch zoning and multi-pathing HBA drivers are all useful tools to ensure that all channels are kept active.
- Balance I/O across the dual controllers of the storage system and strive to keep both controllers and all back-end paths busy. Locate database LUNs and their corresponding log LUNs on separate back-end paths.
- Create volume groups across drive trays to distribute I/O across drive-side loops. See 6.5.6, "Arrays and logical drives" on page 275 for details.
- Choose faster disks. A 15K RPM drive yields ~15% more performance than a 10K RPM drive.
- For the log LUNs, use RAID 10 volumes and enable write cache with mirroring.
- For the database LUNS, enable read caching and enable write caching with mirroring.
- See the following Microsoft documentation for storage based replication best practices and support criteria:

  For Deployment Guidelines for Data Replication, see the following website:

  `http://www.microsoft.com/technet/prodtechnol/exchange/guides/E2k3DataRepl/bedf62a9-dff7-49a8-bd27-b2f1c46d5651.mspx`

  For information about the Multi-site data replication support for Exchange, see the following website:

http://support.microsoft.com/?kbid=895847

7.6.2 Other areas that can affect performance

There are many other areas that can affect the performance of your Exchange Server. For further guidance, see the following websites:

http://go.microsoft.com/fwlink/?LinkId=23454
http://technet.microsoft.com/en-us/library/bb124518.aspx

7.7 Microsoft Exchange 2010

This section covers the Microsoft best practices, providing guidelines based around storage design for deploying Microsoft Exchange 2010 messaging server on the family of Midrange Storage Subsystems.

The following information was taken from Microsoft's documentation on storage requirements for Exchange 2010 and can be read in its entirety at this website:

http://technet.microsoft.com/en-us/library/ee832792%28EXCHG.140%29.aspx

7.7.1 Storage architectures

The Table 7-4 describes supported storage architectures and provides best practice guidance for each type of storage architecture where appropriate.

Table 7-4 Supported Storage architectures

Storage architecture	Description	Best practice
Direct-attached storage (DAS)	DAS is a digital storage system directly attached to a server or workstation, without a storage network in between. For example, DAS transports include Serial Attached SCSI (SAS) and Serial Attached Advanced Technology Attachment (ATA).	Not available
Storage area network (SAN): iSCSI	SAN is an architecture to attach remote computer storage devices (such as disk arrays and tape libraries) to servers in such a way that the devices appear as locally attached to the operating system (for example, block storage). Internet Small Computer System Interface (iSCSI) SANs encapsulate SCSI commands within IP packets and use standard networking infrastructure as the storage transport (for example, Ethernet).	Don't share physical disks backing up Exchange data with other applications. Use dedicated storage networks. Use multiple network paths for stand-alone configurations.

Storage architecture	Description	Best practice
SAN: Fibre Channel (FC)	Fibre Channel SANs encapsulate SCSI commands within FC packets and generally utilize specialized Fibre Channel networks as the storage transport.	Don't share physical disks backing up Exchange data with other applications. Use multiple FC network paths for stand-alone configurations. Follow storage vendor's best practices for tuning FC host bus adapters (HBAs), for example, Queue Depth and Queue Target.

7.7.2 Physical disk types

In this section, we explain considerations regarding the various supported disk types.

Supported disk types

Table 7-5 provides a list of supported physical disk types, as well as best practice guidance for each physical disk type where appropriate.

Table 7-5 Supported Physical Disk Types

Physical disk type	Description	Best practice
SATA	Serial ATA (SATA) is a serial interface for ATA and integrated device electronics (IDE) disks. SATA disks are available in a variety of form factors, speeds, and capacities. In general, choose SATA disks for Exchange 2010 mailbox storage when you have the following design requirements: ► High capacity ► Moderate performance ► Moderate power utilization	Supported: 512 byte sector disks only. 4KB sector disks, including those which use 512-byte emulation, are currently not supported. Requires battery backed caching array controller for optimal data reliability and I/O performance. Physical disk-write caching must be disabled when used without an uninterruptible power supply (UPS). When considering SATA disks, we suggest considering Enterprise class SATA disks, which generally have better heat, vibration, and reliability characteristics.

Physical disk type	Description	Best practice
SAS	SAS is a serial interface for Small Computer System Interface (SCSI) disks. SAS disks are available in a variety of form factors, speeds, and capacities. In general, choose SAS disks for Exchange 2010 mailbox storage when you have the following design requirements: ► Moderate capacity ► High performance ► Moderate power utilization	Supported: 512 byte sector disks only. 4KB sector disks, including those which use 512-byte emulation, are currently not supported. Physical disk-write caching must be disabled when used without a UPS.
Fibre Channel (FC)	FC is an electrical interface used to connect disks to Fibre Channel-based SANs. FC disks are available in a variety of speeds and capacities. In general, choose FC disks for Exchange 2010 mailbox storage when you have the following design requirements: ► Moderate capacity ► High performance ► SAN connectivity	Supported: 512 byte sector disks only. 4KB sector disks, including those which use 512-byte emulation, are currently not supported. Physical disk-write caching must be disabled when used without a UPS.
Solid-state drive (SSD) (flash disk)	An SSD is a data storage device that uses solid-state memory to store persistent data. An SSD emulates a hard disk drive interface. SSD disks are available in a variety of, speeds (different I/O performance capabilities) and capacities. In general, choose SSD disks for Exchange 2010 mailbox storage when you have the following design requirements: ► Low capacity ► Extremely high performance	Supported: 512 byte sector disks only. 4KB sector disks, including those which use 512-byte emulation, are currently not supported. Physical disk-write caching must be disabled when used without a UPS. In general, Exchange 2010 Mailbox servers don't require the performance characteristics of SSD storage.

Factors to consider when choosing disk types

There are several trade-offs when choosing disk types for Exchange 2010 storage. The correct disk is one that balances performance (both sequential and random) with capacity, reliability, power utilization, and capital cost. Table 7-6, which lists the supported physical disk types, provides information to help you when considering these factors.

Table 7-6 Factors in disk type choice

Disk speed (RPM)	Disk form factor	Interface/ Transport	Capacity	Random I/O performance	Sequential I/O performance	Power utilization
5,400	2.5-inch	SATA	Average	Poor	Poor	Excellent
5,400	3.5-inch	SATA	Excellent	Poor	Poor	Above Average
7,200	2.5-inch	SATA	Average	Average	Average	Excellent
7,200	2.5-inch	SAS	Average	Average	Above Average	Excellent
7,200	3.5-inch	SATA	Excellent	Average	Above Average	Above Average
7,200	3.5-inch	SAS	Excellent	Average	Above Average	Above Average
7,200	3.5-inch	FC	Excellent	Average	Above Average	Average
10,000	2.5-inch	SAS	Below Average	Excellent	Above Average	Above Average
10,000	3.5-inch	SATA	Average	Average	Above Average	Above Average
10,000	3.5-inch	SAS	Average	Above Average	Above Average	Below Average
10,000	3.5-inch	FC	Average	Above Average	Above Average	Below Average
15,000	2.5-inch	SAS	Poor	Excellent	Excellent	Average
15,000	3.5-inch	SAS	Average	Excellent	Excellent	Below Average
15,000	3.5-inch	FC	Average	Excellent	Excellent	Poor
SSD: enterprise class	N/A	SATA/SAS/FC	Poor	Excellent	Excellent	Excellent

7.7.3 Best practices for supported storage configurations

This section provides best practice information about supported disk and array controller configurations.

Redundant Array of Independent Disks (RAID) is often used to both improve the performance characteristics of individual disks (by striping data across several disks) as well as to provide protection from individual disk failures. With the advancements in Exchange 2010 high availability, RAID is no longer a required component for Exchange 2010 storage design. However, RAID is still an essential piece to Exchange 2010 storage design for stand-alone servers as well as high availability solutions that require either additional performance or greater storage reliability.

Table 7-7 provides guidance for the common RAID types that can be used with the Exchange 2010 Mailbox server.

Table 7-7 Supported RAID types for the Exchange 2010 Mailbox server role

Data type	Stand-alone: supported/best practices	High availability: supported/best practices
OS/System/Pagefile Volume	All RAID types supported. Best practice: RAID 1/10.	All RAID types supported. Best practice: RAID 1/10.
Exchange Mailbox Database File (EDB) Volume	All RAID types supported. Best practice: 5,400/7,200 disks = RAID1/10 only.	All RAID types supported. Just a bunch of disks (JBOD)/RAIDless supported (3 or more database copies). Best practice: 5,400/7,200 disks = RAID1/10 only or JBOD. Best practice: When lagged, database copies should have either two or more lagged copies or lagged copies ought to be protected with RAID.
Exchange Mailbox Database Log Volume	All RAID types supported. Best practice = RAID1/10.	All RAID types supported. JBOD/RAIDless supported (3 or more database copies). Best practice = RAID1/10. Best practice: When lagged database copies should have either two or more lagged copies or lagged copies ought to be protected with RAID.

Table 7-8 provides guidance about storage array configurations for Exchange 2010.

Table 7-8 Supported RAID types for the Exchange 2010 Mailbox server role

RAID type	Description	Stand-alone: supported/best practices
Disk Array RAID Stripe Size (kb)*	The stripe size is the unit of data distribution within a RAID set.	Best practice: 256 kilobytes (KB) or greater.
Storage Array Cache Settings	The cache settings provided by a battery-backed caching array controller.	Best practice: 75 percent write cache and 25 percent read cache (battery-backed cache).
Physical Disk Write Caching	The settings for the cache are on each individual disk.	Supported: Physical disk write caching must be disabled when used without a UPS.

Table 7-9 provides guidance about database and log file choices.

Table 7-9 Database and log file choices for the Exchange 2010 Mailbox server role

Database and log file options	Description	Stand-alone: supported/best practices	High availability: supported/best practices
File placement: Database/log isolation	Database/log isolation refers to placing the database file and logs from the same mailbox database onto different volumes backed by different physical disks.	Best practice: For recoverability, move database file (.edb) and logs from the same database to different volumes backed by different physical disks.	Isolation of logs and databases isn't required.
File placement: Database files/volume	Database files/volume refers to how you distribute database files within or across disk volumes.	Best practice: Based on your backup methodology.	Supported: When using JBOD, divide a single disk into two volumes (one for database, one for log stream).
File placement: Log streams/volume	Log streams/volume refers to how you distribute database log files within or across disk volumes.	Best practice: Based on your backup methodology.	Supported: When using JBOD, divide a single disk into two volumes (one for database, one for log stream). Best practice: When using JBOD, single database per log per volume.
Database size	The on disk database file size (.edb).	Supported: Approximately 16 terabytes (TB) Best practice: ► 100 gigabytes (GB) or less. ► Provision for 120 percent of calculated maximum database size.	Supported: Approximately 16 TB Best practice: ► 2 TB or less. ► Provision for 120 percent of calculated maximum database size.

Database and log file options	Description	Stand-alone: supported/best practices	High availability: supported/best practices
Log truncation method	The process for truncating and deleting old database log files. There are two mechanisms: ► Circular logging, in which Exchange deletes the logs. ► Log truncation, which occurs after a successful full or incremental Volume Shadow Copy Service (VSS) backup.	Best practice: ► Use backups for log truncation (for example, circular logging disabled). ► Provision for three days of log generation capacity.	Best practice: ► Enable circular logging for deployments that use Exchange 2010 data protection features. ► Provision for three days beyond replay lag setting of log generation capacity.

8

Storage Manager Performance Monitor

When implementing a storage solution, whether it is directly attached to a server, connected to the enterprise network (NAS), or on its own network (Fibre Channel SAN or iSCSI SAN), it is important to know just how well the storage server and its components perform. If this information is not available or collected, growth and capacity planning along with management becomes very difficult.

There are many utilities and products that can help measure and analyze performance. In this chapter, we use the Storage Manager (SM) Performance Monitor utility to view the performance of an IBM System Storage DS5000 Storage Server in real time and show how to collect part of the data. We also look at two other products at the end of the chapter; first, one that uses data gathered by Performance Monitor, and second, one where DS5000 is compliant and allows data gathering and management functions using methods other than Storage Manager and Performance Monitor.

8.1 Analyzing performance

To determine where a performance problem exists, it is important to gather data from all the components of the storage solution. It is not uncommon to be misled by a single piece of information and lulled into a false sense of knowing the cause of a poor system performance, only to realize that another component of the system is truly the cause.

In this section, we look at the Storage Manager Performance Monitor and how it can be used to analyze and monitor the environment.

Storage applications can be categorized according to two types of workloads: transaction based or throughput based.

- Transaction performance is generally perceived to be poor when the following conditions occur:
  - Random reads/writes are exceeding 20 ms, without write cache.
  - Random writes are exceeding 2 ms, with cache enabled.
  - I/Os are queuing up in the operating system I/O stack (due to bottleneck).
- Throughput performance is generally perceived to be poor when the disk capability is not being reached. Causes of this condition can occur from the following situations:
  - With reads, read-ahead is being limited, preventing higher amounts of immediate data availability.
  - I/Os are queuing up in the operating system I/O stack (due to bottleneck).

We consider the following topics:

- Gathering host server data
- Gathering fabric network data
- Gathering DS5000 Storage Server data

8.1.1 Gathering host server data

When gathering data from the host systems to analyze performance, it is important to gather the data from all attached hosts, even though a few might not see any slow performance. Indeed, a normally unrelated host might be impacting others with its processing.

Gather all the statistics possible from the operating system tools and utilities. Data from these will help when comparing it with what is seen by SM Performance Monitor and other measurement products. Utilities vary from operating system to operating system, so check with administrators or the operating system vendors.

Many UNIX type systems offer utilities that report disk I/O statistics and system statistics, such as iostat, sar, vmstat, filemon, nmon, and iozone, to mention a few. All are very helpful with determining where the poor performance originates. With each of these commands, you gather a sample period of statistics that need to be collected. Gathering one minute to 15 minutes worth of samples during the slow period can give a fair sampling of data to review.

In the example shown in Figure 8-1, we see the following information:

- Interval time = (894 + 4 KB)/15 KBps = 60 sec (hdisk1)
- Average I/O size = 227.8 KBps/26.9 tps = 8.5 KB (hdisk4)
- Estimated average I/O service time = 0.054/26.9 tps = 2 ms (hdisk4)
- tps < 75 No I/O bottleneck!
- Disk service times good: No I/O bottleneck
- Disks not well balanced: hdisk0, hdisk1, hdisk5?

```
tty:      tin         tout    avg-cpu: % user    % sys    % idle    % iowait
          24.7         71.3                8.3      2.4      85.6       3.6

Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0            2.2     19.4       2.6        268       894
hdisk1            1.0     15.0       1.7          4       894
hdisk2            5.0    231.8      28.1       1944     11964
hdisk4            5.4    227.8      26.9       2144     11524
hdisk3            4.0    215.9      24.8       2040     10916
hdisk5            0.0      0.0       0.0          0         0
```

Figure 8-1 Example AIX iostat report

The information shown here does not necessarily indicate a problem. Hdisk5 might be a new drive that is not yet being used. All other indications appear to be within expected levels.

Windows operating systems offer device manager tools that can have performance gathering capabilities in them. Also, many third-party products are available and frequently provide greater detail and graphical presentations of the data gathered.

8.1.2 Gathering fabric network data

To ensure that the host path is clean and operating as desired, gather any statistical information available from the switches or fabric analyzers for review of the switch configuration settings, and any logs or error data that might be gathered, which can be critical in determining problems that might cross multiple switch fabrics or other extended network environments. Again, as with gathering host data, fabric network data will help when comparing it with what is seen by SM Performance Monitor.

IBM and Brocade switches offer a supportshow tool that enables you to run a single command and gather all support data at one time. Cisco switches offer this with their **`show tech`** command.

Additionally, gather the host bus adapter parameter settings to review and ensure that they are configured for the best performance for the environment. Many of these adapters have BIOS type utilities that will provide this information. Certain operating systems (such as AIX) can also provide much of this information through system attribute and configuration setting commands. As shown in Example 8-1, the AIX HBA and associated SCSI controller is displayed with the tunable parameters.

Example 8-1 AIX HBA tunable parameters marked with "TRUE" in right hand column.

```
root@itsop630:/root
# lsattr -El fscsi0
attach       switch       How this adapter is CONNECTED         False
dyntrk       no           Dynamic Tracking of FC Devices        True
fc_err_recov delayed_fail FC Fabric Event Error RECOVERY Policy True
scsi_id      0x70800      Adapter SCSI ID                       False
sw_fc_class  3            FC Class for Fabric                   True
root@itsop630:/root
# lsattr -El fcs0
bus_intr_lvl  105        Bus interrupt level                                False
bus_io_addr   0x2ec00    Bus I/O address                                    False
bus_mem_addr  0xec020000 Bus memory address                                 False
init_link     al         INIT Link flags                                    True
intr_priority 3          Interrupt priority                                 False
lg_term_dma   0x800000   Long term DMA                                      True
max_xfer_size 0x100000   Maximum Transfer Size                              True
num_cmd_elems 200        Maximum number of COMMANDS to queue to the adapter True
pref_alpa     0x1        Preferred AL_PA                                    True
sw_fc_class   2          FC Class for Fabric                                True
root@itsop630:/root
```

8.1.3 Gathering DS5000 Storage Server data

When gathering data from the DS5000 Storage Server for analysis, there are two major functions that can be used to document how the system is configured and how it performs:

- Performance Monitor:

  Using the performance Monitor, the DS5000 Storage Server can provide a point in time presentation of its performance at a specified time interval for a specified number of occurrences, which is useful to compare with the host data collected at the same time and can be very helpful in determining hot spots or other tuning issues to be addressed.

  We provide details about the Performance Monitor in 8.2, "Storage Manager Performance Monitor" on page 347.

- Collect All Support Data:

  This function can be run from the IBM TotalStorage DS5000 Storage Manager Subsystem Management window by selecting **Advanced** → **TroubleShooting** → **Collect All Support Data** (see Figure 8-2).

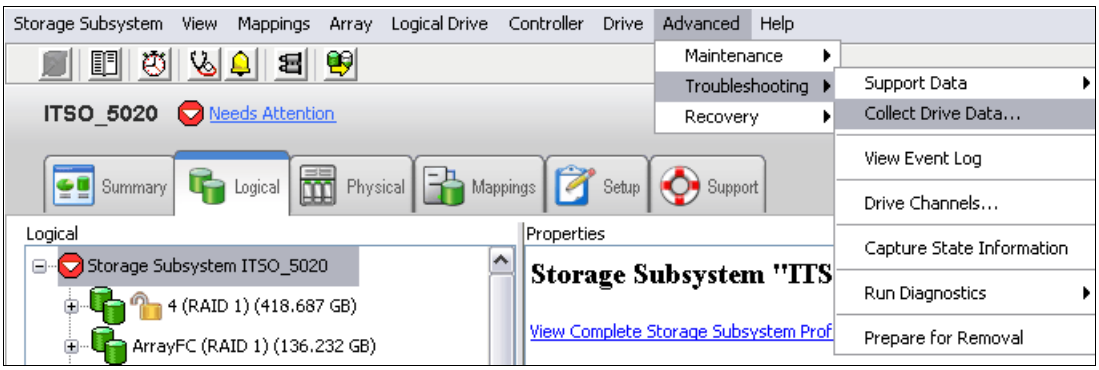


Figure 8-2 Collecting support data

This selection creates a .zip file of all the internal information of the DS5000 Storage Server for review by the support organization, which includes the storage Subsystems Profile, majorEventLog, driveDiagnosticData, NVSRAM data, readLinkStatus, performanceStatistics, and many others. This information, when combined and compared to the Performance Monitor data, can give a good picture of how the DS5000 Storage Server sees its workload being handled, and any areas it sees that are having trouble.

The performanceStatistics file provides a far greater amount of coverage of data gathering, and a further breakdown of the I/O details of what the storage server workload has been. In this spreadsheet based log, the read and write ratio can be seen for all the logical drives, controllers, and the storage server's workload. Also, the cache hits and misses for each type (read and write) can be viewed.

8.2 Storage Manager Performance Monitor

The Storage Performance Monitor is a tool built into the DS5000 Storage Manager client. It monitors performance on each logical drive, and collects the following information:

- Total I/Os
- Read percentage
- Cache hit percentage
- Current KBps and maximum KBps
- Current I/O per sec and maximum I/O per sec

In the remainder of this chapter, we describe in detail how to use data from the Performance Monitor, as well as monitoring the effect of a tuning option parameter in the DS5000 Storage Server that affects the storage server performance.

8.2.1 Starting the Performance Monitor

To launch the Performance Monitor from the SMclient Subsystem Management window, do these steps:

1. Select the **Monitor Performance** icon.
2. Select **Storage Subsystem** → **Monitor Performance**.
3. Select the storage subsystem node in the Logical View or Mappings View, right-click, and select **Monitor Performance**.

The Performance Monitor window opens with all logical drives displayed, as shown in Figure 8-3.

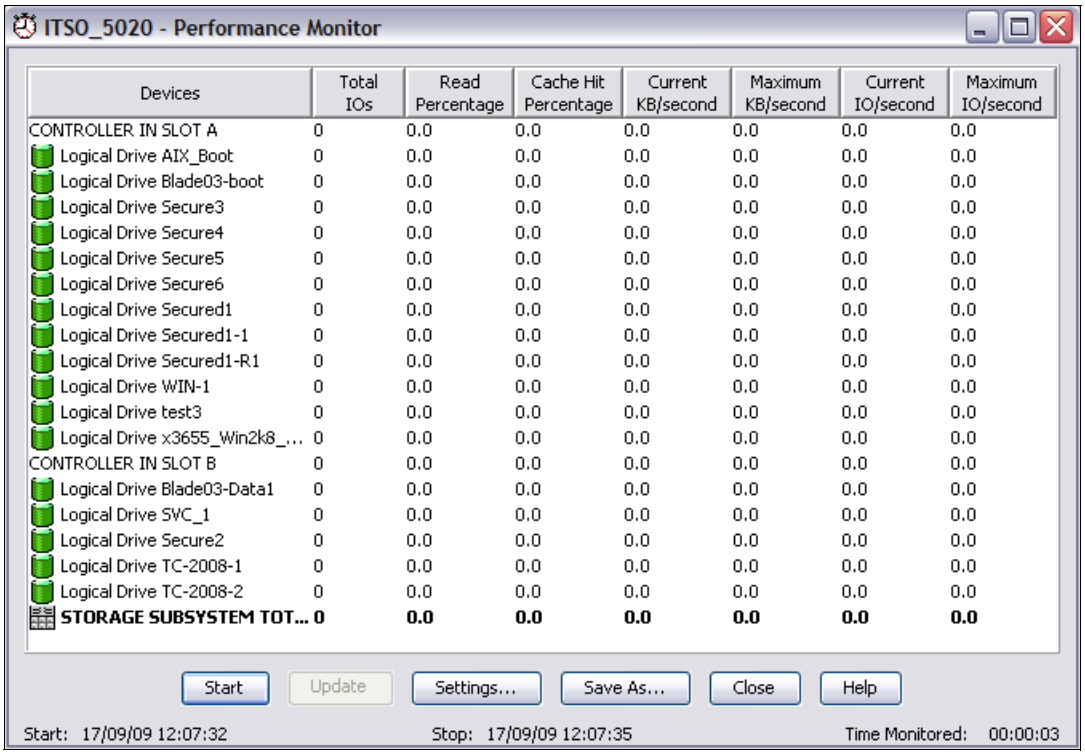


Figure 8-3 Performance Monitor

Table 8-1 describes the information collected by the Performance Monitor.

Table 8-1 Information collected by Performance Monitor

Data field	Description
Total I/Os	Total I/Os performed by this device since the beginning of the polling session.
Read percentage	The percentage of total I/Os that are read operations for this device. Write percentage can be calculated as 100 minus this value.
Cache hit percentage	The percentage of reads that are processed with data from the cache rather than requiring a read from disk.
Current KBps	Average *transfer rate* during the polling session. The transfer rate is the amount of data in kilobytes that can be moved through the I/O data connection in a second (also called *throughput*).
Maximum KBps	The maximum transfer rate that was achieved during the Performance Monitor polling session.
Current I/O per second	The average number of I/O requests serviced per second during the current polling interval (also called an *I/O request rate*).
Maximum I/O per second	The maximum number of I/O requests serviced during a one-second interval over the entire polling session.

Here we describe the following fields:

- Total I/Os:

  This data is useful for monitoring the I/O activity of a specific controller and a specific logical drive, which can help identify possible high-traffic I/O areas.

  If I/O rate is slow on a logical drive, try increasing the array size.

  There might be a disparity in the Total I/Os (workload) of controllers, for example, the workload of one controller is heavy or is increasing over time, whereas that of the other controller is lighter or more stable. In this case, consider changing the controller ownership of one or more logical drives to the controller with the lighter workload. Use the logical drive Total I/O statistics to determine which logical drives to move.

  If the workload across the storage subsystem (Storage Subsystem Totals Total I/O statistic) continues to increase over time, while application performance decreases, this might indicate that you need to add additional storage subsystems to the installation so that the application needs can be met at an acceptable performance level.

- Read percentage:

  Use the read percentage for a logical drive to determine actual application behavior. If there is a low percentage of read activity relative to write activity, consider reviewing the characteristics of the host and application using the logical drive with a view of changing tuning parameters for improved performance.

- Cache hit percentage:

  The percentage of reads that are fulfilled by data from the cache rather than requiring an actual read from disk. A higher percentage is desirable for optimal application performance. There is a positive correlation between the cache hit percentage and I/O rates.

  The cache hit percentage of all of the logical drives might be low or trending downward. This value might indicate inherent randomness in access patterns, or, at the storage subsystem or controller level, it can indicate the need to install more controller cache memory if the maximum amount of memory is not yet installed.

  If an individual logical drive is experiencing a low cache hit percentage, consider enabling cache read-ahead for that logical drive. Cache read ahead can increase the cache hit percentage for a sequential I/O workload.

  To determine if your I/O has sequential characteristics, try enabling a conservative cache read-ahead multiplier (four, for example). Then, examine the logical drive cache hit percentage to see if it has improved. If it has, indicating that your I/O has a sequential pattern, and enable a more aggressive cache read-ahead multiplier (eight, for example). Continue to customize logical drive cache read-ahead to arrive at the optimal multiplier (in the case of a random I/O pattern, the optimal multiplier is zero).

- Current KB/sec and maximum KB/sec:

  The *Current KB/sec* value is the average size of the amount of data that was transferred over one second during a particular *interval period* (or since the “update” was selected) that was monitored. The *Maximum KB/sec* value is the highest amount that was transferred over any one second period during all of the interval periods in the *number of iterations* that we ran for a specific command. This value can show when a peak transfer rate period was detected during the command run time.

  Tip: If Maximum KB/sec is the same as the last interval’s Current KB/sec, extend the number of iterations to see when the peak rate is actually reached. Because it might be on the rise, the aim is to determine the maximum rate.

The transfer rates of the controller are determined by the application I/O size and the I/O rate. Generally, small application I/O requests result in a lower transfer rate, but provide a faster I/O rate and shorter response time. With larger application I/O requests, higher throughput rates are possible. Understanding the typical application I/O patterns can help determine the maximum I/O transfer rates for a given storage subsystem.

Consider a storage subsystem equipped with Fibre Channel controllers that supports a maximum transfer rate of 100 MBps (100,000 KBps). The storage subsystem typically achieves an average transfer rate of 20,000 Kbps. (The typical I/O size for applications is, for example, 4 KB, with 5,000 I/Os transferred per second for an average rate of 20,000 Kbps.) In this case, I/O size is small. Because there is an impact associated with each I/O, the transfer rates will not approach 100,000 Kbps. However, if the typical I/O size is large, a transfer rate within a range of 80,000 to 90,000 Kbps can be achieved.

- Current I/O per second and maximum I/O per second:

  The *Current IO/sec* value is the average number of I/Os serviced in one second during a particular *interval period* that was monitored. The *Maximum IO/sec* value is the highest number of I/Os serviced in any one second period, during all of the interval periods in the *number of iterations* that were ran for a specific command. This value can show when the peak I/O period was detected during the command run time.

  Tip: If Maximum I/Ops is the same as the last interval's Current I/Ops, extend the number of iterations to see when the peak is actually reached, as this might be on the rise, the aim is to determine the maximum rate.

  Factors that affect I/Os per second include access pattern (random or sequential), I/O size, RAID level, segment size, and number of drives in the arrays or storage subsystem. The higher the cache hit rate, the higher the I/O rates.

  Performance improvements caused by changing the segment size can be seen in the I/Os per second statistics for a logical drive. Experiment to determine the optimal segment size, or use the file system or database block size.

  Higher write I/O rates are experienced with write caching enabled compared to disabled. In deciding whether to enable write caching for an individual logical drive, consider the current and maximum I/Os per second. You can expect to see higher rates for sequential I/O patterns than for random I/O patterns. Regardless of the I/O pattern, write caching must be enabled to maximize I/O rate and shorten application response time.

8.2.2 Using the Performance Monitor

The Performance Monitor queries the storage subsystem at regular intervals. To change the polling interval and to select only the logical drives and the controllers required to be monitored, click the **Settings** button.

To change the polling interval, choose a number of seconds in the spin box as shown in Figure 8-4. Each time the polling interval elapses, the Performance Monitor re-queries the storage subsystem and updates the statistics in the table. If monitoring the storage subsystem is in real time, update the statistics frequently by selecting a short polling interval, for example, five seconds. If results are being saved to a file for later review, choose a slightly longer interval, for example, 30 to 60 seconds, to decrease the performance impact.

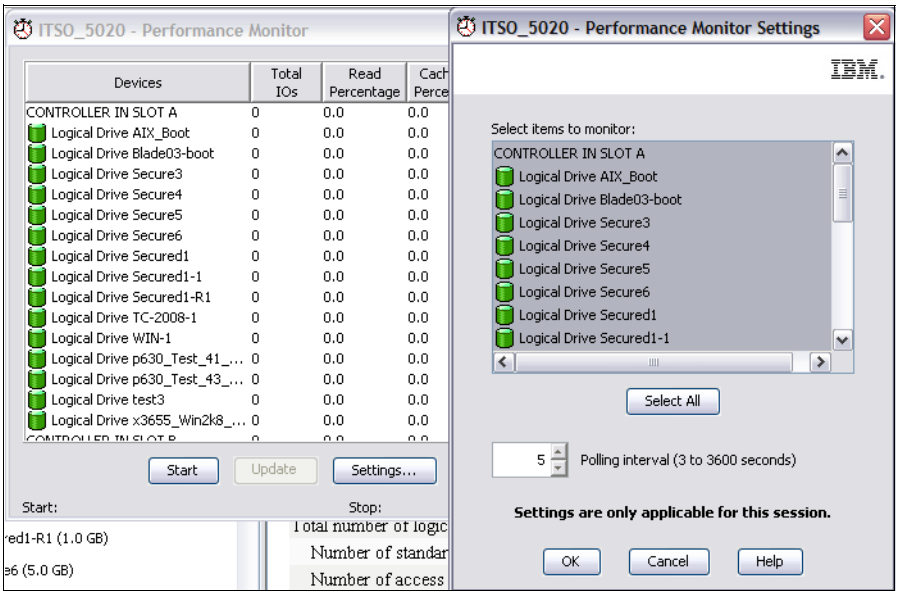


Figure 8-4 Performance Monitor Settings

The Performance Monitor will not dynamically update its display if any configuration changes occur while the monitor window is open (for example, creation of new logical drives, change in logical drive ownership, and so on). The Performance Monitor window must be closed and then reopened for the changes to appear.

Tip: Using the Performance Monitor to retrieve performance data can affect the normal storage subsystem performance, depending on how many items you want to monitor and the refresh interval.

If the storage subsystem monitored begins or transitions to an unresponsive state, a window opens stating that the Performance Monitor cannot poll the storage subsystem for performance data.

The Performance Monitor is a real time tool; it is not possible to collect performance data over time with the Storage Manager GUI. However, a simple script can be used to collect performance data over a period of time for later analysis.

The script can be run from the command line using SMcli or from the Storage Manager Script Editor GUI. From the *Storage Manager Enterprise management* window, select **Tools** → **Execute Script**. Data collected while executing the script is saved in the file and directory specified in the storage Subsystem file parameter.

Using the script editor GUI

A sample of the script editor GUI is shown in Figure 8-5.

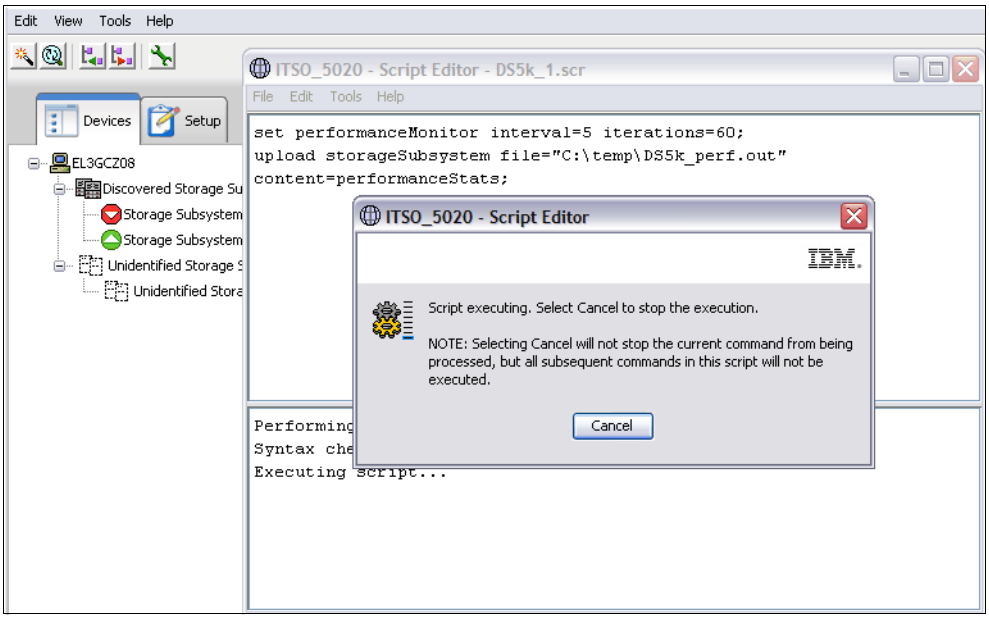


Figure 8-5 Script to collect Performance Monitor data over time

A sample of the output file is shown in Example 8-2.

Example 8-2 Script output file

```
"Performance Monitor Statistics for Storage Subsystem: ITSO_5020 - Date/Time: 21/09/09
10:39:56 - Polling interval in seconds: 5"


"Storage Subsystems ","Total IOs ","Read Percentage ","Cache Hit Percentage ","Current
KB/second ","Maximum KB/second ","Current IO/second ","Maximum IO/second"
"Date/Time: 21/09/09 10:45:06","","","","","","","",""
"CONTROLLER IN SLOT A","12560.0","16.6","12.9","5453.6","26440.6","213.2","354.2"
"Logical Drive AIX_Boot","4537.0","45.8","12.0","3506.4","7444.6","197.4","197.4"
"Logical Drive Blade03-boot","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secure3","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secure4","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secure5","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secure6","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secured1","8022.0","0.1","100.0","1947.2","18996.0","15.8","296.8"
"Logical Drive Secured1-1","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secured1-R1","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive TC-2008-1","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive WIN-1","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive test3","1.0","100.0","0.0","0.0","0.8","0.0","0.2"
"Logical Drive x3655_Win2k8_F_iscsi","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"CONTROLLER IN SLOT B","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Blade03-Data1","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive SVC_1","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive Secure2","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"Logical Drive TC-2008-2","0.0","0.0","0.0","0.0","0.0","0.0","0.0"
"STORAGE SUBSYSTEM TOTALS","12560.0","16.6","12.9","5453.6","26440.6","213.2","354.2"
"Capture Iteration: 59","","","","","","","",""
```

Using the command line interface (CLI)

The **SMcli** command can also be used to gather Performance Monitor data over a period of time. It requires scripting skills in order to set up. Example 8-3 shows a test_script file for execution under AIX, using the ksh test_script:

```
# cat SMcli_1_scr
```

Example 8-3 Test script

```
#!/bin/ksh
#The information is captured from a single Linux/AIX server by running the
# following "Storage Manager Command Line Interface Utility" Linux/AIX command
CMD='set session performanceMonitorInterval=60 performanceMonitorIterations=2; \
show allLogicalDrives performanceStats;'
/usr/SMclient/SMcli -e -S 9.1.39.26 9.1.39.27 -p xxxxxx -c "$CMD"
#(Note; this will run every minute for 2 times; if run every 10 minutes for 10
# times set the
# "performanceMonitorinterval=600"
# "performanceMonitorIterations=10"
```

The first executable line sets the CMD variable; the second executable line invokes the **SMcli** command. Note that for the -S parameter, it is necessary to specify the IP address of both DS5000 controllers (A and B).

The output resulting from the script execution can be redirected to a file by typing the following command:

```
./SMcli_1_scr > script1.o 2>&1
```

This test script collects information for all logical drives, but it is possible to select specific logical drives. Example 8-4 shows how to select only two drives, p630_Test_42_512KB p630_Test_43_128KB.

Example 8-4 Test script for two logical drives for a UNIX platform

```
#!/bin/ksh
#The information is captured from a single Linux/AIX server by running the
# following "Storage Manager Command Line Interface Utility" Linux/AIX command
CMD='set session performanceMonitorInterval=20 performanceMonitorIterations=9;
show LogicalDrives [p630_Test_42_512KB p630_Test_43_128KB ] performanceStats;'
/usr/SMclient/SMcli -e -S 9.11.218.182 9.11.218.183 -c "$CMD"
```

For Windows platform users who want to run a similar script from a Windows server or their own PCs, the script shown in Example 8-5 performs the same function.

Example 8-5 Similar test script for four logical drives on WIN platform

```
rem Run Performance Manager
rem replacement for CLI2 for all logical drives
rem CLI2=show AllLogicalDrives performanceStats;"


set CMD="C:\Program Files\IBM_DS\client\SMcli.exe"
set IP=9.11.218.182 9.11.218.183
set CLI1="set session performanceMonitorInterval=60
performanceMonitorIterations=20;
set CLI2=show LogicalDrives [p630_Test_42_512KB p630_Test_43_128KB
p630_Test_40_64KB p630_Test_43_128KB] performanceStats;"
set OUT="C:\temp\WIN_SCR.txt"

rem cd %CD%
%CMD% -e -S %IP% -p xxxxxxx -c %CLI1% %CLI2% -o %OUT%
```

We created a file called script2.o:

```
./SMcli_2_scr > script2.o 2>&1
```

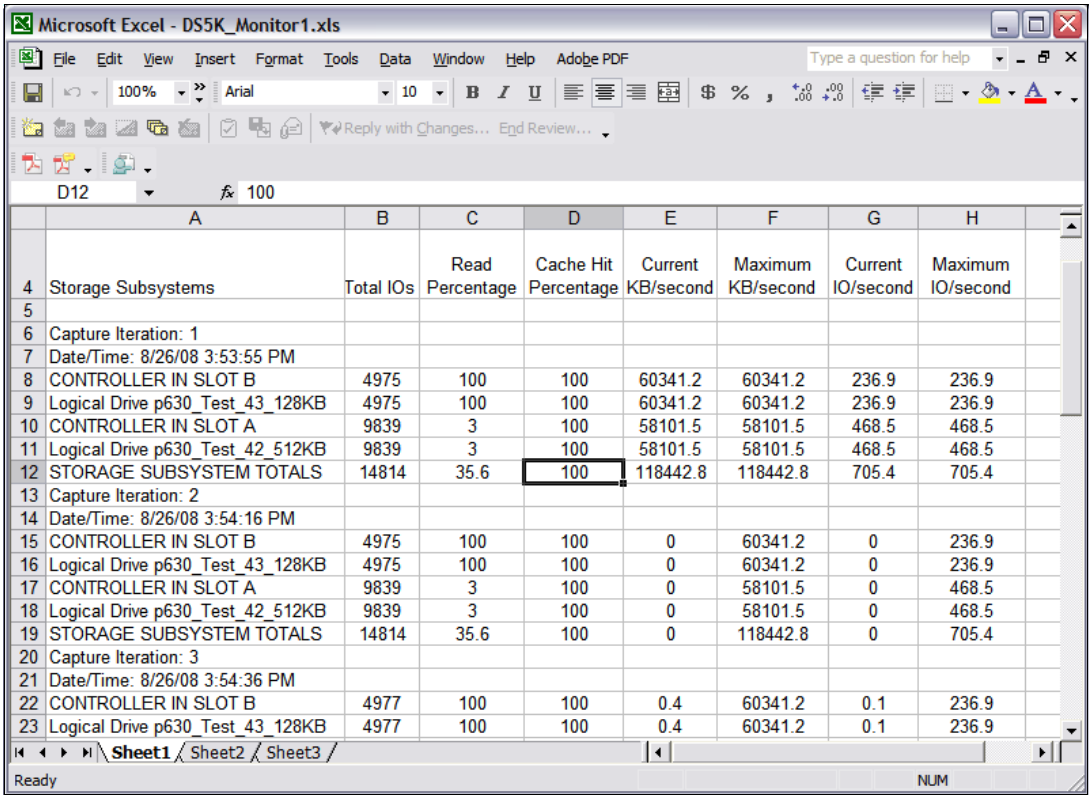
The output file (called `script2.o`) of this script is shown in Example 8-6.

Example 8-6 Output file: script2.o

```
"Performance Monitor Statistics for Storage Subsystem: DS5000 - Date/Time: 8/26/11 3:53:54
PM - Polling interval in seconds: 20"

"Storage Subsystems ","Total IOs ","Read Percentage ","Cache Hit Percentage ","Current
KB/second ","Maximum KB/second ","Current IO/second ","Maximum IO/second"
"Capture Iteration: 1","","","","","","","",""
"Date/Time: 8/26/08 3:53:55 PM","","","","","","","",""
"CONTROLLER IN SLOT B","4975.0","100.0","100.0","60341.2","60341.2","236.9","236.9"
"Logical Drive
p630_Test_43_128KB","4975.0","100.0","100.0","60341.2","60341.2","236.9","236.9"
"CONTROLLER IN SLOT A","9839.0","3.0","100.0","58101.5","58101.5","468.5","468.5"
"Logical Drive
p630_Test_42_512KB","9839.0","3.0","100.0","58101.5","58101.5","468.5","468.5"
"STORAGE SUBSYSTEM TOTALS","14814.0","35.6","100.0","118442.8","118442.8","705.4","705.4"
"Capture Iteration: 2","","","","","","","",""
"Date/Time: 8/26/08 3:54:16 PM","","","","","","","",""
"CONTROLLER IN SLOT B","4975.0","100.0","100.0","0.0","60341.2","0.0","236.9"
"Logical Drive p630_Test_43_128KB","4975.0","100.0","100.0","0.0","60341.2","0.0","236.9"
"CONTROLLER IN SLOT A","9839.0","3.0","100.0","0.0","58101.5","0.0","468.5"
"Logical Drive p630_Test_42_512KB","9839.0","3.0","100.0","0.0","58101.5","0.0","468.5"
"STORAGE SUBSYSTEM TOTALS","14814.0","35.6","100.0","0.0","118442.8","0.0","705.4"
```

All the data saved in the file is comma delimited so that the file can be easily imported into a spreadsheet for easier analysis and review (Figure 8-6).



	A	B	C	D	E	F	G	H
4	Storage Subsystems	Total IOs	Read Percentage	Cache Hit Percentage	Current KB/second	Maximum KB/second	Current IO/second	Maximum IO/second
5								
6	Capture Iteration: 1							
7	Date/Time: 8/26/08 3:53:55 PM							
8	CONTROLLER IN SLOT B	4975	100	100	60341.2	60341.2	236.9	236.9
9	Logical Drive p630_Test_43_128KB	4975	100	100	60341.2	60341.2	236.9	236.9
10	CONTROLLER IN SLOT A	9839	3	100	58101.5	58101.5	468.5	468.5
11	Logical Drive p630_Test_42_512KB	9839	3	100	58101.5	58101.5	468.5	468.5
12	STORAGE SUBSYSTEM TOTALS	14814	35.6	100	118442.8	118442.8	705.4	705.4
13	Capture Iteration: 2							
14	Date/Time: 8/26/08 3:54:16 PM							
15	CONTROLLER IN SLOT B	4975	100	100	0	60341.2	0	236.9
16	Logical Drive p630_Test_43_128KB	4975	100	100	0	60341.2	0	236.9
17	CONTROLLER IN SLOT A	9839	3	100	0	58101.5	0	468.5
18	Logical Drive p630_Test_42_512KB	9839	3	100	0	58101.5	0	468.5
19	STORAGE SUBSYSTEM TOTALS	14814	35.6	100	0	118442.8	0	705.4
20	Capture Iteration: 3							
21	Date/Time: 8/26/08 3:54:36 PM							
22	CONTROLLER IN SLOT B	4977	100	100	0.4	60341.2	0.1	236.9
23	Logical Drive p630_Test_43_128KB	4977	100	100	0.4	60341.2	0.1	236.9

Figure 8-6 Importing results into a spreadsheet

8.2.3 Using the Performance Monitor: An illustration

To illustrate the use of the Performance Monitor, we compare the difference in throughput of four logical devices that have been created on the same array; each LUN is the same size but with varying segment sizes. These logical devices have been mapped to an AIX platform running AIX V6.1. Two identical host bus adapters were attached to the IBM POWER AIX machine and a single SAN fabric switch was used. Example 8-7 shows the configuration where the AIX hdisks are distributed across both HBAs and DS5000 controllers.

Example 8-7 LUN attachment to AIX

```
# mpio_get_config  -Av
Frame id 0:
    Storage Subsystem worldwide name: 608e50017b5bc00004a955e3b
    Controller count: 2
    Partition count: 1
    Partition 0:
    Storage Subsystem Name = 'ITSO_5020'
        hdisk      LUN #   Ownership          User Label
        hdisk2         4   A (preferred)      test3
        hdisk4         2   A (preferred)      Secured1
        hdisk5         3   A (preferred)      AIX_Boot
        hdisk6        40   B (preferred)      p630_Test_40_64KB
        hdisk7        41   A (preferred)      p630_Test_41_8KB
        hdisk8        42   B (preferred)      p630_Test_42_512KB
        hdisk9        43   A (preferred)      p630_Test_43_128KB
```

The LUN naming convention is indicative of the segment size used when creating it on the DS5000 Storage Server, for example, LUN “p630_Test_42_512KB” has a segment size of 512 KB. All LUNs are the same size, as shown in Figure 8-7.

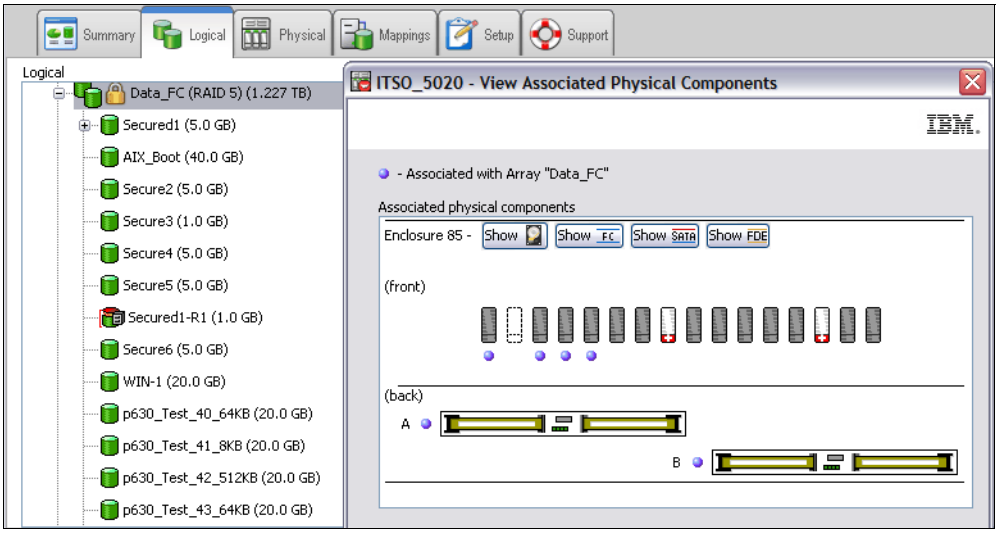


Figure 8-7 Logical/Physical view of test LUNs

The LUNs are used in a test volume group where a single logical volume of equal size is created on each LUN (see Example 8-8 and Example 8-9).

Example 8-8 Creating logical volumes

```
root@itsop630:/root
# mklv -t jfs2 -y test64lv ds5kvg 159 hdisk6
test64lv
root@itsop630:/root
# mklv -t jfs2 -y test8lv ds5kvg 159 hdisk7
test8lv
root@itsop630:/root
# mklv -t jfs2 -y test512lv ds5kvg 159 hdisk8
test512lv
root@itsop630:/root
# mklv -t jfs2 -y test128lv ds5kvg 159 hdisk9
test128lv
root@itsop630:/root
```

Example 8-9 AIX logical volumes and volume group

```
# lspv
hdisk0          00000000f1b495eb                    rootvg          active
hdisk1          000000006d598dee                    rootvg          active
hdisk2          none                                None
hdisk3          0007041aa50076cf                    ds5kvg          active
hdisk4          0007041aa5007563                    ds5kvg          active
hdisk5          0007041ab945d714                    rootvg          active
hdisk6          0007041adddca404                    ds5kvg          active
hdisk7          0007041adddca5a1                    ds5kvg          active
hdisk8          0007041adddca6d8                    ds5kvg          active
hdisk9          0007041adddca80f                    ds5kvg          active
root@itsop630:/root
# lsvg -l ds5kvg
ds5kvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
test1_lv            jfs2       50      50      1    open/syncd    /test1
test2_lv            jfs2       26      26      1    open/syncd    /test2
test64lv            jfs2       159     159     1    closed/syncd  N/A
test8lv             jfs2       159     159     1    closed/syncd  N/A
test512lv           jfs2       159     159     1    closed/syncd  N/A
test128lv           jfs2       159     159     1    closed/syncd  N/A
root@itsop630:/root
# lsvg ds5kvg
VOLUME GROUP:       ds5kvg                  VG IDENTIFIER:
0007041a00004c0000000123a500779e
VG STATE:           active                  PP SIZE:        128 megabyte(s)
VG PERMISSION:      read/write              TOTAL PPs:      914 (116992
megabytes)
MAX LVs:            256                     FREE PPs:       202 (25856 megabytes)
LVs:                6                       USED PPs:       712 (91136 megabytes)
OPEN LVs:           2                       QUORUM:         4 (Enabled)
TOTAL PVs:          6                       VG DESCRIPTORS: 6
STALE PVs:          0                       STALE PPs:      0
ACTIVE PVs:         6                       AUTO ON:        yes
MAX PPs per VG:     32512
MAX PPs per PV:     1016                    MAX PVs:        32
LTG size (Dynamic): 256 kilobyte(s)         AUTO SYNC:      no
HOT SPARE:          no                      BB POLICY:      relocatable
root@itsop630:/root
```

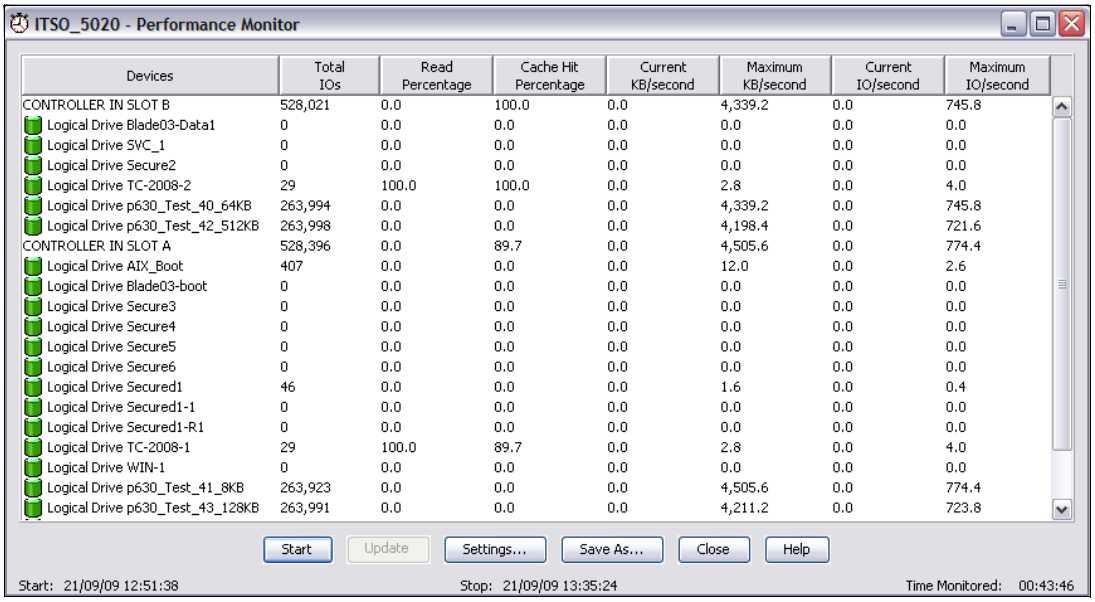
For this test, we disable cache write on the DS5000 Storage Server for all the LUNs and run the following test with the **dd** command, where a sequential write is simulated to the target logical volumes. First, the Performance Monitor is enabled and then commands are run, as shown in Example 8-10.

Example 8-10 dd command

```
time dd if=/dev/zero bs=128k count=12000 of=/dev/test8lv
time dd if=/dev/zero bs=128k count=12000 of=/dev/test64lv
time dd if=/dev/zero bs=128k count=12000 of=/dev/test128lv
time dd if=/dev/zero bs=128k count=12000 of=/dev/test512lv
```

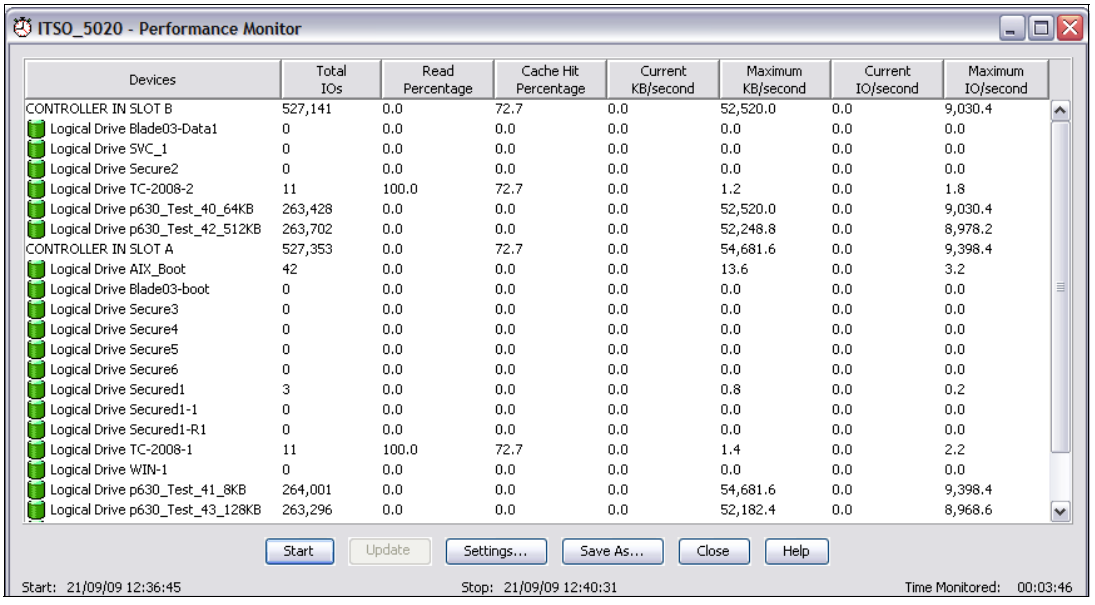
The Performance Monitor that ran for the duration of the commands is shown in Figure 8-8.

The test was then repeated with cache write turned on by the DS5000 Storage Server for each of the LUNs, and the Performance Monitor was stopped and restarted between tests. Figure 8-9 shows the results of this test.



Devices	Total IOs	Read Percentage	Cache Hit Percentage	Current KB/second	Maximum KB/second	Current IO/second	Maximum IO/second
CONTROLLER IN SLOT B	528,021	0.0	100.0	0.0	4,339.2	0.0	745.8
Logical Drive Blade03-Data1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive SVC_1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure2	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive TC-2008-2	29	100.0	100.0	0.0	2.8	0.0	4.0
Logical Drive p630_Test_40_64KB	263,994	0.0	0.0	0.0	4,339.2	0.0	745.8
Logical Drive p630_Test_42_512KB	263,998	0.0	0.0	0.0	4,198.4	0.0	721.6
CONTROLLER IN SLOT A	528,396	0.0	89.7	0.0	4,505.6	0.0	774.4
Logical Drive AIX_Boot	407	0.0	0.0	0.0	12.0	0.0	2.6
Logical Drive Blade03-boot	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure3	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure4	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure5	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure6	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secured1	46	0.0	0.0	0.0	1.6	0.0	0.4
Logical Drive Secured1-1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secured1-R1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive TC-2008-1	29	100.0	89.7	0.0	2.8	0.0	4.0
Logical Drive WIN-1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive p630_Test_41_8KB	263,923	0.0	0.0	0.0	4,505.6	0.0	774.4
Logical Drive p630_Test_43_128KB	263,991	0.0	0.0	0.0	4,211.2	0.0	723.8

Figure 8-8 Observe with Performance Monitor (Cache Write Off)



Devices	Total IOs	Read Percentage	Cache Hit Percentage	Current KB/second	Maximum KB/second	Current IO/second	Maximum IO/second
CONTROLLER IN SLOT B	527,141	0.0	72.7	0.0	52,520.0	0.0	9,030.4
Logical Drive Blade03-Data1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive SVC_1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure2	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive TC-2008-2	11	100.0	72.7	0.0	1.2	0.0	1.8
Logical Drive p630_Test_40_64KB	263,428	0.0	0.0	0.0	52,520.0	0.0	9,030.4
Logical Drive p630_Test_42_512KB	263,702	0.0	0.0	0.0	52,248.8	0.0	8,978.2
CONTROLLER IN SLOT A	527,353	0.0	72.7	0.0	54,681.6	0.0	9,398.4
Logical Drive AIX_Boot	42	0.0	0.0	0.0	13.6	0.0	3.2
Logical Drive Blade03-boot	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure3	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure4	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure5	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secure6	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secured1	3	0.0	0.0	0.0	0.8	0.0	0.2
Logical Drive Secured1-1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive Secured1-R1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive TC-2008-1	11	100.0	72.7	0.0	1.4	0.0	2.2
Logical Drive WIN-1	0	0.0	0.0	0.0	0.0	0.0	0.0
Logical Drive p630_Test_41_8KB	264,001	0.0	0.0	0.0	54,681.6	0.0	9,398.4
Logical Drive p630_Test_43_128KB	263,296	0.0	0.0	0.0	52,182.4	0.0	8,968.6

Figure 8-9 Observe with Performance Monitor (Cache Write On)

Both Figure 8-8 and Figure 8-9 show the results after the identical tests were run. Looking at the “Maximum IO/Second” and “Maximum KB/Second” columns for each figure, we see quite a difference after making a single (albeit major) performance tuning parameter change on each LUN. The maximum IO/second and KB/second columns indicate that, with write cache disabled, the transfer is much slower than when it is enabled.

Each command in Example 8-10 on page 357 that is entered on the AIX host has an output that measures the length of time the command took, as shown in Example 8-11. Measured times and Performance Monitor results from the tests for the AIX host can be compared.

Example 8-11 Output example of dd command

```
# time dd if=/dev/zero bs=128k count=12000 of=/dev/test512lv
12000+0 records in.
12000+0 records out.

real    6m29.15s
user    0m0.12s
sys     0m7.22s
root@itsop630:/root
```

In Table 8-2, we can see when the host and Performance Monitor results are tabled, and the test shows the performance increase when the write cache is enabled. It also highlights the marginally poorer performance of a LUN with a very small segment size in both tests.

Table 8-2 Host and Performance Monitor results

LUN	Total I/Os	Max KB/Sec	Max I/Os/Sec	Time (mm:ss.xx)
Write Cache Disabled				
p630_Test_40_64KB	263,994	4,339.2	745.8	06:24.36
p630_Test_41_8KB	263,923	4,505.6	774.4	06:53.49
p630_Test_42_512KB	263,998	4,198.4	721.6	06:29.15
p630_Test_40_128KB	263,991	4,211.2	723.8	06:29.13
Write Cache Enabled				
p630_Test_40_64KB	263,428	52,520	9,030.4	00:31.19
p630_Test_41_8KB	264,001	54,681	9,398.4	00:38.11
p630_Test_42_512KB	263,702	52,248.8	8,978.2	00:32.96
p630_Test_40_128KB	263,296	52,182.4	8,968.6	00:33.99

With the SM Performance Monitor, LUN and controller workload can be monitored in real time with a summary of performance over the period when the Performance Monitor is stopped. Using **SMcli** and scripts, performance data can be collected and put into a file for later review and used to build historical performance information. Based on the results of the monitoring, volumes can be moved from one controller to the other in order to balance the performance between controllers.

SM Performance Monitor can also be used in conjunction with a specific host in order to confirm any performance tuning parameter changes made on the host, application, or database, and monitor any workload differences of the LUNs.

Tip: If you do not regularly collect and analyze DS5000 performance data using the previous methods, at least collect Performance Monitor data samples periodically even if you do not intend to use or analyze them. They can be used at a point in the future when there are changes planned, whether in performance or capacity. The data can then be used and analyzed as a comparison to the present workloads.

8.3 Use of Performance Monitor Data

The data from Performance Monitor can be used in other performance related tools; it can be imported into these tools to be analyzed. Data generated from Performance Monitor output file (as described in 8.2.2, “Using the Performance Monitor” on page 350 and “Using the command line interface (CLI)” on page 353) can be used by tools so that their meaning can be interpreted easier and analyzed more efficiently.

8.3.1 Disk Magic

Disk Magic uses the Performance Monitor “raw data” file by first creating a comma separated variable (CSV) file and then importing selected data into the tool along with the DS5000 profile. We explain how to do it in Chapter 10, “Disk Magic” on page 425, where data from logical volume and storage controller summaries are manually copied from the CSV file into the Disk Magic tool base line model. Data such as caching statistics including cache hit figures are particularly useful when using Disk Magic.

8.3.2 Tivoli Storage Productivity Centre (TPC) for Disk

TPC for Disk has the ability to collect information from the DS5000 independently of the DS5000 Performance Monitor and is an example of an external tool which, among other functions, can monitor DS5000 performance. TPC is able to store and report both actual and historical data to the same detail as the Performance Manager. The method that TPC employs varies from that used by Disk Magic in that it can collect the data as the DS5000 reports it continuously after the DS5000 is defined to TPC and connected. TPC can also provide logical volume creation and management as well as performance data gathering and reporting.

TPC for Disk uses a number of industry standards supported on an increasing range of equipment that can be managed, thus replacing proprietary software with a common standard. The DS5000 is compliant with the Common Interface Module (CIM) communicating with a CIM Object Module (CIMOM) which will interface with TPC. The CIMOM will have a specific plug-in device handler that manages instructions to and from the DS5000.

The CIM standard for storage management has been integrated into the Storage Network Industry Association (SNIA) in their Storage Management Initiative Specification (SMI-S) allowing a universal open interface for managing storage devices. TPC uses these standards to manage and report on the DS5000 performance and thus becomes a very versatile tool in providing both performance data and historical performance data on the DS5000.

For further details about TPC for disk, see Chapter 9, “IBM Tivoli Storage Productivity Center for Disk” on page 361 and *Tivoli Storage Productivity Center V4.2 Release Guide*, SG24-7894.

9

IBM Tivoli Storage Productivity Center for Disk

In this chapter, we describe how to use IBM Tivoli Storage Productivity Center for Disk (TPC for Disk) to manage and monitor performance on the DS5000 Storage Servers.

The first part of this chapter is a brief overview of the overall TPC offering that includes components such as TPC for Data, TPC for Disk, and TPC for Replication.

The second part of the chapter focuses on TPC for Disk and the steps required to manage and monitor the DS5000 from a TPC server. It also includes examples of performance reports that TPC for Disk can generate.

For more detailed information about TPC, see *IBM Tivoli Storage Productivity Center V4.2 Release Guide*, SG24-7894.

9.1 IBM Tivoli Storage Productivity Center

IBM Tivoli Storage Productivity Center is an integrated set of software components that provides end-to-end storage management. This software offering provides disk and tape library configuration and management, performance management, SAN fabric management and configuration, and host-centered usage reporting and monitoring from the perspective of the database application or file system.

IBM Tivoli Storage Productivity Center offers the following benefits:

- It simplifies the management of storage infrastructures.
- It configures, and provisions SAN-attached storage.
- It monitors and tracks performance of SAN-attached devices.
- It monitors, manages, and controls (through zones) SAN fabric components.
- It manages the capacity utilization and availability of file systems and databases.

9.1.1 Tivoli Storage Productivity Center structure

In this section, we look at the Tivoli Storage Productivity Center structure from the logical and physical view.

Logical structure

The logical structure of IBM Tivoli Storage Productivity Center has three layers, as shown in Figure 9-1.

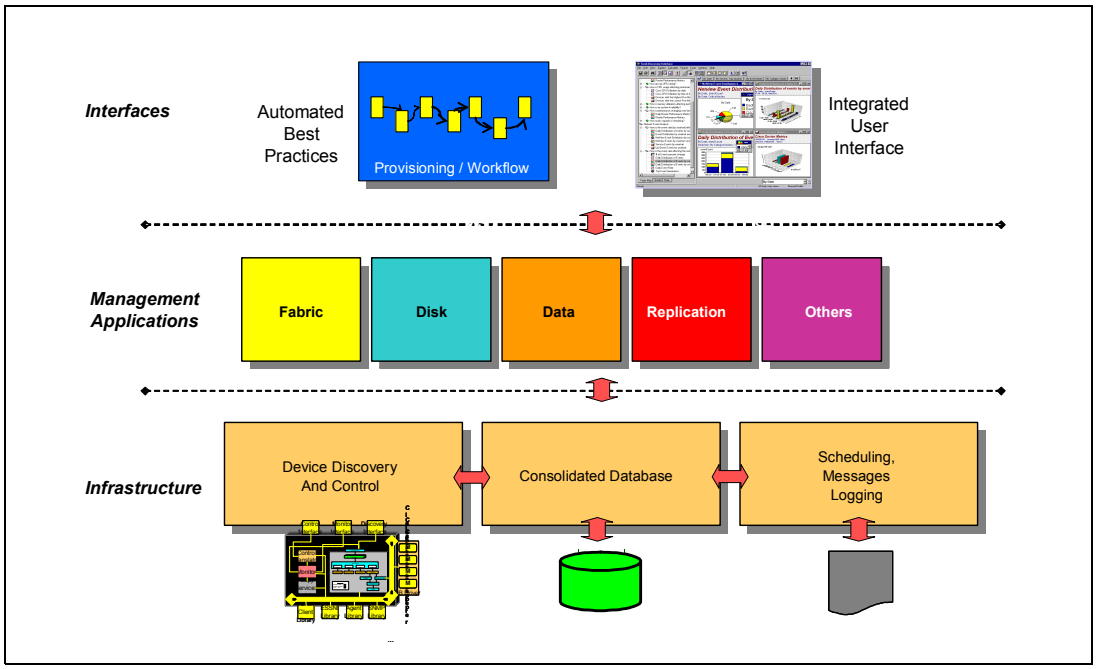


Figure 9-1 IBM Tivoli Storage Productivity Center logical structure

The *infrastructure layer* consists of basic functions such as messaging, scheduling, logging, device discovery, and a consolidated database shared by all components of Tivoli Storage Productivity Center to ensure consistent operation and performance.

The *application layer* consists of core Tivoli Storage Productivity Center management functions, that rely on the common base infrastructure to provide various options of storage or

data management. These application components are most often associated with the product components that make up the product suite, such as fabric management, disk management, replication management, and data management.

The *interface layer* presents integration points for the products that make up the suite. The integrated graphical user interface (GUI) brings together product and component functions into a single representation that seamlessly interacts with the components to centralize the tasks for planning, monitoring, configuring, reporting, topology viewing, and problem resolving.

Physical structure

IBM Tivoli Storage Productivity Center is comprised of the following components:

- A data component: IBM Tivoli Storage Productivity Center for Data
- A disk component: IBM Tivoli Storage Productivity Center for Disk
- A replication component: IBM Tivoli Storage Productivity Center for Replication

IBM Tivoli Storage Productivity Center includes a centralized suite installer.

Figure 9-2 shows the Tivoli Storage Productivity Center physical structure.

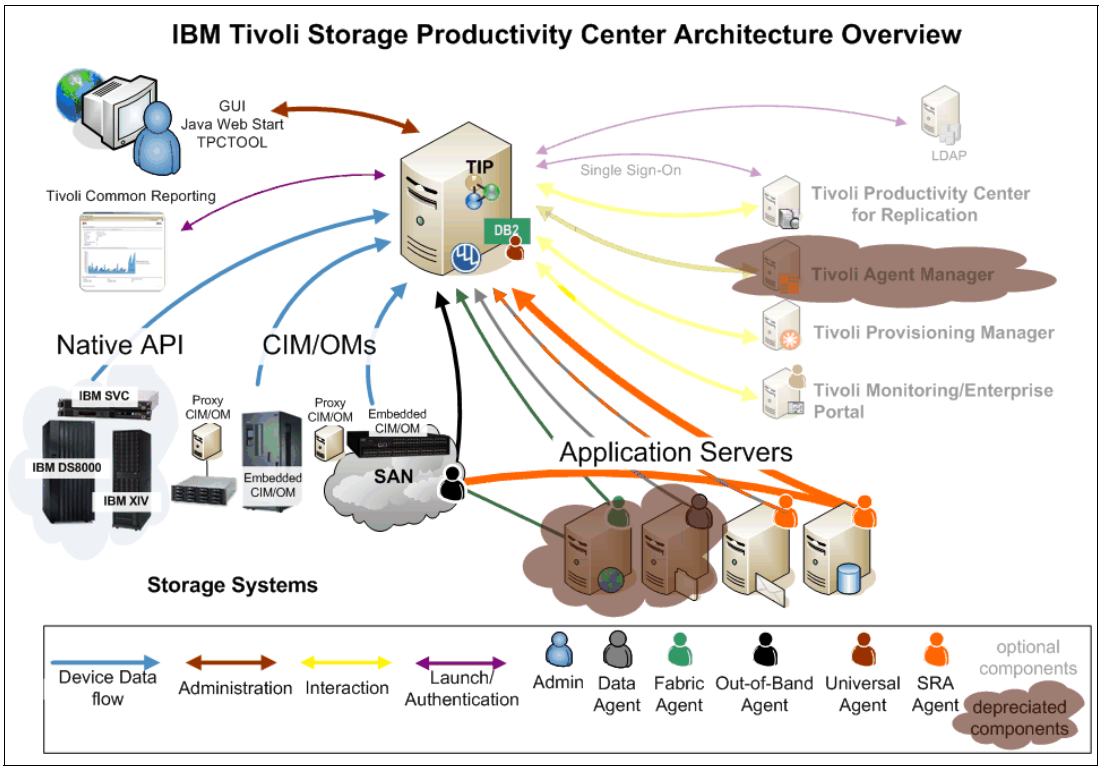


Figure 9-2 IBM Tivoli Storage Productivity Center architecture

The *Data server* is the control point for product scheduling functions, configuration, event information, reporting, and GUI support. It coordinates communication with agents and data collection from agents that scan file systems and databases to gather storage demographics and populate the database with results. Automated actions can be defined to perform file system extension, data deletion, and backup or archiving or event reporting when defined thresholds are encountered. The data server is the primary contact point for GUI user interface functions. It also includes functions that schedule data collection and discovery for the Device server.

The *Device server* component discovers, gathers information from, analyzes performance of, and controls storage subsystems and SAN fabrics. It coordinates communication with agents and data collection from agents that scan SAN fabrics.

IBM Tivoli Storage Productivity Center is integrated with *IBM Tivoli Integrated Portal*. This integration provides functionalities such as single sign-on and the use of Tivoli Common Reporting.

The single DB2 *database instance* serves as the repository for all Tivoli Storage Productivity Center components.

Outside of the server, several interfaces are used to gather information about the environment. The most important sources of information are the Tivoli Storage Productivity Center agents - *Storage Resource agent, Data agent, and Fabric agent,* and SMI-S-enabled storage devices that use a *Common Information Model object manager* (CIMOM) agent (embedded or as a proxy agent). Storage Resource agents, Common Information Model (CIM) agents, and Out-of-Band fabric agents gather host, application, storage system, and SAN fabric information and send that information to the Data server or Device server.

Tip: Data agents and Fabric agents are supported in Tivoli Storage Productivity Center Version 4.2. However, no new functions were added to those agents for this release. For optimal results when using Tivoli Storage Productivity Center, migrate the Data agents and Fabric agents to Storage Resource agents.

Native storage system interfaces are provided in Tivoli Storage Productivity Center Version 4.2 for IBM System Storage DS8000, IBM System Storage SAN Volume Controller, IBM Storwize V7000, and IBM XIV Storage System to improve the management capabilities and performance of data collection. The native interface (also referred to as *native application programming interface* (NAPI)) replace the CIM agent (SMI-S agent) implementation for these storage systems.

The GUI allows you to enter information or receive information for all Tivoli Storage Productivity Center components.

The command line interface (CLI) allows you to issue commands for major Tivoli Storage Productivity Center functions.

IBM Tivoli Storage Productivity Center offerings

Tivoli Storage Productivity center family of products includes:

- Tivoli Storage Productivity Center Basic Edition
- Tivoli Storage Productivity Center for Data
- Tivoli Storage Productivity Center for Disk
- Tivoli Storage Productivity Center for Disk Select
- Tivoli Storage Productivity Center Standard Edition
- Tivoli Storage Productivity Center Select
- Tivoli Storage Productivity Center for Replication Two-site BC
- Tivoli Storage Productivity Center for Replication Three-site BC
- Tivoli Storage Productivity Center for Replication for System z

IBM Tivoli Storage Productivity Center for Disk Select (formerly Productivity Center for Disk Midrange Edition) and Tivoli Storage Productivity Center Select offer equivalent functionality to Productivity Center for Disk and Productivity Center Standard Edition respectively, but are designed to license by the number of storage devices managed, and support Storwize V7000, XIV, DS3000, DS4000, DS5000 as stand-alone devices or when attached to an IBM SAN

Volume Controller. The Select series also supports any storage devices that are attached to Storwize V7000.

Tip: Productivity Center for Disk Select and Productivity Center Select are set apart from the rest of the portfolio because they are:

- Specifically packaged to support System Storage DS3000, DS4000, DS5000, Storwize V7000 and XIV.
- Licensed per storage device, such as disk controllers and their respective expansion units.

Figure 9-3 shows the overview of Tivoli Storage Productivity Center family of products and their key features:

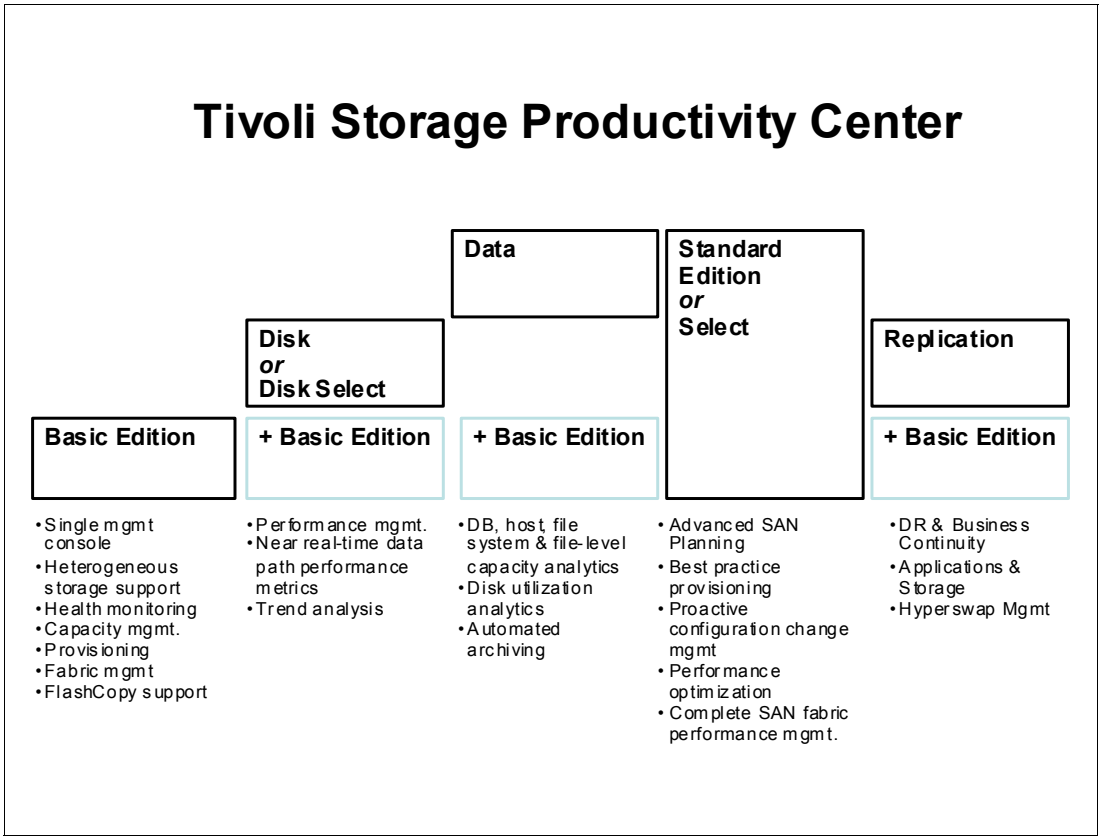


Figure 9-3 Tivoli Storage Productivity Center modular approach

9.1.2 Standards and protocols used in IBM Tivoli Storage Productivity Center

TPC was built upon storage industry standards. In this section, we present an overview of the standards used within the various TPC components.

Common Information Model/Web-Based Enterprise Management

Web-Based Enterprise Management (WBEM) is an initiative of the Distributed Management Task Force (DMTF) with the objective to enable the management of complex IT environments. It defines a set of management and Internet standard technologies to unify the management of complex IT environments.

The main conceptual elements of the WBEM initiative are as follows:

- Common Information Model (CIM): A formal object-oriented modeling language that is used to describe the management aspects of systems
- xmlCIM: The syntax used to describe CIM declarations and messages used by the CIM protocol.
- Hypertext Transfer Protocol (HTTP)
- Hypertext Transfer Protocol over Secure Socket Layer (HTTPS)

HTTP and HTTPS are used as a way to enable communication between a management application and a device that both use CIM.

The CIM Agent provides a means by which a device can be managed by common building blocks rather than proprietary software. If a device is CIM-compliant, software that is also CIM-compliant can manage the device. Using CIM, you can perform tasks in a consistent manner across devices and vendors.

The CIM/WBEM architecture defines the following elements:

- Agent code or CIM Agent:

  An open-systems standard that interprets CIM requests and responses as they transfer between the client application and the device. The Agent is normally embedded into a device, which can be hardware or software. When not embedded (which is the case for devices that are not CIM-ready such as the DS5000 Storage Server), a device provider (usually provided by the device manufacturer) is required.
- CIM Object Manager (CIMOM):

  The common conceptual framework for data management that receives, validates, and authenticates the CIM requests from the client application (such as TPC for disk). It then directs the requests to the appropriate component or a device provider.
- Client application or CIM Client:

  A storage management program, such as Tivoli Storage Productivity Center, that initiates CIM requests to the CIM Agent for the device. A CIM Client can reside anywhere in the network, because it uses HTTP to talk to CIM Object Managers and Agents.
- Device or CIM Managed Object:

  A Managed Object is a hardware or software component that can be managed by a management application by using CIM (for example, a DS5000 Storage Server).
- Device provider:

  A device-specific handler that serves as a plug-in for the CIMOM. That is, the CIMOM uses the handler to interface with the device. There is a device provider for the DS5000.

Tip: The terms *CIM Agent* and *CIMOM* are often used interchangeably. At this time, few devices come with an integrated CIM Agent. Most devices need a external CIMOM for CIM to enable management applications (CIM Clients) to talk to the device.

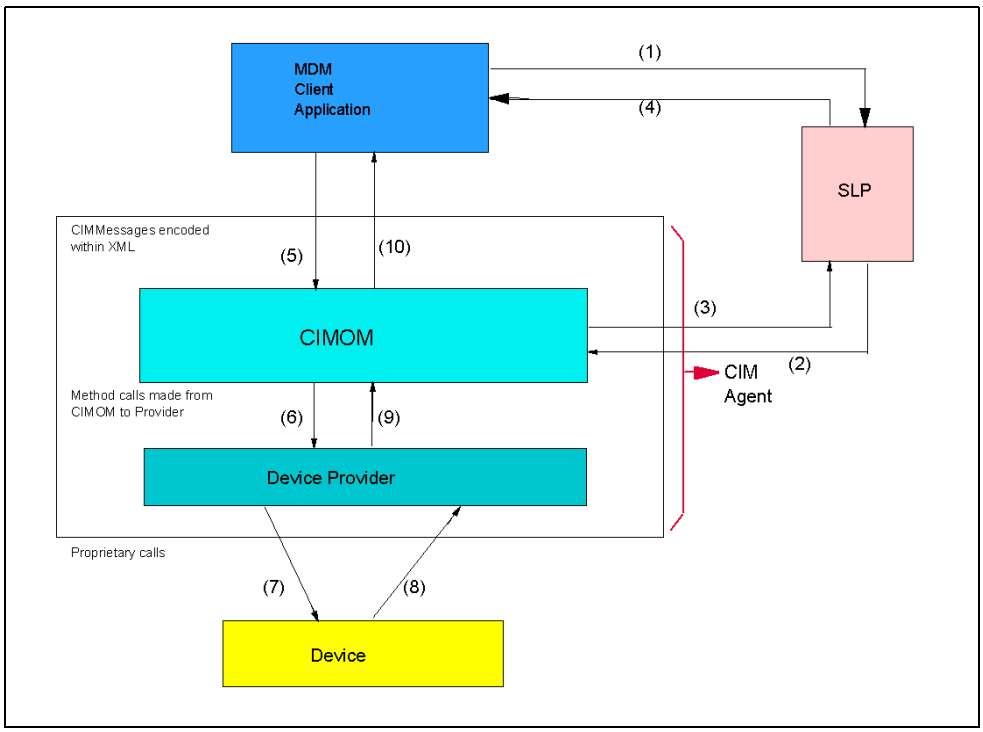


Figure 9-4 CIM architecture elements

For more information, see the following site:

`http://www.dmtf.org/standards/wbem/`

Storage Management Initiative - Specification

The Storage Networking Industry Association (SNIA) has fully adopted and enhanced the CIM for Storage Management in its Storage Management Initiative - Specification (SMI-S). SMI-S was launched in mid-2002 to create and develop a universal open interface for managing storage devices, including storage networks.

The idea behind SMI-S is to standardize the management interfaces so that management applications can use these and provide cross-device management, which means that a newly introduced device can be immediately managed as it conforms to the standards. TPC for disk uses that standard.

Service Location Protocol

The Service Location Protocol (SLP) is an IETF standard that SLP provides as a scalable framework for the discovery and selection of network services.

SLP enables the discovery and selection of generic services, which can range in function from hardware services such as those for printers or fax machines, to software services such as those for file servers, email servers, Web servers, databases, or any other possible services that are accessible through an IP network.

Traditionally, to use a particular service, an end-user or client application needs to supply the host name or network IP address of that service. With SLP, however, the user or client no longer needs to know individual host names or IP addresses (for the most part). Instead, the user or client can search the network for the desired service type and an optional set of qualifying attributes.

The SLP architecture includes three major components:

- Service agent (SA):

  A process working on the behalf of one or more network services to broadcast the services.

- User agent (UA):

  A process working on the behalf of the user to establish contact with a network service. The UA retrieves network service information from the service agents or directory agents.

- Directory agent (DA):

  A process that collects network service broadcasts.

The SA and UA are required components in an SLP environment, where the SLP DA is optional.

The SMI-S specification introduces SLP as the method for the management applications (the CIM clients) to locate managed objects. In SLP, an SA is used to report to UAs that a service that has been registered with the SA is available.

Native API

Tivoli Storage Productivity Center Version 4.2 provides a new access method to gather information from devices. This method is called the Native API (NAPI) and is at this time available for only a limited number of disk storage subsystems.

The Native API does not replace CIM, SNMP or the in-band fabric interfaces. Although it is an addition to the ways Tivoli Storage Productivity Center can get information, you cannot decide which interface you want to use, because the support of NAPI is currently available for the following items:

- IBM DS8000
- IBM SAN Volume Controller
- IBM Storwize V7000
- IBM XIV

9.1.3 IBM Tivoli Storage Productivity Center publications

The publications listed in this section are considered particularly suitable for a more detailed description of the topics covered in this chapter:

- *IBM Tivoli Storage Productivity Center V4.2 Release Guide*, SG24-7894
- *IBM System Storage Productivity Center Deployment Guide*, SG24-7560
- *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364

Online resources: More information can be found at these websites:

- IBM Storage Software support website:

  `http://www.ibm.com/servers/storage/support/software/`

- Tivoli Storage Productivity Center:

  `http://publib.boulder.ibm.com/infocenter/tivihelp/v4r1/index.jsp`

9.2 Managing DS5000 using IBM TPC for Disk

Tivoli Storage Productivity Center for Disk is designed to provide storage device configuration and management from a single console. It includes performance capabilities to help monitor and manage performance, and measure service levels by storing received performance statistics into database tables for later use. Policy-based automation enables event action based on business policies. It sets performance thresholds for the devices based on selected performance metrics, generating alerts when those thresholds are exceeded. Tivoli Storage Productivity Center for Disk helps simplify the complexity of managing multiple SAN-attached storage devices.

TPC for Disk also includes performance monitoring and reporting capabilities such as:

- Provides reporting across multiple arrays from a single console.
- Helps monitor metrics such as throughput, input and output (I/O), data rates, and cache utilization.
- Receives timely alerts that can enable event action based on your policies when thresholds are exceeded.
- Offers continuous and proactive performance analysis with comprehensive real-time monitoring and fault identification to help improve SAN availability.
- Helps you improve storage return on investment by helping to keep SANs reliably and dependably operational.
- Helps reduce storage administration costs by simplifying the management of complex SANs.
- Supports the performance reporting capabilities on the IBM System Storage SAN Volume Controller (SVC) with attached DS3000, DS4000 and/or DS5000 devices.
- Supports performance reporting capabilities for any storage virtualized by the IBM Storwize V7000.

The performance function starts with the data collection task, responsible for capturing performance statistics for the devices and storing the data in the TPC database.

Thresholds can be set for certain performance metrics depending on the type of device. Threshold checking is performed during data collection. When performance is outside the specified boundaries, alerts can be generated.

After performance data has been collected, you can configure Tivoli Storage Productivity Center for Disk to present graphical or text reports on the historical performance behavior of specified devices.

For DS5000 Storage Servers, you can use TPC for Disk to perform storage provisioning, logical drive (LUN) creation and assignment, performance management, and reporting. TPC for Disk can monitor the disk subsystem ports, arrays, and measure logical drives throughput, IOPS, and cache rates.

Device discovery is performed by the SLP, and configuration of the discovered devices is possible in conjunction with CIM agents associated with those devices, using the standard mechanisms defined in SMI-S.

For devices that are not CIM ready, as the DS5000, the installation of a proxy application (CIM Agent) is required.

To monitor and manage a DS5000 through TPC, perform the following task sequence:

1. Install the CIM agent for DS5000.
2. Register CIMOM in TPC.
3. Probe discovered DS5000 to gather device information.

4. Create Performance Monitor job.

9.2.1 Installing the CIM agent for DS5000

In this section, we describe the implementation of the CIM agent for a DS5000, assuming that a TPC environment is already in place (we used the latest TPC Version 4.2.2.78 for our samples throughout this chapter).

The DS5000 is not a CIM-ready device. Therefore, a DS5000 device provider (acting as the CIM Agent) must be installed on a host system to bridge communications between the DS5000 and the TPC server. The device provider (agent) for the DS5000 is available by IBM product support and it is called the SMI-S Provider. The open source CIMOM OpenPegasus (CIM) is installed as part of the installation procedure.

Tip: It is considered mandatory to run the CIM Agent software on a separate host from the Tivoli Storage Productivity Center server. Attempting to run a full Tivoli Storage Productivity Center implementation on the same host as the CIM agent results in dramatically increased wait times for data retrieval. You might also experience resource contention and port conflicts.

Installation procedure

To install the SMI-S Provider on supported Windows platform, follow these steps:

1. Follow the instructions on IBM website to download the SMI-S Provider installation code:

   `http://www-01.ibm.com/support/docview.wss?rs=40&q1=support+matrix&uid=swg21386446`

   Choose the version intended for the use with IBM Tivoli Storage Productivity Center and download the correct operating system of the server on which the SMI-S Provider is going to be installed. At the time of writing, the latest SMI-S Provider version for DS5000 Storage Server is 10.19.GG.21, which is suggested specifically for use with IBM TPC 4.2.1 (fixpack 5) and later.

   Download and read the `readme` file as well to make sure that your current storage subsystem firmware level is supported.

2. Extract the downloaded file and launch the installer to start the installation process. In our example, the installer file name is `LSI_ARRAY-WS32-10.19.6021.exe`. Launching the executable starts the Installshield Wizard and opens the Welcome window, as shown in Figure 9-5 on page 371. Click **Next** to continue.

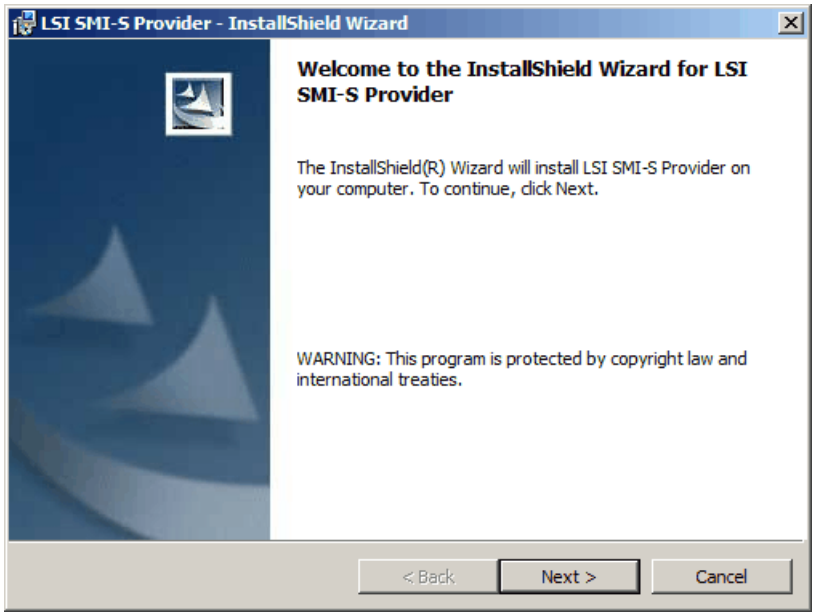


Figure 9-5 Welcome window of DS5000 SMI-S Provider InstallShield Wizard

The License Agreement window opens.

3. If you agree with the terms of the license agreement, click **Yes** to accept the terms and continue the installation.Change the destination folder or keep the default, and click **Next** (Figure 9-6).

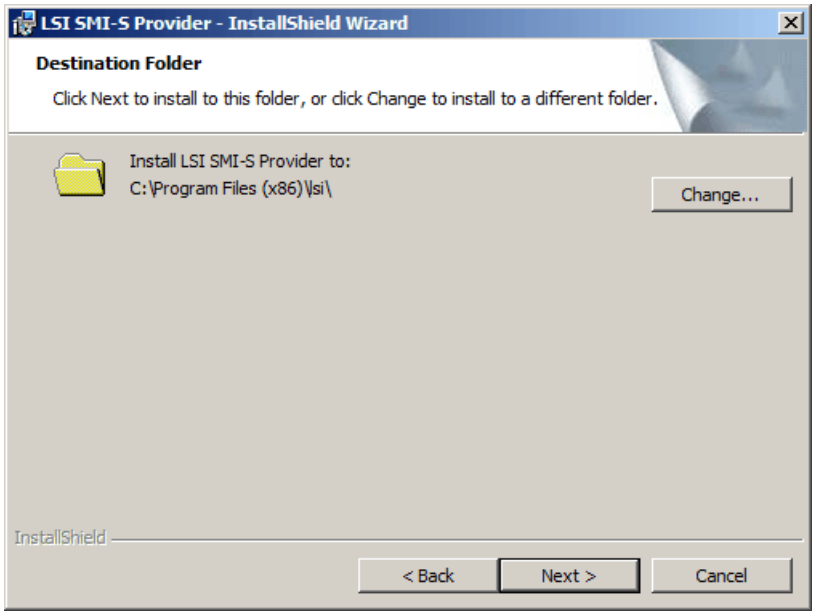


Figure 9-6 Select the Destination Folder

4. Confirm the installation and click **Install**.
5. The installation opens a text file in Notepad. Follow the instructions in the Notepad file to add the storage subsystems IP addresses. (Figure 9-7 on page 372).

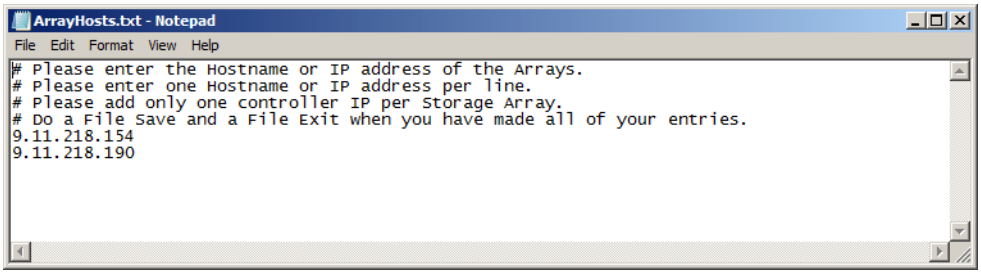


Figure 9-7 SMI-S Provider configuration file

Caution: Do not enter the controller management IP addresses of a DS5000 Storage Server in multiple DS5000 SMI-S Provider configuration files within the same subnet. Doing it can cause unpredictable results on the Tivoli Storage Productivity Center for Disk server and can cause a loss of communication with the DS5000 devices.

6. After you have added the storage arrays and closed the Notepad, the installation program continues normally. Follow the instructions and prompts on the screen. The installation prompts you to choose the CIMOM authentication type (Figure 9-8). The default is Disable Authentication, which allows unrestricted access to the CIMOM server. To restrict access to the CIMOM server to authorized users only, select Enable Authentication, then follow the instructions and prompts to enter the username and password of the authorized users of the CIMOM server.

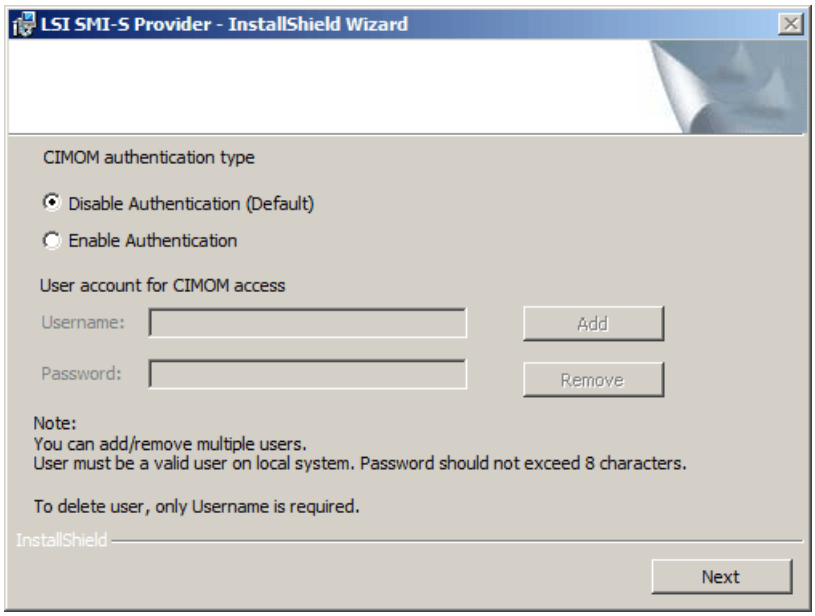


Figure 9-8 Select CIMOM authentication type

You can always add a CIMOM user authentication later by creating a CIMOM server user through command-line utility **`cimuser.`** See Example 9-1:

Example 9-1 Registering a new CIMOM user

```
C:\Program Files (x86)\lsi>cimuser -a -u cimuser1
Please enter your password: ********
Please re-enter your password: ********
User added successfully..
```

```
C:\Program Files (x86)\lsi>cimuser -l
cimuser1
```

7. Click **Next** to finish the installation of SMI-S Provider. When the installation of the SMI-S Provider is complete, the **Pegasus CIM Object Manager** Service is started. Click **Finish** on the InstallShield Wizard Complete window (Figure 9-9).

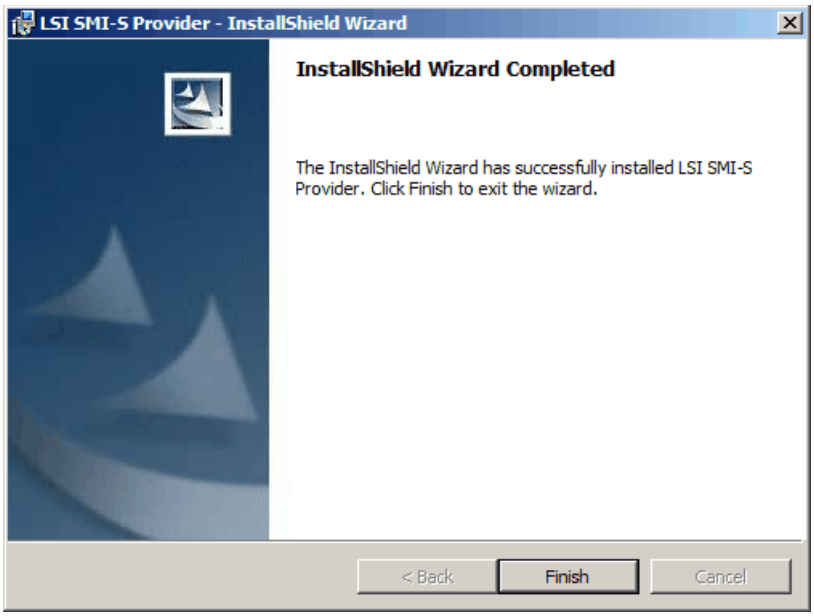


Figure 9-9 InstallShield Wizard Completed window

During the start of the service, the SMI-S Provider processes all of the entries in the `ArrayHosts.txt` file. The configuration is stored in a file named:

```
C:\Program Files (x86)\lsi\pegasus\provider\array\ArrayHosts.txt
```

Every time you change anything with the registered DS5000 controllers and restart the CIMOM server, and after you make a new discovery, the `AarrayHosts.txt` is updated with a new time stamp.

Verifying the CIMOM Service

To verify that the service has started, open the Windows Services window (**Start** → **All Programs** → **Administrative Tools** → **Services**) and checking the status of the CIMOM service, as shown in Figure 9-10 on page 374.

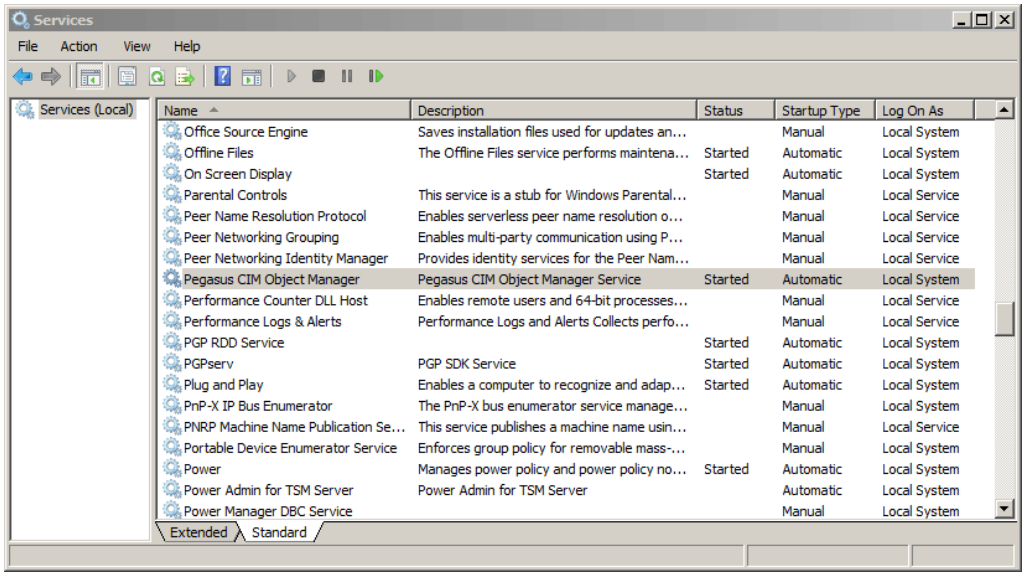


Figure 9-10 Verify the Pegasus CIM Object Manager Service

Use this interface to start/stop/restart the service. Note that you need to restart the service after any modification to the `ArrayHost.txt` file.

9.2.2 Registering the DS5000 SMI-S Provider in TPC

There are more than one ways to register the CIMOM agent with the TPC server. In previous versions of IBM Tivoli Storage Productivity Center, setting up a device for monitoring required performing tasks in different locations of the user interface, and each type of device required different steps. In IBM Tivoli Storage Productivity Center Version 4.2 and later, all the task for setting up and configuring a device are consolidated in the **Configure Devices** wizard.

Suggestion: In most environments, we find that using CIMOM discovery does not have a large advantage, simply because most CIMOMs have security turned on, which means Tivoli Storage Productivity Center is unable to get a list of devices from the CIMOM. To disable SLP-DA and automatic discovery, do not add any IP of SLP agents and uncheck the box for local subnet scanning as shown in Figure 9-11 on page 375.

In this section, we focus on using the Configure Devices wizard to show how to register DS5000 SMI-S Provider with TPC server. For other methods of adding additional data sources in TPC, see the *IBM Tivoli Storage Productivity Center V4.2 Release Guide*, SG24-7894.

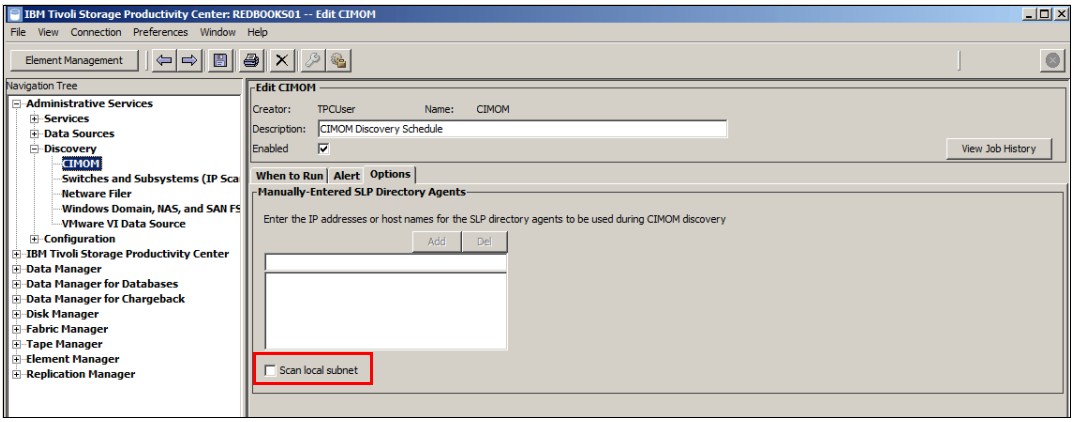


Figure 9-11 Disable automatic CIMOM discovery

Registration procedure

Follow these steps to complete the registration procedure:

1. To register the DS5000 CIMOM and add DS5000 Storage Servers as data sources in IBM Tivoli Storage Productivity Center, open the TPC console GUI, and click **Configure Devices** icon in the toolbar or select the Configure Devices node in the navigation tree (Figure 9-12).

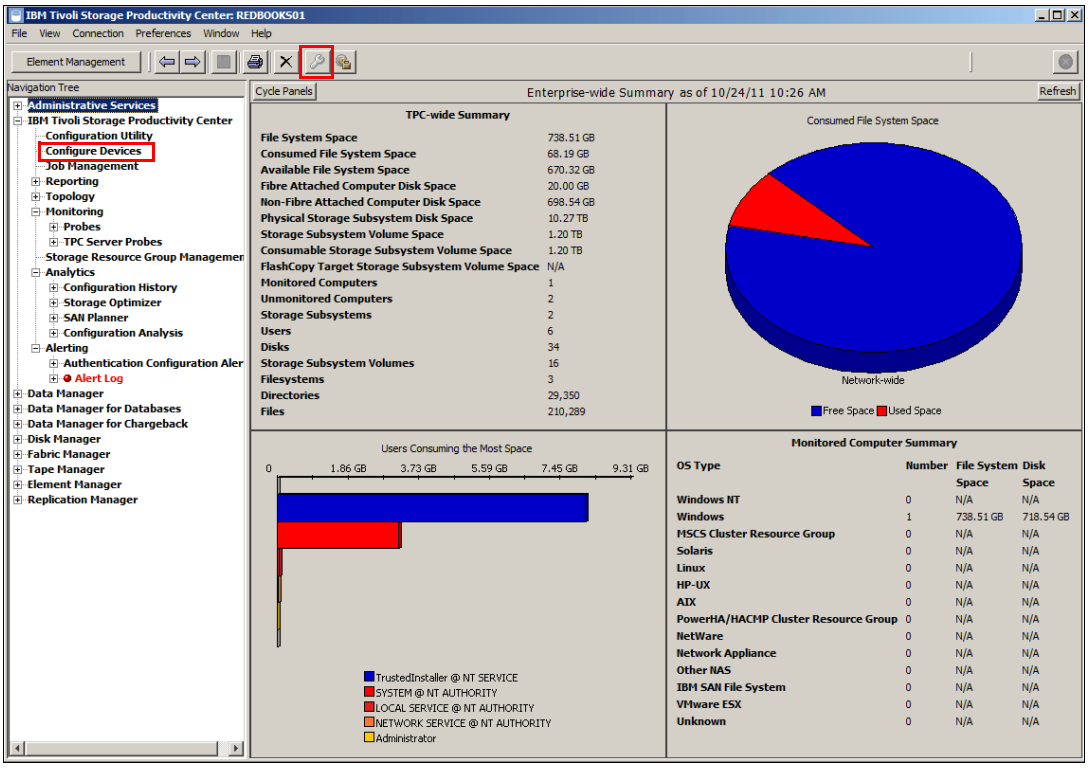


Figure 9-12 Launching the Configure Devices Wizard

2. Select the type of device to be added, in this case, **Storage Subsystem** (Figure 9-13 on page 376) and click **Next**.

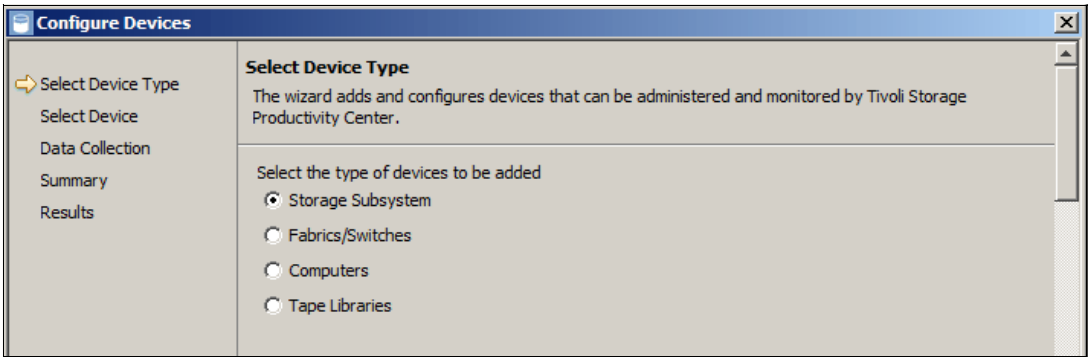


Figure 9-13 Select Device Type

3. Select **Add and configure new storage subsystems** (Figure 9-14) and click **Next**.

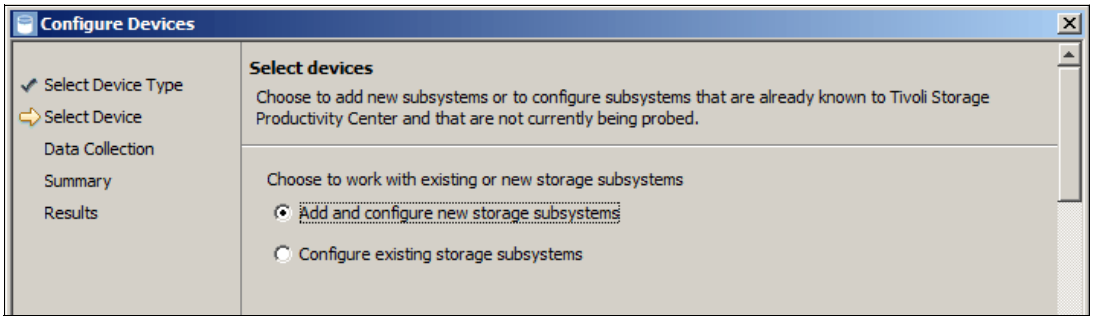


Figure 9-14 Add new storage subsystems

4. On the next page (Figure 9-15 on page 377), configuration panel of new storage subsystem is displayed. For DS5000, you need to choose **Other** as Device Type, as it does not use NAPI. TPC uses this information to discover the storage subsystems that are managed by CIMOM. You can then configure those storage subsystems as part of the wizard session. Following configuration fields need to be filled out:

- Host: IP address or fully qualified domain name of the server where you installed the DS5000 SMI-S Provider (*not* the IP address of your DS5000 Storage Server).
- Protocol: HTTP (The SMI-S Provider does not provide secure communication at the time of writing.)
- Username: any username (The SMI-S Provider does not require any specific user ID and password to authenticate.)
- Password: any password (not null)
- Port: 5988
- Interoperability Namespace: /interop
- Display name: any name to identify the CIM agent
- Description: optional

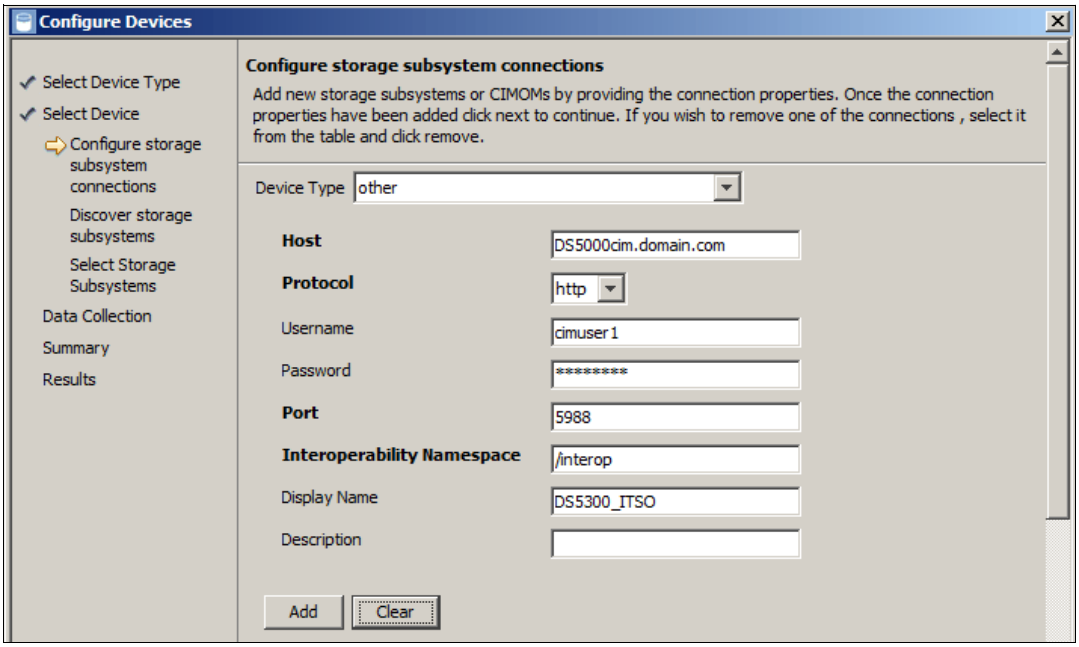


Figure 9-15 Configure storage subsystem connections

After you provide all the information, click the **Add** button. This initiates a test connection to the CIMOM and if successful, it adds the CIMOM into the device list. You can repeat this procedure to add additional CIMOMs if needed.

Tip: If the connection test failed, you must verify that your DS5000 SMI-S Provider is properly installed on the server and that the service is running. Also, check for any connectivity or possible firewall problems.

5. After you add all data sources successfully to the configured devices list, clicking **Next** on this page begins the discovery process of the storage subsystem managed by CIMOM. TPC collects initial configuration information for these entities. Figure 9-16 shows the discovery process log.

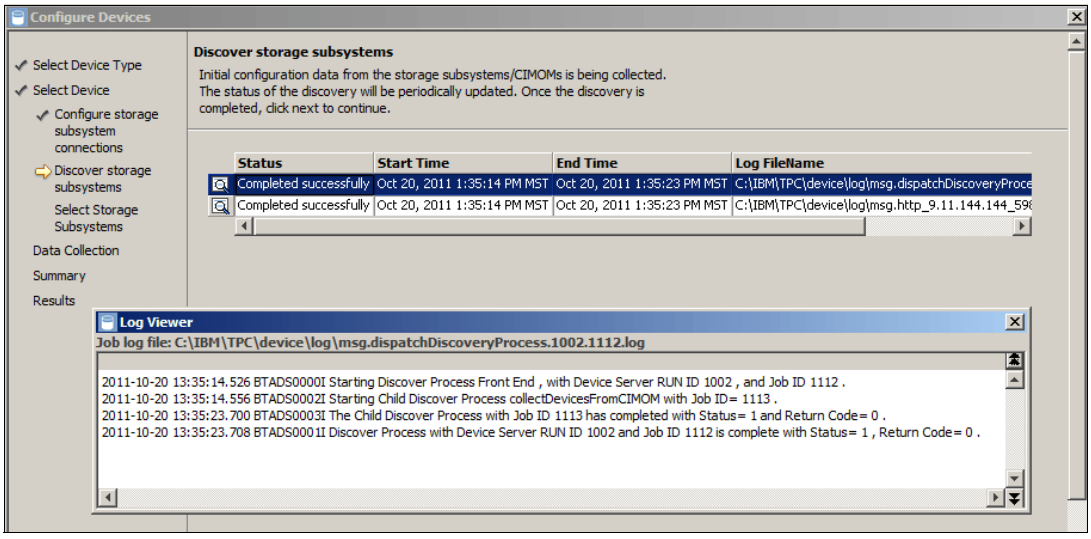


Figure 9-16 Discover storage subsystems

Click **Next** to continue to the next page.

6. Select the storage subsystems that you want to configure. Any storage subsystems that were added in previous steps are automatically selected (Figure 9-17). Deselect storage subsystems that you do not want to configure and proceed to **Next** page.

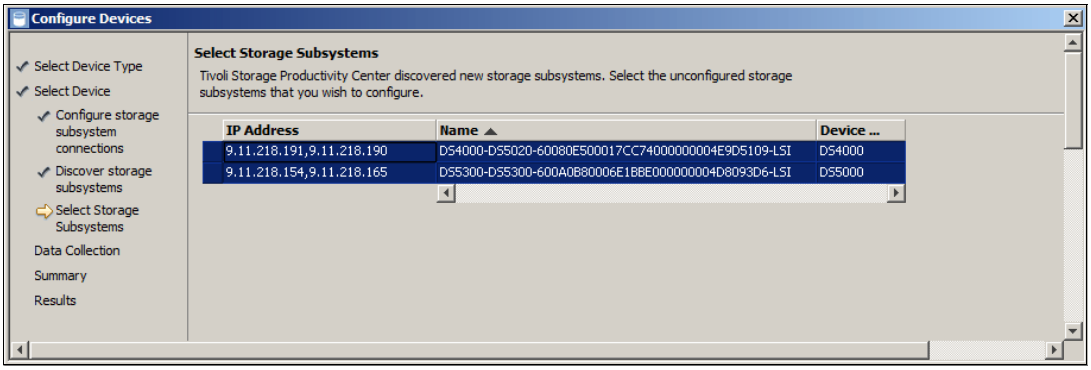


Figure 9-17 Select storage subsystems to configure with TPC

7. Here you can directly configure the storage subsystem monitoring (Probes) by assigning them to monitoring groups or monitoring templates. Use monitoring groups and templates to determine when data is collected and what alert conditions are checked for the storage systems that you are configuring. The scheduling details are displayed for better overview (Figure 9-18):

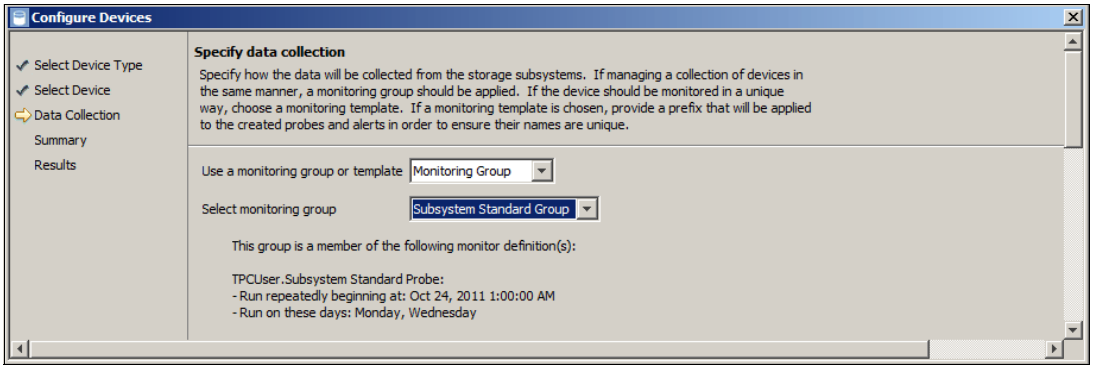


Figure 9-18 Specify data collection

The storage subsystems are automatically included in a probe schedule to which the group is associated. For example, if you select Subsystem Standard Group, the storage systems are included in the Subsystem Standard Probe schedule.Click **Next**.

8. Based on your configuration choices on previous page, the summary of configuration changes is displayed. The Review user selection page includes following information:
   - A list of devices that you are configuring.
   - The name of the monitoring group that you selected. This value is not displayed if you selected a template.
   - The name of the probe schedule that is created based on the template you selected. This value is not displayed if you selected a monitoring group.
   - Information about the probe schedule created for a template.
   - The names of the alerts that are created based on the template you selected. This value is not displayed if you selected a monitoring group.

   Confirm these selections and click **Next**. On the next page (Figure 9-19 on page 379), click **Finish** to close the wizard and probe the devices that you have configured.

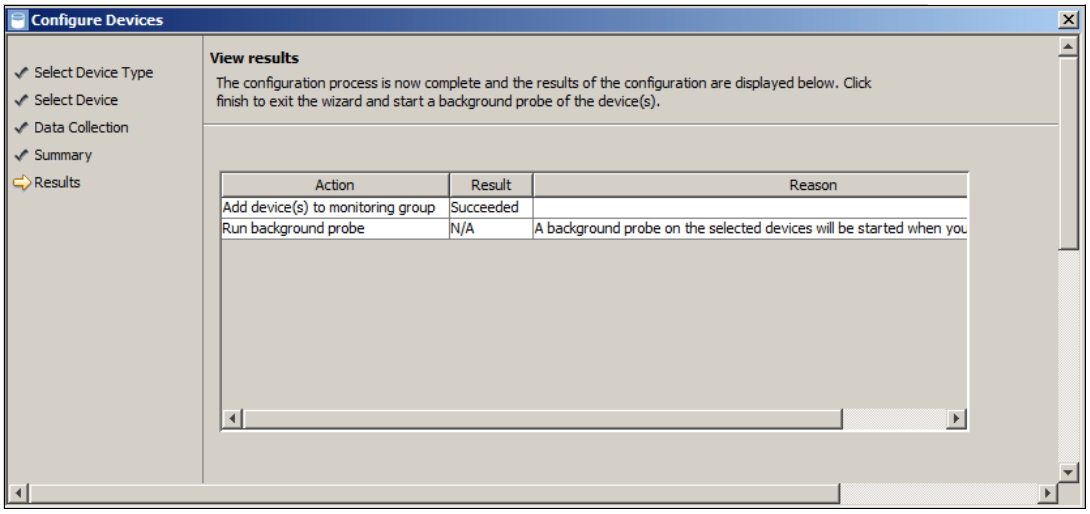


Figure 9-19 View results and launch the probe

This probe is a special job. It is not the job that is defined for your device, because that could also run the probe for other devices defined in the same job. If that probe was running, starting it again results in errors, because you can not run two probes for a single device from a single server at the same time. This special job is the *CLI and Event Driven Job*.

9. After the probe is submitted, a dialog box opens asking you to view the job history. If you click **Yes**, it takes you to Job Management panel of Tivoli Storage Productivity Center where you can check the job status (Figure 9-20):

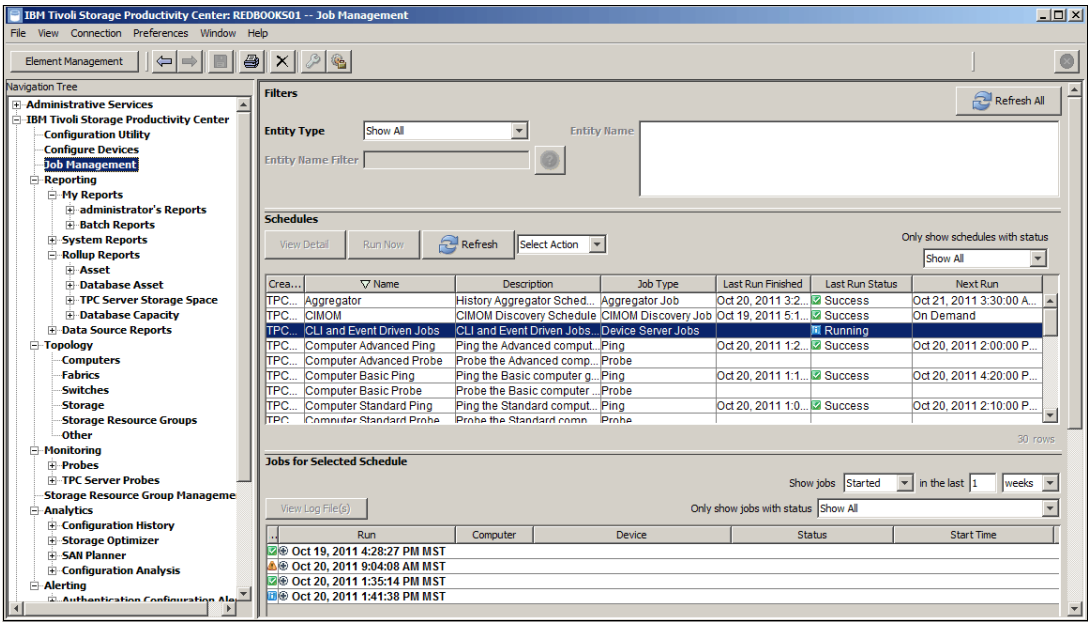


Figure 9-20 CLI and Event Driven Job Probe

10. You can verify that the new monitoring probe was created for DS5000 storage subsystem by going to Navigation tree **IBM Tivoli Storage Productivity Center** → **Monitoring** → **Probes** and selecting the predefined probe that you assigned in step 7 on page 378. In **Current Selections** window on the right side you can see the storage subsystems that

are associated with the probe (Figure 9-21). You can click **View Job History** icon on the right side for more details of the probe job.

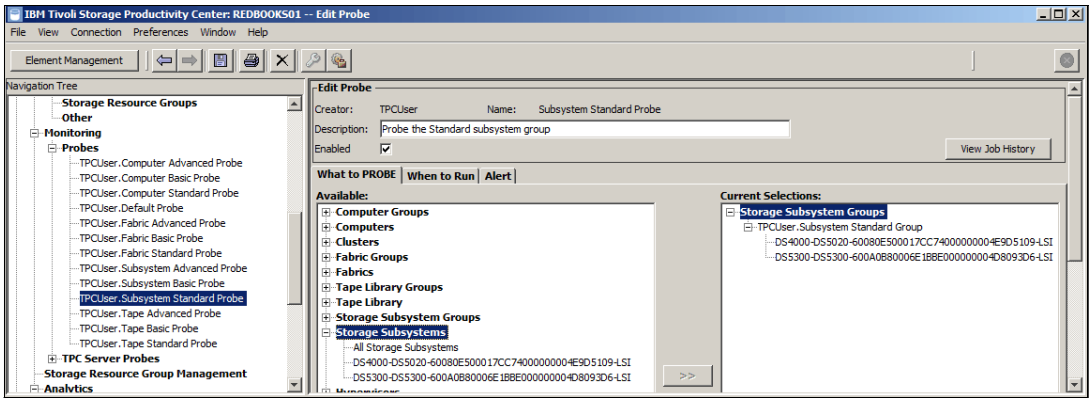


Figure 9-21 Verify the probe for DS5000 storage subsystem

Showing Managed Devices

After the CIM agent was successfully registered, you can view the devices managed by the CIM agent (the DS5000 SMI-S Provider in our case). From the Navigation tree, go to **Administrative Services** → **Data Sources** → **CIMOM Agents,** highlight the CIMOM agent and click **Show Managed Devices**, as shown in Figure 9-22. Note that two DS5000 series Storage Servers have been found, such as a DS5020 and a DS5300 Storage Server.

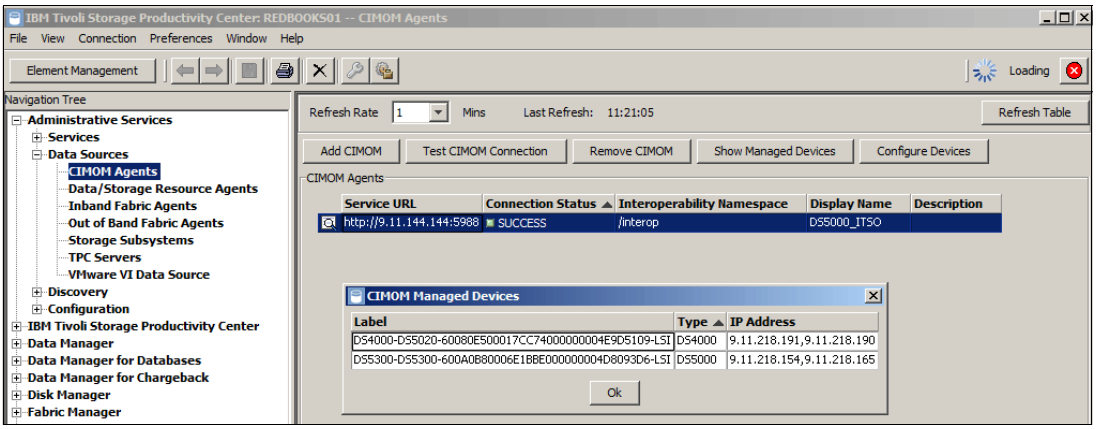


Figure 9-22 Show Managed Devices

9.2.3 Probing the CIM agent

The CIMOM discovery and registration process only detects and recognize devices bridged by the CIMOM agent. TPC needs to get more detailed information about a device to effectively manage and monitor it, which can be accomplished by running a probe job. The probe interacts with the CIM agent by sending a request for information about the configured device. The probe job must in fact be performed before you attempt to manage the device. With the DS5000, a probe job needs to be run before volume (logical drive) management and performance monitoring can be performed on the storage subsystem.

Although we have defined a default monitoring probe in previous section, in real production environment with many storage subsystems configured with IBM Tivoli Storage Productivity Center for Disk it is not ideal to have all the systems use the default TPC settings and scheduling. It is strongly advised to customize data collections in such way, that they do not interfere with each other. It is desired not only to balance the TPC server load, which can get

very intensive as the number of data sources raise, but also because it gives you better control and flexibility in managing variety of jobs inside TPC environment. You can simply disable the default probe by unchecking the **Enabled** box as shown in Figure 9-23.

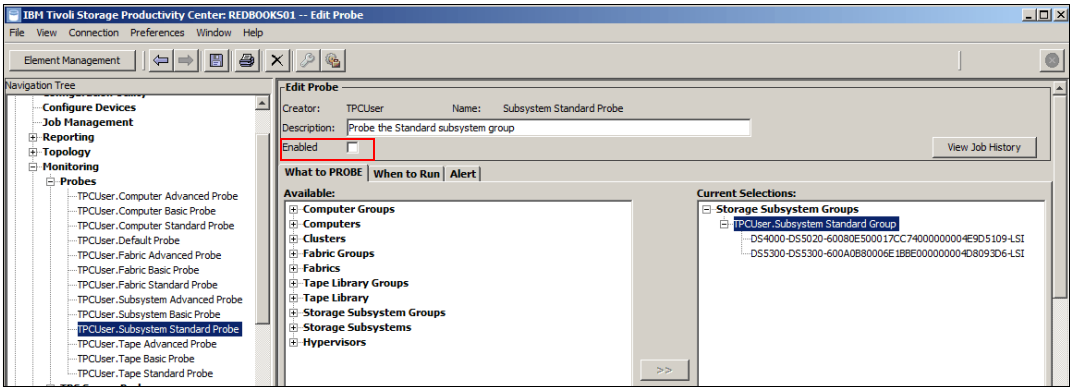


Figure 9-23 Disable the default probe

Configuring a probe job

To configure a probe job, open the TPC console, and expand **IBM Tivoli Storage Productivity Center** → **Monitoring** → **Probes**. Right-click **Probes** and click **Create Probe**, as shown in Figure 9-24.

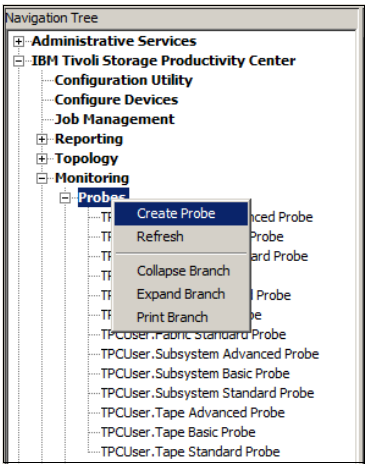


Figure 9-24 Create probe

The Create Probe tab is displayed in the right pane. Expand **Storage Subsystems**, select the DS5000 Storage Server you want to probe, and click the **>>** button, as shown in Figure 9-25 on page 382.

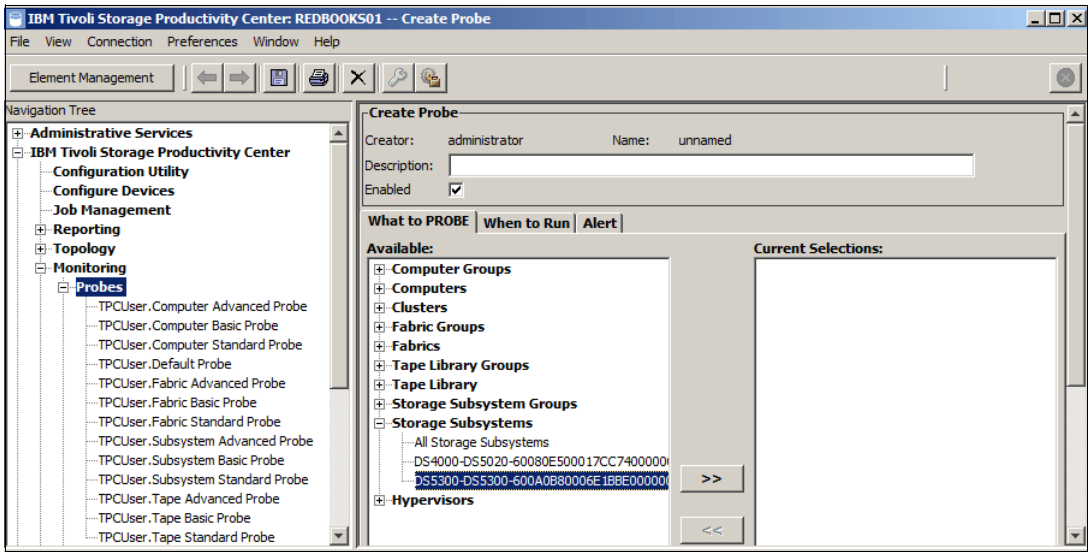


Figure 9-25 Select storage subsystem to probe

Click the **When to Run** tab, as indicated in Figure 9-26.

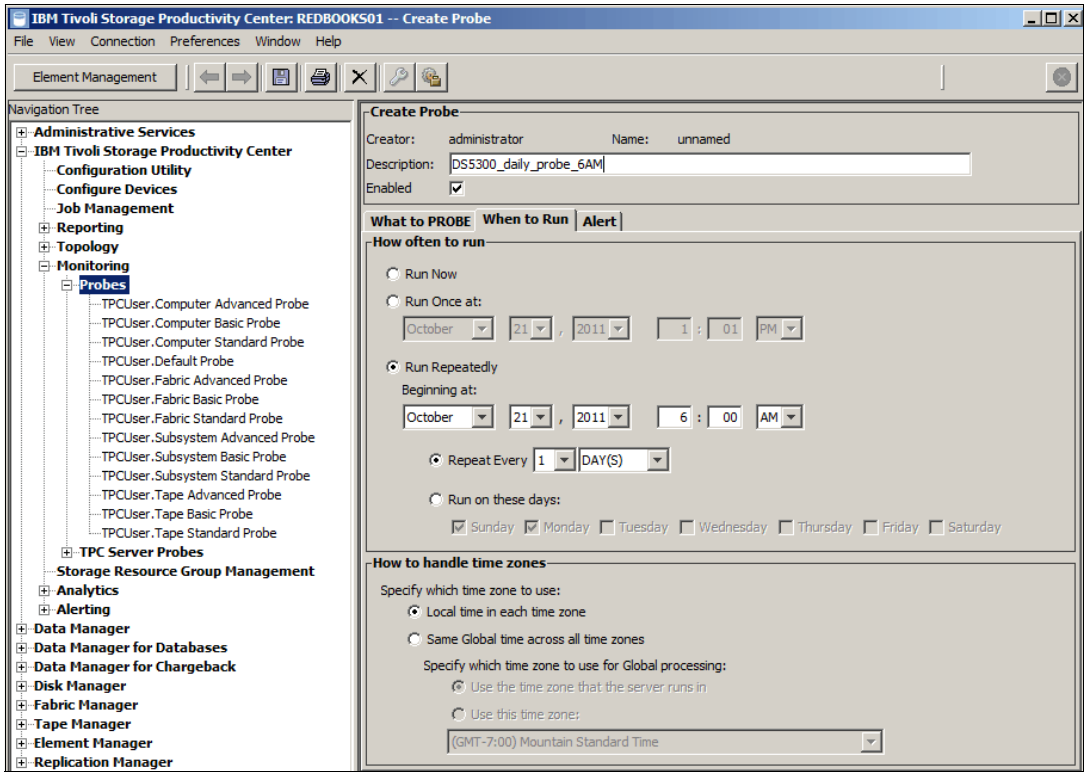


Figure 9-26 When to run the probe

On the **When to Run** tab, you can customize how often you want the probe job to run. You can also specify how to handle time zones, which is especially useful when the storage subsystems to be probed are located in various time zones. Next click the **Alert** tab, as shown in Figure 9-27 on page 383.

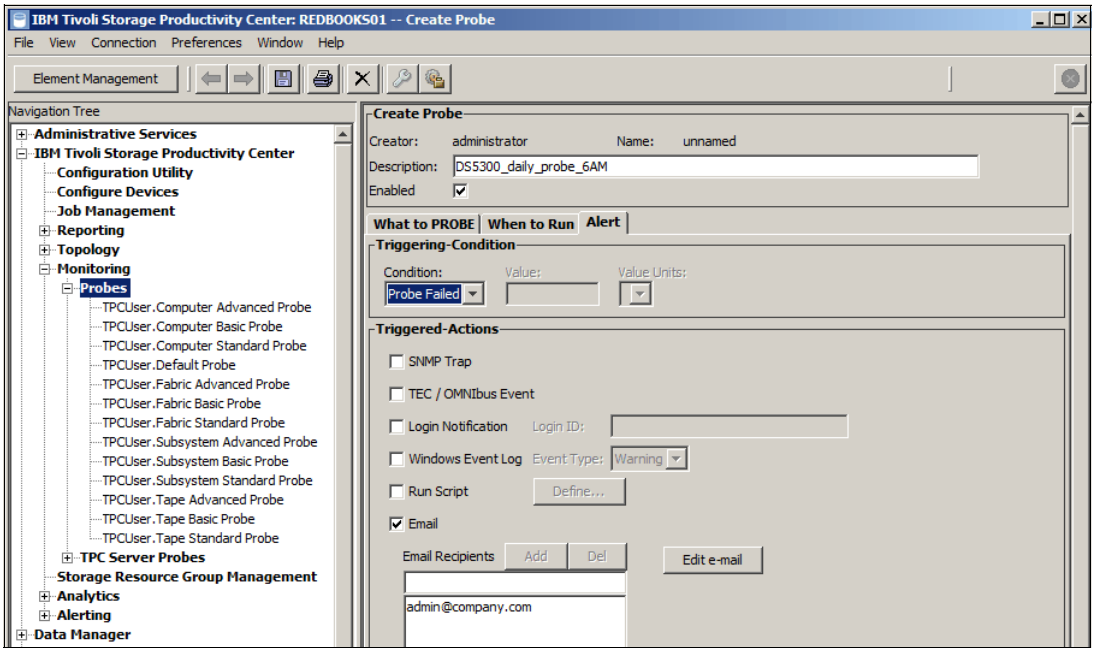


Figure 9-27 Alert when the probe failed

On the Alert tab, you can configure what actions must be automatically taken when the probe fails. There are several options available, including logging the error messages to the Windows Event Log or sending email to specific recipients.

To save the probe job, click **File** → **Save** or use the **Save** icon on the taskbar. You are prompted for a probe job name, as shown in Figure 9-28. Type the probe name of your choice and click **OK** to save and submit the job. The job is activated according to the schedule you have configured earlier.

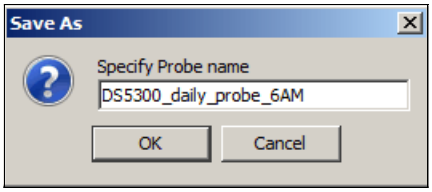


Figure 9-28 Save and name the probe

To check the probe job status, expand **IBM Tivoli Storage Productivity Center** → **Job Management** or click directly on **View Job History** dialog box that pops up after the probe is saved. In this example, the probe job name is *DS5300_daily_probe_6AM*, as shown in Figure 9-29 on page 384.

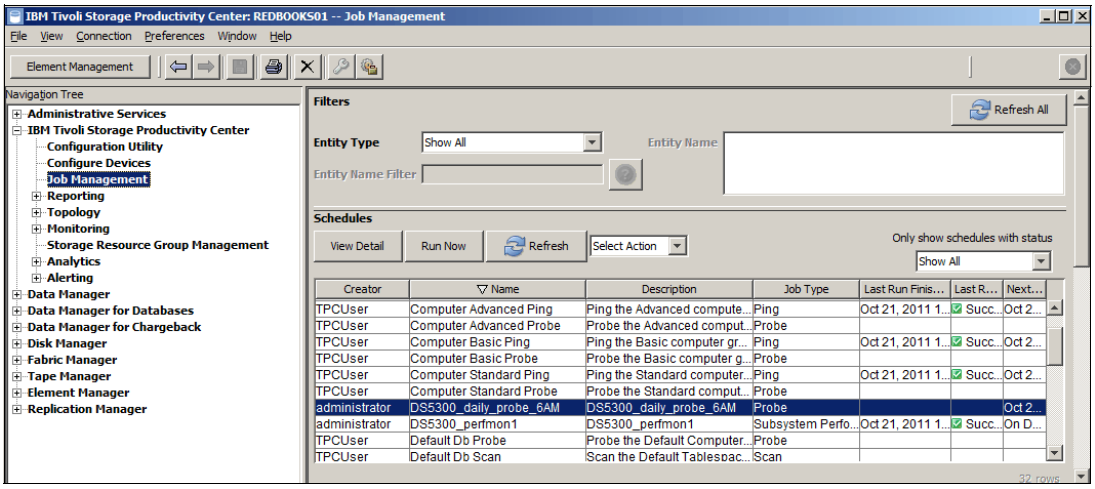


Figure 9-29 Check probe status

When new frames or disks (capacity) are added to an existing array that is already in TPC, you need to make sure the array is probed to reflect the totals in TPC. Make sure the probe job completes. If a regularly scheduled probe job is about to run you can opt to wait on that job to collect the updated probe information.

Tip: When a new Storage Server is discovered by TPC, you need to add the device to a probe job. Although TPC has a default job that probes all storage arrays, we advise you to probe each subsystem individually and disable the default probe job. One guideline is never to have two probe jobs running at the same time against the same CIM or storage subsystem. If possible, try to have one probe job complete before starting another, however, depending on the size of the environment, this practice might not be possible. In this case, try to minimize the time that the job runs overlap.

Viewing storage subsystem information

After the probe job created for the storage subsystem has successfully completed, you can view more detailed information about your storage subsystem and start using TPC to manage the system.

To view detailed information about the storage subsystem, expand **Data Manager** → **Reporting** → **Asset** → **By Storage Subsystem**, as shown in Figure 9-30.

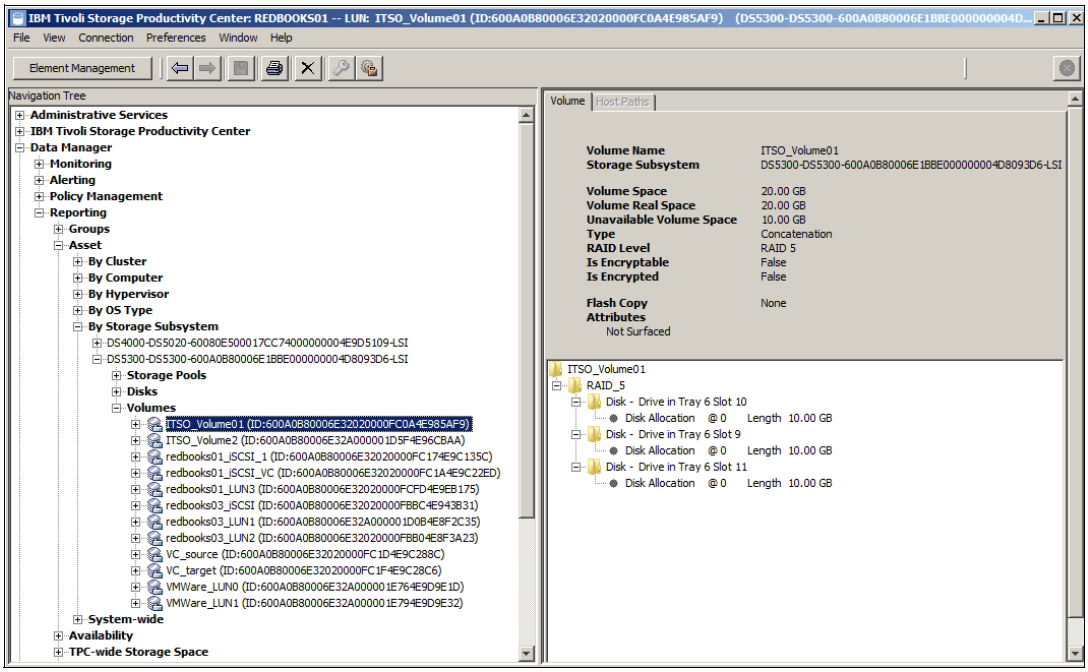


Figure 9-30 Detailed information of storage subsystem

9.2.4 Creating a Performance Monitor job

Before you can monitor storage subsystem performance, you need to create a Performance Monitor job. The job collects performance data for the storage subsystem.

To create a Performance Monitor job for a storage subsystem, follow these steps:

1. Expand **Disk Manager** → **Monitoring**, then right-click **Subsystem Performance Monitors** and click **Create Subsystem Performance Monitor**. In the Available subsystems tab in the right pane, select the storage subsystem to be monitored and move it to the Selected Subsystem tab, as shown in Figure 9-31.

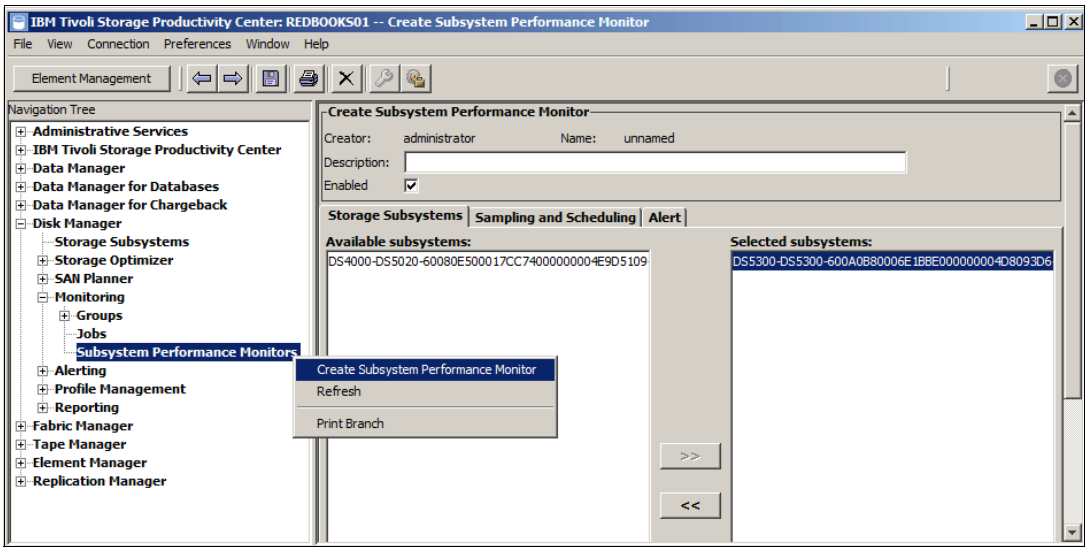


Figure 9-31 Create Performance Monitor job

2. Select the **Sampling and Scheduling** tab to configure the interval length, duration, and scheduling of the Performance Monitor job to be submitted, as shown in Figure 9-32.

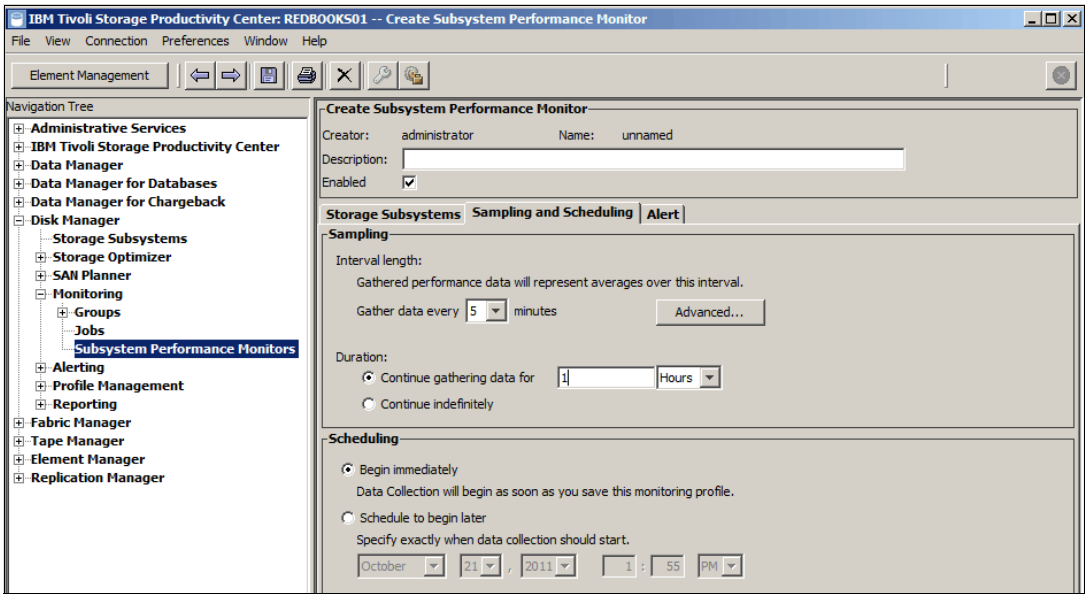


Figure 9-32 Sampling and scheduling

3. Go to the **Alert** tab to select what actions must be triggered if the monitor fails, as shown in Figure 9-33.

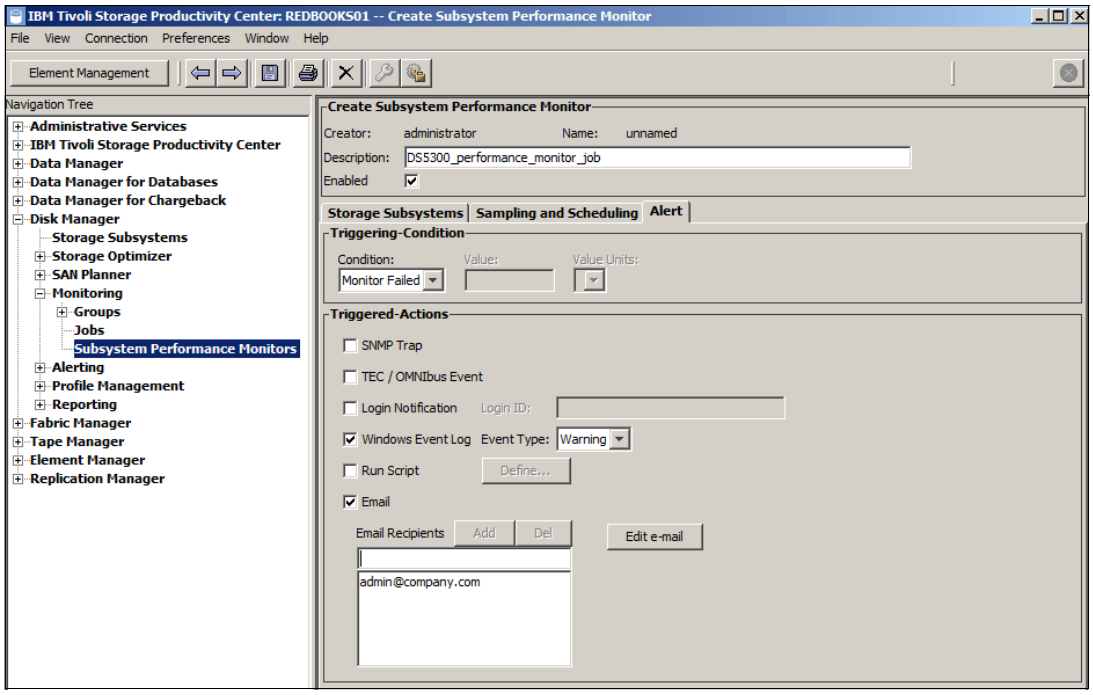


Figure 9-33 Alert tab

4. Select **File** → **Save** or use the **Save** icon on the taskbar to save the job and specify the job name when prompted, as shown in Figure 9-34.

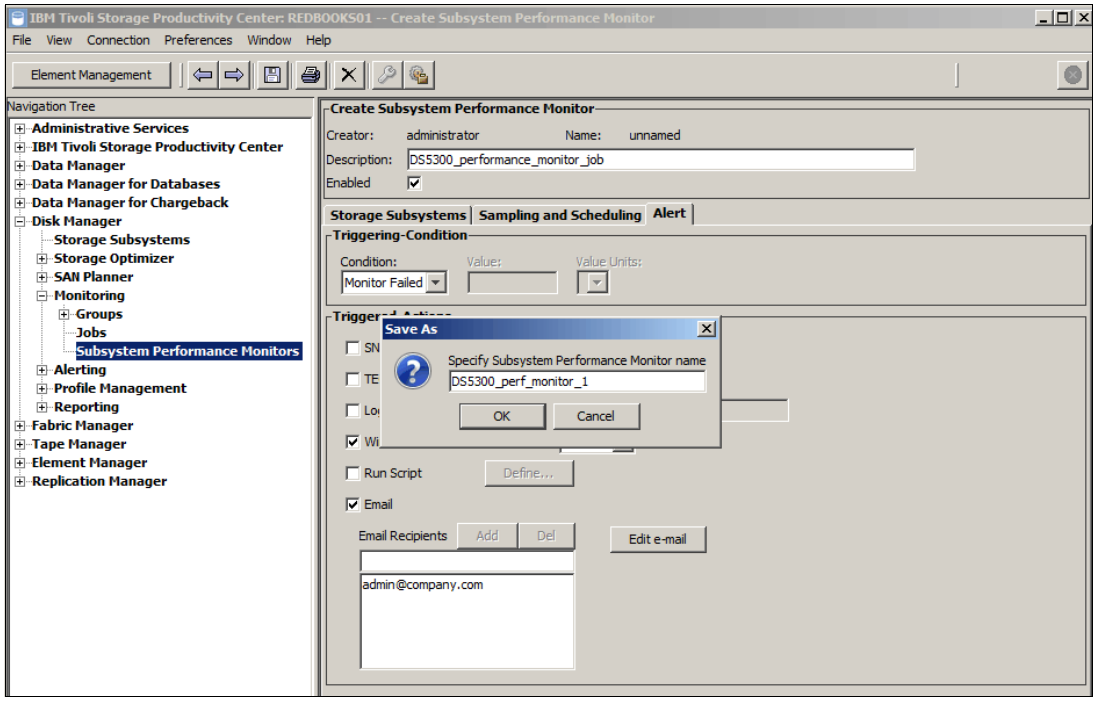


Figure 9-34 Save Performance Monitor job

5. Your Performance Monitor job is submitted now. To verify the job status, expand **IBM Tivoli Storage Productivity Center → Job Management** or click **Yes** directly on confirmation dialog box that pops up after the job is saved, as shown in Figure 9-35.

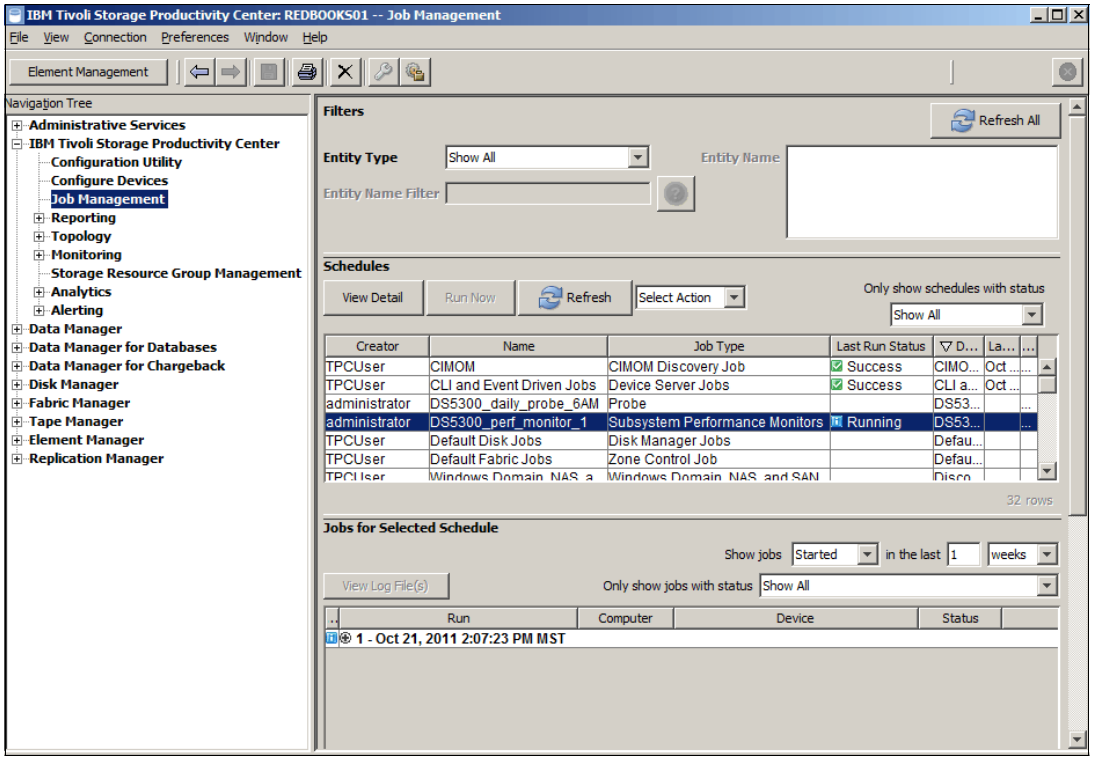


Figure 9-35 Verify Performance Monitor job

During data collection, a CIMOM or NAPI always reports all available data to Tivoli Storage Productivity Center. It includes data collected from different levels for a storage subsystem, for example, from volumes, from arrays, and from the whole storage subsystem. So, the total size of the performance data collected increases multiple times when you include more storage subsystems into performance collection. The granularity of the data collection is either all information from a subsystem, or nothing at all. One way to control the amount of data collected therefore, is to limit the number of subsystems.

Tip: Instead of using a “continue indefinitely” performance monitor, we suggest to run the performance monitor 24 hours a day and restart it again. Through this we can ensure that all changes in the device configuration and also job alerts are noted and included in the performance monitoring job. The reason we now can host a 24 hour performance monitor instead of 23 hours, is that the Tivoli Storage Productivity Center for Disk code was enhanced in TPC Version 4.2 to support the shutdown and restart without having the long delays seen in prior versions.

Another tip is to stagger your probe jobs so that you are not running them all at the same time. See Table 9-1.

Table 9-1 Example of Staggering the TPC jobs

Operation	Frequency	Day of Week	Time of Day
CIMOM Discovery	4-6 times/day	All	Every 4 hours
DS5020 Probe	Twice/week	Sunday, Tuesday	10:00 PM

Operation	Frequency	Day of Week	Time of Day
DS5300 Probe	Twice/week	Monday, Wednesday	11:00 PM
DS5020 PerfMon	23 hours	All	05:00 PM
DS5300 PerfMon	23 hours	All	06:00 PM

9.3 TPC reporting for DS5000

As we have described in Chapter 6, “DS5000 performance tuning” on page 257, the storage subsystem performance is only one piece of the performance puzzle and is impacted by many factors, including type of workload created by the application (transaction or throughput based), the system hosting the application, and performance of other SAN components.

To resolve performance issues with a storage subsystem, the storage administrator must understand the dynamics of the various performance characteristics.

A good approach is to look at current and historical data for the configuration and workloads that are not getting complaints from users, and do a trending from this performance base. In the event of performance problems, look for the changes in workloads and configuration that can cause them. TPC can help you accomplish just that by collecting performance data over time and generating performance reports.

9.3.1 DS5000 performance report

IBM Tivoli Storage Productivity Center provides two types of reports: predefined reports and custom reports. All the predefined reports for storage subsystems are performance related and can be found under **IBM Tivoli Storage Productivity Center** → **Reporting** → **System Reports** → **Disk**.

For other non performance-related reports, you can generate custom reports under **Disk Manager** → **Reporting**, which can also be used to generate performance-related reports.

In TPC, DS5000 performance data can be displayed in a table or graphical report. It can display recent or historical performance data, which can be exported into a file for offline analysis. Using custom reports, performance report for DS5000 can be created by Disks, Volumes, Storage Pools (Arrays), Computer, Storage Subsystem and by Volume to HBA Assignment.

For DS5000, there are several metrics/parameters can be monitored by TPC, as shown in Table 9-2.

Table 9-2 DS5000 performance metrics

Metrics	Description
Read I/O rate Write I/O rate Total I/O rate	Average number of I/O operations per second for both sequential and non-sequential read/write/total operations for a particular component over a time interval.
Read cache hits Write cache hits Total cache hits	Percentage of cache hits for non-sequential read/write/total operations for a particular component over a time interval.

Metrics	Description
Read Data Rate Write Data Rate Total Data Rate	Average number of megabytes (2^20 bytes) per second that were transferred for read/write/total operations for a component over a specified time interval.
Read transfer size Write transfer size Overall transfer size	Average number of KB per I/O for read/write/total operations for a particular component over a time interval.
Total port I/O rate	Average number of I/O operations per second for send and receive operations for a particular port over a time interval.
Total port data rate	Average number of megabytes (2^20 bytes) per second that were transferred for send and receive operations for a port over a specified time interval.
Total port transfer size	Average number of KB transferred per I/O by a particular port over a time interval.

9.3.2 Generating reports

This section provides several examples of how to generate reports using predefined and custom reports.

Example 1: DS5000 predefined performance report

In this example, we compare the overall read I/O rate of DS5020 and DS5300 Storage Servers. Remember that the results are affected by many factors including the workload on each storage subsystem. This example only illustrates the steps required to compare the performance of storage subsystems within a TPC environment.

To generate the report using TPC predefined reports:

1. Expand **IBM Tivoli Storage Productivity Center** → **Reporting** → **System Reports** → **Disk** to see a list of system supplied performance reports.

2. Click one of the listed predefined reports (note that not all predefined performance reports are supported with DS5000 at the time of writing). In the example, shown in Figure 9-36, we click **Subsystem Performance** report to display a list of all storage subsystems monitored by TPC with their performance metrics values. Note that column *Subsystem* shows the machine type and model followed by its serial number.

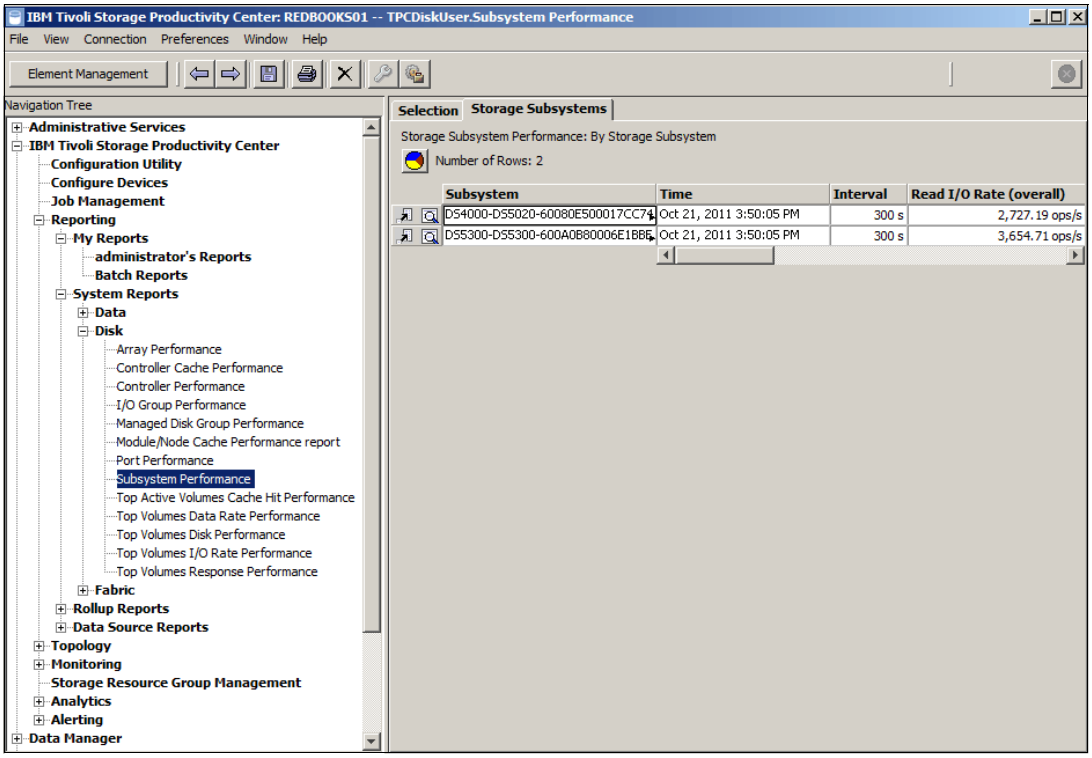


Figure 9-36 Subsystem Performance report

You can click the chart icon to generate a graph or click the **Selection** tab to modify this predefined report and regenerate it.

3. In our example, we click the **Selection** tab because we want to generate a report that compares the overall read I/O rate of the DS5020 and the DS5300. Move to the Included column the performance metrics (Time, Interval and Overall Read I/O rate) you want to include in the report, as shown in Figure 9-37.

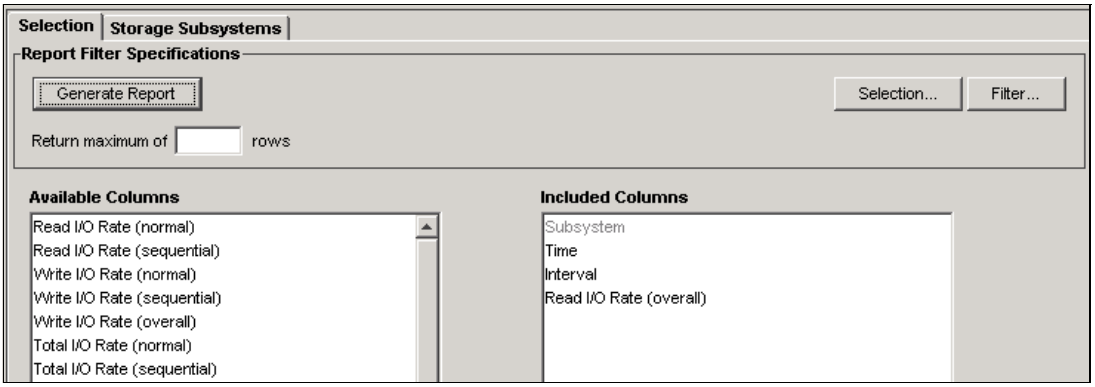


Figure 9-37 Selection of performance metrics

4. Click **Selection** on the Selection tab to select the storage subsystems for which you want to generate the report. In our case, we select **DS5020** and **DS5300**, as shown in Figure 9-38.

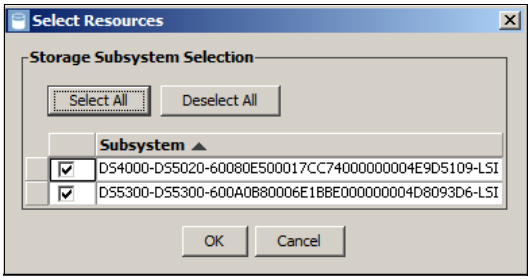


Figure 9-38 Storage subsystem selection

5. Click **Generate Report** (still on the Selection tab) and you get the result shown in Figure 9-39.

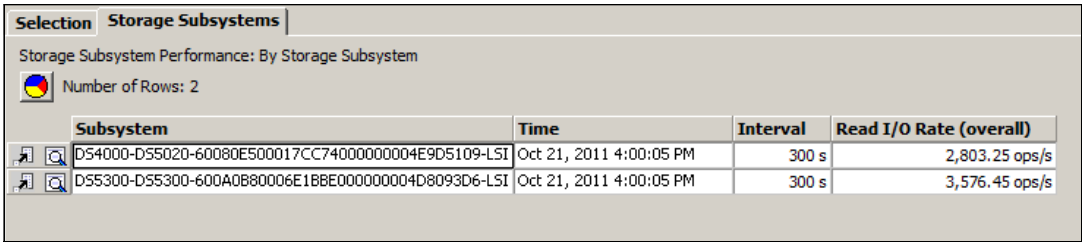


Figure 9-39 Generated report based on selection

6. We can now create a chart based on the generated report. Click chart icon and select the chart type and rows of the report that you want to be included in the chart as well, as the performance metric to be charted (see Figure 9-40).

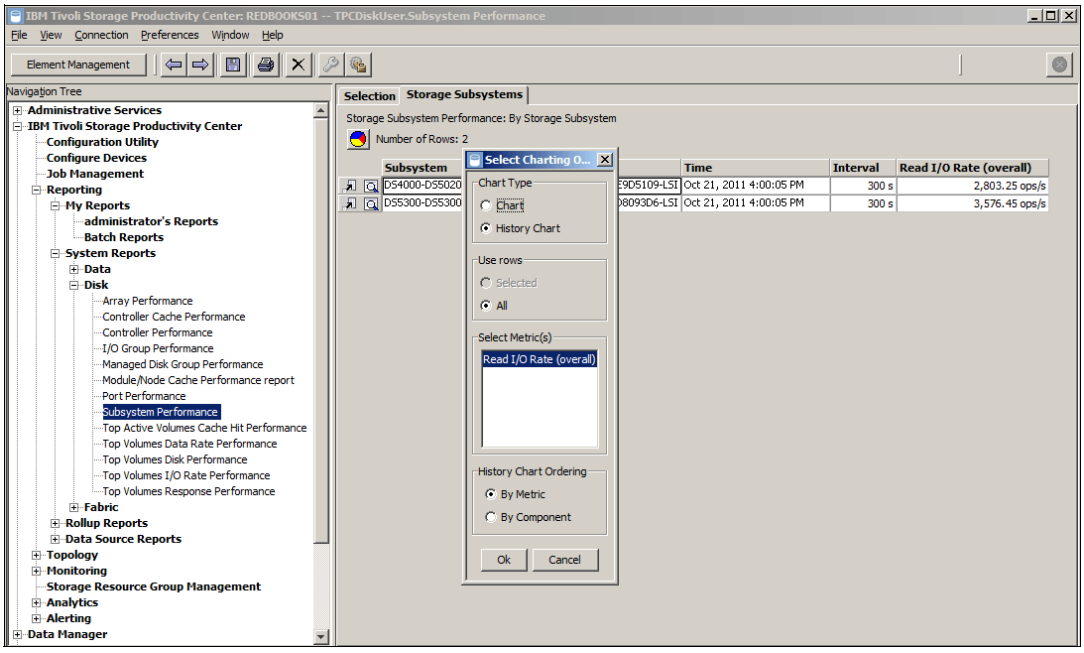


Figure 9-40 Select Charting Option

7. If we select the chart type as Chart in the Select Charting Option, the system generates a bar diagram for the selected storage subsystems, representing the average value of the selected metrics over the complete period for which the samples are available.
8. Here we select **History Chart** to generate the chart as shown in Figure 9-41.

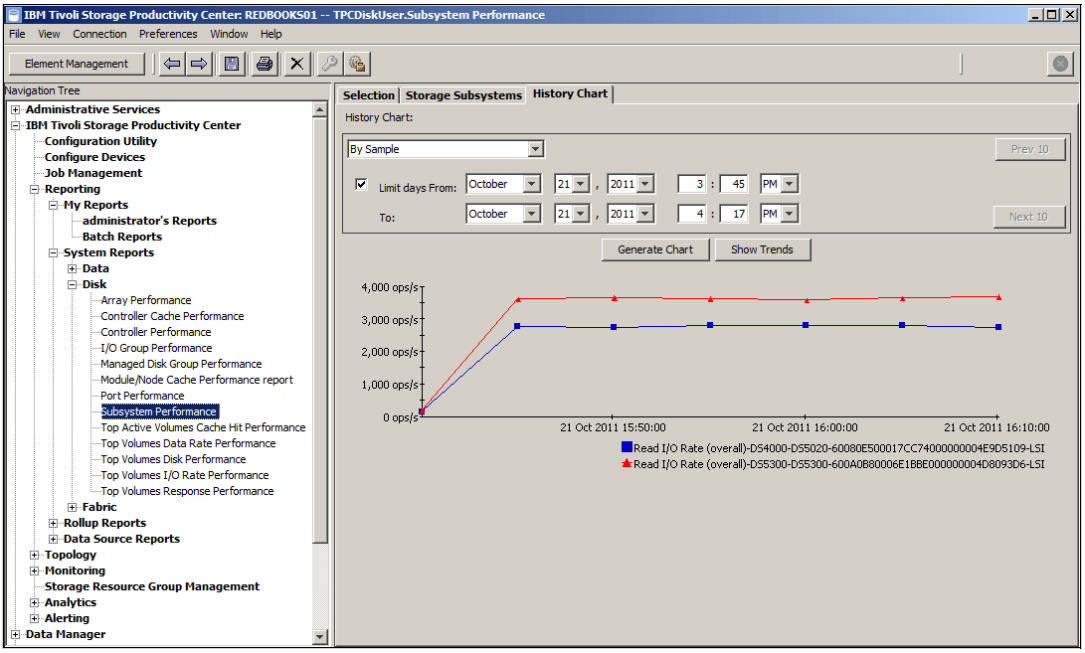


Figure 9-41 Result of historical chart

We can now graphically compare and analyze the Storage Servers performance for the selected metric. In this example, we see that overall read I/O rate of DS5300 is higher than for the DS5020.

You can generate similar reports to compare other Storage Servers and look at various performance metrics.

Note that the generated chart can also show the performance trend lines which are useful to foresee performance bottlenecks and determine appropriate measures to prevent them from occurring.

Example 2: DS5000 custom performance report

In this example, in a TPC custom report, we measure and compare the overall read and write I/O rate of two volumes (logical drives) in a DS5300 Storage Server.

Follow these steps to generate this particular custom performance report in TPC:

1. Expand **Disk Manager** → **Reporting** → **Storage Subsystem Performance**, then click **By Volume**.
2. On the Selection tab, move all performance metrics in the Included Columns into the Available Columns, except for read I/O rate (overall) and write I/O rate (overall) metrics. Check the Display historic performance data check box and select the start and end dates to generate a historical report, as shown in Figure 9-42.

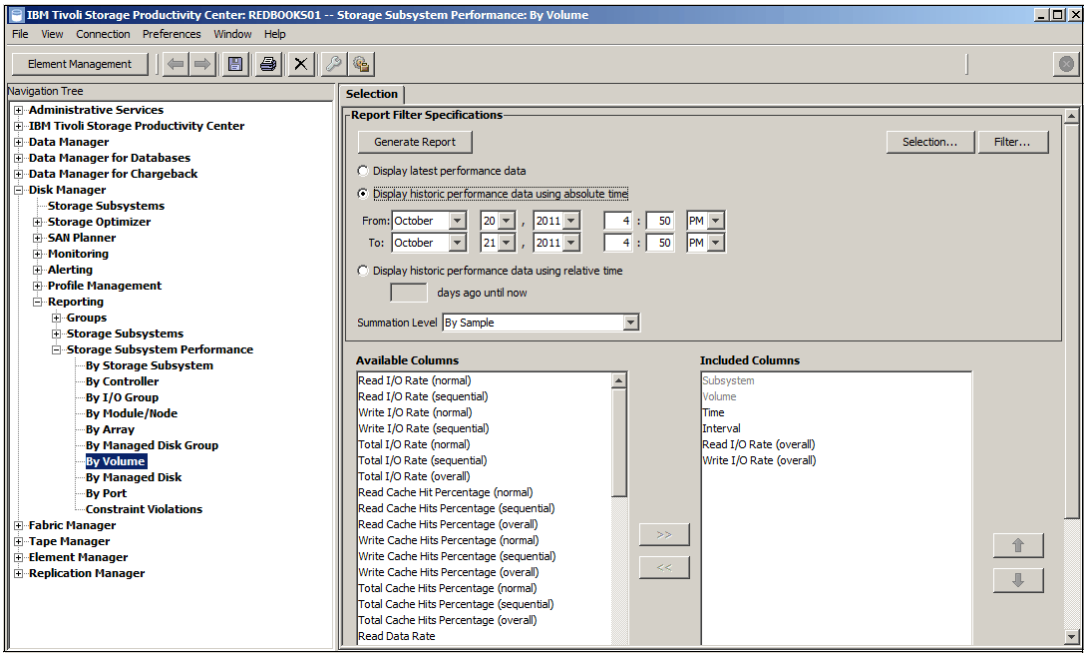


Figure 9-42 Selecting performance metrics and historical data

3. Next select the storage subsystem to report on. Click **Selection** to bring up the Select Resources window. In our example, we select two volumes from a DS5300 storage subsystem, as shown in Figure 9-43.

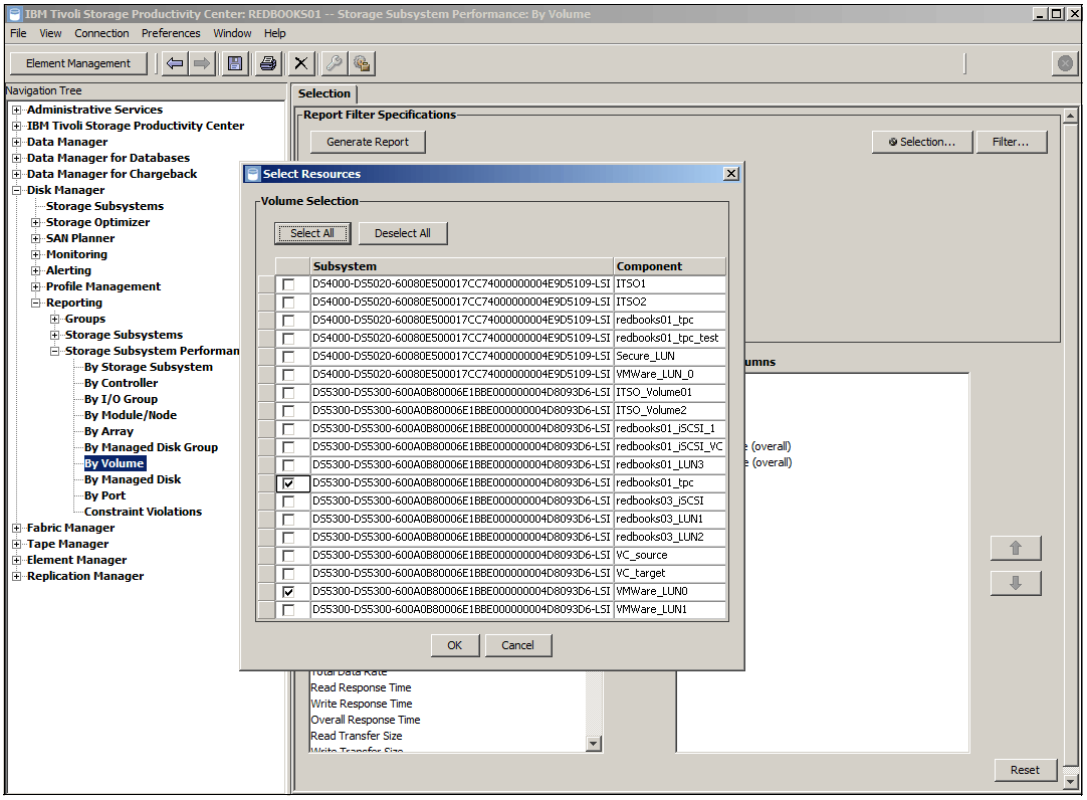


Figure 9-43 Select volumes from Volume Selection

4. Click **Generate Report** to start the query. The result is displayed as shown in Figure 9-44.

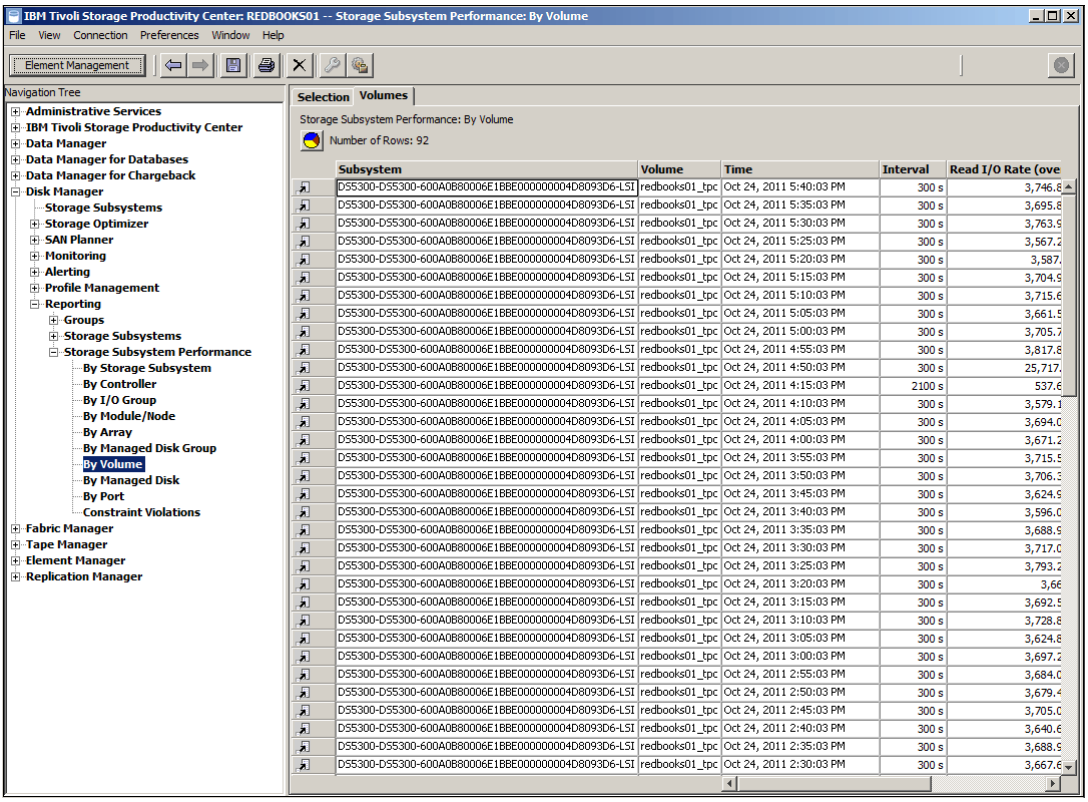


Figure 9-44 Query result

To export the query result into a file for offline analysis, select **File** → **Export Data** from the menu bar.

5. Now we generate a chart. Select all of the results and click the chart icon, then select **History Chart** and all of the metrics, as shown in Figure 9-45.

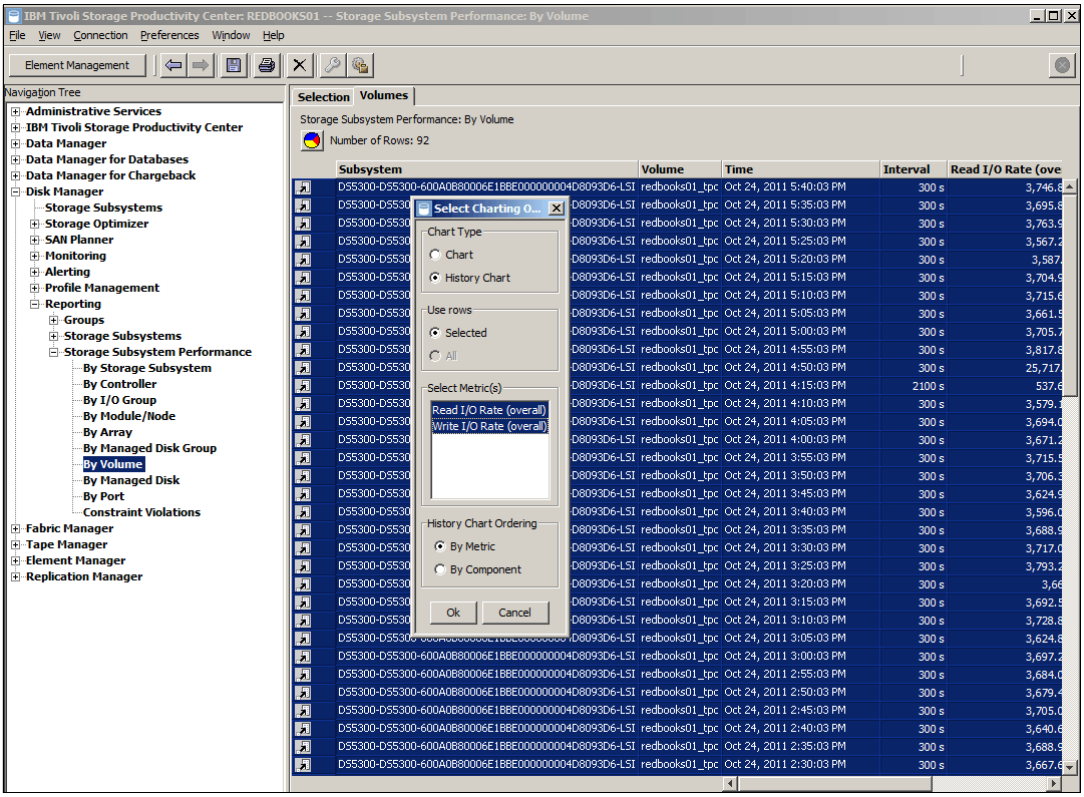


Figure 9-45 Select Charting window

The chart is displayed as shown in Figure 9-46.

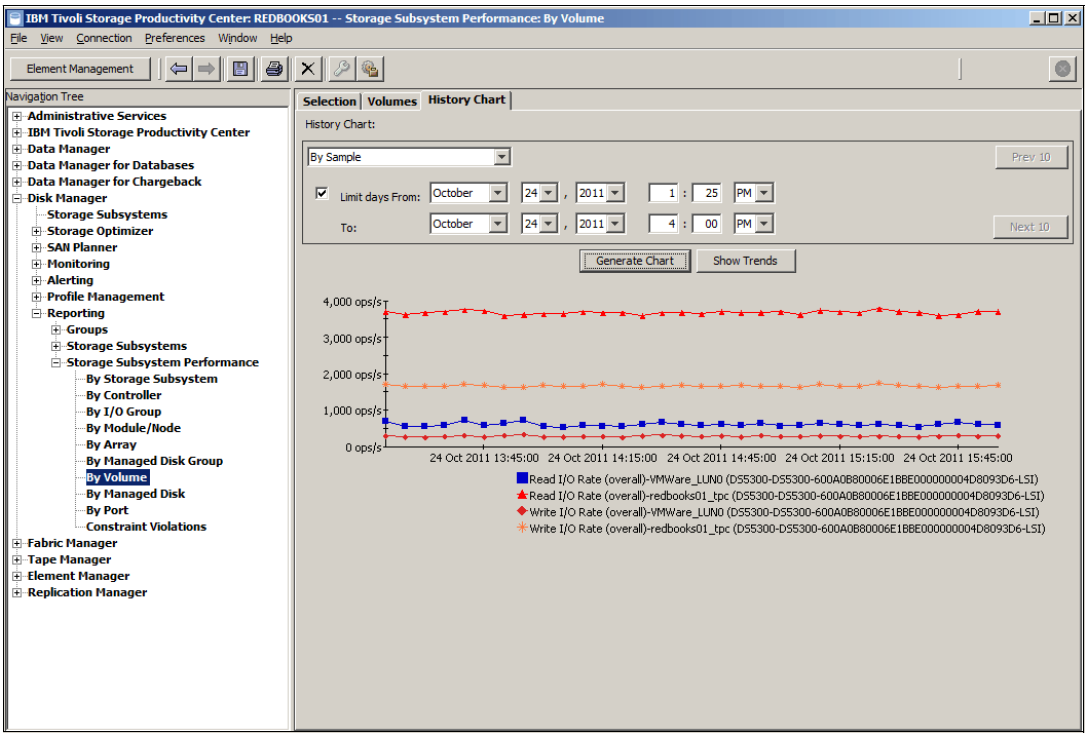


Figure 9-46 Graphical report result

From the chart result, we see that volume *redbooks01_tpc* has both an higher read I/O rate and write I/O rate compared to volume *VMWare_LUN0*, which means that volume *redbooks01_tpc* is sustaining an heavier workload. As we described earlier in this book, this type of information can be used, for example, to do performance tuning from the application, operating system, or the storage subsystem side.

Example 3: DS5000 volume to HBA assignment report

TPC reporting is also able to generate non-performance reports. The following example shows steps required to generate a *Volume to HBA assignment* report for a DS5020 and a DS5300 that were part of our TPC environment. Because there are no predefined non-performance-related reports, we need to generate a custom report.

Follow these steps to generate a Volume to HBA Assignment report By Storage Subsystem in TPC:

1. Expand **Disk Manager** → **Reporting** → **Storage Subsystems** → **Volume to HBA Assignment** → **By Storage Subsystem**. Remove unnecessary information from the Included Columns into Available Columns, and click **Selection** to select the storage subsystems for which you want to create this report. In our example, we only include Volume WWN, HBA Port WWN, and SMI-S Host Alias, and selected both DS5300 and DS5020 storage subsystems known to our TPC environment, which is shown in Figure 9-47.

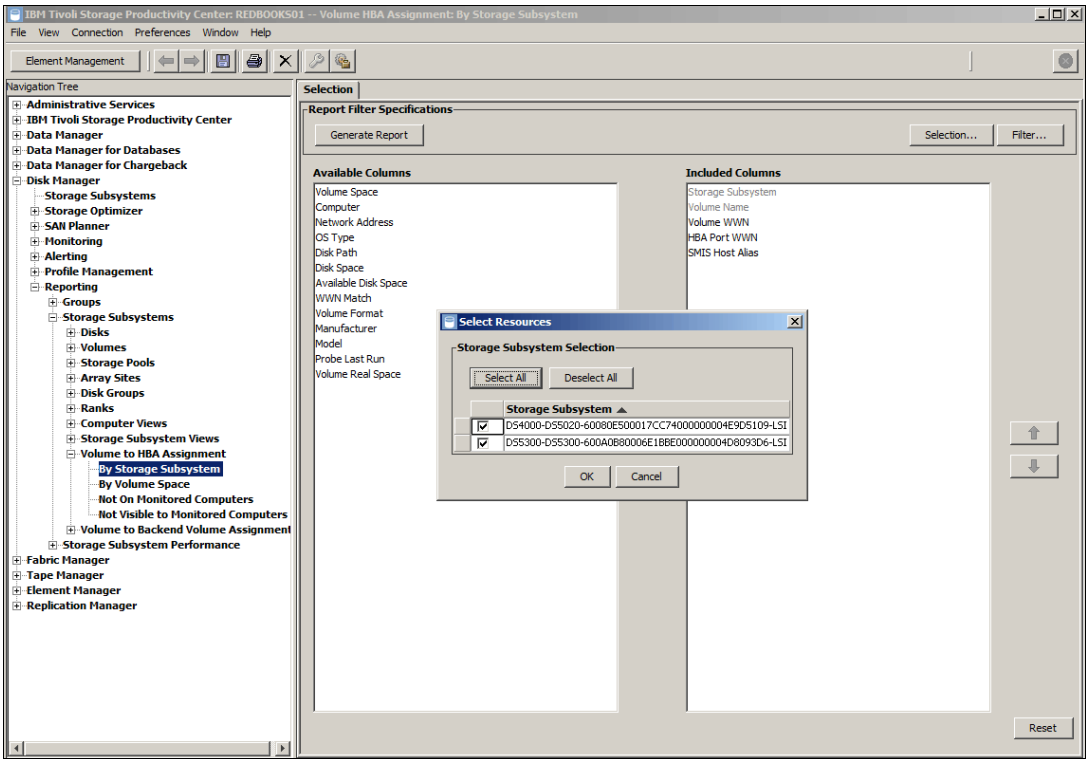


Figure 9-47 Volume to HBA assignment, select storage subsystems

2. Click **Generate Report** to generate the report, or optionally click **Filter** for a more specific query. A report example is shown in Figure 9-48.

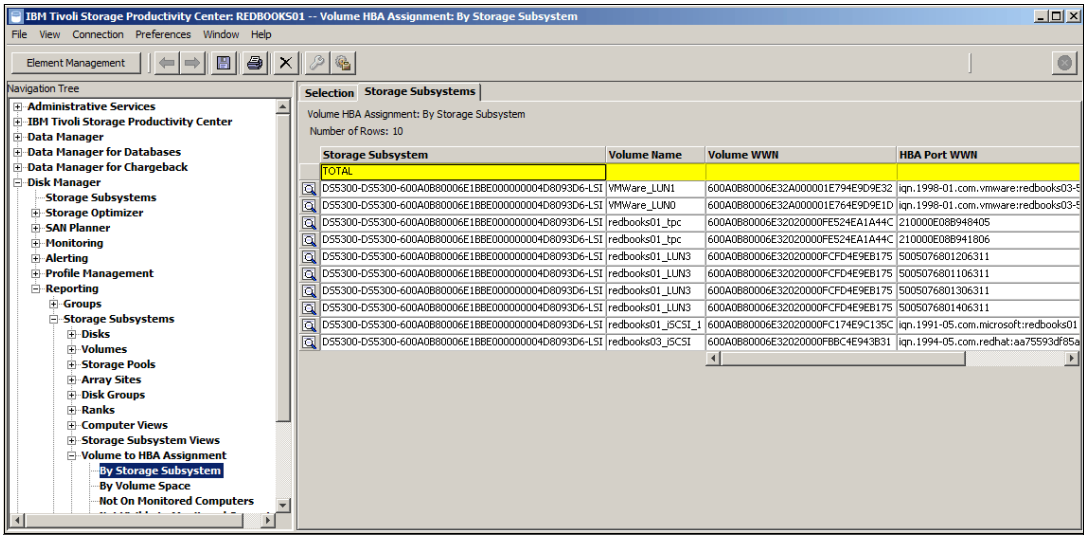


Figure 9-48 Volume to HBA assignment report

The report shows volume name, volume WWN, HBA port WWN, and SMI-S host alias for the selected storage subsystems. For a storage administrator, this report simplifies the tasks required to list volumes to servers assignments, as long as the host aliases reflect the server name.

3. You can click the magnifier icon to the left of each row to see more detailed information about the volume, including the RAID level of the array. In our example, we click the magnifier icon on the left of volume *redbooks01_tpc*, and the result is shown in Figure 9-49.

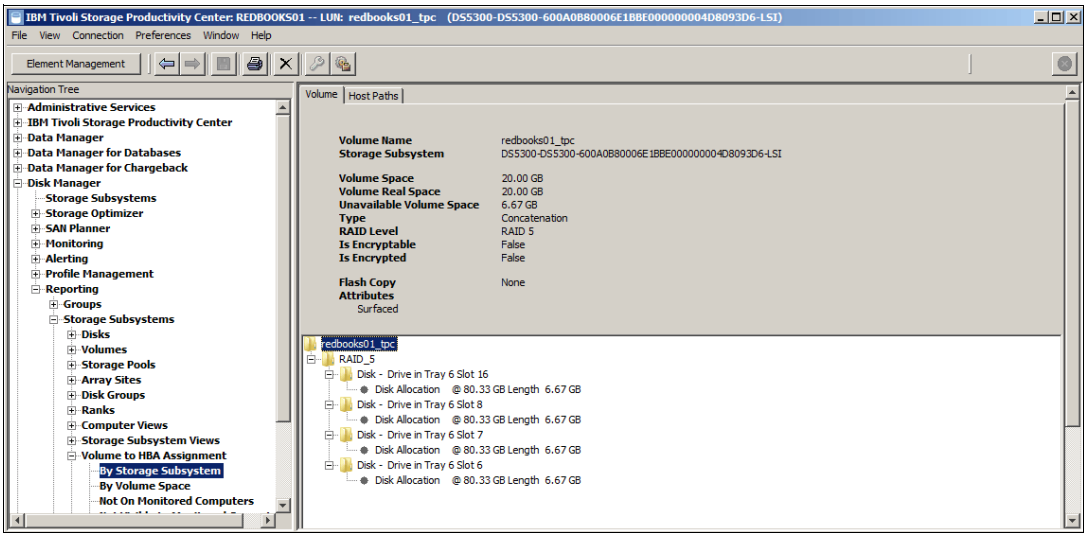


Figure 9-49 More detailed volume information

9.4 TPC Reports and Disk Magic

In this section, we explain a method to import the TPC reports into Disk Magic. We show how to figure out a baseline model that can discover potential hotspots and help you to solve performance issues. Moreover, this baseline model is going to act as a starting point to process the same imported TPC report on other Storage Servers, for example, in case you are interested to try a specific workload actually running on a legacy DS4700 on another Storage Server like a DS5300 for getting a comparison in terms of response time and resource utilization. Disk Magic is able to import a TPC Report with its important performance parameters, such as cache hit, read and write block I/O, IOPS and DataRate (MBps).

9.4.1 TPC and Disk Magic: Overview

Basically, Disk Magic can import a CSV file previously exported from a report created within the TPC server containing sample data gathered at specific time intervals.

If we generate a TPC report at storage subsystem level, each row contains performance data of the *whole* Storage Server and not directly related to a specific Logical drive, so when the CSV file is imported into Disk Magic, it is allowed to choose a peak interval only at Storage Server level without getting the necessary information required at logical drive level, therefore, in this case, only an approximate analysis is possible. This situation can be avoided by creating a TPC report for each volume (logical drive).

As explained in Chapter 10, “Disk Magic” on page 425, a Disk Magic Storage Server model analysis must be composed of a number of Host Servers equal to the number of Logical Drives managed by the Storage Server, so that the Disk Magic fields are going to be correctly filled with specific Logical Drive data.

Note that if a real Host Server has multiple Logical Drives mapped, Disk Magic can emulate this configuration by creating a “Pseudo-Host Server” for each Logical Drive mapped, giving the opportunity of filling specific data at individual Logical Drive level. Within Disk Magic we assume that there always exists a 1:1 mapping between an Host or Pseudo-Host Server and a Logical Drive.

Tip: Keep in mind that each Logical Drives data placed onto Disk Magic, must be gathered using the same sample time interval. This interval is going to be the one in which the whole Storage Server’s performance data get its maximum value. In this way a real aggregate peak analysis can be performed.

Figure 9-50 on page 402 shows a Flow Diagram of the steps we are going to describe in detail in the following paragraph. Each block of the diagram represent a specific procedure to be performed at TPC or Disk Magic Level. From a storage configuration standpoint, note that a DS Storage Subsystem Profile is required for collecting the Storage Server configuration layout, including logical drive capacity, array ownership, number of disks, RAID protection, disk drive type, speed and size.

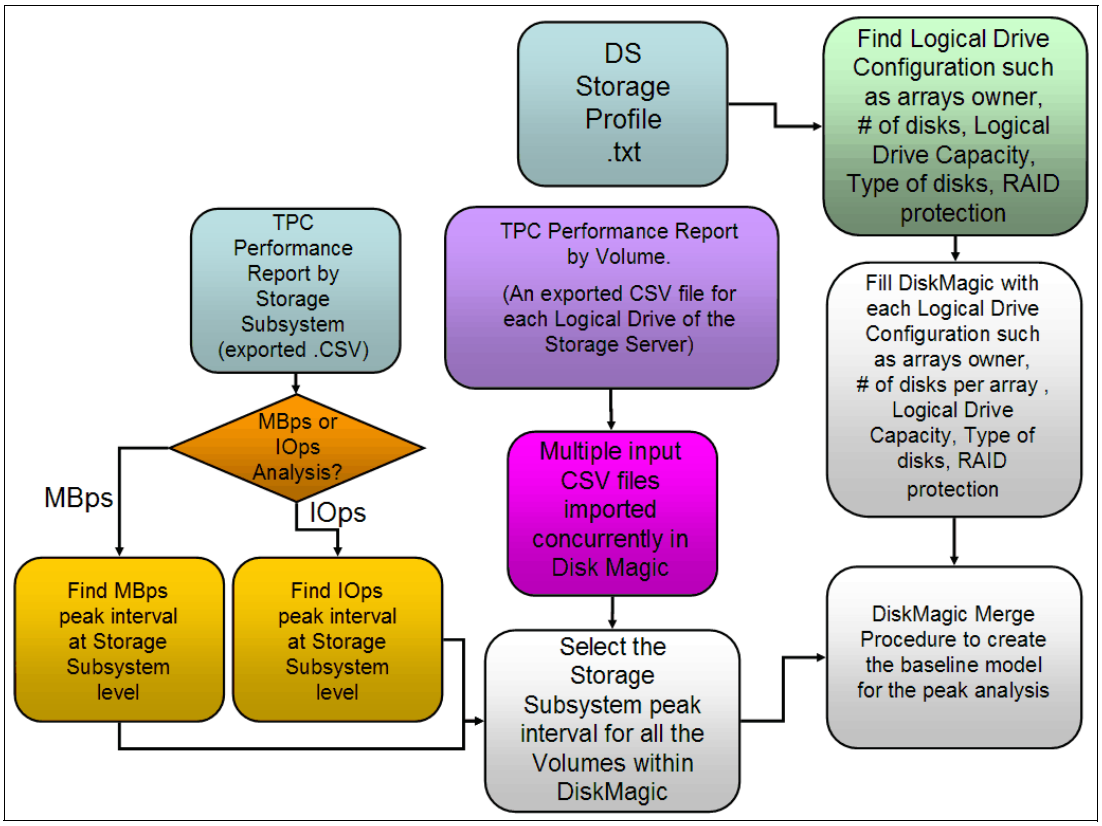


Figure 9-50 Block Diagram of the method

During the import phase of the TPC CSV files, Disk Magic reads the data and automatically sets the most important parameters for the baseline including read cache hit, write I/O blocksize, read I/O blocksize, read percentage, IOps, and DataRate in MBps.

During the processing of the imported Volumes (Logical Drives) CSV files, Disk Magic creates a Host Server and a Storage Server for each Logical Drive parsed, and because of this view, you might be surprised: Disk Magic has a special function that is able to merge multiple Storage Servers, creating a separate Storage Server that serves all the back-end logical drives previously imported.

At this point, on the Disk Magic screen you can discover a number of Host Servers equal to the number of Logical Drives that you intend to process and a merged Storage Server that manages all these Logical Drives as well, which is going to be exactly what we are looking for.

In the following sections, we describe and show the entire procedure by means of a simple example.

9.4.2 TPC and Disk Magic: Analysis example

This example gives you an overview of the procedure required to parse the TPC “by Volume” Reports within Disk Magic. Such a method can be useful for the following scenarios:

- Detect performance issues in your DS5000 Storage Server (for example, too high service time or excessive disk utilization) and simulate various configuration layouts for Storage Servers in order to solve the problem (for example, changing the RAID level, or increasing the number of disks within the array, or even changing the Storage Server model)

- Get the current Storage Servers resource utilization (CPU, FC Bus, Disks Utilization, Processor Utilization)
- Map the peak workloads time interval of your current Storage Server model to another model in order to get a comparison in terms of performance and resource utilization (What-if Analysis)
- Perform a Potential Analysis, increasing the workload (in terms of MBps or IOps) up to the maximum reachable value and take note of which Storage Server hardware resource becomes the bottleneck inhibiting further workload growth.

First we detect the peak time interval for one of the two major performance indicators (IOps or DataRate). In this example we focus on the IOps case. As shown in Figure 9-51, within the TPC console under the left pane, we select **Disk Manager** → **Reporting** → **Storage Subsystem Performance** → **By Storage Subsystem,** whereas under the right pane. we set the time period that we intend to parse, and we choose the required parameters as well. These parameters are **Subsystem (mandatory)**, **Time**, **Interval** and **Total I/O Rate (overall)**. We choose these parameters as we need to capture the time interval that gets the maximum value in terms of IOps at Storage Server level.

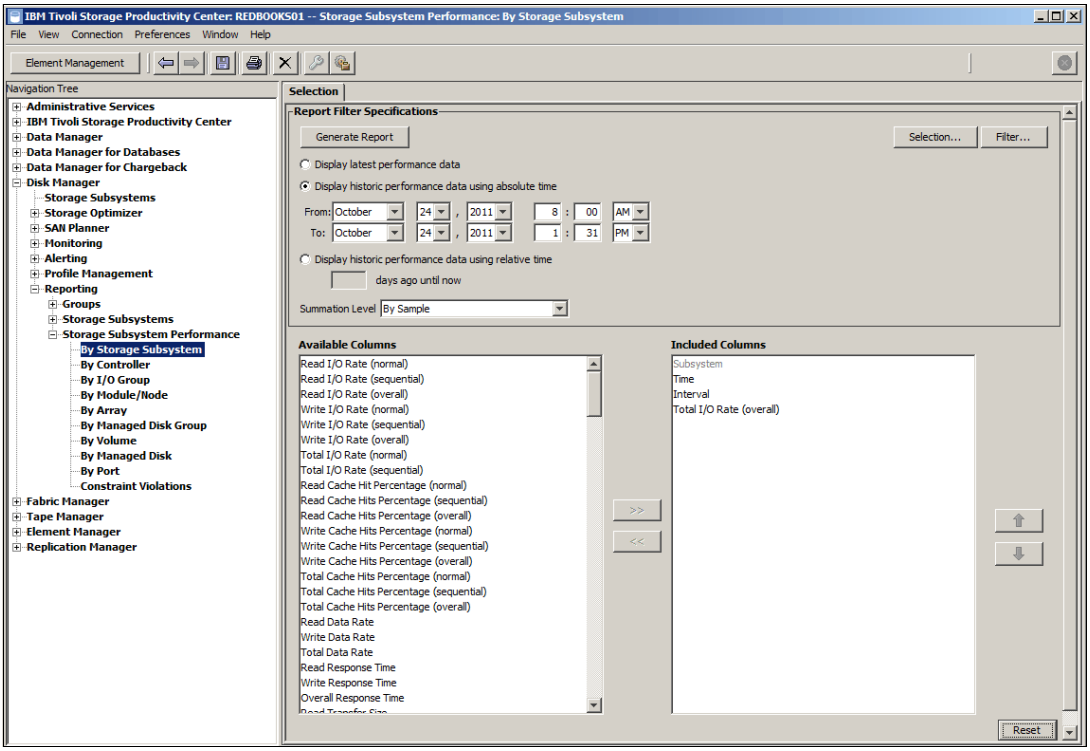


Figure 9-51 Storage Subsystem Report Configuration

Figure 9-52 shows a *Select Resources* window containing the Storage Servers currently monitored by TPC (in our case we get a DS5020 and a DS5300 Storage Server). We select one of the Storage Servers that we intend to process (we proceed with the DS5300 Storage Server).

Note that this procedure is valid for the entire IBM DS series and not only for these two newer models mentioned before (DS5020 and DS5300). You can use this procedure for all the other older DS models, such as DS4300, DS4400 and so on, which is important in case you plan to replace your dated DS Storage Server with one of the newer ones.

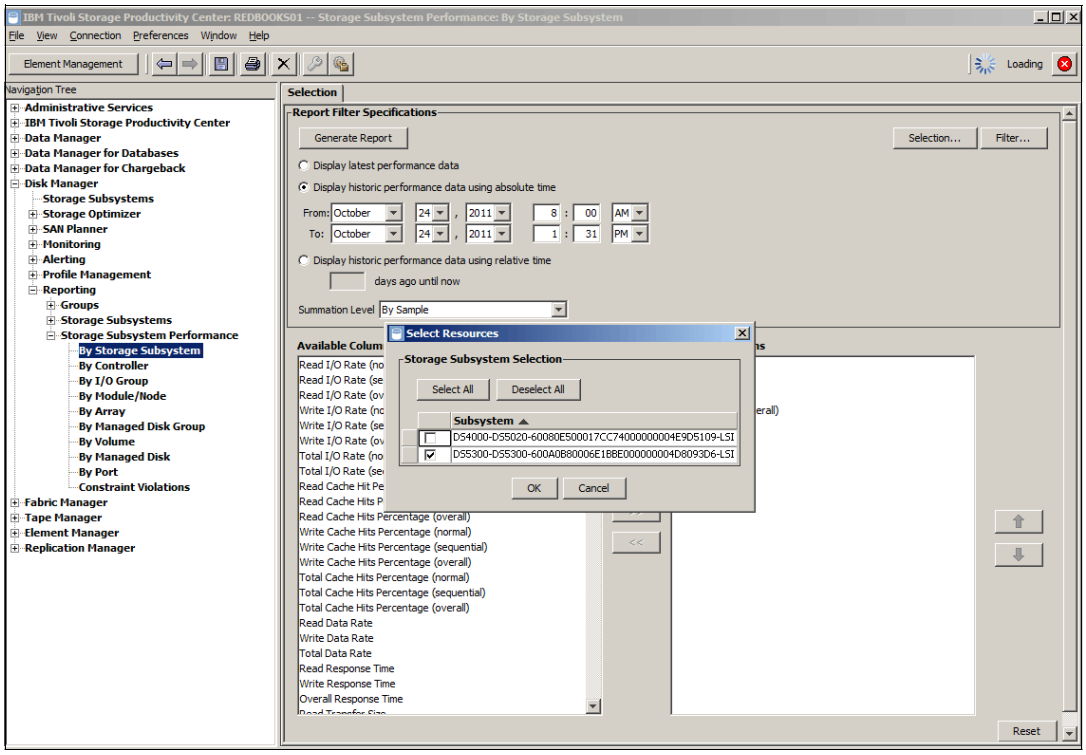


Figure 9-52 Storage Server selection

After selecting the appropriate Storage Server, click **OK** to come back to the main TPC window where you can click **Generate Report** to get to the window shown in Figure 9-53.

The tab **Storage Subsystems** under the right pane contains a table with all the samples gathered during the selected period of time. Note that the columns are exactly the same that we have chosen in Figure 9-51 on page 403. Clicking the column header leads to a sorting in increasing or decreasing order. Obviously, after sorting the Total I/O Rate (overall) column by increasing order, the peak value (at storage subsystem level) moves to the first row of the table. This peak occurs in October, 24th 2011 at 10:30:03 AM and gets a total amount of 13,309.59 operations per second (ops/s).

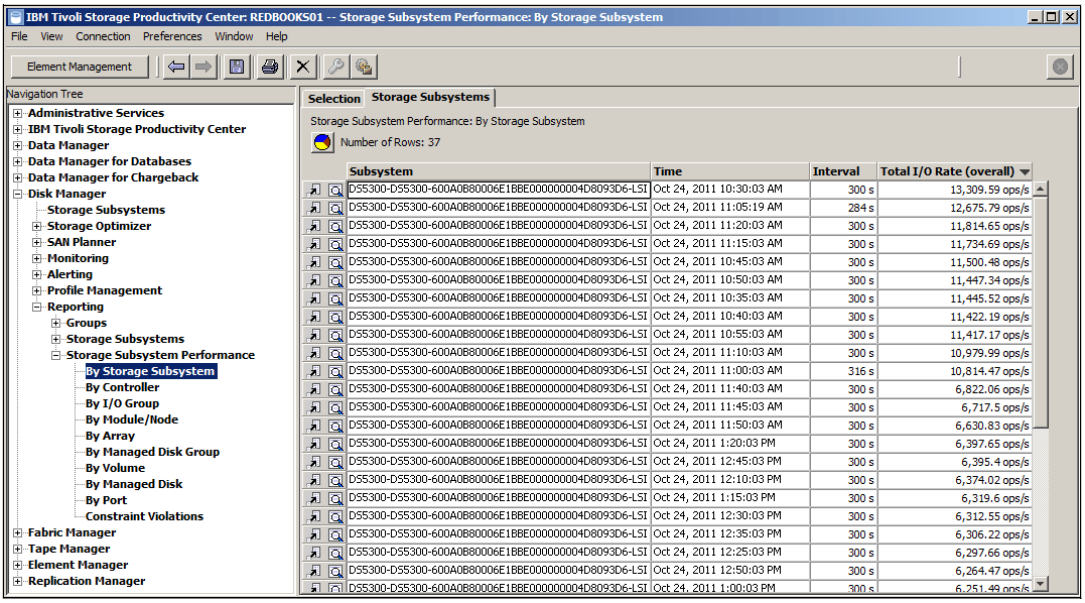


Figure 9-53 Storage Subsystems Report

At this point, we can export the report to CSV format, by clicking **File** → **Export Data** as shown in Figure 9-54.

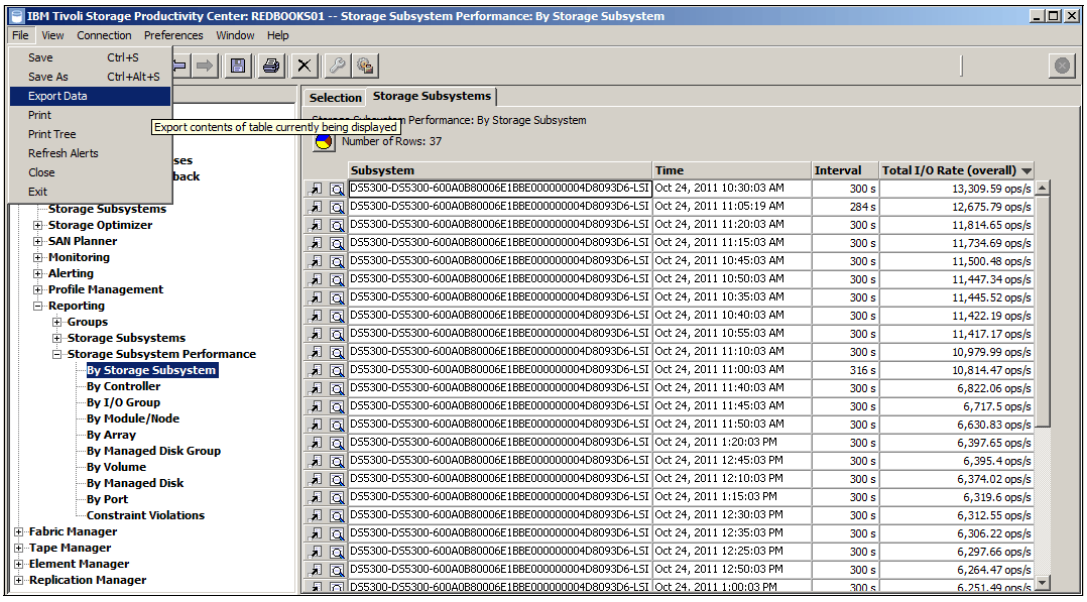


Figure 9-54 Export Data Procedure

As shown in Figure 9-55, a browsing window asks where to save the Report and format to save depending on the file type (see the pull-down menu with text field “Files of type”). We select a **Comma delimited with headers** in order to get a Disk Magic-aware format.

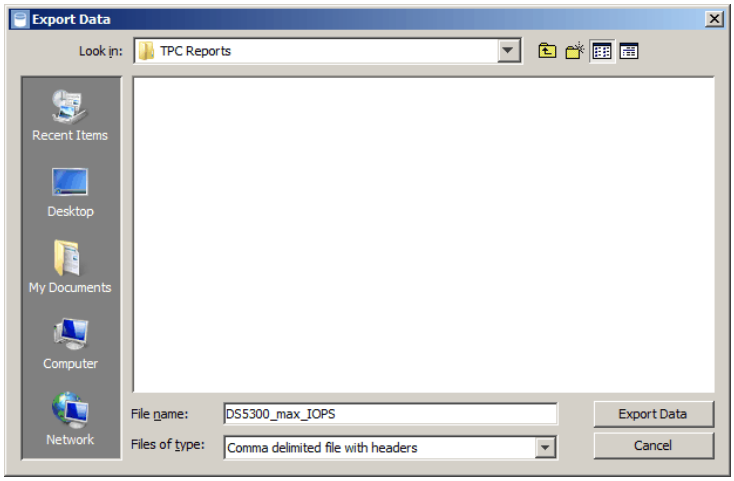


Figure 9-55 TPC Export Data Window

As shown in Figure 9-56, the CSV file looks very similar to the TPC table previously shown in Figure 9-54 on page 405. Note that the *Interval* column indicates the sample time interval in seconds.

	A	B	C	D
1	Subsystem	Time	Interval	Total I/O Rate (overall)
2	DS5300-DS5300-600A0B80006E	10/24/2011 10:30	300	13309.59
3	DS5300-DS5300-600A0B80006E	10/24/2011 11:20	300	11814.65
4	DS5300-DS5300-600A0B80006E	10/24/2011 11:15	300	11734.69
5	DS5300-DS5300-600A0B80006E	10/24/2011 10:45	300	11500.48

Figure 9-56 Exported data on CVS file

At this point, for each Volume (or Logical Drive using the DS5000 terminology) managed by our Storage Server, we are going to generate a report within the same time frame used during the generation of the Storage Subsystem Report. To do that, in the left pane of the main TPC window, we select **Disk Manager** → **Reporting** → **Storage Subsystem Performance** → **By Volume** followed by a specific column selection in the right pane under **Included Columns:**

- Subsytem (mandatory)
- Volume (mandatory)
- Time
- Interval
- Read I/O Rate (overall)
- Write I/O Rate (overall)
- Total I/O Rate (overall)
- Read Cache Hits Percentage (overall)
- Write Cache Hits Percentage (overall)
- Total Cache Hits Percentage (overall)
- Read Data Rate
- Write Data Rate
- Total Data Rate
- Read Transfer Size
- Write Transfer Size
- Overall Transfer Size

Figure 9-57 shows the related TPC window. All the parameters listed in the text pane labeled under “Included Columns” are required by Disk Magic in order to automatically set the workload and the cache profile for each Logical Drive. This Disk Magic capability is a useful time saving task that permits to avoid human errors in case of manual input.

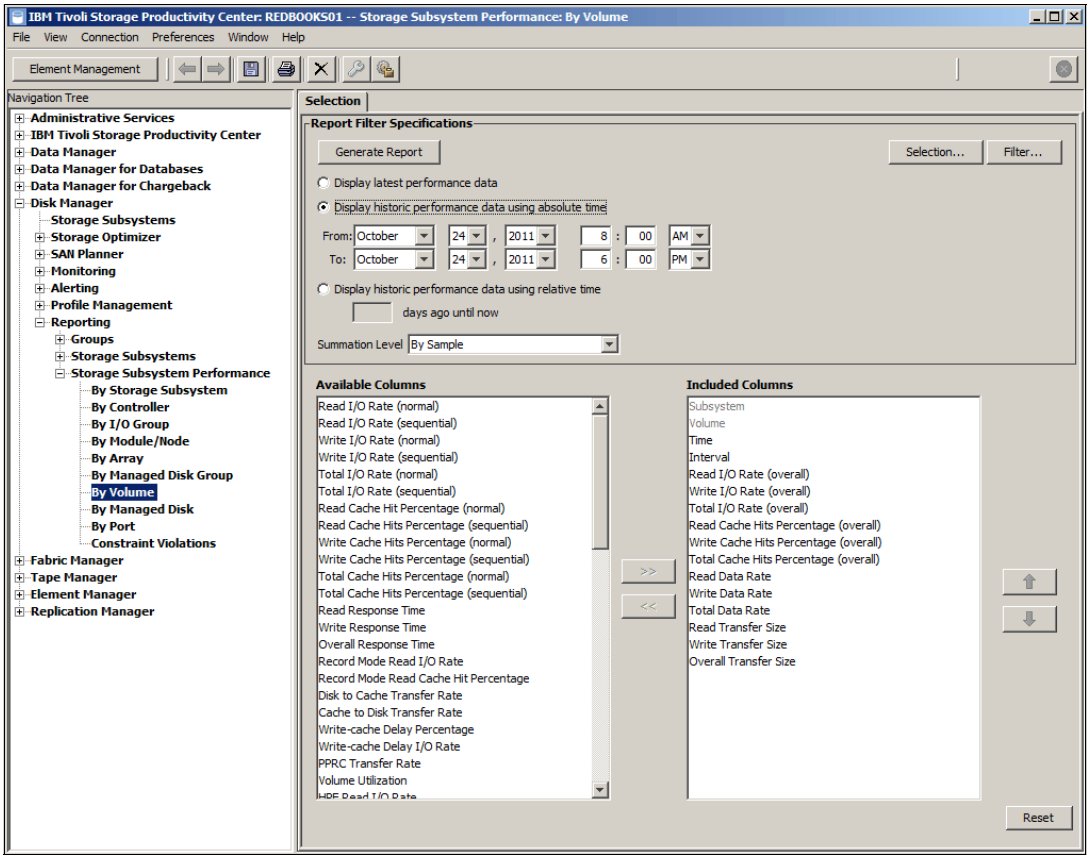


Figure 9-57 Volume Report Configuration

Click the **Selection** button at the top of the right pane in order to select the individual Logical Drive to parse.

The *Select Resources* window opens as shown in Figure 9-58 and all the Logical Drives belonging to all the managed Storage Servers are listed and checkable. Potentially a Report containing all the Logical Drives can be done but it is not useful for our intent, because we need to generate an individual CSV file for each Logical Drive, therefore we must check an individual check box at a time.

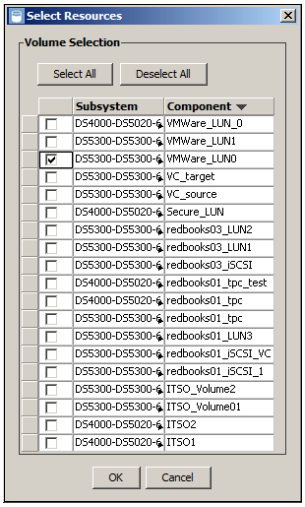


Figure 9-58 Volume Selection

After choosing the Logical Drive (**VMWare_LUN0** as shown in Figure 9-58), come back to the parent window (Figure 9-57 on page 407) and click the **Generate Report** button.

The window in Figure 9-59 opens and in the right pane, under the Volumes tab and you can see a table containing the measured data for the specific VMWare_LUN0 Logical Drive. Again, it is possible to sort the column to get an increasing or decreasing order. In this case, we must detect the time frame that gets the maximum value of the IOps at storage subsystem level (October 24th 2011 at 10:30:03 AM) as mentioned before (see Figure 9-56 on page 406).

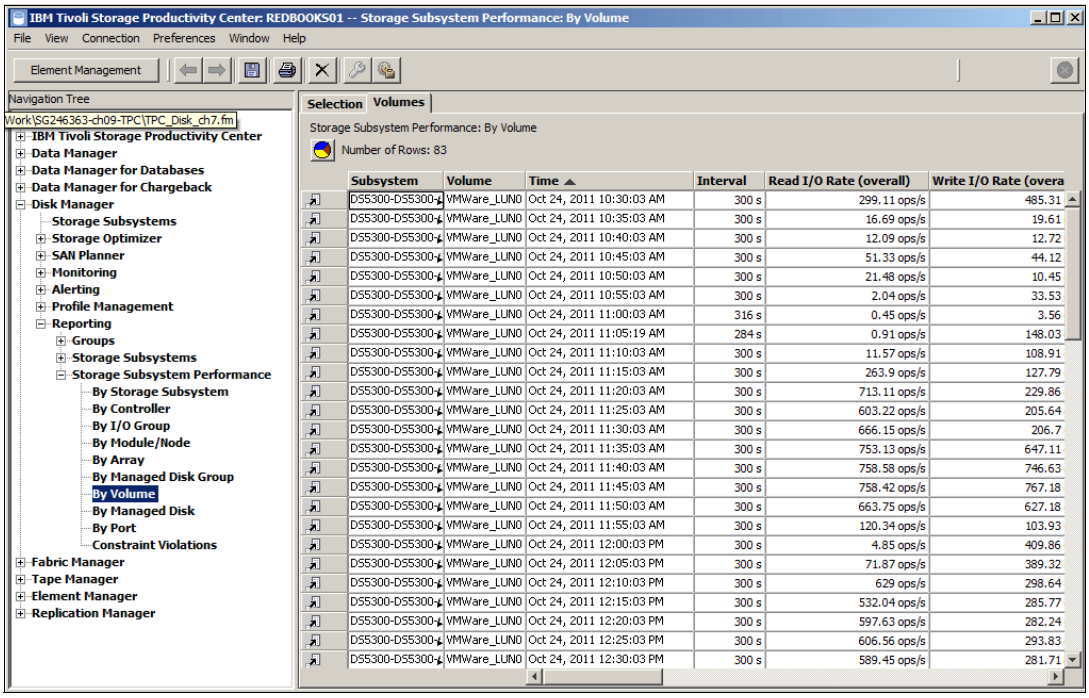


Figure 9-59 Volume Report

As previously done with the Storage Subsystem Report, we export the results to a CSV file by clicking **File** → **Export Data** as shown in Figure 9-60.

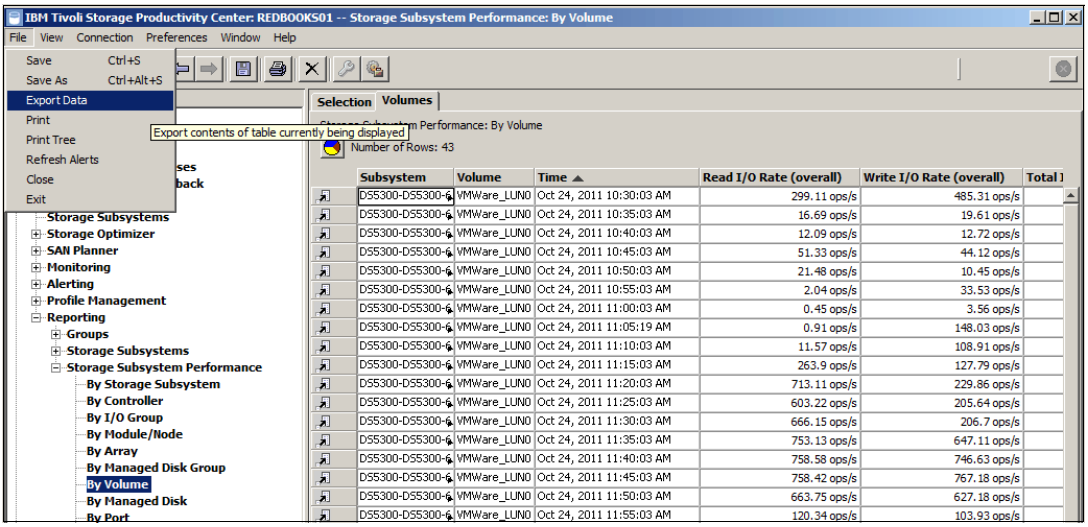


Figure 9-60 Export Data Procedure

Again, we export in **Comma delimited file with headers** format as shown in Figure 9-61. Name each CSV file with the name of the logical drive that it represents, so that you make sure to import into Disk Magic all the Logical Drives required.

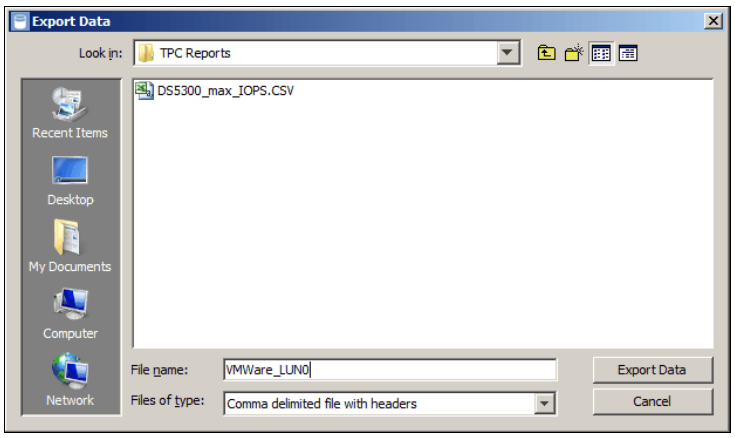


Figure 9-61 TPC Export Data Window

Within the Storage Manager (SM) client, in the left pane of the *Subsystem Management* window under the *Logical* tab, we can see the two Logical Drives that we take into consideration (Figure 9-62). In this phase we are interested to pick up the Storage Server configuration in terms of Logical drive Capacity, array RAID protection, number of disks per array and Hard Disk Drive (HDD) type/speed. We can find all this information in the Storage Manager client or in the Storage Subsystem Profile text file.

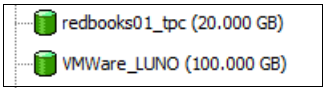


Figure 9-62 Logical Drive size

Figure 9-63 on page 410 shows the CSV Report File for the whole storage subsystem, where we can underline the peak interval row in order to easily verify that the sum of the **Total I/O Rate (overall)** for each Logical Drive is going to match exactly the value underlined in Figure 9-63 (in this example 13 309.59 ops/sec).

	A	B	C	D
1	Subsystem	Time	Interval	Total I/O Rate (overall)
2	DS5300-DS5	10/24/2011 10:30	300	13309.59
3	DS5300-DS5	10/24/2011 11:05	284	12675.79
4	DS5300-DS5	10/24/2011 11:20	300	11814.65
5	DS5300-DS5	10/24/2011 11:15	300	11734.69
6	DS5300-DS5	10/24/2011 10:45	300	11500.48

Figure 9-63 Storage Subsystem CSV Report File

In Figure 9-64, we show the CSV files for one of the two Logical Drives processed (VMVare_LUN0).

	A	B	C	D	E	F
1	Subsystem	Volume	Time	Read I/O Rate (overall)	Write I/O Rate (overall)	Total I/O Rate (overall)
2	DS5300-DS5300-	VMWare_LUN0	10/24/2011 10:30	299.11	485.31	784.42
3	DS5300-DS5300-	VMWare_LUN0	10/24/2011 10:35	16.69	19.61	36.29
4	DS5300-DS5300-	VMWare_LUN0	10/24/2011 10:40	12.09	12.72	24.81
5	DS5300-DS5300-	VMWare_LUN0	10/24/2011 10:45	51.33	44.12	95.45
6	DS5300-DS5300-	VMWare_LUN0	10/24/2011 10:50	21.48	10.45	31.93

Figure 9-64 Volume CSV Report File

At this point, a work around is required to correctly import all the CSV Logical Drive files into Disk Magic. While TPC generates Reports by Volume including two mandatory columns, that

are the **Subsystem** and the **Volume** columns (as you can see in Figure 9-64), this format cannot be imported into Disk Magic. Disk Magic import only supports Reports at storage subsystem level, which obviously does not have the **Volume** column.

As mentioned before, a single storage subsystem analysis can be done but it is not as accurate as we need, so we assume that each Logical Drive acts as a pseudo Storage Server overwriting the rows belonging to the **Subsystem** column with the rows belonging to the **Volume** column and subsequently removing the column **Volume**.

See Figure 9-65 to view the new Logical Drive CSV format.

	A	B	C	D	E
1	Subsystem	Time	Read I/O Rate (overall)	Write I/O Rate (overall)	Total I/O Rate (overall)
2	redbooks01_tpc	10/24/2011 10:30	7734.2	4790.98	12525.17
3	redbooks01_tpc	10/24/2011 10:35	8347.95	3061.28	11409.23
4	redbooks01_tpc	10/24/2011 10:40	8331.98	3065.4	11397.38
5	redbooks01_tpc	10/24/2011 10:45	8330.93	3074.09	11405.02
6	redbooks01_tpc	10/24/2011 10:50	8341.14	3074.27	11415.41

Figure 9-65 Modify CSV Report File for Disk Magic

Finally, before importing on Disk Magic, we intend to figure out the RAID arrays that host the appropriate Logical Drives as shown in Storage Subsystem Profile in Example 9-2.

Example 9-2 Logical Drives Summary

```
NAME                 STATUS   CAPACITY    RAID LEVEL  ARRAY       LUN         ACCESSIBLE BY    MEDIA TYPE
INTERFACE TYPE
redbooks01_tpc       Optimal  20.000 GB   5           Data        0           Host REDBOOKS01  Hard Disk Drive
Fibre Channel
VMWare_LUN0          Optimal  100.000 GB  5           Data        3           Host VMWare_5    Hard Disk Drive
Fibre Channel
```

Keep in mind that our test DS5300 has many Logical Drives, but we are going to arrest the I/O activity on all but the two Logical Drives of interest in order to simplify our example. The RAID arrays used in this example is a (3+P) RAID 5 array with FC Hard Disk Drives (HDD) of 450 GB capacity and 15000 revolutions per minute (RPM). The *iometer* tool has been used to generate specific I/O activity on these two Logical Drives. For more details about iometer, see the following website:

http://www.iometer.org/

At this point we have all the information required to concurrently import all the Logical Drive CSV files into Disk Magic. To do that, open the Disk Magic program and in the *Welcome to Disk Magic* window, select the radiobox **"Open and iSeries Automated input (*.IOSTAT, *.TXT, *.CSV)"** and click the "**OK**" button (Figure 9-66).

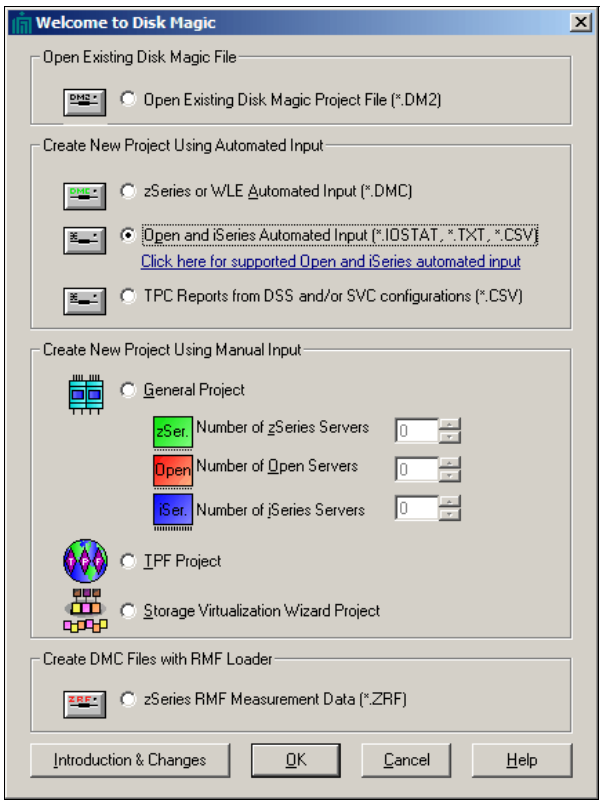


Figure 9-66 Disk Magic welcome window

Then a *Select multiple Disk Magic input files using the Shift or Ctrl Key* window opens and you are allowed to select multiple CSV files as shown in Figure 9-67 on page 412. It is easy to determine your Logical Drive CSV files if you name them with the name of the Logical Drive that they represent.

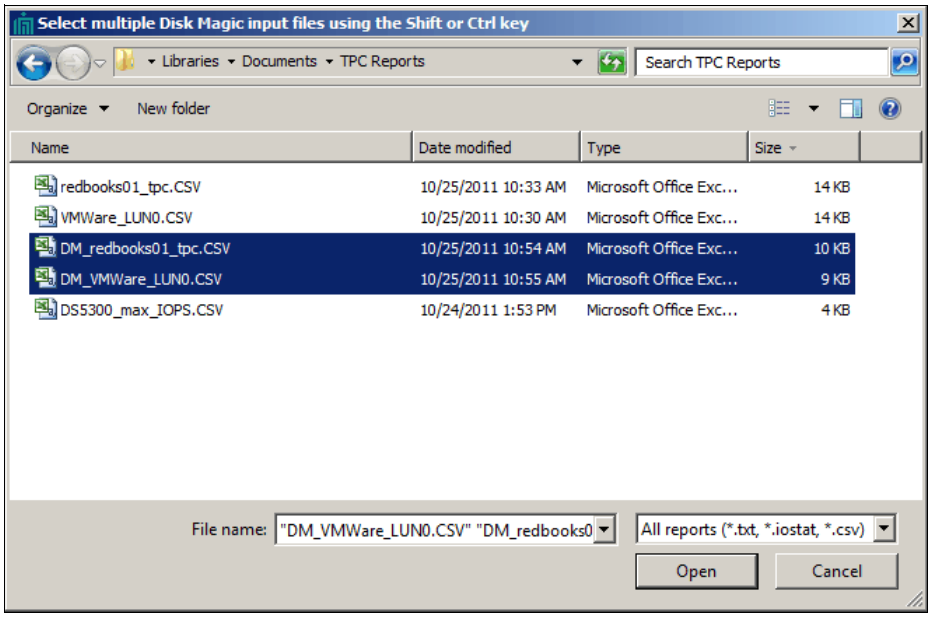


Figure 9-67 TPC Volume Report's CSV files selection

After the multiple CSV selection, the *Multiple File Open - File Overview* window is displayed, and here you must select all the files by clicking the **Select All** button. After having selected all the Logical Drive CSV files, click the **Process** button in order to start the file processing (Figure 9-68).

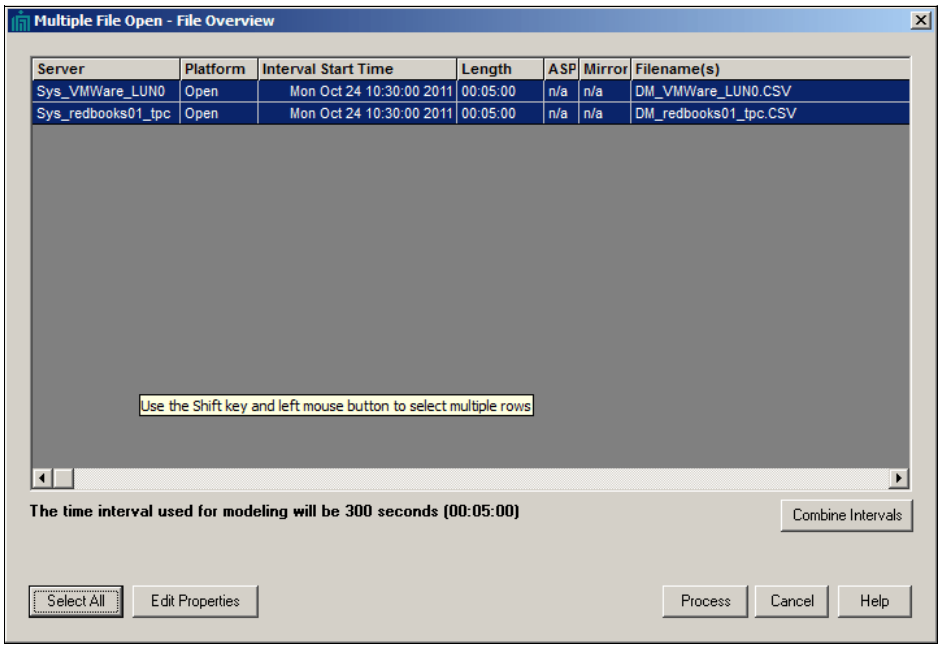
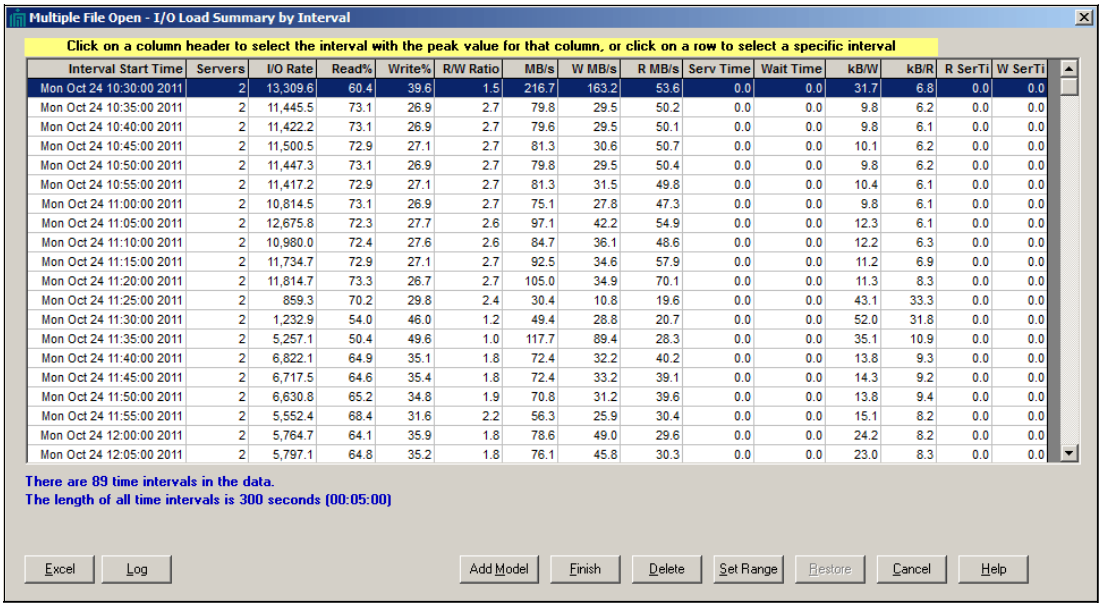


Figure 9-68 Disk Magic Multiple File Open Window

In Figure 9-69 on page 414 you can see all the Time Intervals, the number of servers parsed (two in our example) and all the data gathered and properly summed. For example, the first outlined row of the table shows two servers and an I/O rate that equals the sum of the I/O rates experienced on the two Logical Drives (again 13 309.6 IOps).

At this point in time, we select this row because it contains the aggregate data that we expect to parse. After the row selection, we click the **Add Model** button to generate the baseline model followed by clicking the **Finish** button.



Interval Start Time	Servers	I/O Rate	Read%	Write%	R/W Ratio	MB/s	W MB/s	R MB/s	Serv Time	Wait Time	kB/W	kB/R	R SerTi	W SerTi
Mon Oct 24 10:30:00 2011	2	13,309.6	60.4	39.6	1.5	216.7	163.2	53.6	0.0	0.0	31.7	6.8	0.0	0.0
Mon Oct 24 10:35:00 2011	2	11,445.5	73.1	26.9	2.7	79.8	29.5	50.2	0.0	0.0	9.8	6.2	0.0	0.0
Mon Oct 24 10:40:00 2011	2	11,422.2	73.1	26.9	2.7	79.6	29.5	50.1	0.0	0.0	9.8	6.1	0.0	0.0
Mon Oct 24 10:45:00 2011	2	11,500.5	72.9	27.1	2.7	81.3	30.6	50.7	0.0	0.0	10.1	6.2	0.0	0.0
Mon Oct 24 10:50:00 2011	2	11,447.3	73.1	26.9	2.7	79.8	29.5	50.4	0.0	0.0	9.8	6.2	0.0	0.0
Mon Oct 24 10:55:00 2011	2	11,417.2	72.9	27.1	2.7	81.3	31.5	49.8	0.0	0.0	10.4	6.1	0.0	0.0
Mon Oct 24 11:00:00 2011	2	10,814.5	73.1	26.9	2.7	75.1	27.8	47.3	0.0	0.0	9.8	6.1	0.0	0.0
Mon Oct 24 11:05:00 2011	2	12,675.8	72.3	27.7	2.6	97.1	42.2	54.9	0.0	0.0	12.3	6.1	0.0	0.0
Mon Oct 24 11:10:00 2011	2	10,980.0	72.4	27.6	2.6	84.7	36.1	48.6	0.0	0.0	12.2	6.3	0.0	0.0
Mon Oct 24 11:15:00 2011	2	11,734.7	72.9	27.1	2.7	92.5	34.6	57.9	0.0	0.0	11.2	6.9	0.0	0.0
Mon Oct 24 11:20:00 2011	2	11,814.7	73.3	26.7	2.7	105.0	34.9	70.1	0.0	0.0	11.3	8.3	0.0	0.0
Mon Oct 24 11:25:00 2011	2	859.3	70.2	29.8	2.4	30.4	10.8	19.6	0.0	0.0	43.1	33.3	0.0	0.0
Mon Oct 24 11:30:00 2011	2	1,232.9	54.0	46.0	1.2	49.4	28.8	20.7	0.0	0.0	52.0	31.8	0.0	0.0
Mon Oct 24 11:35:00 2011	2	5,257.1	50.4	49.6	1.0	117.7	89.4	28.3	0.0	0.0	35.1	10.9	0.0	0.0
Mon Oct 24 11:40:00 2011	2	6,822.1	64.9	35.1	1.8	72.4	32.2	40.2	0.0	0.0	13.8	9.3	0.0	0.0
Mon Oct 24 11:45:00 2011	2	6,717.5	64.6	35.4	1.8	72.4	33.2	39.1	0.0	0.0	14.3	9.2	0.0	0.0
Mon Oct 24 11:50:00 2011	2	6,630.8	65.2	34.8	1.9	70.8	31.2	39.6	0.0	0.0	13.8	9.4	0.0	0.0
Mon Oct 24 11:55:00 2011	2	5,552.4	68.4	31.6	2.2	56.3	25.9	30.4	0.0	0.0	15.1	8.2	0.0	0.0
Mon Oct 24 12:00:00 2011	2	5,764.7	64.1	35.9	1.8	78.6	49.0	29.6	0.0	0.0	24.2	8.2	0.0	0.0
Mon Oct 24 12:05:00 2011	2	5,797.1	64.8	35.2	1.8	76.1	45.8	30.3	0.0	0.0	23.0	8.3	0.0	0.0

Figure 9-69 Disk Magic I/O Load Summary by Interval

A *Close dialog and return to the main window* dialog window is displayed asking if you added a Model for all the intervals you need. We added a model for the selected row, because all we need is the peak interval and nothing else, even if it is possible to parse multiple intervals (Figure 9-70). Finally, click the **Yes** button.

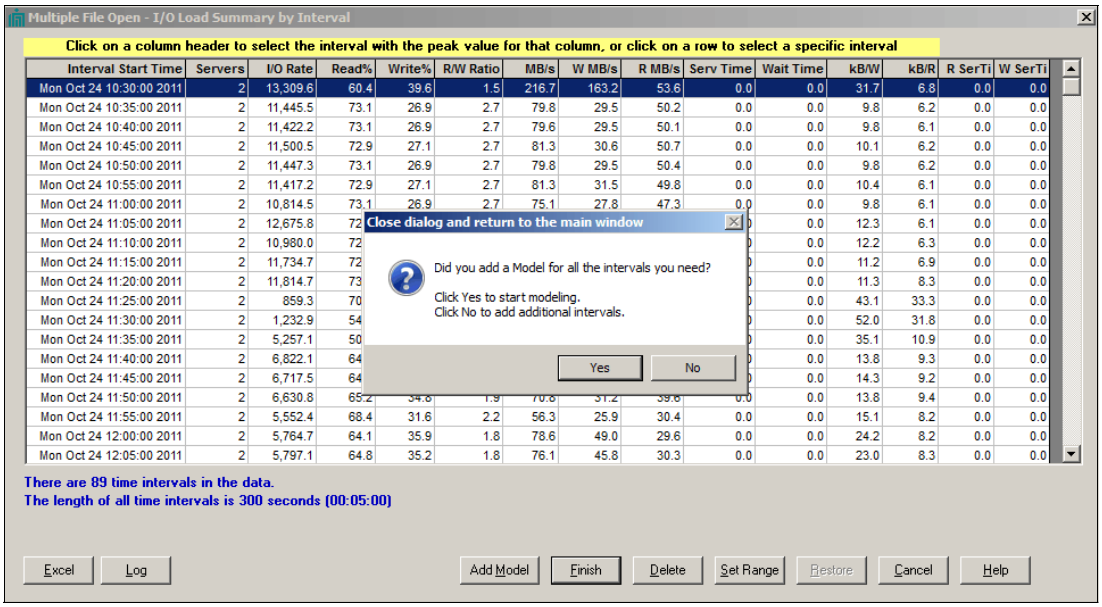


Figure 9-70 Start Modeling Dialog Window

As mentioned before, Disk Magic creates a model with two Host Server and two (pseudo) Storage Servers as you can see in Figure 9-71 on page 415. Each (pseudo) Storage Server is named with the Logical Drive name and contains only its performance and cache data. It is as if a single Host Server generates I/O activity on an individual Logical Drive managed by an individual Storage Server.

The subsequent Disk Magic merge operation is going to consolidate all the Logical Drives under an individual (and no more “pseudo”) Storage Server.

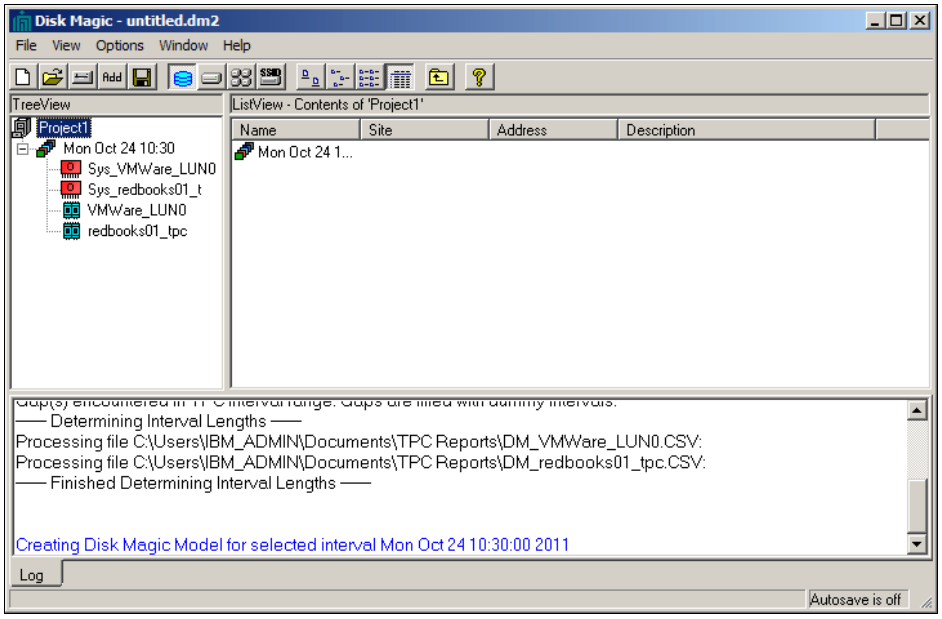


Figure 9-71 Disk Magic main window after the import process

In Figure 9-72, in the right pane, we can see that by default Disk Magic generates a model with a DS8100 Storage Servers. Obviously you can change the Storage Server model, selecting the one you own (a DS5300 in this example) in case you plan to have a point in time analysis of your current environment. Otherwise, you can select another Storage Server model in order to re-map your point in time peak workload on it (for example, in case you intend to replace your own dated Storage Server).

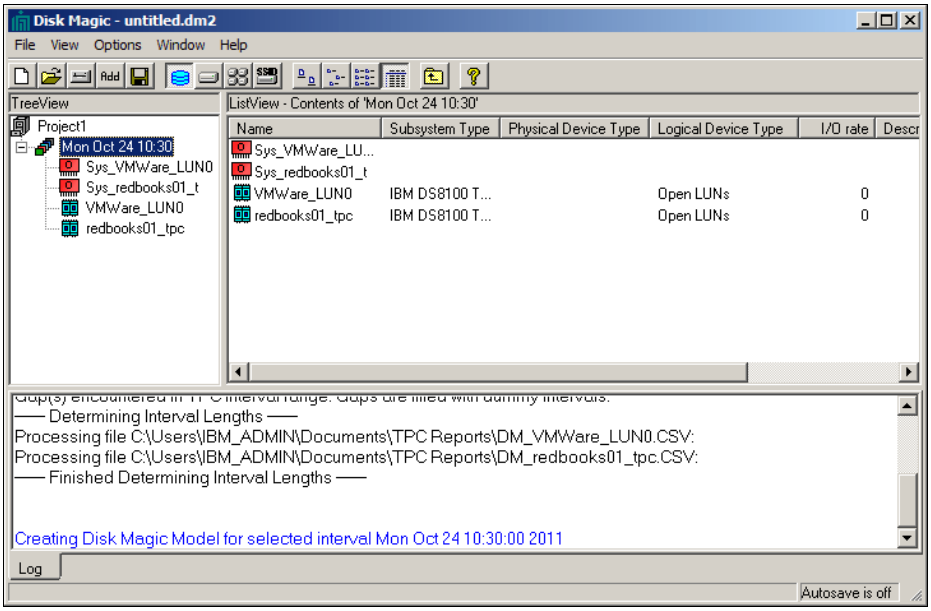


Figure 9-72 Disk Magic main window after the import process

Double-clicking the pseudo Storage Server (VMWare_LUN0 or redbooks01_tpc), a window representing the pseudo Storage Server opens as shown in Figure 9-73. This window is composed of four tabs, that are General, Interfaces, Open Disk, and Open Workload.

The Open Workload tab is automatically filled with the performance and cache data of the specific Logical Drive with which the pseudo Storage Server is called (redbooks01_tpc in Figure 9-73), whereas the other Logical Drives, although present in the Open Workload tab, get all the performance and cache data equal to zero. This configuration makes sense because each pseudo Storage Server must sustain only the I/O activity for the Logical Drives that represents.

Before the merging operation, in the General tab, you can change the model and the cache of the pseudo Storage Server, whereas in the Interfaces tab you can change the front-end connectivity in terms of number of FC host ports, number of FC HBA ports per Host Server and speed (1,2, 4 or 8 Gbps).

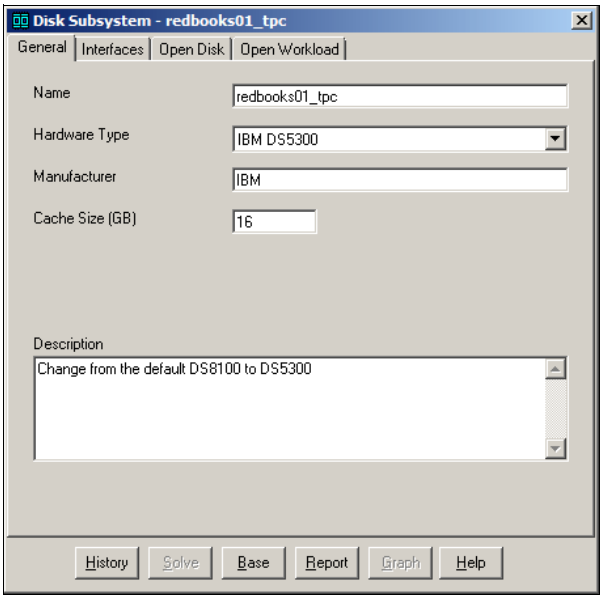


Figure 9-73 Pseudo Storage Server window - Changeable parameters

The Open Disk tab contains a number of tabs equals to the number of Host Servers (therefore equals to the number of Logical Drives because of the 1:1 mapping) imported into Disk Magic from TPC. Starting from the DS Storage Subsystem Profile or simply looking at the Storage Manager, you can capture all the information required for each Logical Drive in order to fill the mandatory fields for this tab:

- Capacity allocated to the Logical Drive
- Type of Physical Hard Disk Drive (HDD)
- RAID protection
- Number of HDD per array

Figure 9-74 on page 417 shows the tab Sys_redbooks01_tpc filled with its Logical Drive data. Select the **Distribute Open Server workloads over arrays of same disk type** check box because when selected, capacity is distributed over all available RAID arrays with the requested HDD type. Now RAID arrays will be used to their capacity and the workload will be equally distributed over all available arrays, which is the default setting.

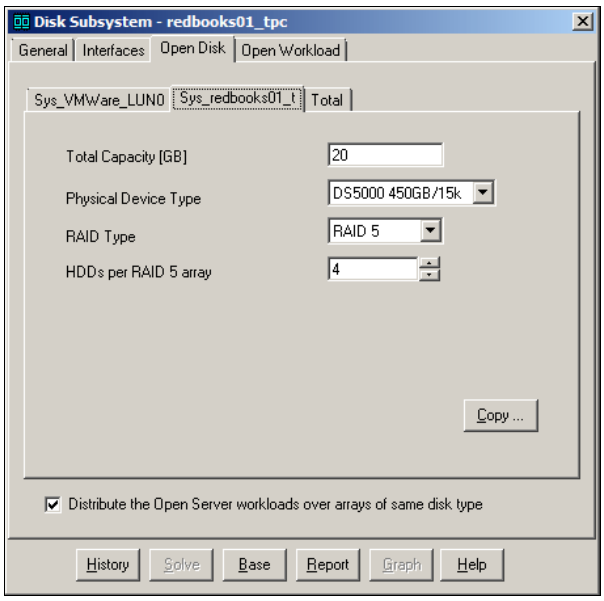


Figure 9-74 Pseudo Storage Server window - Changeable parameters

Figure 9-75 shows the Logical Drive configuration data for the other Host Server present in this analysis, that is, Sys_VMWare_LUN0.

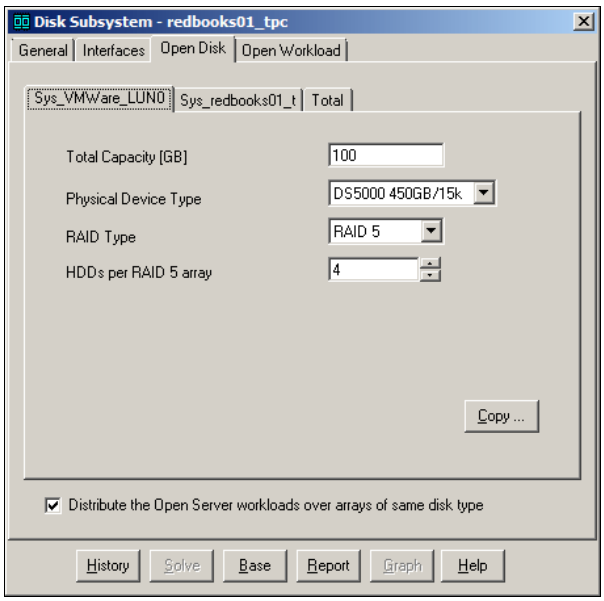


Figure 9-75 Pseudo Storage Server window - Changeable parameters

Figure 9-76 shows the performance data automatically filled during the parsing of the Logical Drive CVS files. Keep in mind that the Open Workload of each pseudo Storage Server contains as many tabs as the number of Host Servers (and therefore Logical Drives) imported from TPC. But within each pseudo Storage Server Open Workload tab, only the tab of the Logical Drive that it represents is automatically filled with valid data. The other tabs related to other Logical Drives contain zero values within this pseudo Storage Server. Of course the same is valid for each pseudo Storage Server. Note that the data automatically filled includes I/Os per sec, throughput in MB per sec and read/write transfer size in KB and Cache statistics.

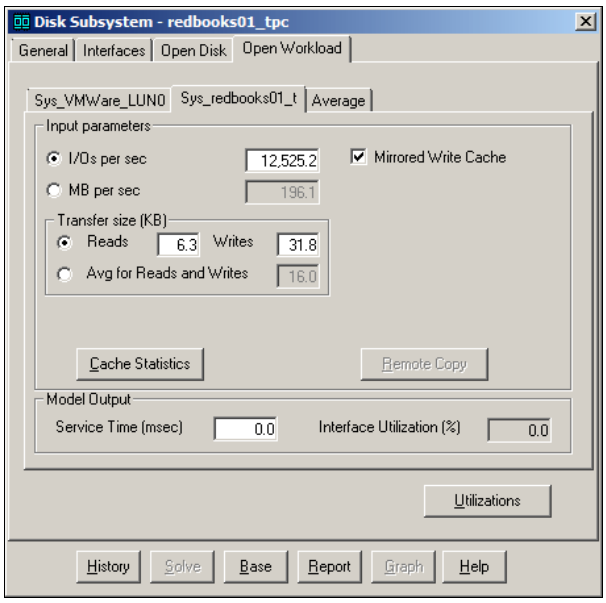


Figure 9-76 Pseudo Storage Server window - Performance and Cache data

Figure 9-77 shows all the cache data automatically filled during the processing of the imported TPC files.

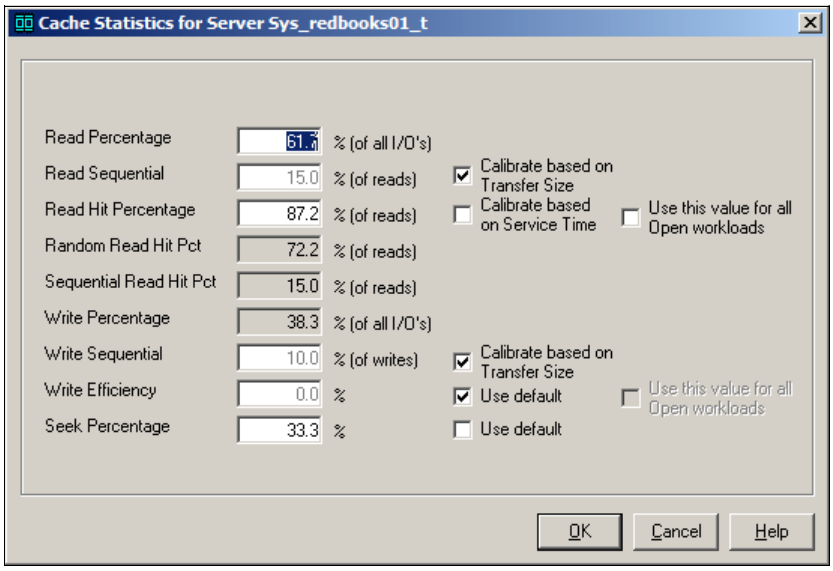


Figure 9-77 Cache parameters

Figure 9-76 and Figure 9-77 are meant to prove that each pseudo Storage Server contains only the performance data of the Logical Drive that it represent. The Average tab always contains the sum of the I/Os per sec of all the Host Servers. In this case, you can verify that the I/Os per sec performed by the Sys_redbooks01_tpc Host Server equals the value shown in the Average tab (Figure 9-78).

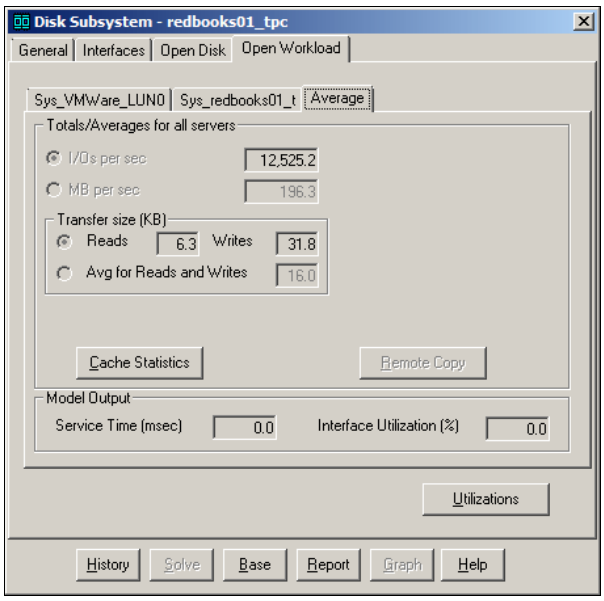


Figure 9-78 Average Performance data profile

Before proceeding with the merge of all the pseudo Storage Servers in a separate Storage Server, for each pseudo Storage Server, a base creation is required; to do that, click the **Base** button at the bottom of the window. The Disk Magic Base process establishes an internal set of values that are used as a starting/reference point for any subsequent modeling, which is true for IBM zSeries, IBM iSeries, UNIX systems, and Windows systems workloads. The reference values are collectively referred to as the Base or the baseline model.

The term base is used because in each subsequent modeling step Disk Magic starts from the values in the base to predict the effect of changes to either the configuration or workload, such as an increase of the overall I/O rate or cache size, or a change of hardware or device type.

You must have Disk Magic establish a base after you have completed the initial input of the configuration and workload data, which is accomplished by clicking the Base button in the Disk Subsystem dialog. Note that establishing a base automatically includes the calibration process.

Tip: Disk Magic automatically establishes a new base when you merge existing Disk Subsystems together or when a new Host Server is added to the model.

Figure 9-79 shows how to merge the pseudo Storage Servers into the target Storage Server. You need to select all the pseudo Storage Servers on the right pane, then right-click and select **Merge** → **Add Merge Source Collection and create New Target.**

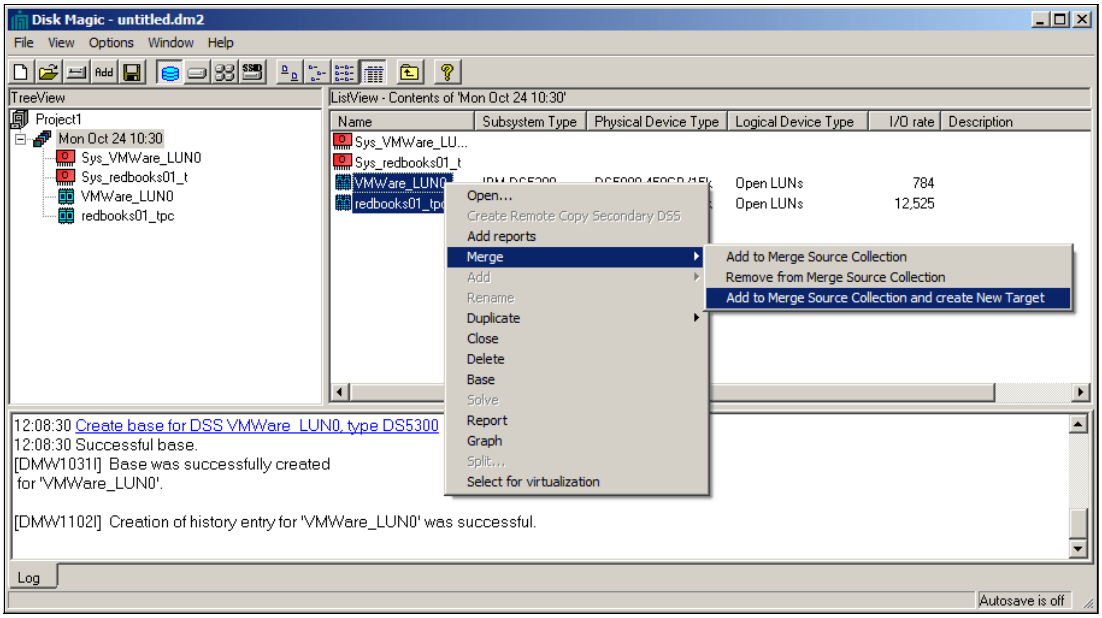


Figure 9-79 Disk Magic Merge Operation

In Figure 9-80, you can see the window that merges the pseudo Storage Servers. The default name assigned by Disk Magic is “MergeTarget1”. You need to click the **Start Merge** button to start the merge operation.

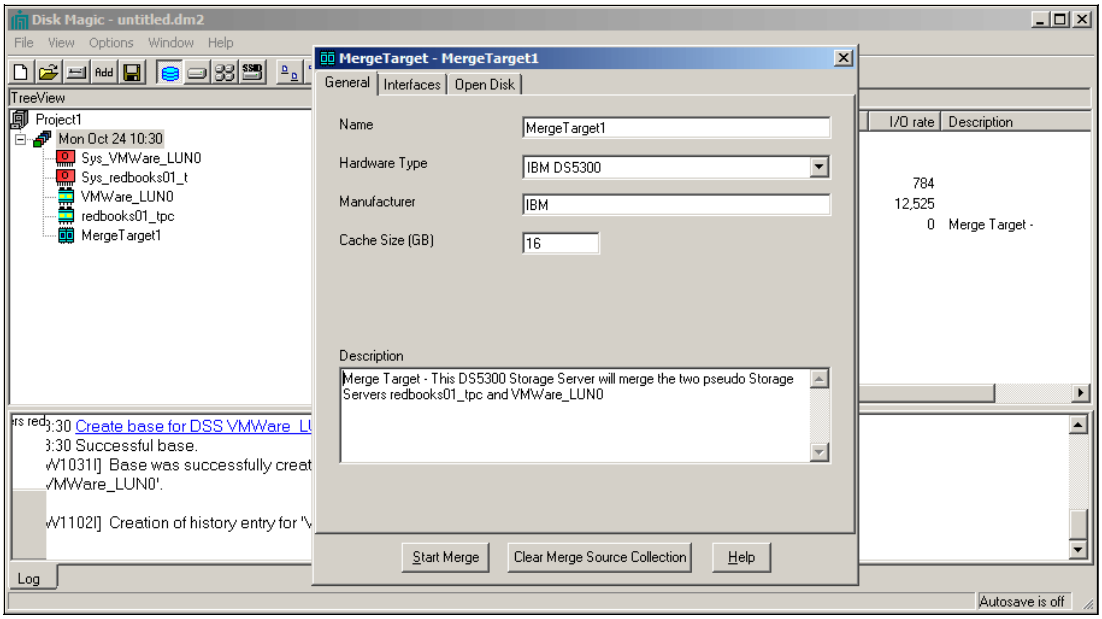


Figure 9-80 The Merge Target Window

As shown in Figure 9-81, you can select the radio button, **I want to merge all workloads on the selected DSSs**.

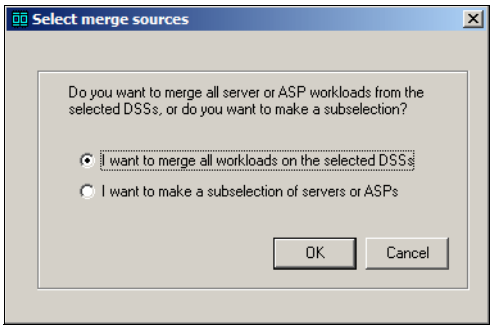


Figure 9-81 Pseudo Storage Servers selection for merging process

As shown in Figure 9-82, the target Storage Server is going to inherit all the characteristics of the pseudo Storage Server members.

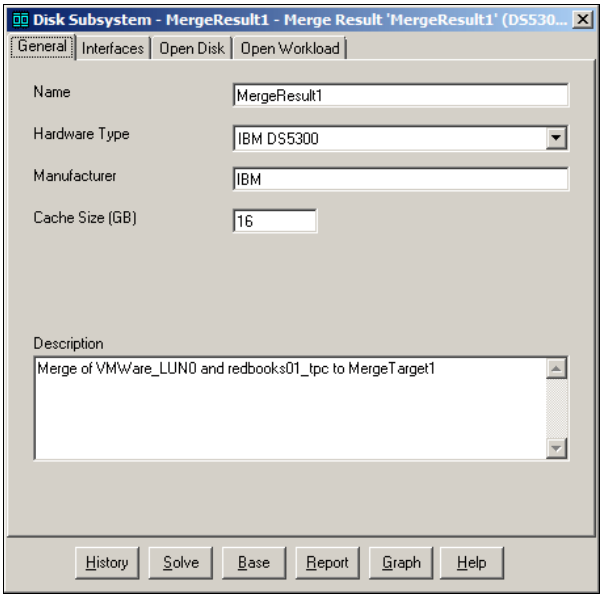


Figure 9-82 Changing the Merged Storage Server characteristics

Figure 9-83 shows the From Servers tab under the Interfaces tab. Note that each of the two Host Servers connected to the Storage Server has mounted two 4 Gbps FC HBA ports (see Count and Server Side columns).

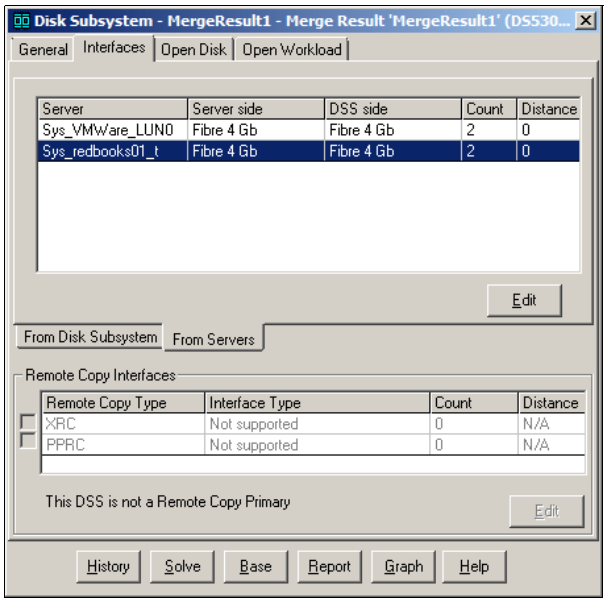


Figure 9-83 Changing the Host Server connectivity toward the Merged Storage Server

Figure 9-84 shows the From Disk Subsystem tab under the Interfaces tab, where you can see the active host ports of the DS5300 Storage Server (two ports per controller).

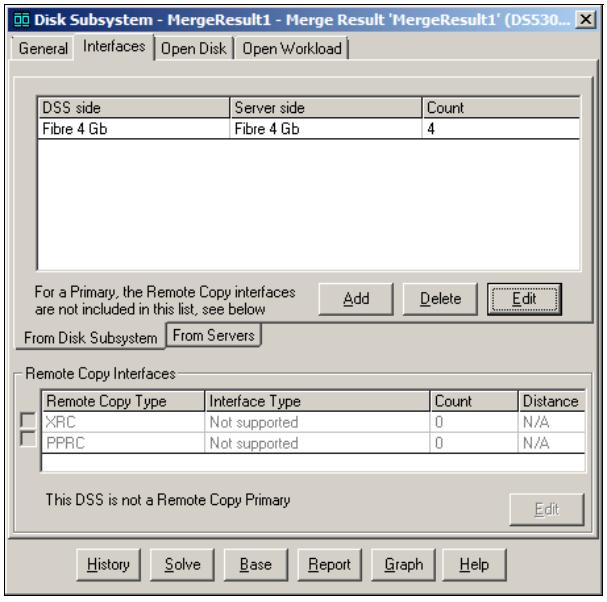


Figure 9-84 Changing the Host Port connectivity of the Merged Storage Server

Finally, in Figure 9-85, under the Open **Workload** → **Average** tab, you can see the performance data related to all the Logical Drives included in the analysis. By clicking the **Solve** button, you can get the outcomes in terms of the following variables:

- Service Time (msec)
- Interface Utilization (%)
- Utilizations

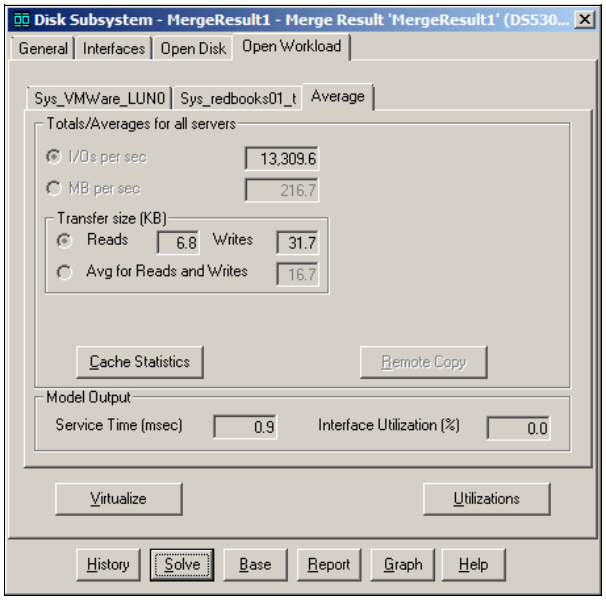


Figure 9-85 Summary of the consolidated performance data

Finally, by clicking the **Utilizations** button in Figure 9-85, the window *Utilizations IBM DS5300* is displayed as shown in Figure 9-86. Here you can check the utilization of the Storage Servers resources. See Chapter 10., “Disk Magic” on page 425 for more detailed information about Disk Magic parameters and HOWTOs. Briefly, in this example, you can see that the DS5300 resources are under utilized, except for the Highest HDD Utilization (%).

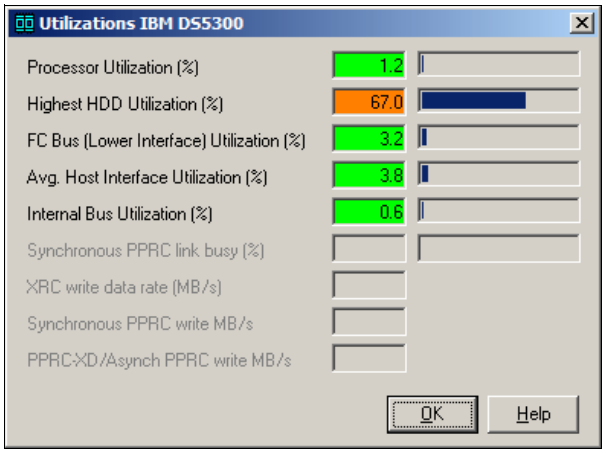


Figure 9-86 Storage Server’s resources utilization

10

Disk Magic

This chapter describes and illustrates the use of Disk Magic for Windows, a product developed by the IntelliMagic and licensed to IBM and IBM Business Partners.

Disk Magic is a tool for sizing and modelling storage subsystems for various open systems environments, as well as IBM iSeries and IBM zSeries platforms. It provides accurate performance and capacity analysis and planning for IBM System Storage DS5000, DS8000, IBM SAN Volume Controller (SVC), IBM Storwize V7000, SAN attached N series, and other vendors storage subsystems.

Disk Magic allows for in-depth environment analysis work to be undertaken prior to the purchase of the new equipment.

10.1 Disk Magic overview

Disk Magic is a flexible and powerful tool to model the performance and capacity requirements of each storage server belonging to the midrange DS5000 series. The Disk Magic model of a storage subsystem is a mathematical abstraction of the actual storage subsystem and the workload running on this subsystem. Components represented in this model include storage technologies like channel interfaces, host bus adapters, cache, Copy Services implementations, RAID types, or disk drive technologies.

Disk Magic helps evaluate important considerations such as which disk type to use, which RAID levels are appropriate for your applications, the cache size requirements, and the utilization level of HBAs. Disk Magic shows current and expected response times, utilization levels, and throughput limits for your own installation's I/O load and server configuration.

10.1.1 Data collection and modeling

In a Disk Magic study, you start by collecting data about you current server and storage environment. The collected data is entered (automated or manual input) in Disk Magic and is used by Disk Magic to establish a baseline. You must have Disk Magic establish a *base* after you have completed the initial entering of the configuration and workload data. The term base is used because in each subsequent modeling step Disk Magic starts from the values in the base to predict the effect of changes to either the configuration or workload, such as an increase of the overall I/O rate or cache size, or a change of hardware and/or device type.

With the data, Disk Magic can create a simulated model of your environment. Disk Magic allows for *what-if analysis* on items such as disk upgrades, moving workloads from one storage subsystem to another, or using another RAID level. Disk Magic keeps a history of all configuration changes made, which allows for the restoration of the project to a known stage.

Disk Magic can produce reports and graphs showing utilization figures on CPU and disk loads. The Report function creates a report of the current step in the model, as well as the base from which this model was derived. The report can be used to verify that all data was entered correctly. It also serves as a summary of the model results. The report lists all model parameters, including the ones that Disk Magic generates internally. You can also produce various graph types.

Disk magic runs in a Microsoft Windows environment. It does not require a connection with the storage subsystem to perform any of its functions because it processes manually entered data or input data coming from other useful tools such as UNIX/AIX/Linux iostat or Windows Performance Monitor log files.

At the time of writing, with DS5000 series, Disk Magic software does not support the features such as iSCSI connectivity, Solid State Disks, and Full Disk Encryption. See the IntelliMagic website in order to determine the latest enhancements:

`http://www.intellimagic.net`

10.1.2 Disk Magic functional program enhancements

New storage technologies are constantly improving their key performance features, form factors, reliability, scalability, or ability to provide for demanding service level agreements. Disk Magic modeling uses the mechanisms to describe and relate the real performance inputs to the predefined performance baselines of individual device types. It is essential for the tool to regularly update its data sources for these performance baselines so that they contain latest valid data. See the following list of new features of latest Disk Magic Release.

Here are the new features in Disk Magic Release 8.9.8:

- Support for SVC and V7000 R6.3.
- SVC and V7000 IBM Easy Tier® for internal and external drives.
- New disk types for IBM Storwize V7000: 200GB SSD, 400GB SSD, 3TB 7.2k (3.5").
- New drives for DS8800: 300GB 15k (2.5"), 900 10k (2.5"), 3TB 7.2k (3.5").
- Support for a third expansion frame to DS8800.
- Support for Easy Tier v3 in DS8000, including up to 3 tiers in each extent pool.
- Support for Extended Address Volumes II. The maximum number of cylinders for logical type “3390-A” has been increased from 262,668 to 1,182,006.
- Updated the TPC Loader to v6.2.1.
- Change default number of node pairs of V7000 stand-alone to 1.
- Support of 8Gb Front end channels for DS5000 models.
- Expiration date of November 15, 2011.

10.2 Information required for DS5000 modeling with Disk Magic

The importance of gathering accurate performance data for the servers and the storage subsystem cannot be stressed enough. This information must reflect the environment as a good base line in itself.

To collect the data, identify one or two peak periods with a workload that is critical to your installation. Indeed, because you need to make sure that the storage server will be able to handle peak workloads, it makes sense to feed Disk Magic with peak workload data. However, in the selection of peak periods, you must avoid extremes (that is, an I/O workload that is unusually high or unusually low).

Make sure as well that the collection time is bounded to the peak period, as to not skew the statistics by periods of low activity within the same interval, which means that if you choose a long sampling interval there is a risk that I/O workload peaks will be cushioned by periods of low activity within the same interval.

Finally, make sure that your statistics are complete and include all the servers that use the storage subsystem, whereas in case of San Volume Controller (SVC) or Storwize V7000 that manages more storage subsystems, you must gather the input data for all storage subsystems to be attached to the virtualization clustered system.

You can automate the gathering of data by using *iostat* (in a Linux, AIX, or UNIX environment) or *perfmon* log files (in a Windows environment) for all of the servers that use the storage subsystem. In this case, Disk Magic automatically summarizes in a table all the input data by interval (each column header indicates a performance counter specified in the *iostat* or *perfmon* log file), and across all servers that contribute workload to the storage subsystem. Then, it is possible to select the interval with the peak value for a specific performance counter and use it as a baseline for your analysis.

The measurement data generated with iostat or Windows perfmon is limited because Storage Server layout configuration data (Total capacity, Physical Device Type, RAID type, HDDs per RAID array) and cache statistics are not specified, which makes it more difficult for Disk Magic to establish a fine-tuned calibrated model. However, Disk Magic can use the available statistics to compute a reasonable cache read hit ratio and to set the percent of read and write sequential activity.

- The *cache Read Hit Ratio* is computed based on the measured service time in combination with the configuration information, such as storage server type and physical disk type. Note that this only works if the data is collected on a storage subsystem supported by Disk Magic, and not when it is collected on internal disks. For internal disks, Disk Magic uses a default Read Hit Percentage.

- The *Read Sequential and Write Sequential percentages* are selected based on their transfer sizes. A large transfer size indicates a higher sequential content. and a small transfer size indicates a higher random access content.

Tip: As a best practice, you can (and must) choose to enter the sequential percentages yourself if they are known to you from another source.

Next we briefly describe how to use *perfmon* and *iostat* to collect statistics for Disk Magic.

If these files are unavailable, you can manually enter the following data: blocksize, read/write ratio, read hit ratio, and I/O rate. Note that, except for the I/O rate, Disk Magic will provide defaults for all other parameters when you do not know or cannot easily collect this data.

10.2.1 Windows: perfmon and Disk Magic

Windows Performance Monitor (or perfmon) is a useful tool for gathering performance statistics for Disk Magic. Disk Magic can automatically process the performance data in a Windows perfmon file.

To start the Performance Monitor on Windows Server 2008 and set up the logging, go to the task bar and click **Start** → **Control Panel** → **Administrative Tools** → **Performance Monitor** → **Data Collector Sets** → **User Defined**, then right-click, select **New** and then **Data Collector Set**. The **Create new Data Collector Set** opens, and here you can enter the name of the collector set (in this case, *MyDataCollector*) and then select the radiobox **Create manually (Advanced)** as shown in Figure 10-1. The manual selection allows you to select the counters of interest in your analysis.

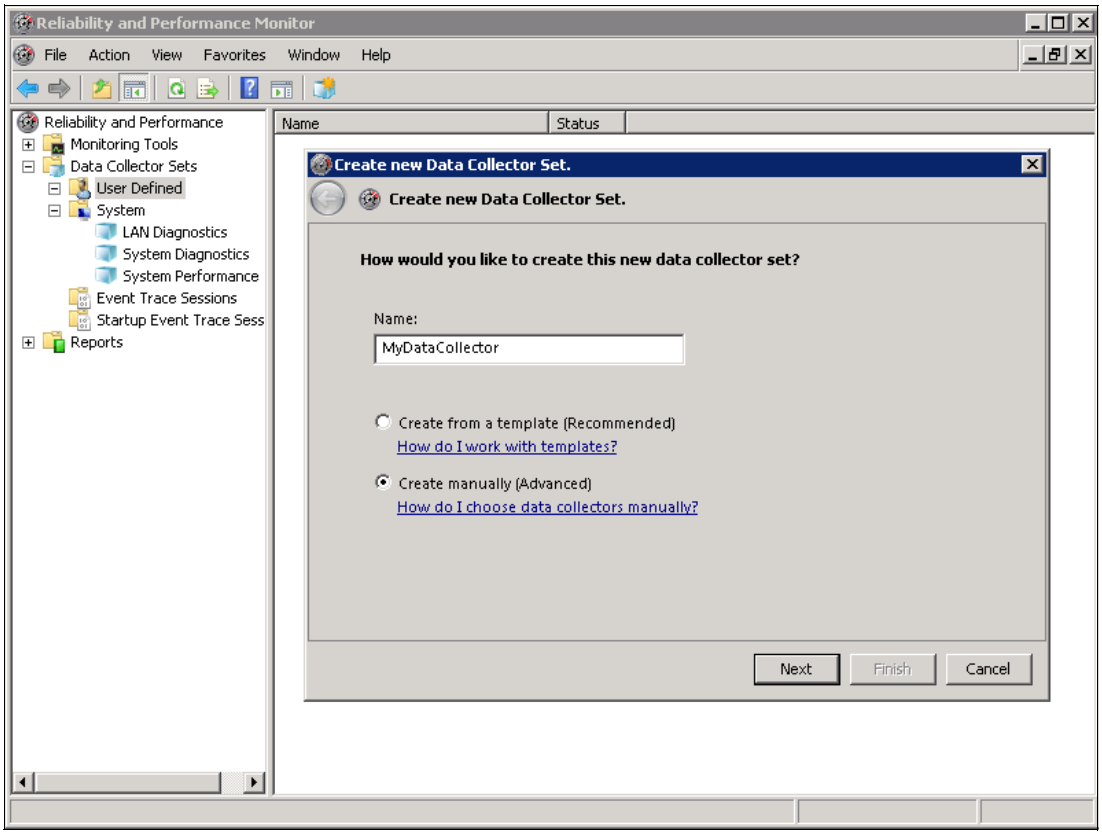


Figure 10-1 Create new Data Collector Set

In the following window, select **Create data logs** and flag the check box **Performance counter** in order to gather performance data, as shown in Figure 10-2. Then click the **Next** button.

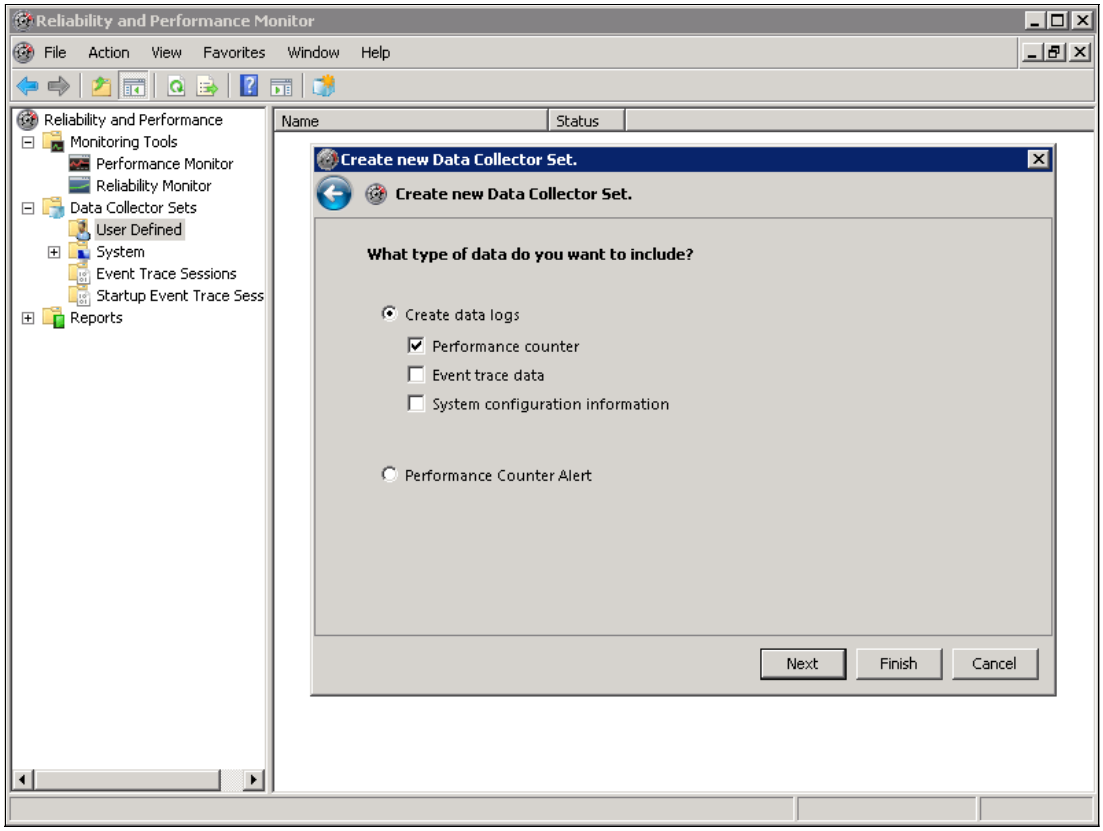


Figure 10-2 Data log creation

At this point, you select the **Sample interval** and the time measurement unit (seconds, minutes, hours, days, or weeks) as shown in Figure 10-3. Specify a logging interval that is enough to ensure that detailed accurate information is to be gathered, and then click **Add** to choose the performance counters.

Tip: Specify a logging interval of at least 5 minutes and no more than 15 minutes.

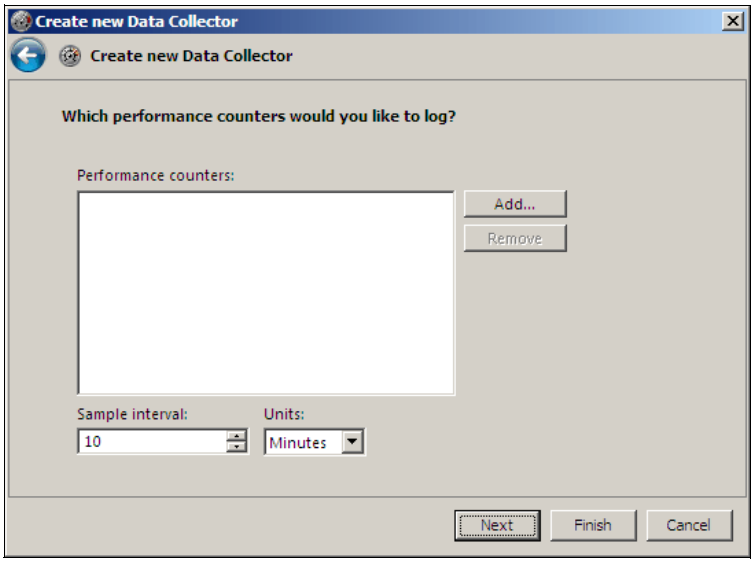


Figure 10-3 Set sample interval

In Figure 10-4 in the top left pane, you can see the list of all the available counters for your local computer or for another network computer (click the button **Browse** to discover another computer on the network).

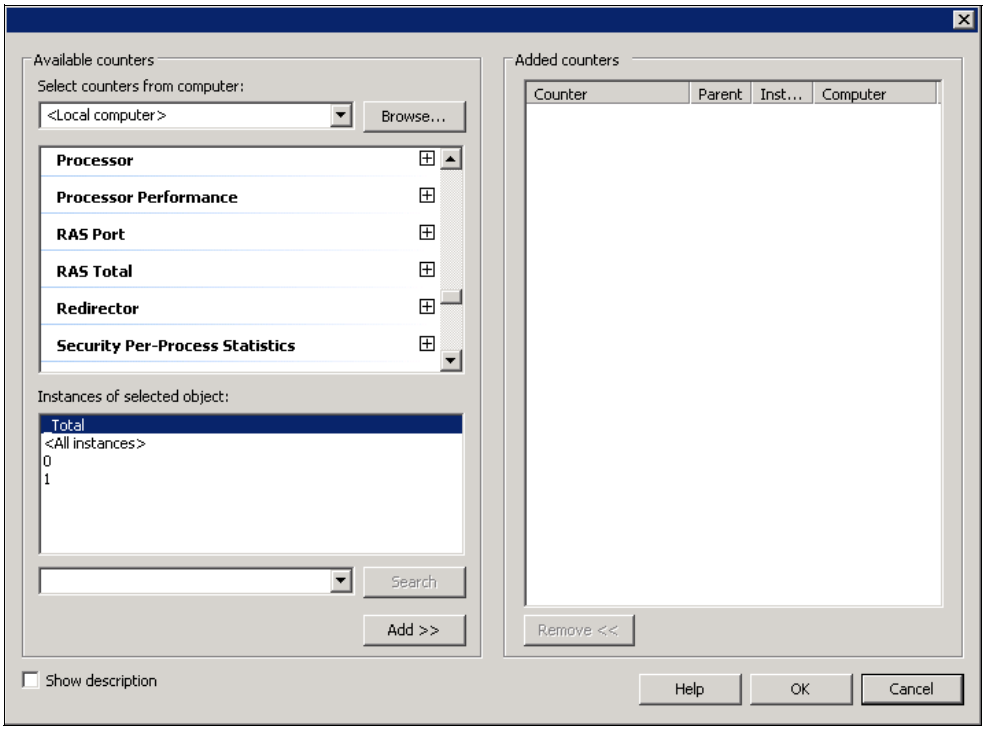


Figure 10-4 Add counters

At this point you can search the counters of concern in the top left pane.

In Figure 10-5 you can see the selection of all of the counters for the PhysicalDisk parameter. Note that we are going to choose only the external disks on the DS Storage Server and not the internal system disk.

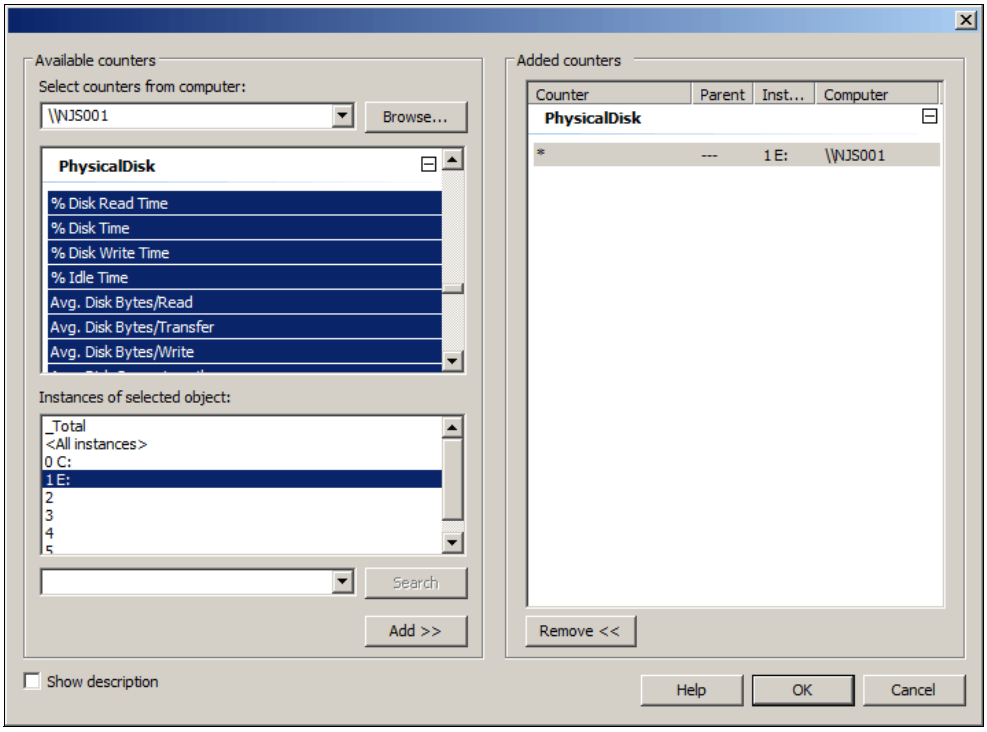


Figure 10-5 Add all counters only for external PhysicalDisk parameter

Tip: The entity that can be modeled by Disk Magic is the physical disk, that is, a Logical Drive on an external Storage Server. So when you need to use perfmon to gather I/O load statistics, you must make sure to run the perfmon procedure on each host server that accesses the Storage Server to be analyzed and to start and stop the perfmon procedure at the same time on each host server.

On the bottom left pane named **Instances of selected object** (Figure 10-5), you can select the physical disks for which the performance must be logged. They are only the disks that are managed by the external Storage Server. This selection must not include the system drive, which is normally the server's internal disk drive. In case of SAN boot, you must select the boot Logical Drive in order to check a potentially high paging activity.

You can select more than one disk, but keep in mind that only the SAN-based disks (Logical Drives) must be selected. You can also include the total, but it will be disregarded by Disk Magic because it performs its own total for the selected disks. Click **add** to include selected object instance (disk) into counters and then **OK** to return to the **Create New Data Collector Set** window. Click **Finish**.

Next, on the right pane in Figure 10-6, you select the **DataCollector01** entry, right-click and then click **Properties** to open the **DataCollector01 Properties** window (Figure 10-7).

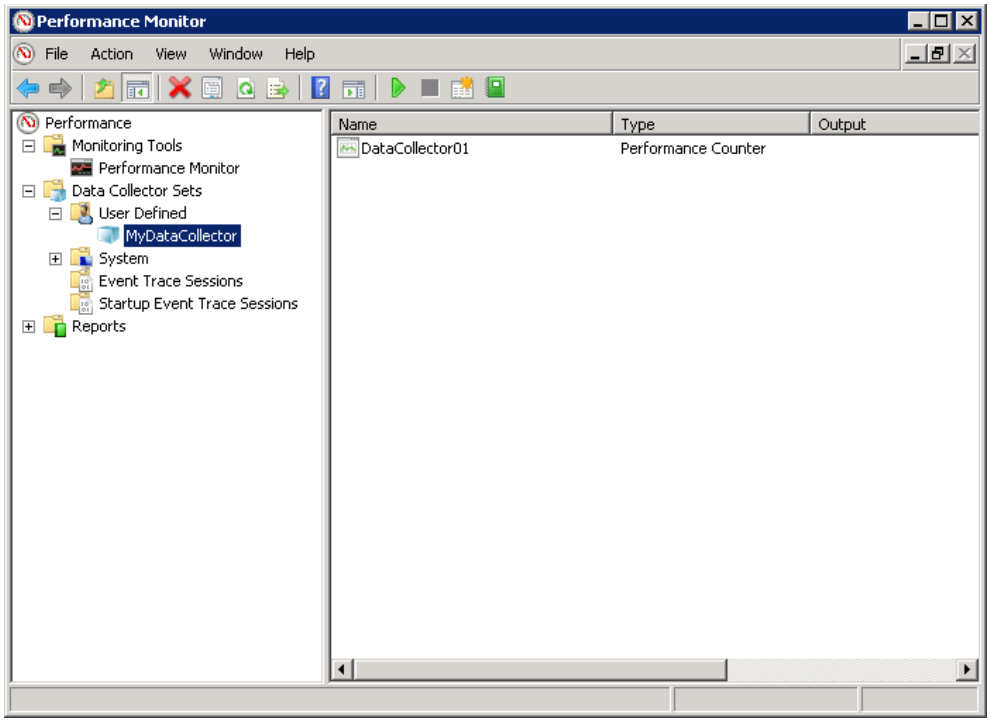


Figure 10-6 New User Defined Data Collector

In this window, under the **Performance Counters** tab, you specify the logging file format by selecting the item **Comma Separated** in the **Log format** pull-down menu as shown in Figure 10-7. This procedure is going to produce a comma delimited file (.CSV file).

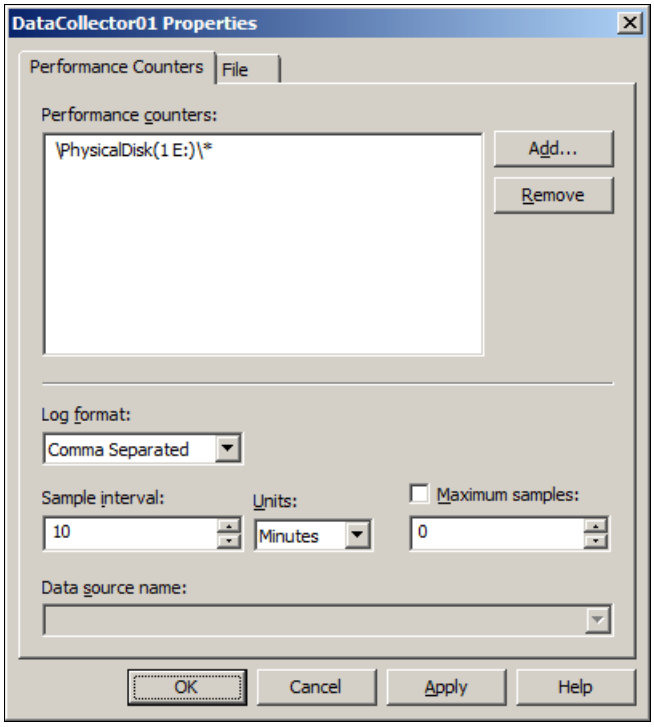


Figure 10-7 Log file format selection

Click the **File** tab and select the name of your CSV file and if necessary, the file format. Choose a meaningful format and naming convention for your CSV files (Figure 10-8).

Tip: Check the **Prefix file with computer name** check box to easily recognize the Host Server when the file is going to be imported within Disk Magic.

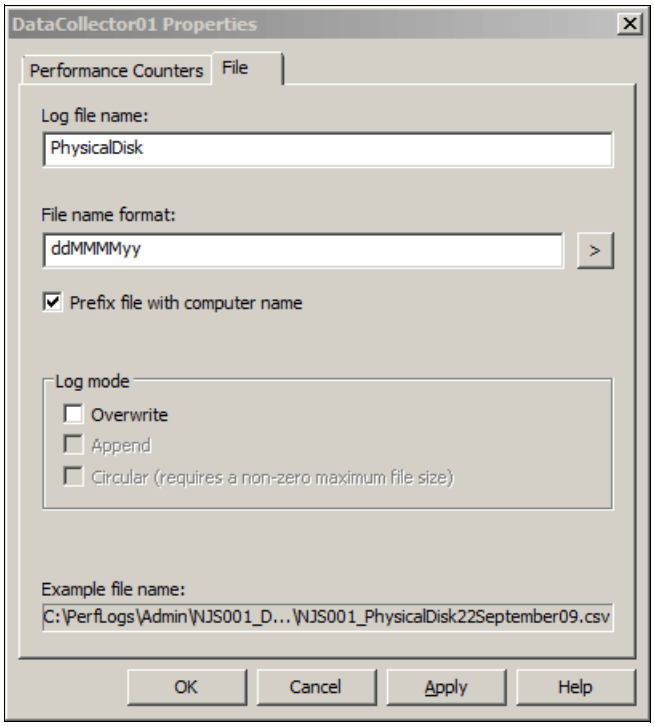


Figure 10-8 Log file format selection

Note that if a Host Server has multiple Logical Drives mapped, Disk Magic can emulate this configuration by creating a “pseudo-Host Server” for each Logical Drive mapped, thus giving the opportunity of filling specific performance data for each Logical Drive individually mapped to each pseudo-Host Server.

Tip: Within Disk Magic, we assume that there always exists a 1:1 mapping between a Host Server or Pseudo-Host Server and a Logical Drive.

We can concurrently generate a CSV file for each Logical Drive mapped onto an individual Host Server in order to have a more accurate scenario from the Disk Magic standpoint. Then you must rename the host name column inside each CSV file belonging to the same Host Server in order to stick to the real scenario in terms of Logical Drives.

Naming convention: Suppose that you have the following situation:

- You have a Windows 2008 Server with hostname NJS001 and with two LUNs mapped (named Data1 and Data2).
- You can create two CSV files named NJS001_Data1_22September2011.csv and NJS001_Data2_22September2011.csv, respectively.
- Then, within these two CSV files, you must rename each hostname’s reference with NJS001_Data1 and NJS001_Data2.

Finally, you must define the date and time the logging is to be started and stopped and which days of the week must be included in the collection, which can be easily done on the **Schedule** tab, as you can see in Figure 10-9. Ensure that perfmon is set to gather I/O load statistics on each server that accesses the storage subsystem to be analyzed. Also, make sure to start and stop the perfmon procedure at the same time on each Windows server that gets its Logical Drives from the external Storage Server.

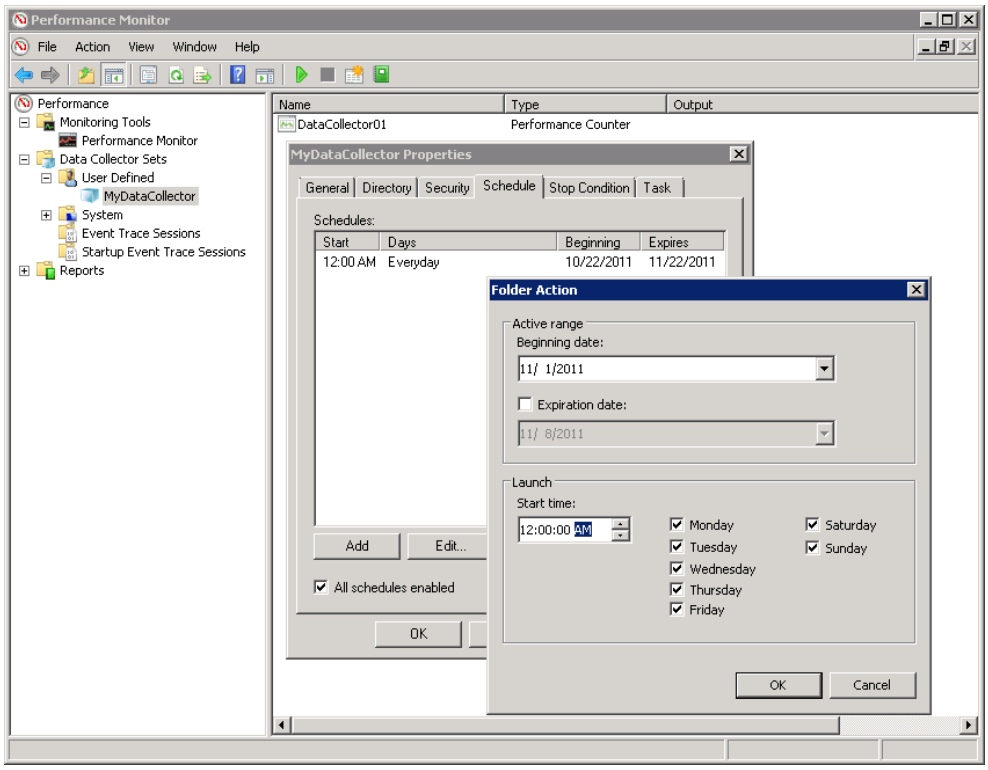


Figure 10-9 Set schedule logging range

Finally, you are ready to start the collecting process moving in the left pane of the **Performance Monitor** window, by right-clicking the name of your personal Data Collector Set (in this case, **MyDataCollector**) and then clicking the **Start** item.

In order to use this perfmon file in Disk Magic, start Disk Magic and select the radiobox **Open and iSeries Automated Input (*.IOSTAT, *.TXT, *.CSV)** as shown in Figure 10-10. As an alternative to the previous procedure, use **File > Open input for iSeries and/or Open...** from the Disk Magic main window and set the file type to *Perfmon Reports (*.csv)*. This method can be used when Disk Magic is already running and you want to start a new project.

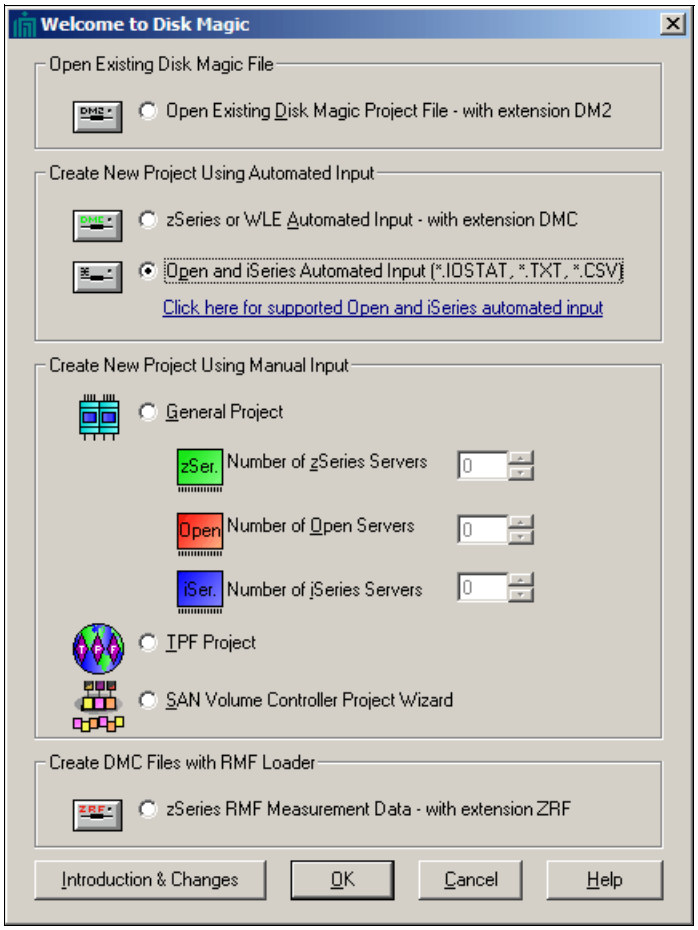


Figure 10-10 Importing Windows performance data into Disk Magic

At this point, click the **OK** button to open the *Select multiple Disk Magic input files using the Shift or Ctrl Key* window (Figure 10-11 and Figure 10-12).

This browsing window allows you to explore the local file system in order to determine and upload the Windows perfmon log files of your concern. After the files are selected, you click the **Open** button to create a new project with one or more servers (Windows, Linux, or UNIX) and one Storage Server.

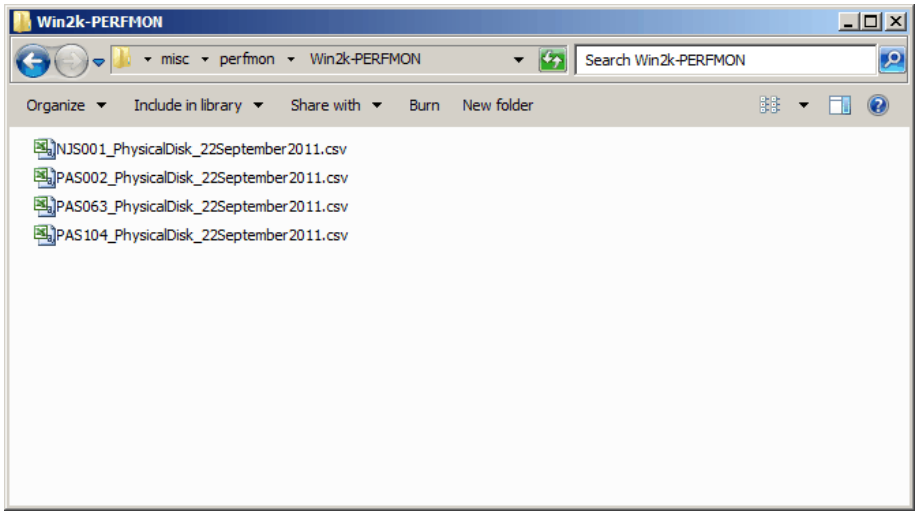


Figure 10-11 Browsing Window to select input data

Tip: Within the current Disk Magic version, it is always possible to upload multiple Windows perfmon log files or even a mix of log files coming from UNIX, Linux, and Windows. It allows you to do a comprehensive analysis of the workload sustained by all the heterogeneous Host Servers attached to the Storage Server.

It is common for a Storage Server to be attached to multiple Windows servers. To get a full view of the Storage Server's activity, the workload statistics of all servers must be taken into consideration. Disk Magic supports this through its *Multi-File Open feature* (Figure 10-12), where you can make it process the complete collection of perfmon files that relate to a single storage server. It shows a summary of the perfmon files that you requested Disk Magic to process. Using the dialog, you can verify that each of the servers is represented and see the start times and interval lengths for your servers.

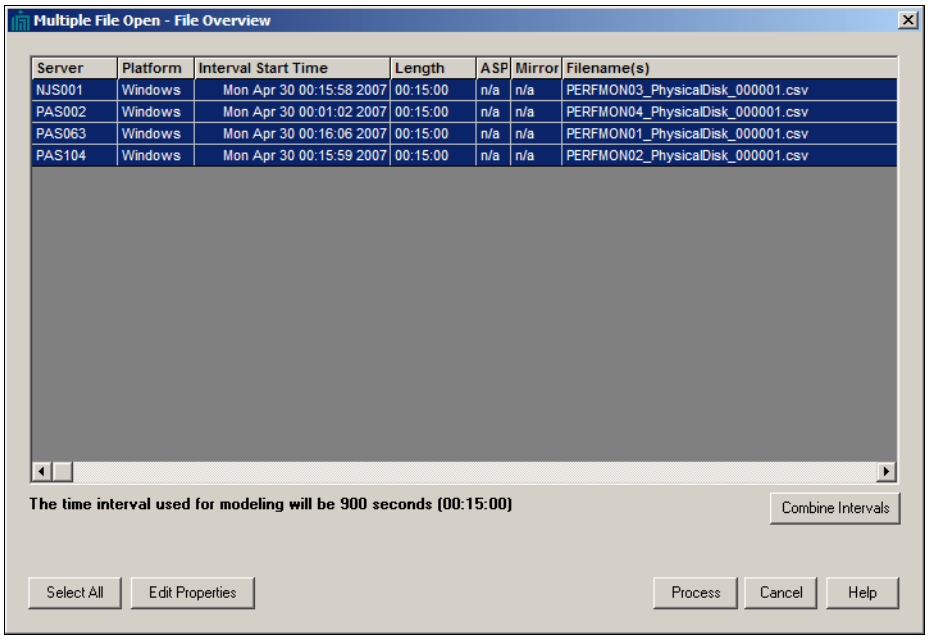


Figure 10-12 Multiple File Open

At this stage, Disk Magic did not read every file in full; it just reads enough to pick up the start time of the period for which the file was created, and the interval length. At this time, it does not know yet how many intervals are included in each file. You must use the next dialog, that is, the **I/O Load Summary by Interval** (Figure 10-13), to make sure that each file contains data for each interval. The interval lengths will typically be round values (00:10:00 or 00:15:00 or something similar). You might see unique interval lengths for particular files, for instance, 10 minutes on the UNIX servers and 5 minutes on the Windows servers. This difference will be handled automatically by Disk Magic by aggregating all data to the least common multiple of all intervals. The least common multiple is shown in the left bottom area of the dialog.

You might see start times that are slightly mismatched, for instance one measurement period started at the hour (20:00:00) and another starts at a little after the hour (20:01:03). Again, it is something that Disk Magic will handle automatically, no action from you side is required.

Disk Magic then summarizes the data and allows you to identify the interval you intend to create a model for with Disk Magic. The challenge in summarizing the data generated by multiple servers is to find the most relevant overall interval: Disk Magic finds the peak of the sums, rather than the sum of the peaks.

After you have selected the files and clicked the **Open** button, a summary table of all the gathered data samples is displayed as shown in Figure 10-13. At this point, you can click a column header to select the interval with the peak value for that column, or click a row to select a specific interval to enter.

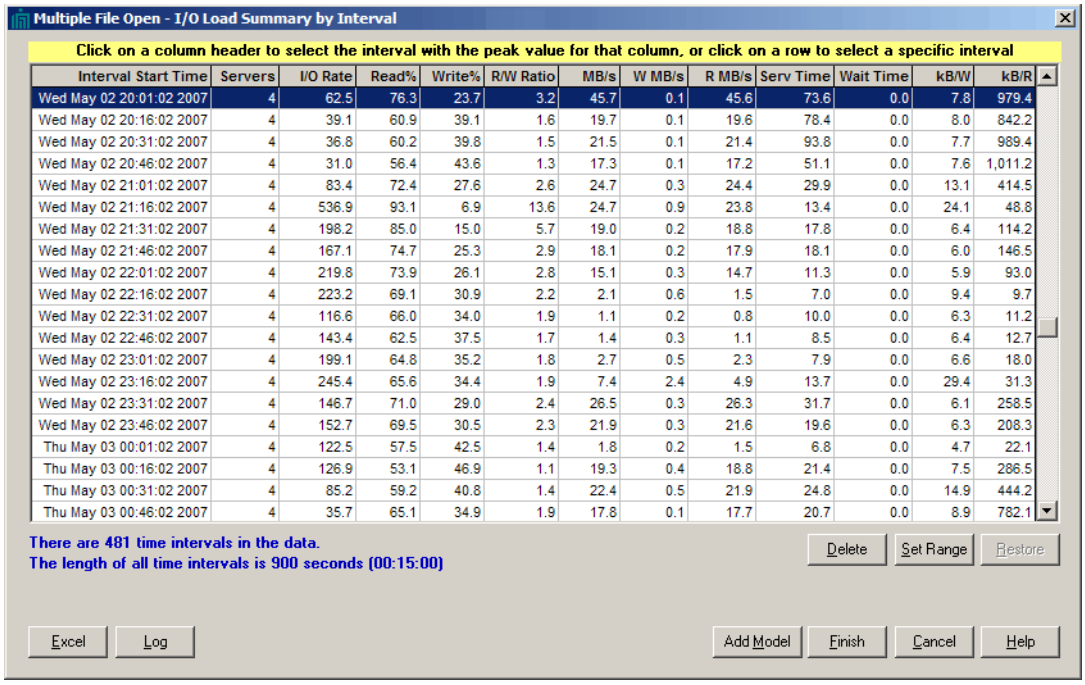


Figure 10-13 I/O Load Summary by Interval

By clicking the button **Add Model**, the perfmon file is processed and Disk Magic displays a pop-up a message: **"**`DMW1311A Is this workload currently on a Disk Subsystem type that is supported in Disk Magic? (If it is, Disk Magic will use the measured service times for calibration purposes.)`**"** where you answer **yes**, because DS5000 is fully supported in Disk Magic.

At this point, you have completed the first part of any Disk Magic study, the creation of a Baseline model. You can now proceed to model the effects of changing the hardware, interfaces, or I/O load.

Tip: Disk Magic uses the term Disk Storage Subsystem (DSS) to identify a Storage Server and, in the baseline model creation, it works out with a default Storage Server *IBM System Storage DS8100 Turbo* model. After having created the baseline model, you can easily change the Storage Server model with the model that you need.

In Figure 10-14, you can see in the left pane (called TreeView), four servers (represented by a red icon followed by the server host name) and a storage server (represented by a pale blue icon followed by DSS1). You can change the name of the Storage Server from DSS1 to a more suitable name, right-clicking the pale blue icon and selecting **rename**. As you can see in Figure 10-14, we have selected a Hardware Type IBM DS5300 and we have renamed the storage server as DS5300.

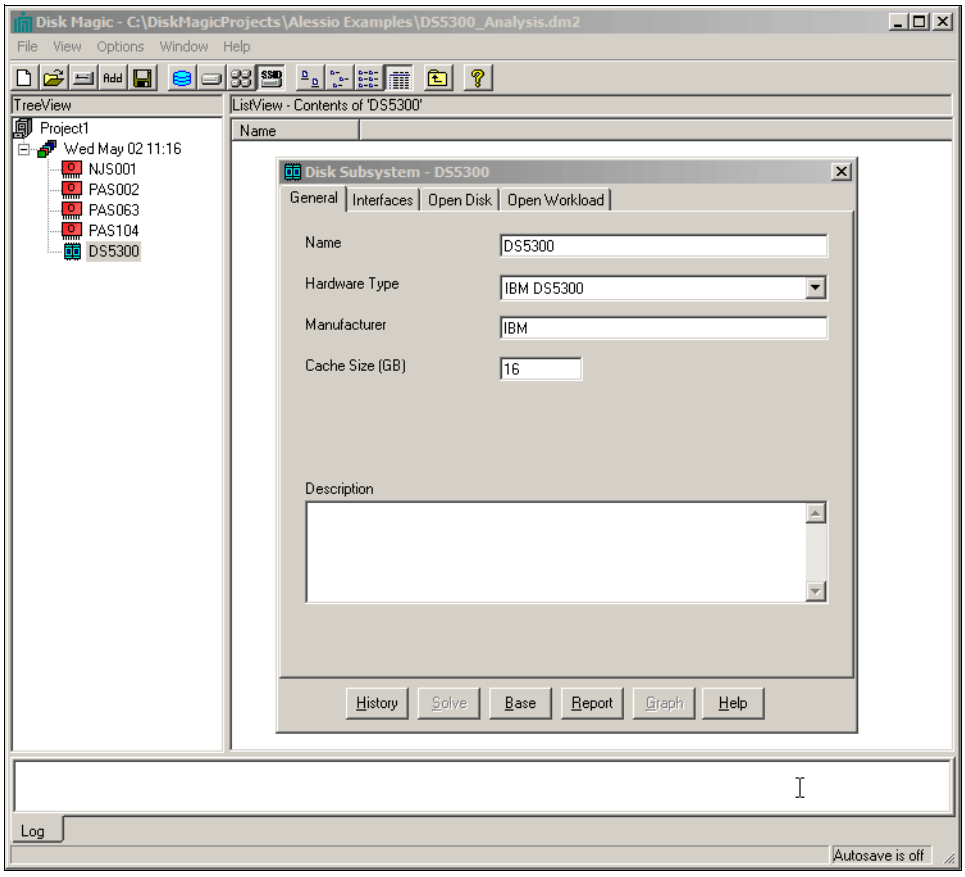


Figure 10-14 Disk Magic view showing the imported servers from PerfLog files

Because you have changed the Storage Server model from default to a DS5000 series Storage Server model, you also need to adjust the host connectivity both from the Storage Server side and from the Host Server side, in order to set the FC connectivity layout that you have in mind. For example, in Figure 10-15 under the column header “Count”, we have used sixteen 4 Gbps host ports for the Storage Server.

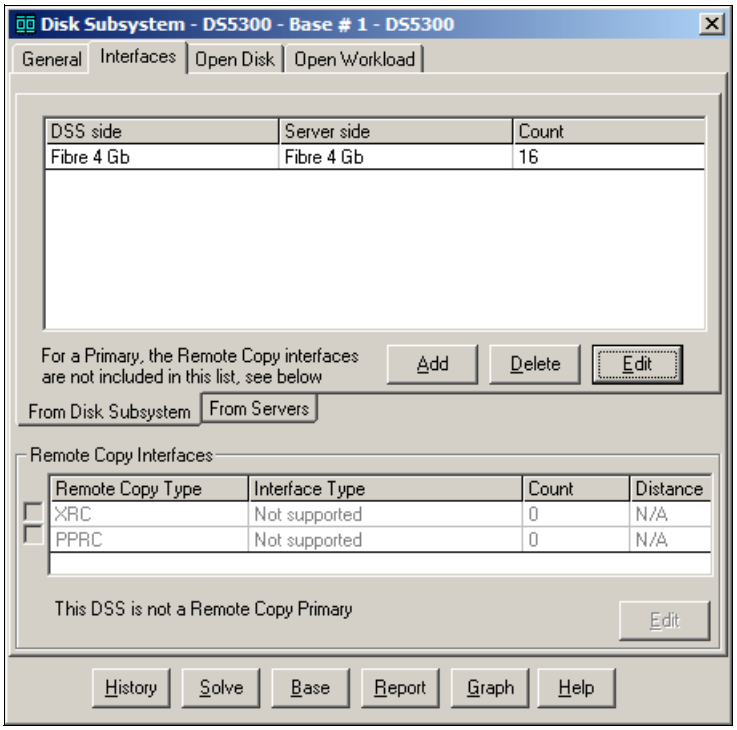


Figure 10-15 Storage server host ports

We select two 4 Gbps ports HBAs for each of the four hosts attached to the Storage Server as well (Figure 10-16).

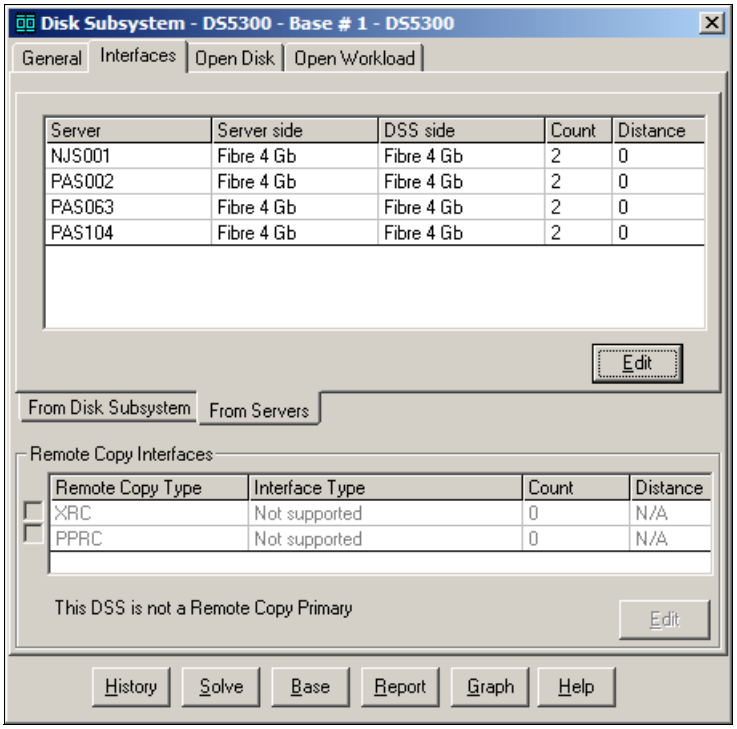


Figure 10-16 Server HBA ports

The measurement data generated with iostat or Windows perfmon is limited because it does not contain Storage Server configuration information such as Total capacity, Physical Device Type, RAID type, and HDDs per RAID array. At this point, this configuration data can be picked up from the DS storage profile or directly from the DS Storage Manager (SM). For example, Figure 10-17 shows that the Host Server *PAS104* has been mapped to a Logical drive of 250GB in a RAID 5 array composed of seven FC 300 GB 15k Hard Disk Drives (HDD).

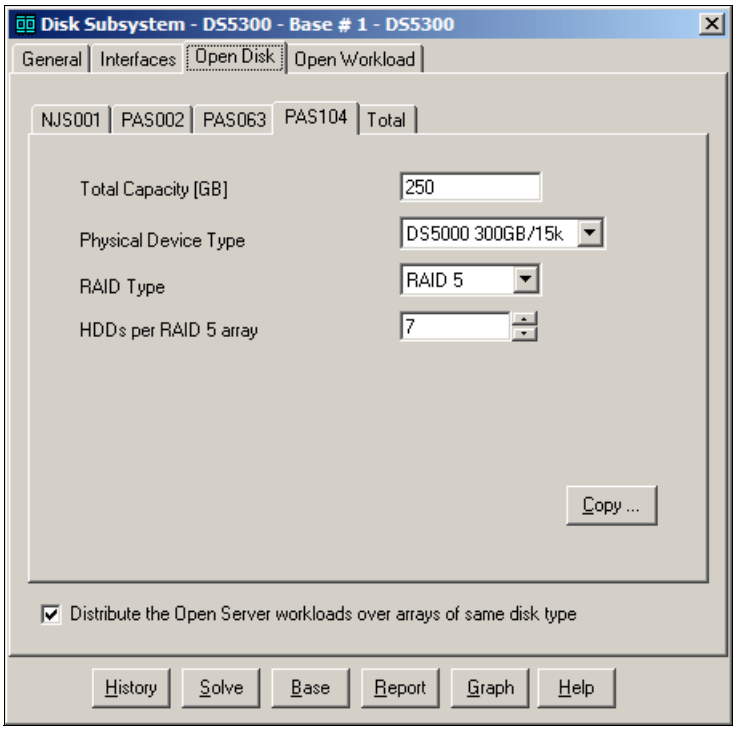


Figure 10-17 Storage server disk configuration

10.2.2 Linux and UNIX: iostat and Disk Magic

For the Linux and UNIX environments, performance data can be captured using iostat. The resulting output file (report) can be processed for use with Disk Magic.

Automated input for UNIX and Linux is supported for the following environments:

- AIX
- HP UNIX
- Sun Solaris
- Linux (Red Hat and SUSE)

The **`iostat`** command produces I/O load statistics, including MBps read and write.

Cache statistics or information about the type of storage subsystem is not included in the iostat report. The cache size must be entered manually into Disk Magic before Disk Magic can create the base line.

The iostat reports do not include information that can allow Disk Magic to identify the entity of a storage server, and therefore it is not possible for Disk Magic to separate the I/O load statistics by storage subsystem. Consequently, you must not create an iostat file that covers more than one storage subsystem.

You must make sure to run the iostat procedure on each server that accesses the storage subsystem to be analyzed. Make sure as well to start and stop the iostat procedure at the same time on each server.

The iostat automated input process is performed as follows:

1. Enter the command depending upon the operating system:
   - For AIX:

     ```
     iostat i n > servername.iostat
     ```

   - For HP UNIX:

     ```
     iostat i n > servername.iostat
     ```

   - For Sun Solaris:

     ```
     iostat -xtc i n > servername.iostat or iostat -xnp il n > systemname.iostat
     ```

   - For Linux:

     ```
     iostat -xk i n > servername.iostat
     ```

   Where:

   - *i* is the interval in seconds.
   - *n* is the number of intervals, which must always be greater than 1.
   - *servername*.iostat is the output file. Its file type must always be iostat.

   Here is a sample Linux command:

   ```
   iostat -xk 600 10 > linuxserver.iostat.
   ```

   Tip: With automated input, the number can be quite large if necessary in order to try to capture a peak time or IO activity. A day or a week at most is common (when input had to be calculated manually from the data, a guideline for the number of intervals collected was only 10 intervals at 10 minutes each). Moreover, an interval length of 10 minutes (600 seconds) works well, however, it can be anything between 5 minutes and 30 minutes.

   Make sure that the person who collects the data confirms what interval length and number of intervals were used in the command, as well as what time of day the data collection was started. The commands do not document this information in the output file.

2. When the data collection has finished, edit servername.iostat to insert two lines of header information:

   ```
   os iostat system=servername interval=i count=n
   ddmonyyyy hh:mm
   ```

   - *os* is the operating system on which the iostat performance data was collected, such as AIX, HP-UX, Sun Solaris, or Linux. It does not matter whether you use upper-case or lower-case, or a mix. Disk Magic will also accept the following permutations of the UNIX / Linux names: hp ux, hpux, sunsolaris, sun-solaris, solaris, sun, redhat, and suse.
   - *i* and *n* must be the same as in the original **iostat** command.
   - *dd*, *yyyy*, *hh*, and *mm* are numerical, and 'mon' is alphabetic and reflects the month of the date. They must reflect date and time of the first interval in the iostat gathering period.

   For example, the line to add for an Iostat Linux file is:

   ```
   redhat iostat system=linuxserver interval=600 count=10
   13nov2011 10:10
   ```

To use the resulting iostat file in Disk Magic, start Disk Magic and select the radiobox **Open and iSeries Automated Input (*.IOSTAT, *.TXT, *.CSV)** as shown in Figure 10-18, which will create a new project with one or more servers and one disk system. Later, you can add more iostat data from other servers. As an alternative to the previous option, use **File → Open input for iSeries and/or Open...** from the Disk Magic main window and set the file type to *iostat Reports (*.iostat)*. This last method can be used when Disk Magic is already running and you want to start a new project.

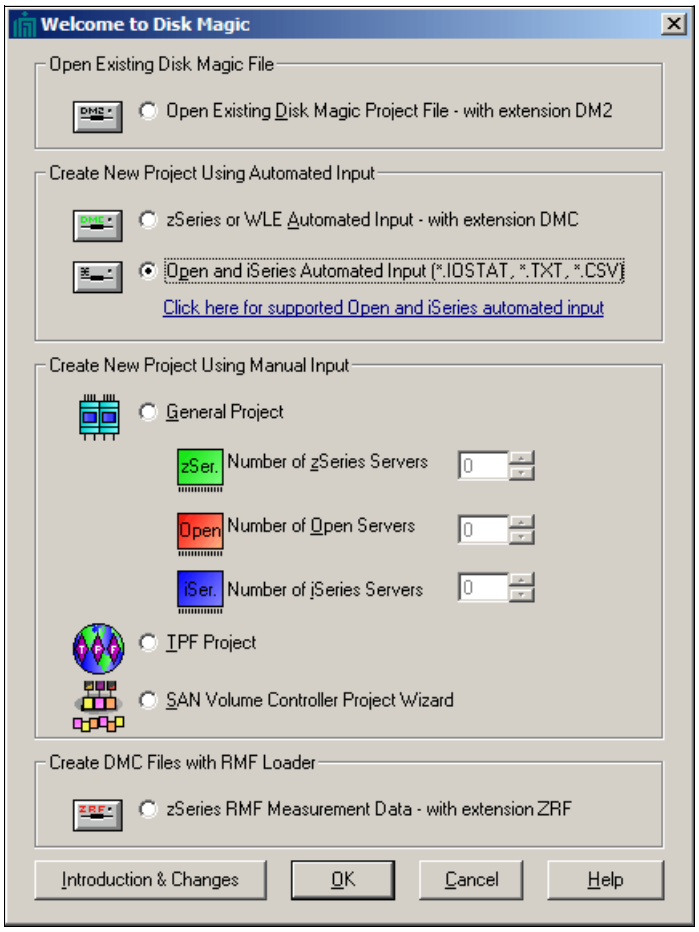


Figure 10-18 Importing an iostat file into Disk Magic

Tip: Cache statistics or information about the type of storage subsystem are not included in iostat. Disk Magic is aware of this restriction and when creating a Base from iostat data, it knows that the iostat data is accurate, but the configuration information is not. So the algorithms are designed to disregard configuration information, except for one parameter: the cache size of the Disk Subsystem. Therefore, in particular when you will enter measured cache statistics, you should make sure to enter the correct value on the Disk Subsystem dialog, General page before invoking the Base function.

10.2.3 Mixed platforms and Disk Magic

As mentioned before, it is common for a storage server to be attached to multiple UNIX, Linux, or Windows servers. To get a full view of the storage subsystem's activity, the workload statistics of all servers (Linux, UNIX, Windows) need to be taken into consideration.

Disk Magic supports this activity through its Multi-File Open feature. You can make it process the complete collection of iostat files and perfmon files that relate to a single storage server. The Multiple File Open-File Overview shows a summary of the iostat and perfmon files that you requested Disk Magic to process (Figure 10-19).

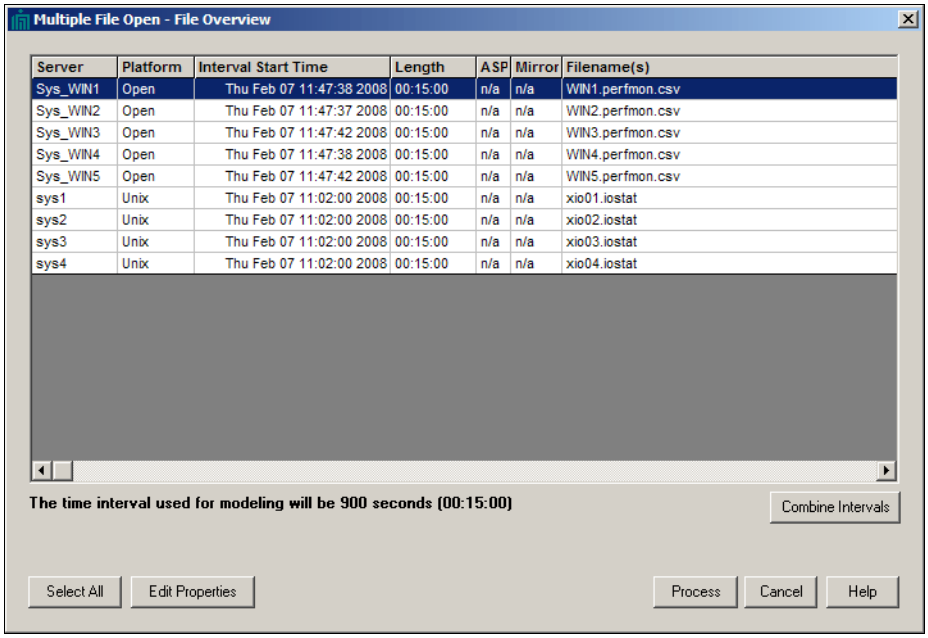


Figure 10-19 Multiple File Open for Mixed Environment

In the Multiple File Open - File Overview window shown in Figure 10-19, you can verify that each of the servers is represented and see the start times and interval lengths for your servers.

At this stage, Disk Magic did not read every file in full; it just reads enough to pick up the start time of the period for which the file was created, and the interval length. At this time, it does not know yet how many intervals are included in each file. You must use the next dialog, the I/O Load Summary by Interval, to make sure that each file contains data for each interval.

The interval lengths will typically be round values (00:10:00 or 00:15:00 or something similar). You might see unique interval lengths for particular files, for instance, 10 minutes on the UNIX servers and 5 minutes on the Windows servers. This difference will be handled automatically by Disk Magic by aggregating all data to the least common multiple of all intervals. The least common multiple is shown in the left bottom area of the dialog.

You might see start times that are slightly mismatched, for instance one measurement period started at the hour (20:00:00) and another starts at a little after the hour (20:01:03). Again, it is something that Disk Magic will handle automatically, no action from your side is required.

For perfmon data, there might be a deviation in the interval length, because of the way perfmon writes the intervals. You must verify that you are seeing only round values for all lines in the File Overview. When a server has an interval length that is not a round value, such as 00:14:21 instead of 00:15:00, then select that line and press Edit Properties to change it to the normalized value. For instance, when there are 5 Windows servers and 4 of them show an interval length of 00:15:00 but one shows an interval length of 00:14:43, then you need to change the last one to 00:15:00 (Figure 10-20).

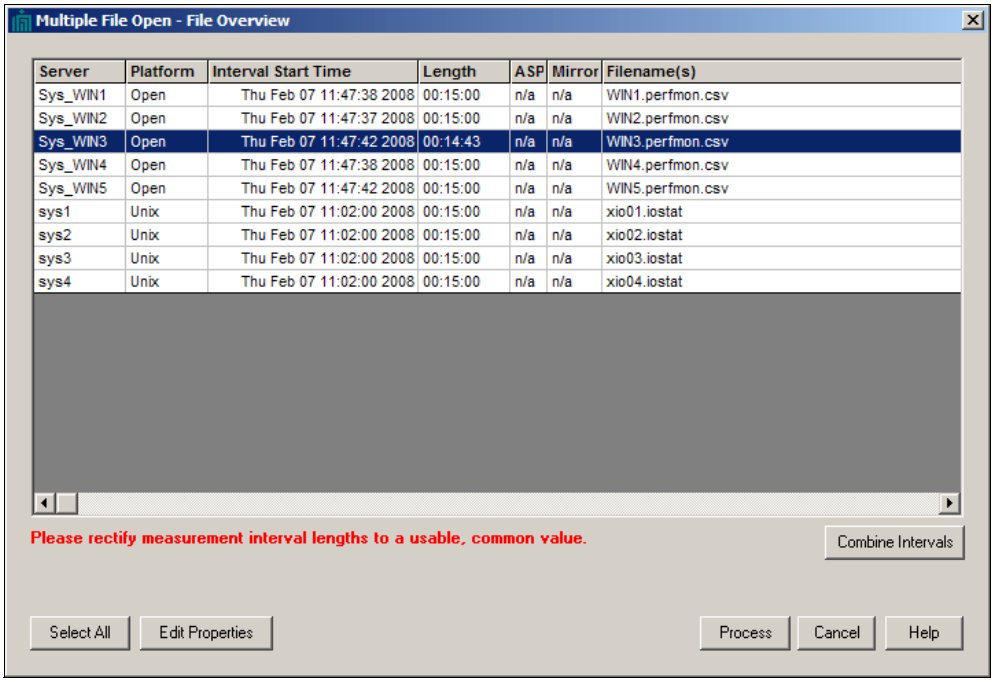


Figure 10-20 Windows Server with a wrong interval length value

To rectify the wrong interval length value, you need to double-click the related row in the **Multiple File Open - File Overview windows**. Then a new pop-up window will appear as shown in Figure 10-21, and now you can round up the Interval Length to the right value, that is, 15 minutes in our case.

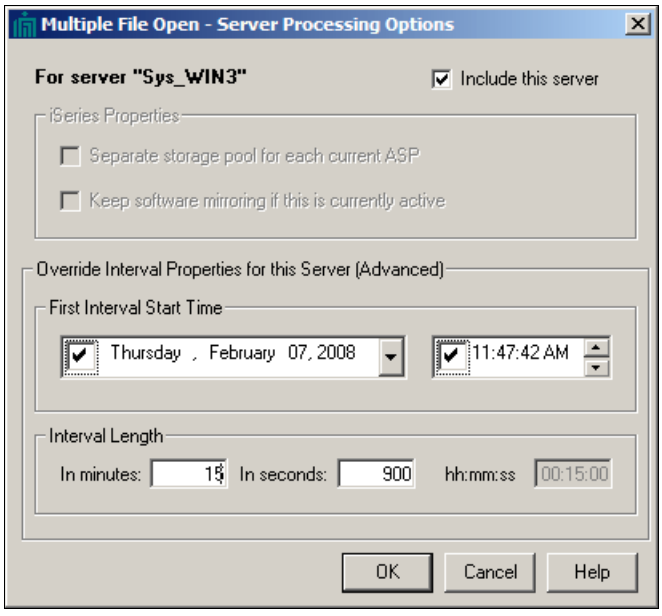


Figure 10-21 Server Processing Options

10.3 Disk Magic configuration example

To illustrate the use of Disk Magic, consider the following environment:

- A DS5300, with EXP5000 enclosures, is used.
- All arrays need to have enclosure loss protection.
- Five separate hosts (three of them are Windows based, whereas the other two are Linux based) can access the DS5300 through SAN, and each host is equipped with two 4 Gbps HBAs.

As they are Windows hosts, the host type will be defined as *open* in Disk Magic. The statistics gained from the perfmon files are used in Disk Magic as a base line.

The hosts are as follows:

- Host 1: database server
  - It requires 350 GB of RAID 5 with 146 GB high-speed drives.
  - It has an expected workload of 100 I/Os per second with a 16 K transfer size.
  - Read percentage is expected to be 63%.
- Host 2: file and print server
  - It requires at least 1 TB of storage of RAID 5 with 450 GB on 15K drives.
  - It has an expected workload of 50 I/Os per second with an 8 K transfer size.
  - Read percentage is expected to be 60%.
- Host 3: database server
  - It requires 500 GB of RAID 5 with 146 GB high-speed drives.
  - It has an expected workload of 150 I/Os per second with a 4 K transfer size.
  - Read percentage is expected to be 60%.
- Host 4: email server
  - It requires 350 GB of RAID 10 with73 GB high-speed drives.
  - It has 500 users, and an expected I/O load of 1 I/O per second per mailbox.
  - Its expected workload is 550 I/Os per second with a 4 K transfer size.
  - Read percentage is expected to be 50%.
- Host 5: database server
  - It requires 400 GB of RAID 10 with 146 GB high-speed drives.
  - It has an expected workload of 150 I/Os per second with an 8 K transfer size.
  - Read percentage is expected to be 65%.

We create a new project in Disk Magic (Figure 10-22). We initially upload only one Windows server using the procedure described in “Windows: perfmon and Disk Magic” on page 428.

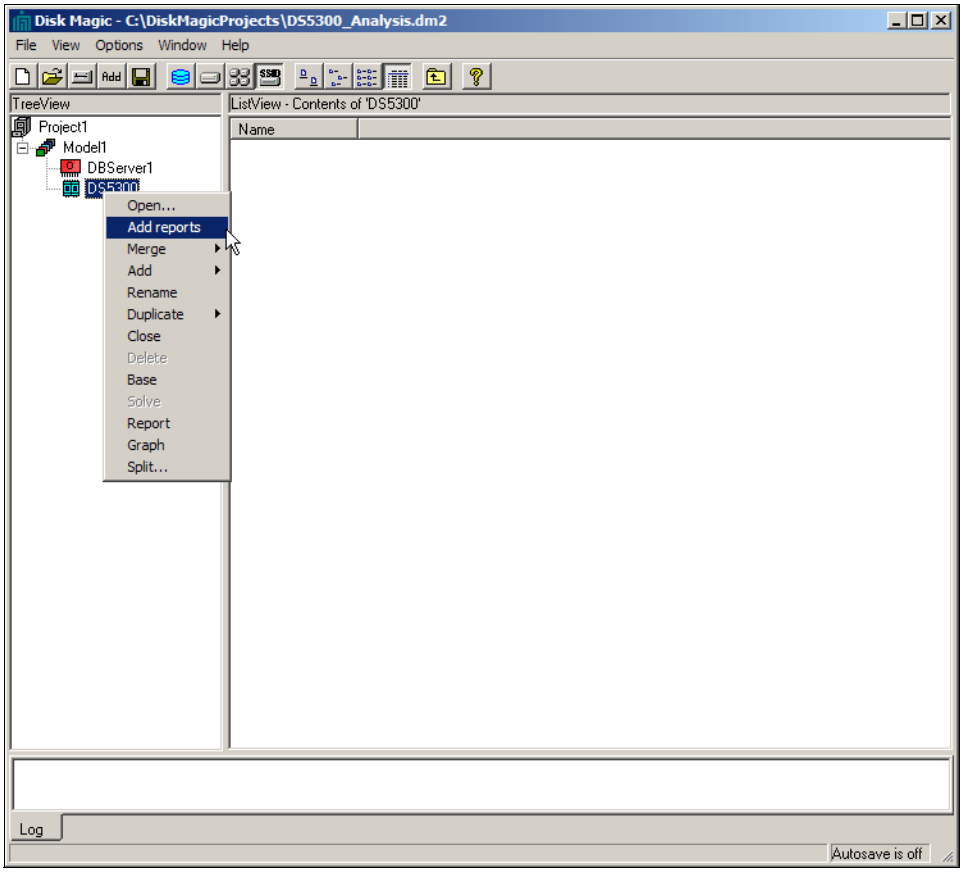


Figure 10-22 Add additional performance counters

As you can see in Figure 10-22, the main window is divided in three panes:

- The *TreeView* displays the structure of a project with the servers and hosts that are included in a model. We started by specifying just one storage subsystem (DS5300) and one open server (DBServer1).
- The *ListView* shows the content of any entity selected in the TreeView.
- The Browser is used to display reports and also informational and error messages.

To add the remaining servers and their collected performance data, right-click the storage subsystem. From the drop-down menu, select **Add Reports** (Figure 10-22 on page 447). A browsing window allows to select multiple Disk Magic input files (perfmon and iostat) using the Shift or Ctrl key (Figure 10-23).

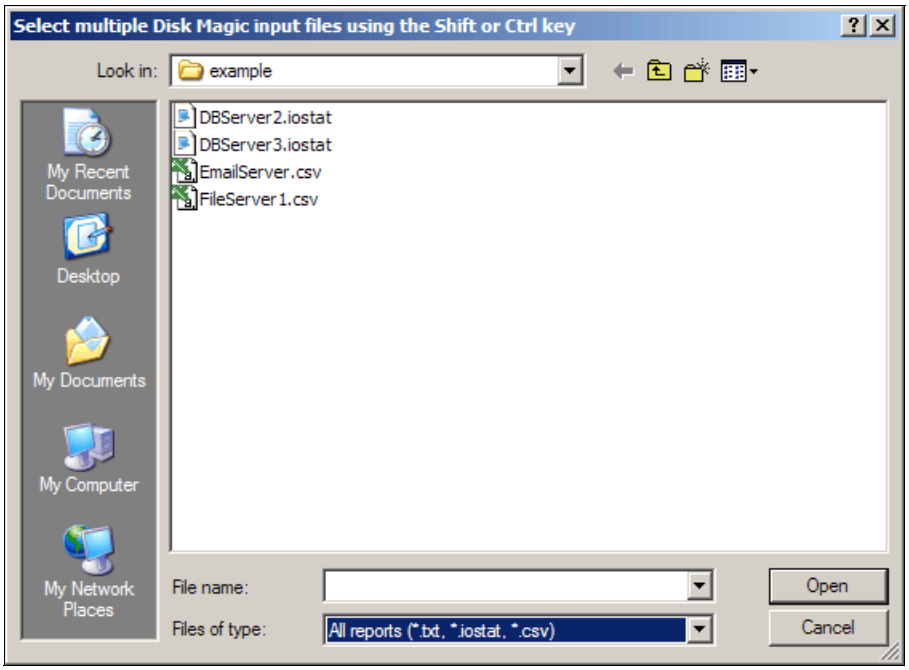
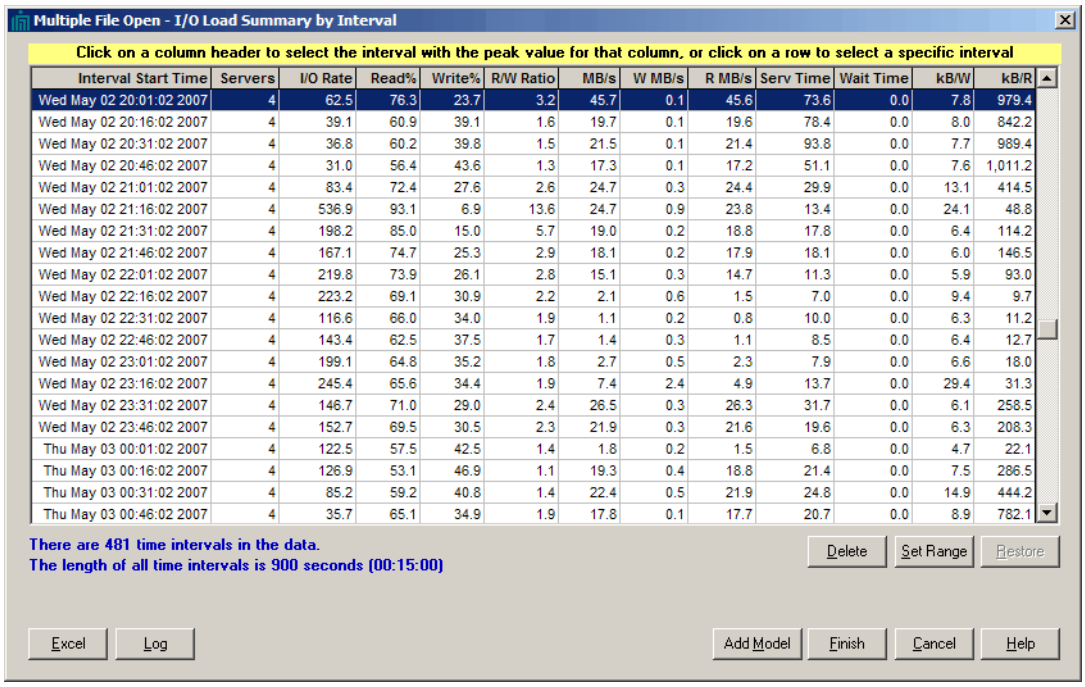


Figure 10-23 Add additional performance monitor files (Linux and Windows)

At this point the **Multiple File Open - File Overview windows** opens. Here, you select the four rows, and you click **process** in order to gather all the performance data coming from the four hosts (Figure 10-24).



Interval Start Time	Servers	I/O Rate	Read%	Write%	R/W Ratio	MB/s	W MB/s	R MB/s	Serv Time	Wait Time	kB/W	kB/R
Wed May 02 20:01:02 2007	4	62.5	76.3	23.7	3.2	45.7	0.1	45.6	73.6	0.0	7.8	979.4
Wed May 02 20:16:02 2007	4	39.1	60.9	39.1	1.6	19.7	0.1	19.6	78.4	0.0	8.0	842.2
Wed May 02 20:31:02 2007	4	36.8	60.2	39.8	1.5	21.5	0.1	21.4	93.8	0.0	7.7	989.4
Wed May 02 20:46:02 2007	4	31.0	56.4	43.6	1.3	17.3	0.1	17.2	51.1	0.0	7.6	1,011.2
Wed May 02 21:01:02 2007	4	83.4	72.4	27.6	2.6	24.7	0.3	24.4	29.9	0.0	13.1	414.5
Wed May 02 21:16:02 2007	4	536.9	93.1	6.9	13.6	24.7	0.9	23.8	13.4	0.0	24.1	48.8
Wed May 02 21:31:02 2007	4	198.2	85.0	15.0	5.7	19.0	0.2	18.8	17.8	0.0	6.4	114.2
Wed May 02 21:46:02 2007	4	167.1	74.7	25.3	2.9	18.1	0.2	17.9	18.1	0.0	6.0	146.5
Wed May 02 22:01:02 2007	4	219.8	73.9	26.1	2.8	15.1	0.3	14.7	11.3	0.0	5.9	93.0
Wed May 02 22:16:02 2007	4	223.2	69.1	30.9	2.2	2.1	0.6	1.5	7.0	0.0	9.4	9.7
Wed May 02 22:31:02 2007	4	116.6	66.0	34.0	1.9	1.1	0.2	0.8	10.0	0.0	6.3	11.2
Wed May 02 22:46:02 2007	4	143.4	62.5	37.5	1.7	1.4	0.3	1.1	8.5	0.0	6.4	12.7
Wed May 02 23:01:02 2007	4	199.1	64.8	35.2	1.8	2.7	0.5	2.3	7.9	0.0	6.6	18.0
Wed May 02 23:16:02 2007	4	245.4	65.6	34.4	1.9	7.4	2.4	4.9	13.7	0.0	29.4	31.3
Wed May 02 23:31:02 2007	4	146.7	71.0	29.0	2.4	26.5	0.3	26.3	31.7	0.0	6.1	258.5
Wed May 02 23:46:02 2007	4	152.7	69.5	30.5	2.3	21.9	0.3	21.6	19.6	0.0	6.3	208.3
Thu May 03 00:01:02 2007	4	122.5	57.5	42.5	1.4	1.8	0.2	1.5	6.8	0.0	4.7	22.1
Thu May 03 00:16:02 2007	4	126.9	53.1	46.9	1.1	19.3	0.4	18.8	21.4	0.0	7.5	286.5
Thu May 03 00:31:02 2007	4	85.2	59.2	40.8	1.4	22.4	0.5	21.9	24.8	0.0	14.9	444.2
Thu May 03 00:46:02 2007	4	35.7	65.1	34.9	1.9	17.8	0.1	17.7	20.7	0.0	8.9	782.1

Figure 10-24 I/O Load Summary for the further four hosts added later

By clicking the column header I/O rate in order to select the aggregate I/O peak, clicking **Add Model** and then **Finish** on the **I/O Load summary by Interval** window (Figure 10-24 on page 448), you create a base line model and so are able to do analysis on the whole solution.

Tip: Keep in mind that when you process multiple files and you select a specific column header in the **I/O Load summary by Interval** window, Disk Magic figures out the peak of the sum and not the sum of the peaks for the metric selected.

If you double-click the storage subsystem, the Disk Subsystem dialog opens. It contains the following tabs:

- The *General* tab allows you to enter storage subsystem level configuration data, such as the brand and type of subsystem and cache size. You can rename the storage subsystem and set the appropriate type, as shown in Figure 10-25.

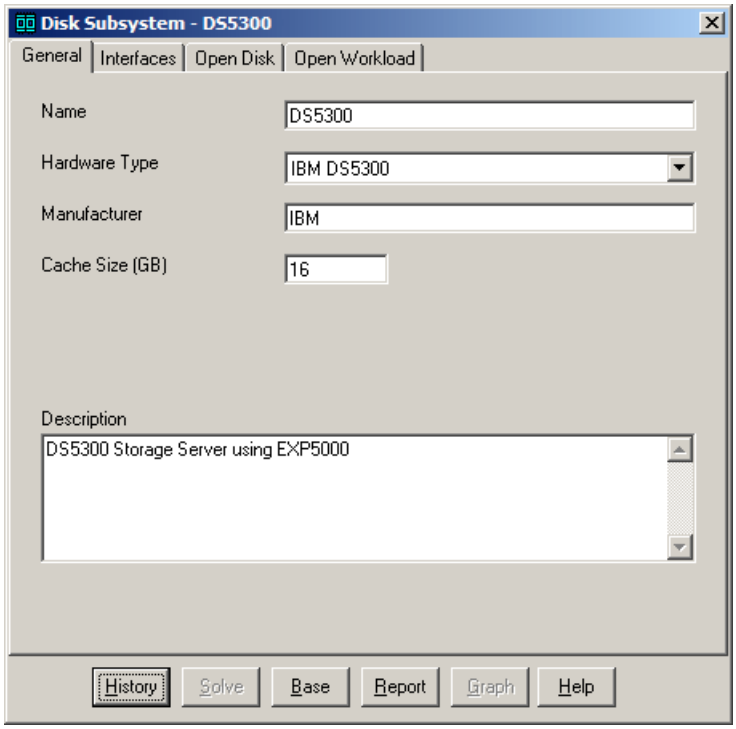


Figure 10-25 Storage subsystem General tab

- The *Interfaces* tab is used to describe the connections from the servers and the storage subsystem. As shown in Figure 10-26, you can specify the speed and the number (Count) of interfaces. In this example, we have selected 4 Gbps as data rate, but you can select other speeds as well, by selecting one or more rows and clicking the **Edit** button.

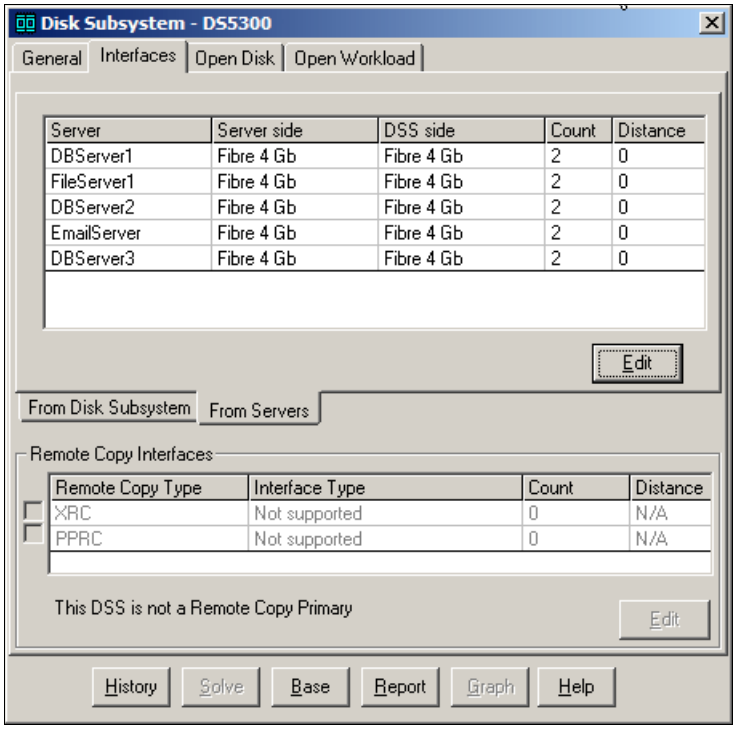


Figure 10-26 Host Interfaces

Then, you might want to change the number of host ports on the storage server connected to the hosts. In this case, you need to select the **From Disk Subsystem** tab and then make the change as desired (Figure 10-27).

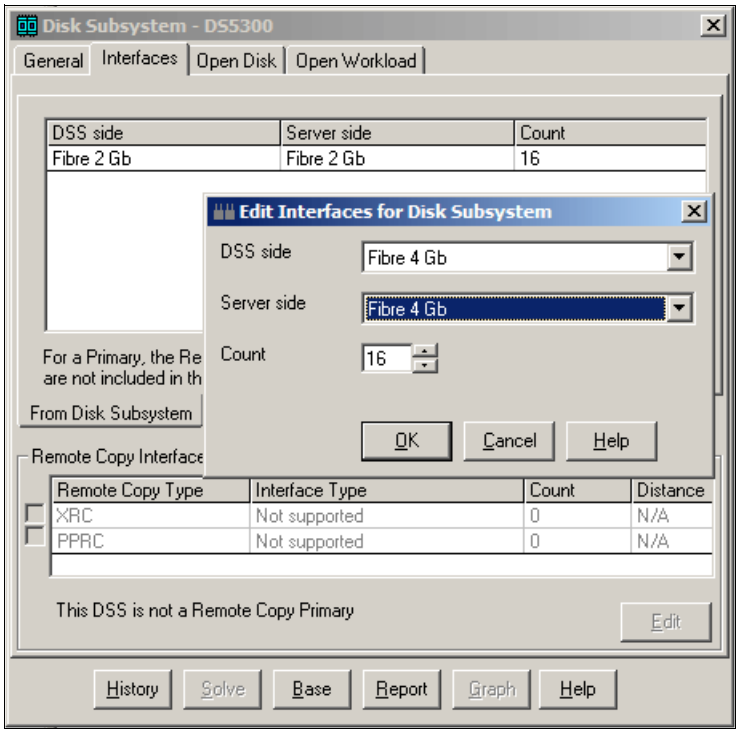


Figure 10-27 Change host-port connectivity in the storage server

- The Open Disk page, shown in Figure 10-28, is used to describe the physical disk and Logical Drive (LUN) configuration for each host.

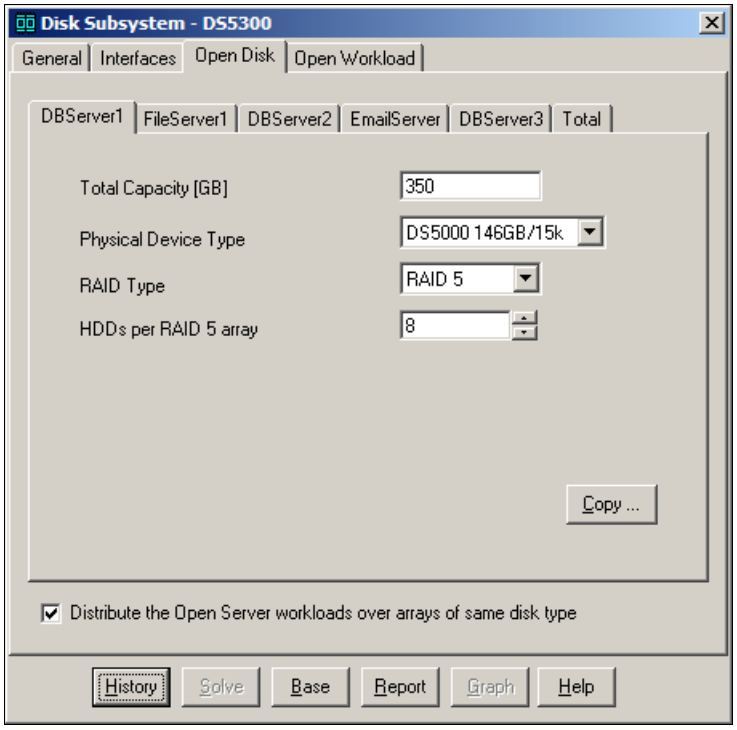


Figure 10-28 Set the disk size, number of disks, and RAID level

As you can see in Figure 10-28, Disk Magic creates a tab in the Open Disk page for each open server that has been defined in the model.

Note that you can only define one type of physical disks and one LUN for each host. If you choose RAID 5, you will be asked to specify the number of devices in each RAID 5 array. For instance, if your RAID 5 arrays will consist of 8 disks, that is, 7 data disks plus 1 parity, you specify 8. Disk Magic computes the total capacity for all servers on the Total page and tells you how many physical disks of which type will be required. This number includes parity (RAID 5) or mirror (RAID 1 / RAID 10) disks, but excludes spare disks.

Tip: For servers with more than one LUN, you will need to define another host for each additional LUN. The workload for these pseudo servers needs to be added manually under the pseudo host name, because an automatic performance file import might add the statistics to the actual existing host.

Tip: The total needed number of disks is computed based on the physical capacity of the disks (for example, 146 GB), from which space used by DS4000/DS5000 model for configuration information, called DACStore, is subtracted. The DACStore space requirements are unique for various DS4000/DS5000 models, so the number of required disks might differ slightly when you move from one DS4000/DS5000 model to another, even if you do not change the disk type.

On the window, you can see the check box, **Distribute Open workload over arrays of same disk type.**

- When not selected, Open capacity will be sequentially allocated per RAID array, starting with a new array for each server and using as many RAID arrays as is required to hold the specified capacity. If less than a full array's capacity is needed, the remaining capacity will be assumed to remain unused. Another effect is that the load on certain arrays is likely to be much higher than on other arrays, which is a major performance disadvantage, which can result in one RAID array being reported as being over 100% busy although other arrays still have a very low load.
- When selected, capacity will be distributed over all available RAID arrays with the requested DDM type. Now RAID arrays will be used to their capacity and the workload will be equally distributed over all available arrays, which is obviously the default.

- The Open Workload page is used to enter the I/O load statistics for each server in the model.

  All of the data elements are entered at the subsystem level. I/O rate or MBps are required. Importing the performance files will automatically populate these fields. Note that it is also possible to select the check box **Mirrored Write Cache** in order to mirror every write operation toward a specific LUN from controller owner's cache to the other controller's cache (Figure 10-29).

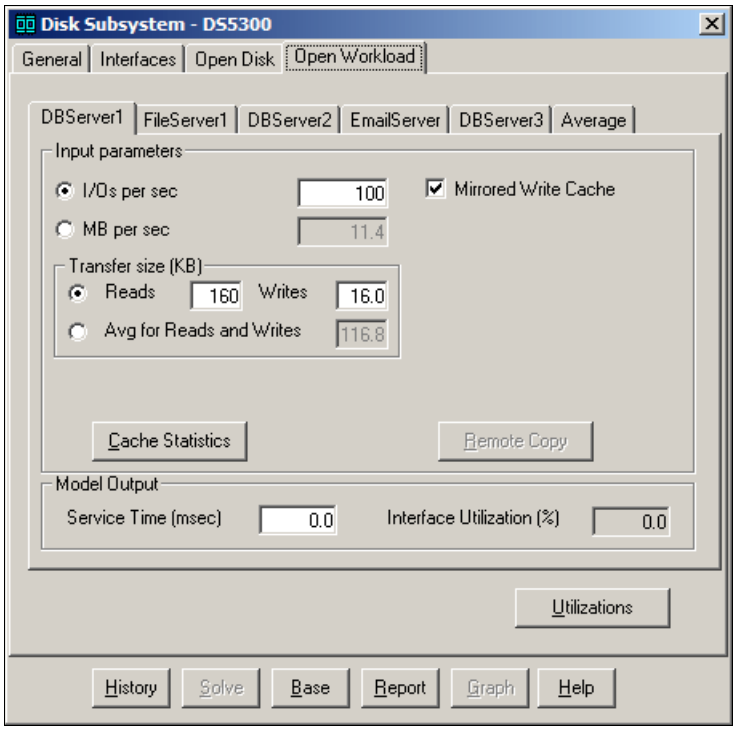


Figure 10-29 Set the workload per host

Optionally, click the **Cache Statistics** button to enter the cache statistics as shown in Figure 10-30.

Attention: You must not change the *Read* and *Write* percentages if you used automated input, because iostat (except HP-UX iostat) and perfmon contain precise information about read and write rates.

If you have measured cache statistics for the storage subsystem being analyzed, de-select **Calibrate based on Service Time** and **Calibrate based on Transfer Size** to enter those statistics manually.

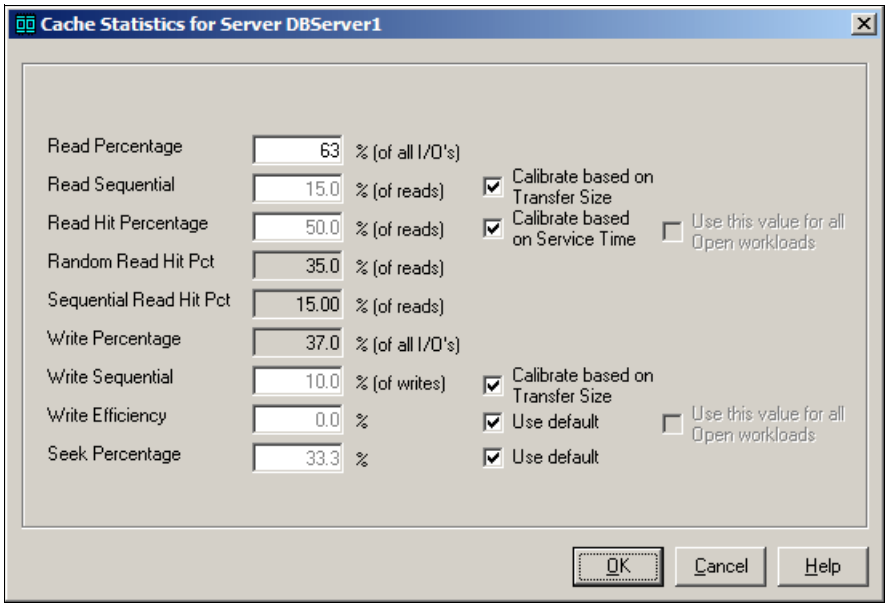


Figure 10-30 Adjust cache statistics

The precision of the model can be substantially improved if you have measured cache statistics that can be entered on the **Cache Statistics for Server** window, which is relevant when you enter all data for the model manually and also when you use an automated input procedure like in this case.

The measurement data generated on these platforms (iostat or perfmon) is limited; for instance, there is virtually no configuration information and there are no cache statistics, which makes it more difficult for Disk Magic to establish a fine-tuned calibrated model. However, Disk Magic can use the available statistics to compute a reasonable cache read hit ratio and to set the percent of read and write sequential activity based on Table 10-1.

Table 10-1 Disk Magic Automatic Percent Sequential

KB per I/O	Percent Sequential
0	
4	5
8	10
12	15
16	20
20	25
24	30
28	35
32	40
36	45
40	50
44	55
48	60
52	65
56	70
60	75
64	80
68	85
72	90
76	95
80	100

The **Write Efficiency** parameter is a number representing the number of times a track is written to before being destaged to the disk. 0% means a destage is assumed for every single write operation, a value of 50% means a destage occurs after the track has been written to twice. 100% write efficiency means that the track is written to over and over again and there are no actual destages to the disk, an unlikely situation. The default is a worst case 0%. This value is probably conservative, if not pessimistic, for an open systems workload. When you enter your own value, rather than accepting the default, then you can request Disk Magic to use your value for all servers by check-marking the selection: **Use this value for all open systems workloads**.

The Write Efficiency is not computed by Disk Magic, simply because iostat and perfmon data do not provide any variables that can be used as input to an algorithm. However, from a modeling perspective, write efficiency is an important parameter because it can have a significant influence on the utilization of the physical disks. The Disk Magic default of 0% will in many cases be too pessimistic, so if Disk Magic cannot calibrate a workload, then you might consider changing the write efficiency to a higher value, which is especially effective when the I/O load incorporates many writes.

After the initial data we just described has been entered into Disk Magic, click **Base** in the Disk Subsystem dialog window to create a baseline for the configuration.

Two other parameters need to be considered for tuning the system as well as possible:

Write Sequential:

This parameter refers to the percentage of back-end writes that is sequential. This percentage can be computed by Disk Magic based on transfer size, in which case Calibrate based on Transfer Size check box must be checked.

Seek percentage (%):

The Seek Percentage indicates for which percentage of the I/O requests the disk arm needs to be moved from its current location. This parameter is mostly a legacy of how storage subsystems used to work a long time ago and has now turned into a tuning option for advanced users. Accept the default.

A message, as shown in Figure 10-31, confirms that the base was successfully created.

Figure 10-31 Base created

After a base is created, you can examine the resource utilizations by clicking **Utilizations** on the Workload page. The *Utilizations IBM DS5300* dialog opens, as shown in Figure 10-32.

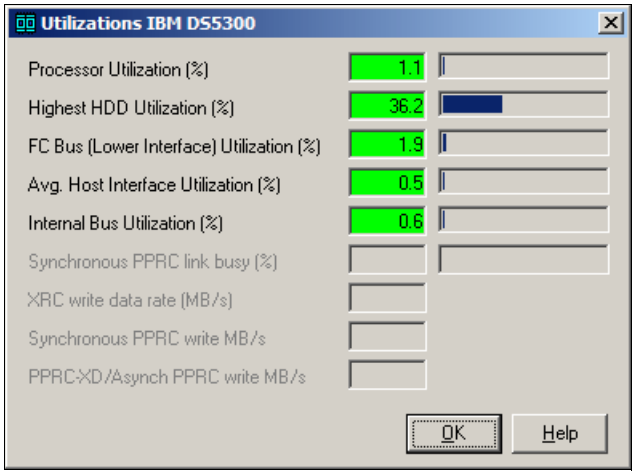


Figure 10-32 Utilization of the DS5300 Storage Server

Note that the utilization of the storage server's resources is low considering the following aggregate measured data (Figure 10-33 and Figure 10-34). Keep in mind that the utilization numbers can be useful to identify hardware elements creating a bottleneck in a particular configuration.

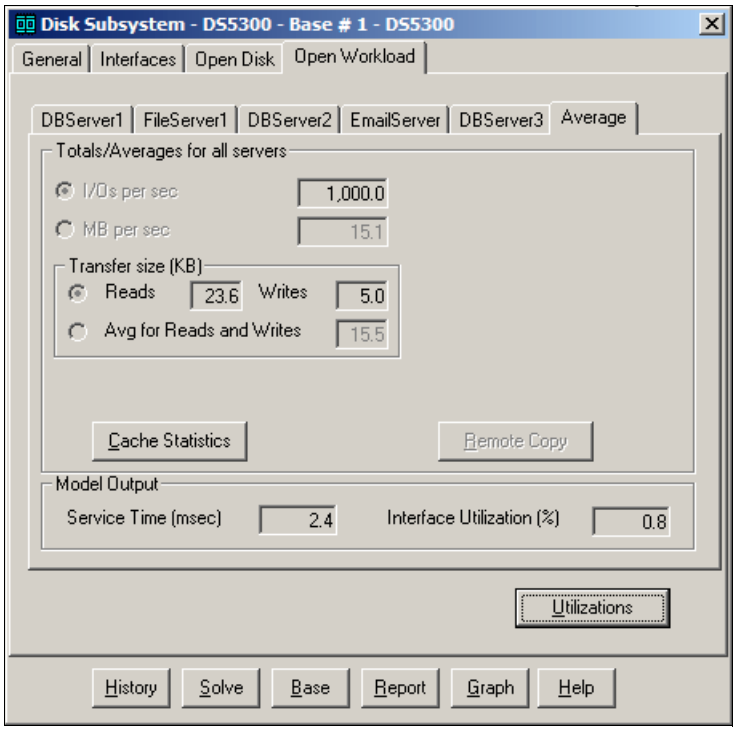


Figure 10-33 Aggregate Performance View

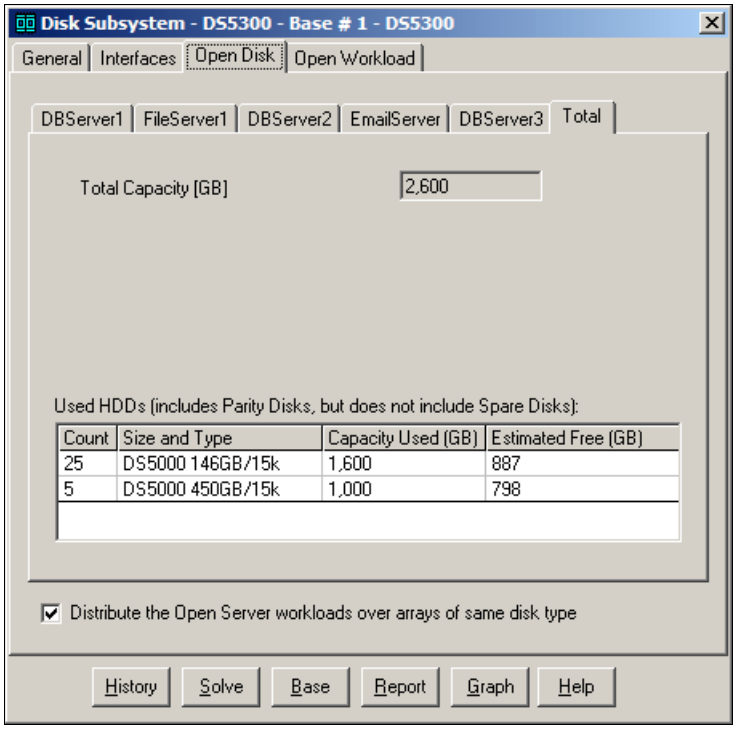


Figure 10-34 Total Number and Type of disk drives used

For example, if hard disk drives (HDD) Utilization is high, it can be useful to rerun the model with a higher number of smaller capacity HDDs, whereas if the Avg. Host interface Utilization (%) is too high, consider to add host ports if they are available on the controllers of the storage servers.

Disk Magic keeps a history of all of the configuration changes made, which allows for the restoration of the project to a known stage. To get the History Control Panel, click **History** in the Disk Subsystem window.

In our case, Figure 10-35 shows that at this point in time only the base has been created.

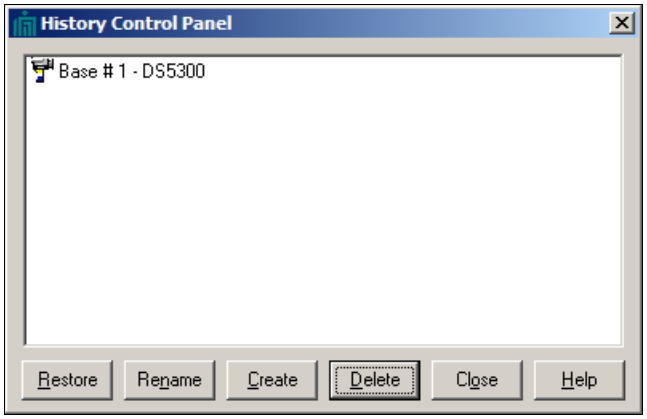


Figure 10-35 Base image - History Control Panel

After the base is established, what-if analysis can be undertaken. For example, you can analyze the effect of:

- Changing the number of disks in an array
- Changing the RAID protection of a specific array
- Analyzing the difference between SATA and Fibre Channel disks
- Comparing various storage servers

As you perform your various *what-if analyses* and select **Solve**, additional reports appear in this History Control Panel. Rename each entry to reflect what the changes represent.

10.3.1 Report

Selecting **Report** in the Disk Subsystem window creates a report of the current step in the model, as well as the base from which this model was derived. The report can be used to verify that all data was entered correctly. It also serves as a summary of the model results. The report lists all model parameters, including the ones that Disk Magic generates internally.

Example 10-1 shows a sample report.

Example 10-1 Base output from the project

```
Base model for this configuration
Project                  Project1

Model                    Model1

Disk Subsystem name:     DS5300

Type:             DS5300 (IBM)
Cache size:        16384 Mbyte
Interfaces:           16 Fibre 4 Gb FAStT, 4 Gb Server Interface Adapters

Total Capacity:   2600.0 GB
  approximately       25 disks of type DS5000 146GB/15k
  approximately        5 disks of type DS5000 450GB/15k

Remote copy type:        No Remote Copy Active or Remote Copy Not Supported

Advanced FAStT Outputs (%):
Processor Utilization:                  1.1
Highest HDD Utilization:               36.2
FC Bus (Lower Interface) Utilization:   1.9
PCI Bus Utilization:                    0.6
Avg. Host Interface Utilization:        0.5

Open Server    I/O  Transfer  Resp  Read Read Read Write Write Chan
               Rate Size (KB) Time  Perc Hit  Seq   Hit  Eff   Util
Average        1000    15.5    2.4    56   50   17  100    0     1
DBServer        100   116.8    2.8    70   50  100  100    0     0
FileServ         50     8.0    1.7    60   50   10  100    0     0
DBServer        150     4.0    2.3    60   50    5  100    0     1
EmailSer        550     4.0    2.3    50   50    5  100    0     2
DBServer        150     4.0    2.9    65   50    5  100    0     1
```

10.3.2 Graph

Graph solves the model and opens a dialog with options to select/define the graph to be created. Various graph types are possible, such as stacked bars and line graphs. By default, data to be charted is added to the spreadsheet and graph, allowing you to compare the results of successive modeling steps for a storage subsystem, or to compare results of various subsystems that are part of the same model.

Figure 10-36 shows the various graph options:

- Service Time in ms
- Read Service Time in ms
- Write Service Time in ms
- Avg Hit (Read and Write) in %
- Read Hit Percentage
- Busy Percentage of Host Interfaces
- Busy Percentage of Disk Interfaces
- I/O in MB per second

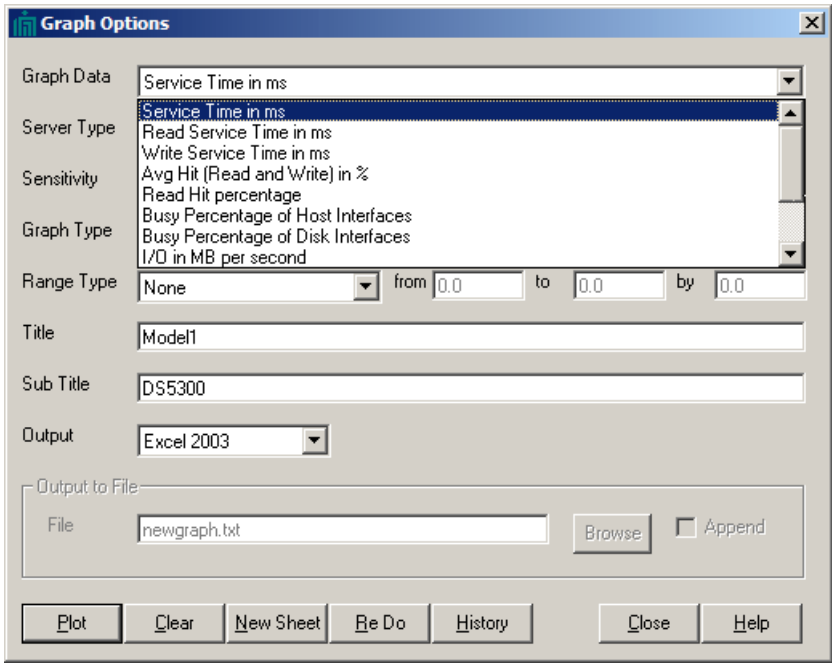


Figure 10-36 Pull-down menu for y-axis selection

Next select a graph type, as shown in Figure 10-37.

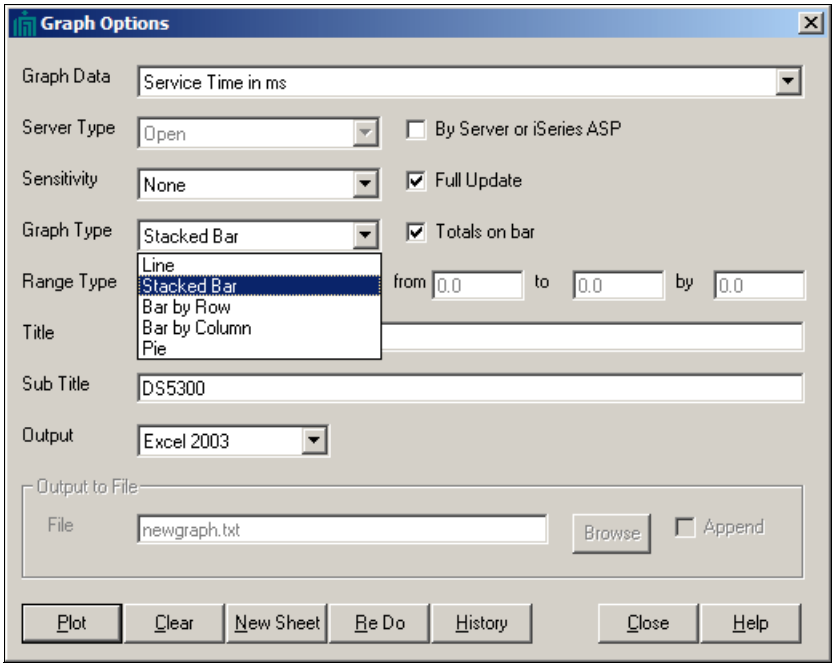


Figure 10-37 Graph Options - Graph Type

Then select the range type and the range values, as shown in Figure 10-38.

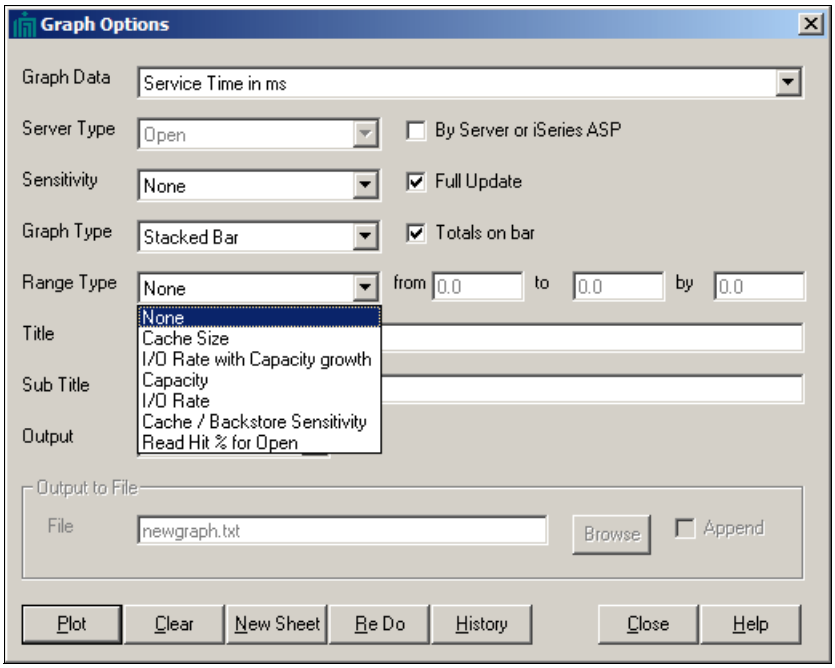


Figure 10-38 Graph Options - Range Type

You must also select an output type as either Excel2003 or text file. In Figure 10-39, you can see the values chosen for the y-axis the Service Time in ms, whereas for the x-axis, we choose the I/O Rate with Capacity grow. Note that it might be extremely useful to select as a Range Type the item “I/O rate” as well, because it can show the potentially sustainable I/O rate of the storage server as is, without increasing the number of spindles.

Additionally, it can show the reason for which the storage server cannot further grow in I/O rate, for example because of an excessive disk utilization. Finally, it might be a matter of interest, observing the trend of specific y-axis items (such as Busy Percentage of Host Interfaces, Busy Percentage of Disk Interfaces, and so on) as the I/O rate change without capacity grows.

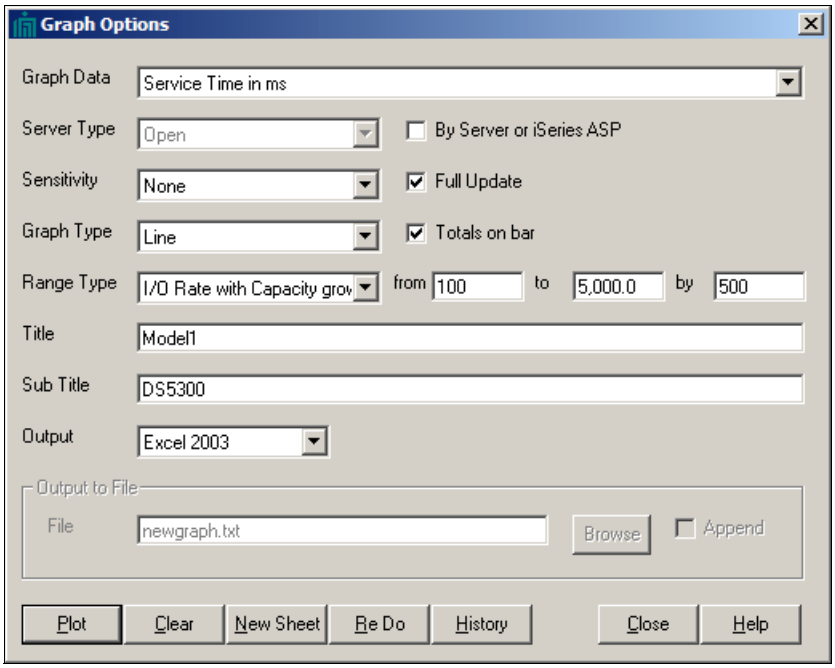


Figure 10-39 Setting for Service Time graph with range type of I/O rate with Capacity grow

Select **Plot** to generate the graph. The resulting graph is shown in Figure 10-40.

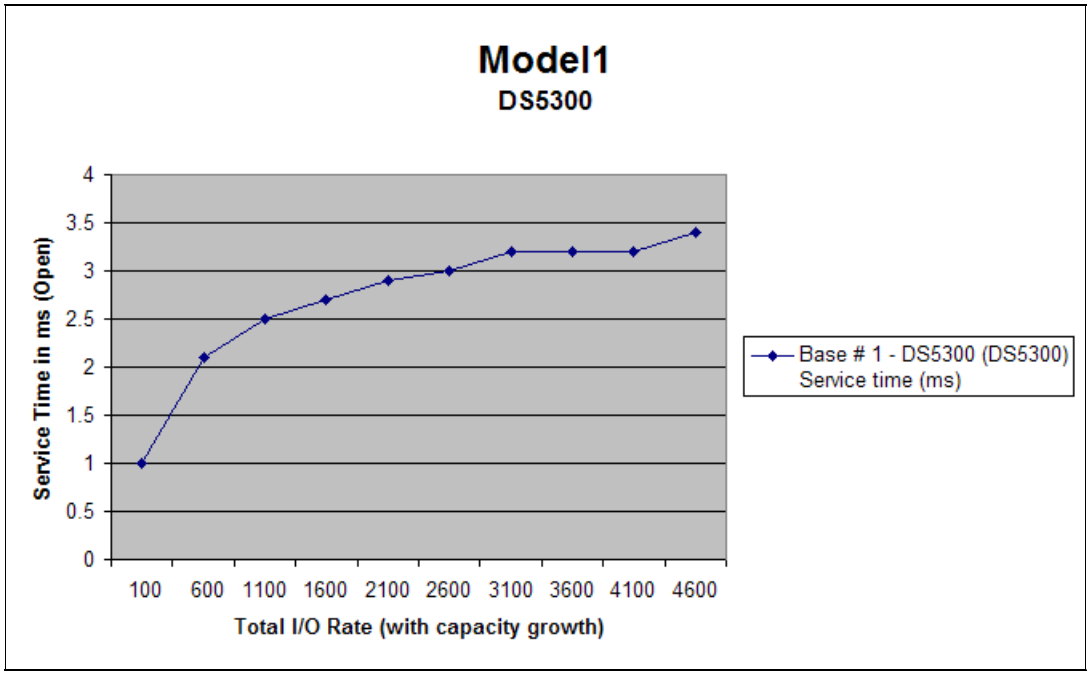


Figure 10-40 Output of Service Time graph with I/O Rate with Growth Capacity

The next graph compares read I/O rate in MB per second to I/O rate. The range was set from 100 to 2700 by 200 steps (Figure 10-41).

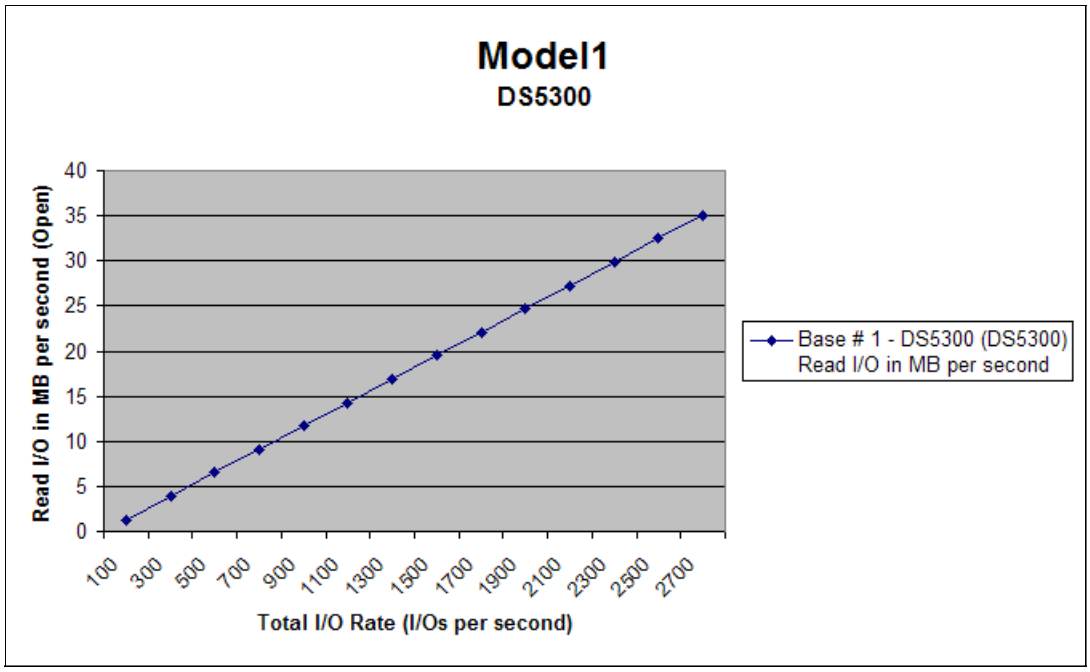


Figure 10-41 Output from read I/O versus I/O rate graph

In Figure 10-42, we add the DS5300 Write I/O in MB per second profile on the same graph shown in Figure 10-41 in order to get a relative comparison between the amount of reads and writes.

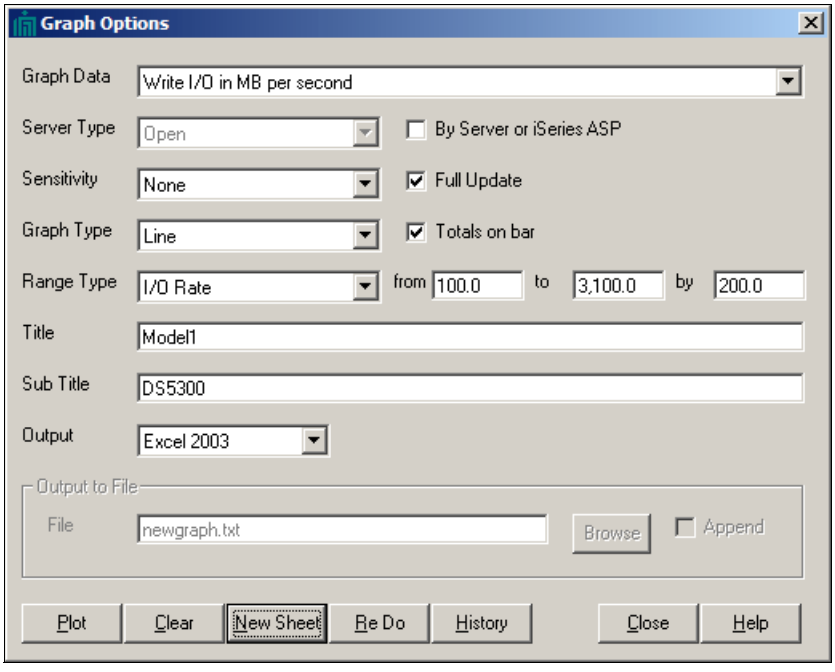


Figure 10-42 Settings for write I/O to compare read I/O graph

A graph with both the read and write I/O versus the I/O rate on the same scale is shown in Figure 10-43.

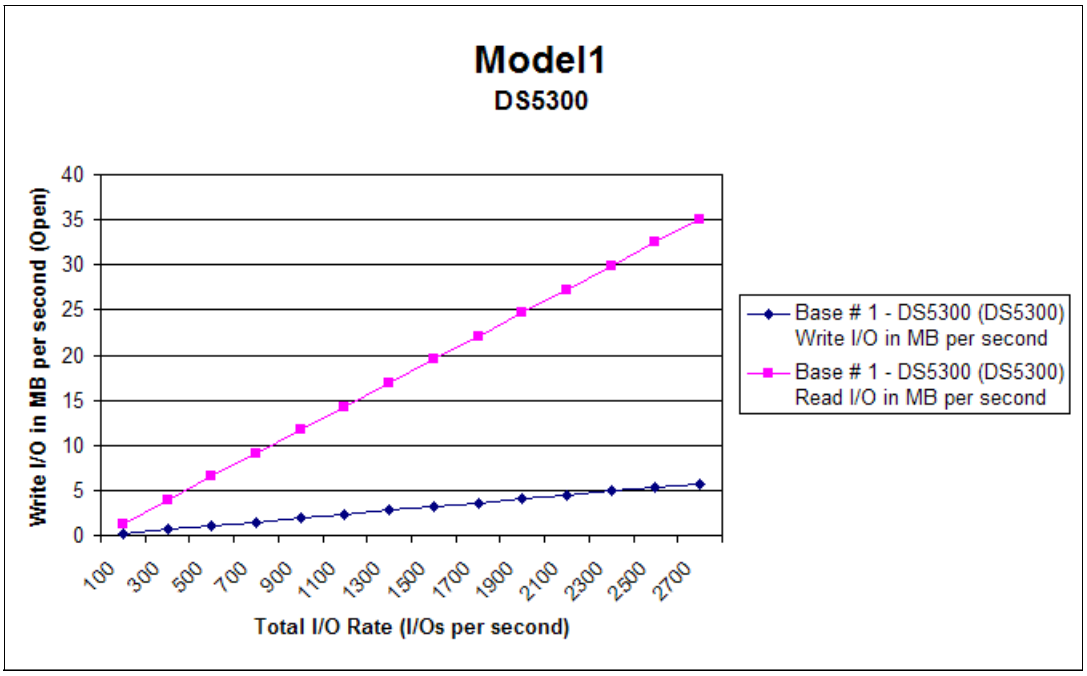


Figure 10-43 Output for read versus write I/O graph

10.3.3 Disk Magic and DS Storage Manager Performance Monitor

In Chapter 8, “Storage Manager Performance Monitor” on page 343, we describe the features of the DS Storage Manager Performance Monitor and we point out the script to collect performance data over a user-defined period of time and the format of the output text file. All the data saved in the file is comma delimited so that the file can be easily imported into a spreadsheet for analysis and review.

This spreadsheet, after proper manipulation, can be used to gather and tabulate a selection of useful data of all the Logical Drives during the same aggregate storage subsystem total peak interval. In this way, the sum of a specific performance data (for example IO per second) of each Logical Drive, during the aggregate storage subsystem peak interval, gives the aggregate storage subsytem IO per second.

Obviously, we must determine and process two Storage Subsystem peak intervals:

- I/O per second (typically meaningful for transaction workloads)
- Throughput in MB per second (typically meaningful for sequential workloads such as backups)

A block diagram can help in understanding the method (Figure 10-44).

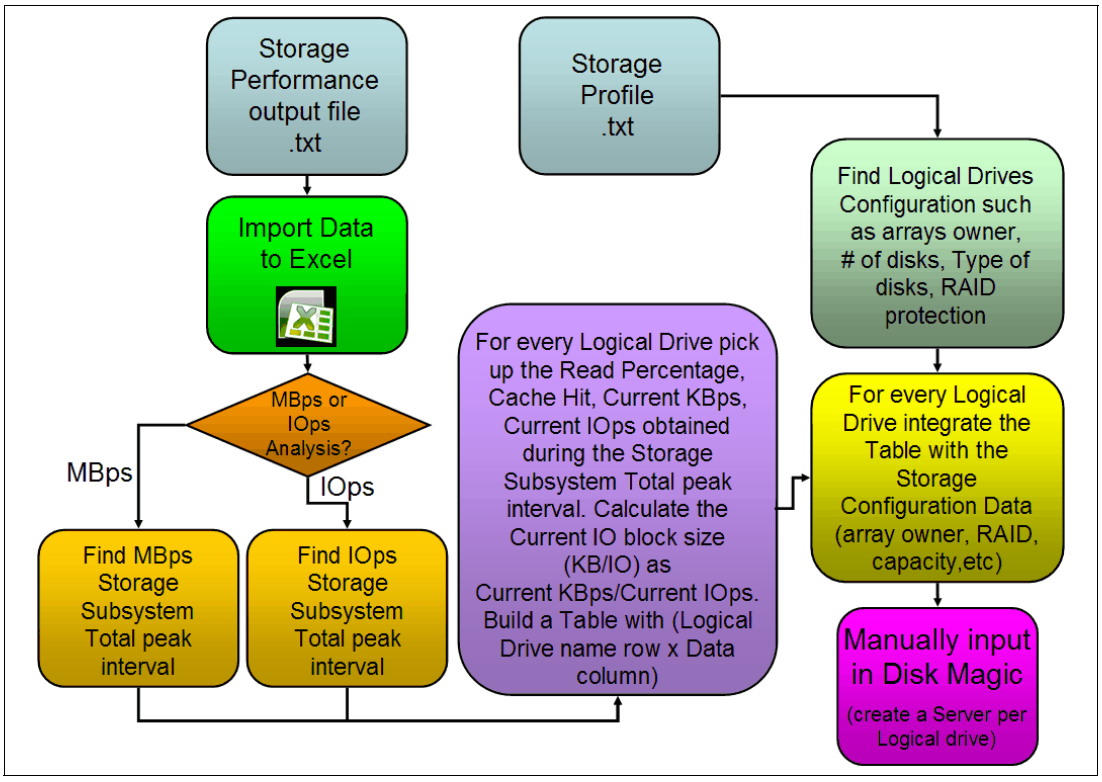


Figure 10-44 Block diagram to manually enter the Storage Performance data in Disk Magic

Assume that we have one Performance text file extrapolated from a script launched on the Storage Manager client (Chapter 8, “Storage Manager Performance Monitor” on page 343) and a Storage Subsystem Profile text file extrapolated from the Storage Manager client through the following procedure:

- Go to IBM System Storage DS ES (Subsystem Management, and in the top menu, select **Storage Subsystem** → **view** → **Profile** as shown in Figure 10-45.

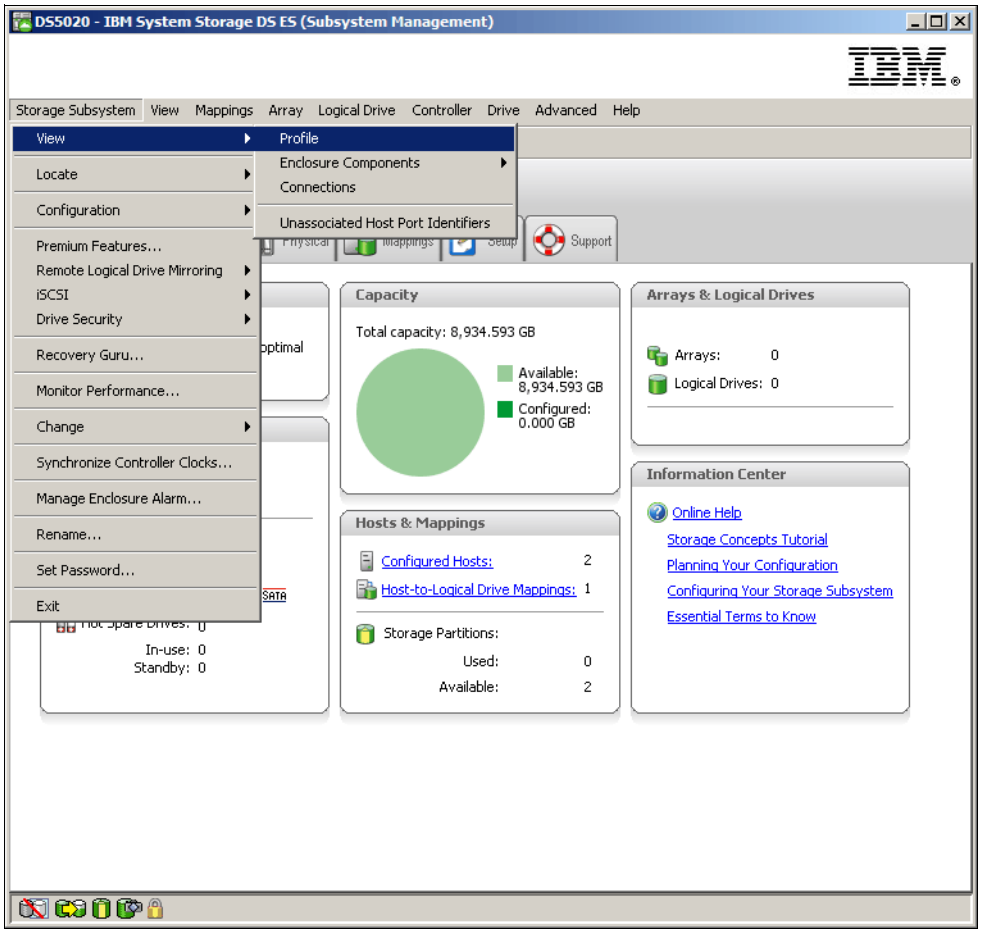


Figure 10-45 Select the profile

- The Storage Subsystem Profile opens as shown in Figure 10-46.

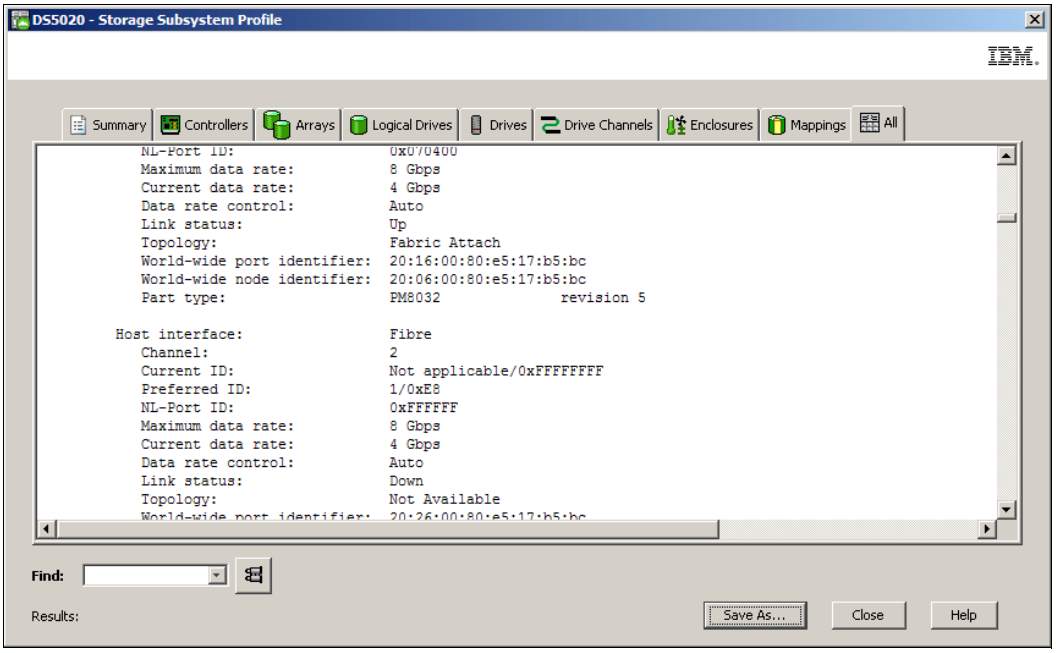


Figure 10-46 Storage Subsystem Profile View

- Save the file in text format, including all the section as shown in Figure 10-47.

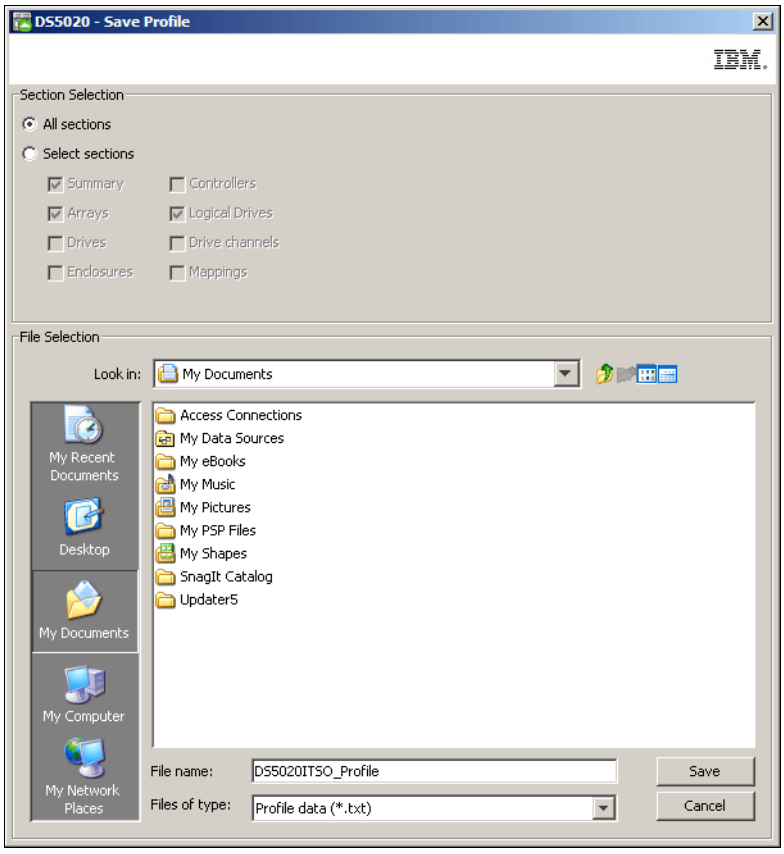
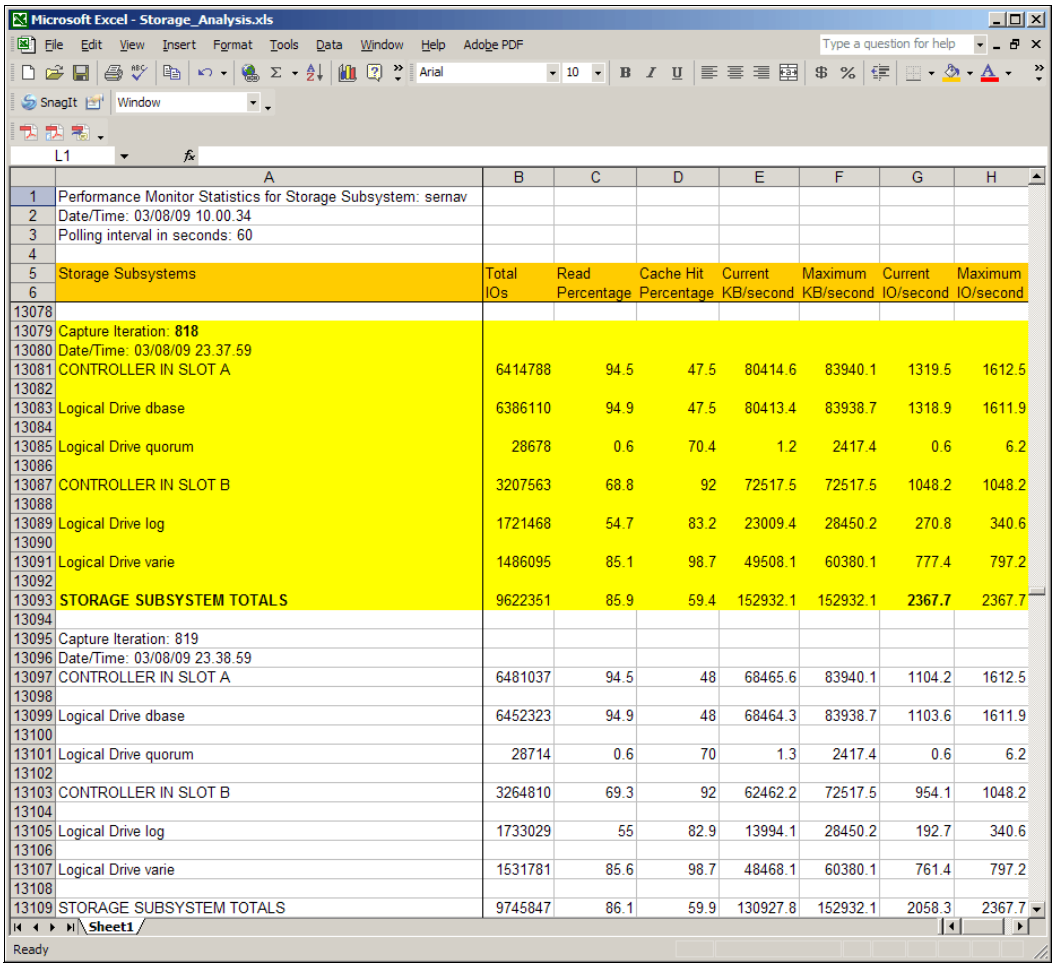


Figure 10-47 Save the profile file in text format including all the sections

In the Excel file in which we have imported the Storage Manager performance file, we are going to determine and point out the interval that get the maximum value of the *IO per second* parameter at **STORAGE SUBSYSTEM TOTALS** level. We take a look at the *IO per second* values for each Logical Drive as well. It is fairly easy to find the *I/O per second* peak value:

- Go to the *Current IO/second* column.
- Set an Excel automatic filter and find the greatest numeric value for that column.
- Note the peak value and determine the Interval Time in which it has occurred, as well as all the Logical Drive performance data that you need for your Disk Magic model during that Interval Time.

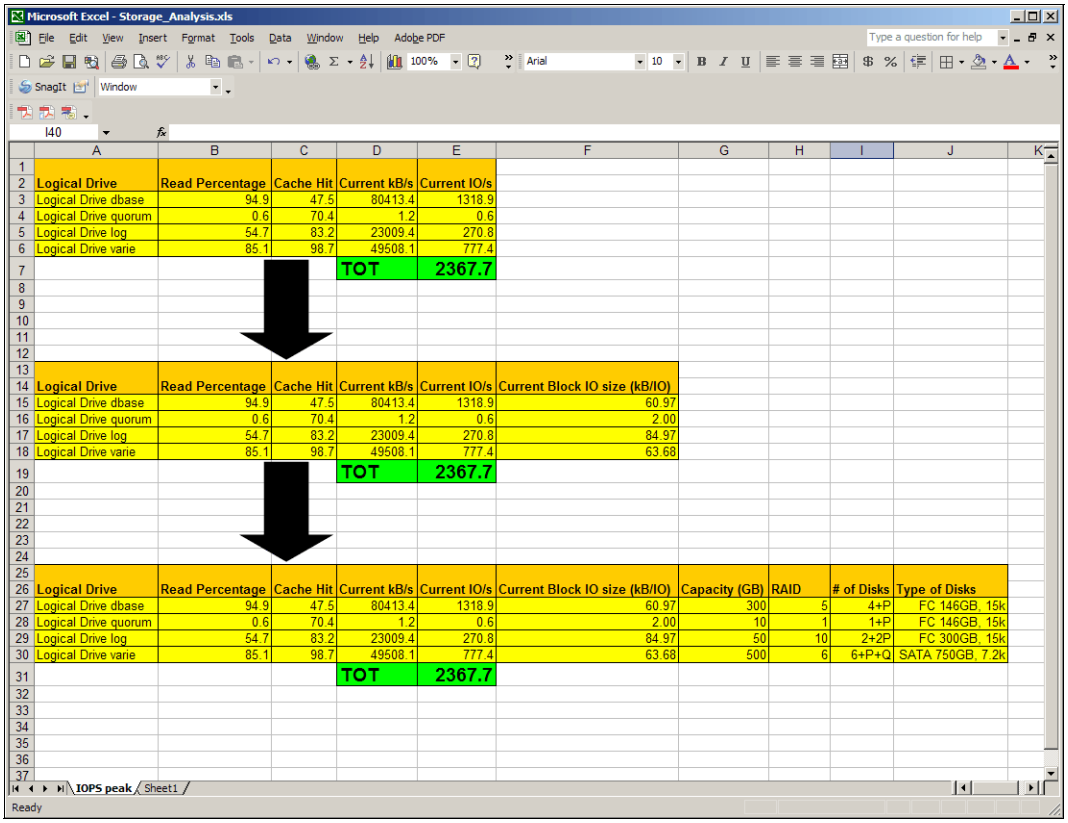
For example, in Figure 10-48, we have found that the Capture Iteration number 818 has produced the maximum Current IO per second at the **STORAGE SUBSYSTEM TOTALS** level, and this value represents the aggregated peak, such as 2367.7 IO per second.


Microsoft Excel - Storage_Analysis.xls

	A	B	C	D	E	F	G	H
1	Performance Monitor Statistics for Storage Subsystem: sernav							
2	Date/Time: 03/08/09 10.00.34							
3	Polling interval in seconds: 60							
4								
5	Storage Subsystems	Total IOs	Read Percentage	Cache Hit Percentage	Current KB/second	Maximum KB/second	Current IO/second	Maximum IO/second
6								
13078								
13079	Capture Iteration: **818**							
13080	Date/Time: 03/08/09 23.37.59							
13081	CONTROLLER IN SLOT A	6414788	94.5	47.5	80414.6	83940.1	1319.5	1612.5
13082								
13083	Logical Drive dbase	6386110	94.9	47.5	80413.4	83938.7	1318.9	1611.9
13084								
13085	Logical Drive quorum	28678	0.6	70.4	1.2	2417.4	0.6	6.2
13086								
13087	CONTROLLER IN SLOT B	3207563	68.8	92	72517.5	72517.5	1048.2	1048.2
13088								
13089	Logical Drive log	1721468	54.7	83.2	23009.4	28450.2	270.8	340.6
13090								
13091	Logical Drive varie	1486095	85.1	98.7	49508.1	60380.1	777.4	797.2
13092								
13093	**STORAGE SUBSYSTEM TOTALS**	9622351	85.9	59.4	152932.1	152932.1	**2367.7**	2367.7
13094								
13095	Capture Iteration: 819							
13096	Date/Time: 03/08/09 23.38.59							
13097	CONTROLLER IN SLOT A	6481037	94.5	48	68465.6	83940.1	1104.2	1612.5
13098								
13099	Logical Drive dbase	6452323	94.9	48	68464.3	83938.7	1103.6	1611.9
13100								
13101	Logical Drive quorum	28714	0.6	70	1.3	2417.4	0.6	6.2
13102								
13103	CONTROLLER IN SLOT B	3264810	69.3	92	62462.2	72517.5	954.1	1048.2
13104								
13105	Logical Drive log	1733029	55	82.9	13994.1	28450.2	192.7	340.6
13106								
13107	Logical Drive varie	1531781	85.6	98.7	48468.1	60380.1	761.4	797.2
13108								
13109	STORAGE SUBSYSTEM TOTALS	9745847	86.1	59.9	130927.8	152932.1	2058.3	2367.7

Figure 10-48 Find the peak interval among the data imported on Excel

Now you can add another spreadsheet and create the table with a row per each Logical drive and with the meaningful data from the Disk Magic point of view (Figure 10-49). Keep in mind that the storage configuration data has been captured from the storage profile text file and that the column **Current Block IO size (KB/IO)** has been calculated with a formula as the ratio between the current KB/s and current IO/s.



Microsoft Excel - Storage_Analysis.xls

Logical Drive	Read Percentage	Cache Hit	Current kB/s	Current IO/s
Logical Drive dbase	94.9	47.5	80413.4	1318.9
Logical Drive quorum	0.6	70.4	1.2	0.6
Logical Drive log	54.7	83.2	23009.4	270.8
Logical Drive varie	85.1	98.7	49508.1	777.4
			TOT	2367.7

Logical Drive	Read Percentage	Cache Hit	Current kB/s	Current IO/s	Current Block IO size (kB/IO)
Logical Drive dbase	94.9	47.5	80413.4	1318.9	60.97
Logical Drive quorum	0.6	70.4	1.2	0.6	2.00
Logical Drive log	54.7	83.2	23009.4	270.8	84.97
Logical Drive varie	85.1	98.7	49508.1	777.4	63.68
			TOT	2367.7	

Logical Drive	Read Percentage	Cache Hit	Current kB/s	Current IO/s	Current Block IO size (kB/IO)	Capacity (GB)	RAID	# of Disks	Type of Disks
Logical Drive dbase	94.9	47.5	80413.4	1318.9	60.97	300	5	4+P	FC 146GB, 15k
Logical Drive quorum	0.6	70.4	1.2	0.6	2.00	10	1	1+P	FC 146GB, 15k
Logical Drive log	54.7	83.2	23009.4	270.8	84.97	50	10	2+2P	FC 300GB, 15k
Logical Drive varie	85.1	98.7	49508.1	777.4	63.68	500	6	6+P+Q	SATA 750GB, 7.2k
			TOT	2367.7					

Figure 10-49 Table building as input for Disk Magic

At this point we have all the information required to create a base line model with Disk Magic, and in particular we have gotten an important parameter to model the cache statistics, such as the **Cache Hit** parameter.

First of all, we create a new project using manual input as shown in Figure 10-50. Note that we have selected a number of open servers equal to the number of Logical Drives.

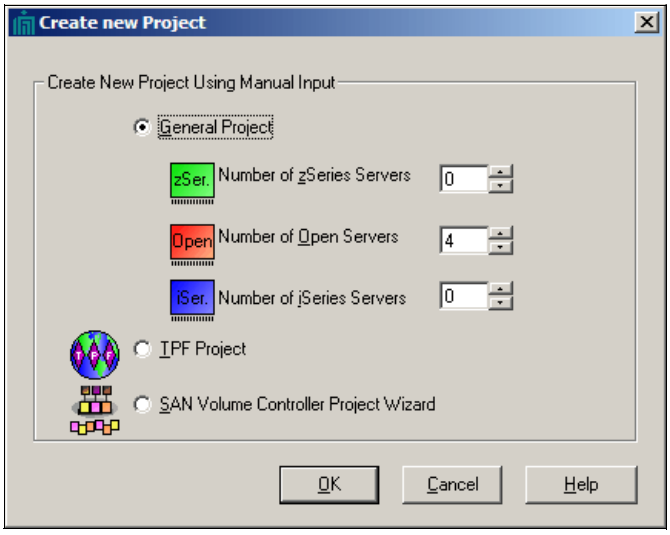


Figure 10-50 Create a manual project

Then we rename the open servers in order to fit the Logical Drive names (Figure 10-51).

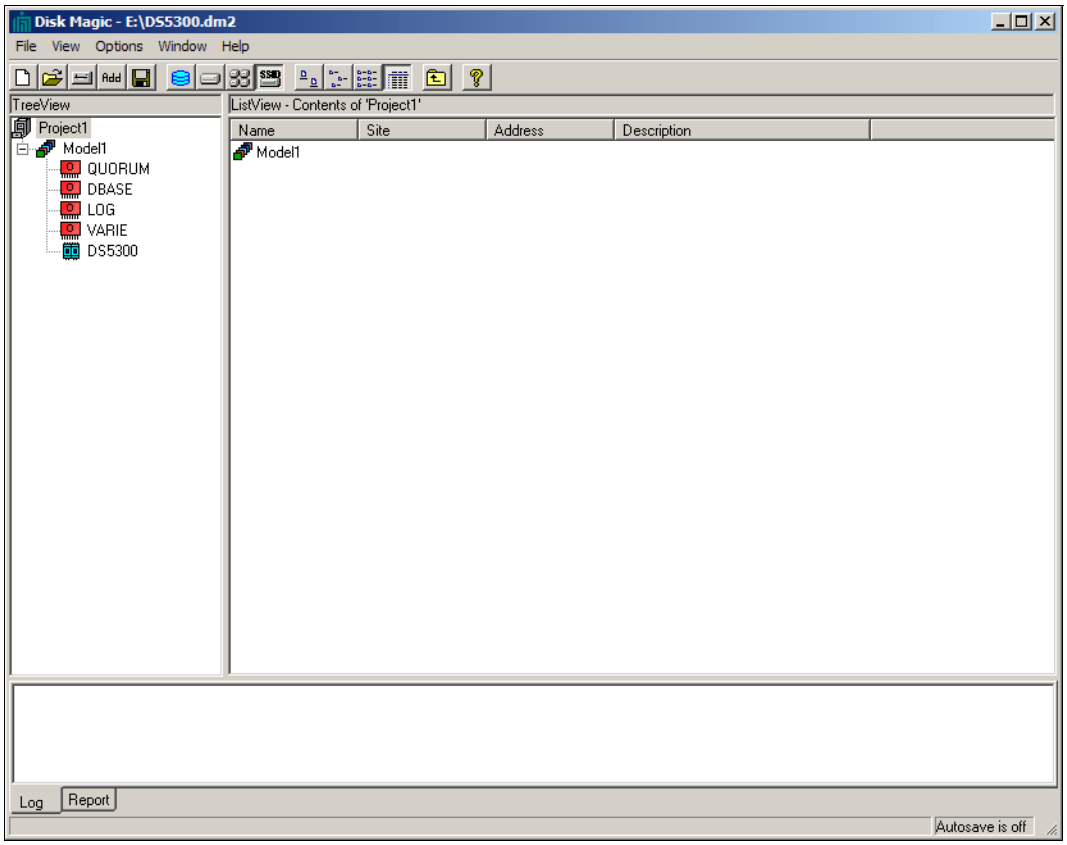


Figure 10-51 Server names match Logical Drive names

Finally, we manually insert in the Disk Magic fields, all the data collected in the tables created in the Excel worksheet (see Figure 10-49 on page 469).

11

Storage virtualization guidelines for DS5000 series

Virtualization has become an essential part of today's storage environments. It helps to optimize storage growth, reduce complexity, offers flexible management, and significantly improves cost management.

In this chapter, we briefly introduce the major virtualization systems: IBM System Storage SAN Volume Controller (SVC) and IBM Storwize V7000. We describe their components and concepts, then compare the SVC and IBM Storwize V7000 Copy Services with those of the DS5000 series. We focus on general best practices and guidelines for the use of SVC or IBM Storwize V7000 with the DS5000 Storage Server, including a configuration example.

For the purpose of this chapter, the term *storage virtualization system, virtualization system,* or *clustered system* refers to both IBM System Storage SAN Volume Controller and IBM Storwize V7000.

11.1 IBM storage virtualization overview

In this section, we briefly explain the basics of storage virtualization, focusing on the DS5000 Storage Server when configured with the IBM System Storage SAN Volume Controller or IBM Storwize V7000.

11.1.1 Storage virtualization concepts

Storage virtualization typically refers to the management of heterogeneous pools of storage and making them accessible to the hosts as volumes through a common management interface. The storage pools can include multiple generations of storage systems from various vendors.

The IBM System Storage SAN Volume Controller (SVC) and IBM Storwize V7000 provide a combined hardware and software solution. They are the virtualization engines referred to as *nodes*. The two nodes form a high availability fault tolerant system so that if one of the nodes fails the surviving node automatically takes over. These pairs are called an *input/output group (I/O group)*. I/O groups are responsible for serving the I/O to a given storage volume. There can be up to four I/O groups (8 nodes) in SVC clustered system, and single I/O group inside control enclosure of IBM Storwize V7000.

The storage LUNs, both internal or external, are presented to the storage virtualization systems as *managed disks* (MDisks), which are then subdivided down into smaller chunks called *extents*. Pools of these extents with similar performance characteristics are bundled together into *storage pools*. The virtualization systems finally present the virtualized LUNs to the hosts as *volumes*. Figure 11-1 shows basic building blocks of a virtualization system.

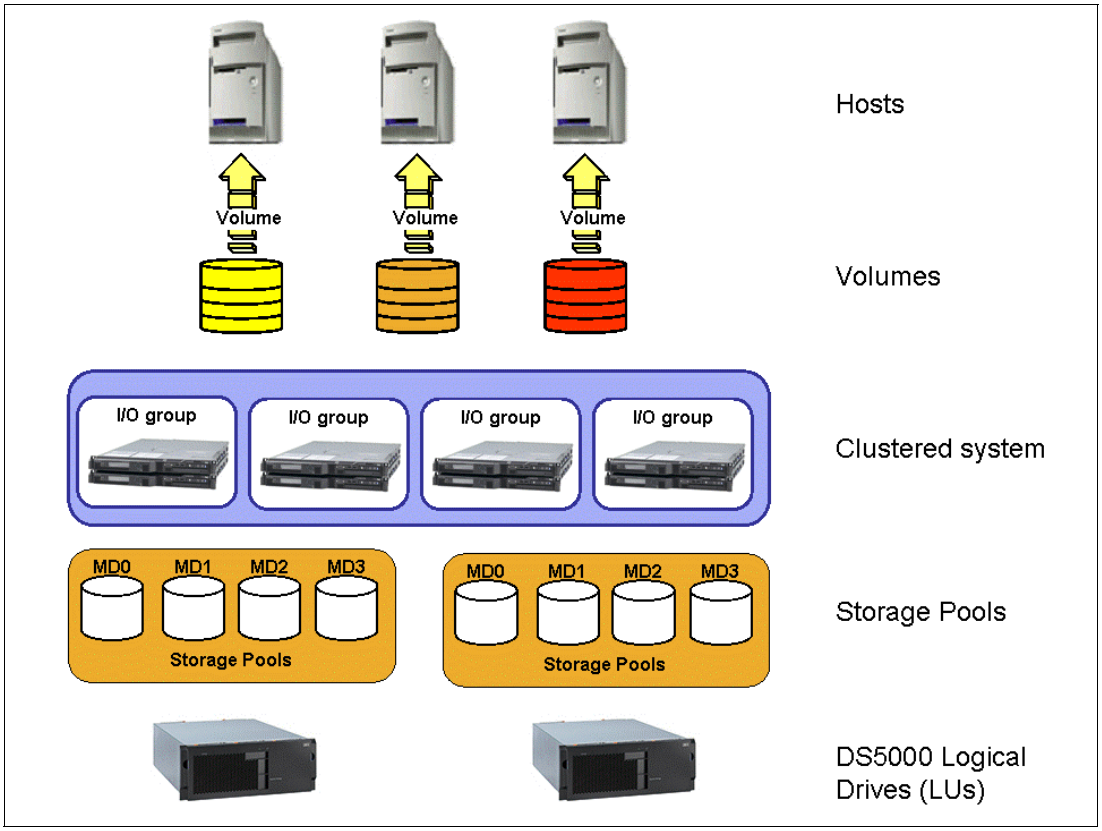


Figure 11-1 Basic virtualization concepts

11.1.2 Storage virtualization glossary of terms

A storage virtualization implementation consists of hardware and software elements. Here we define various concepts and terminology used with storage virtualization systems.

Clustered system

For maximum availability and performance, IBM System Storage SAN Volume Controller and IBM Storwize V7000 consist of nodes which are paired into dual, active/active, fault tolerant units called I/O groups. The SVC can scale up to four I/O groups within single clustered system. The battery back-up power supply units (PSU) in IBM Storwize V7000, and uninterruptible power supply units (UPS) connected to SVC nodes ensure the high availability and resiliency of the storage virtualization system. Both systems run on the same software platform based on IBM SAN Volume Controller code.

Node

A node is a name given to the individual server inside storage virtualization clustered system on which the virualization software runs. Nodes are always installed in pairs to form the I/O group. Each clustered system node can only be a member of one I/O group.

Managed disk

A managed disk (MDisk) is a SCSI disk presented by the storage subsystem (for instance, a DS5000 logical drive) and managed by the storage virtualization system. A managed disk provides usable blocks (or extents) of physical storage to the clustered system. The MDisks are not visible to hosts systems accessing SVC or IBM Storwize V7000.

An LUN (logical drive) presented to a storage virtualization system can be in one of three states: unmanaged, managed, or image mode:

- An *unmanaged* disk is one that has been assigned to the storage virtualization system but not yet managed by it.
- A *managed* disk is one that belongs to, and is managed by storage virtualization system.
- An *image mode* disk is one that has been imported into storage virtualization system and contains data. This technique is used when migrating existing LUNs with data into the storage clustered system. The existing data on the LUN is preserved. To extend the disk after being managed by the storage virtualization system, this type of disk must be migrated to an MDisk.

Extent

An extent is a fixed size unit of data that is used to manage the mapping of data between MDisks and volumes. The extent size choices are 16, 32, 64, 128, 256, 512, 1024, 2048, 4096 or 8192 MB. The choice of extent size affects the total storage capacity that can be managed by the storage virtualization system. For example, a 16 MB extent size supports a maximum capacity of 64 TB, while maximum extent size 8192 MB supports maximum storage capacity of 32 PB.

Storage pool

The storage pool is a collection of MDisks. Each storage pool is composed of a number of extents, which are numbered sequentially, starting with 0. When creating a storage pool, you must choose an extent size. After being set, the extent size stays constant for the life of that storage pool and cannot be changed. Each storage pool can have a unique extent size.

Volume

A volume is a logical entity that represents extents contained in one or more MDisks from a storage pool. Volumes are allocated in a whole number of extents. These extents are then presented to the host as a volume for use (Figure 11-2).

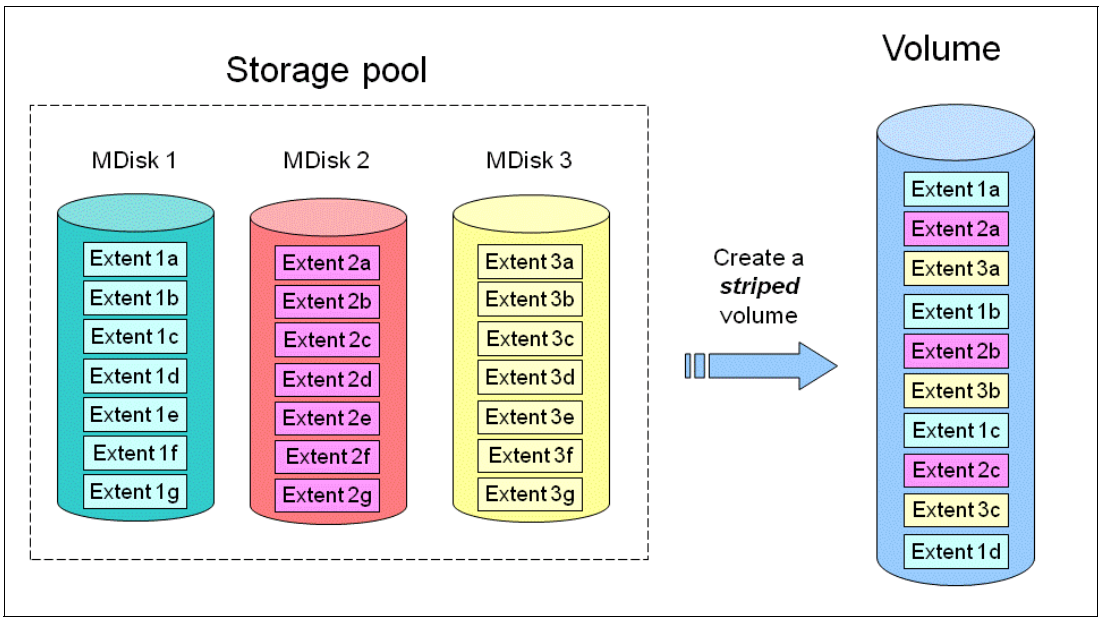


Figure 11-2 Extents used to create a volume

Tip: The SVC and Storwize V7000 are not RAID controllers for their external storage. The disk subsystems attached to SVC and Storwize V7000 must provide the RAID configuration.

I/O group

An I/O group contains two nodes, which are configured as a pair. Each node is associated with only one I/O group. The nodes in the I/O group provide access to the volumes in the I/O group. Each volume is associated with exactly one I/O group. During normal operation of storage virtualization system, each volume has affinity to one of the nodes inside the I/O group (*preferred node*). The preferred node serves all the I/O operations to its own volumes. In case of node failure or scheduled maintenance window when one of the nodes needs to be taken offline, all I/O operations are directed to the working node in the I/O group. In this situation, write caching is disabled to avoid data consistency issues (write-through mode).

Figure 11-3 shows the relationship between individual storage virtualization components.

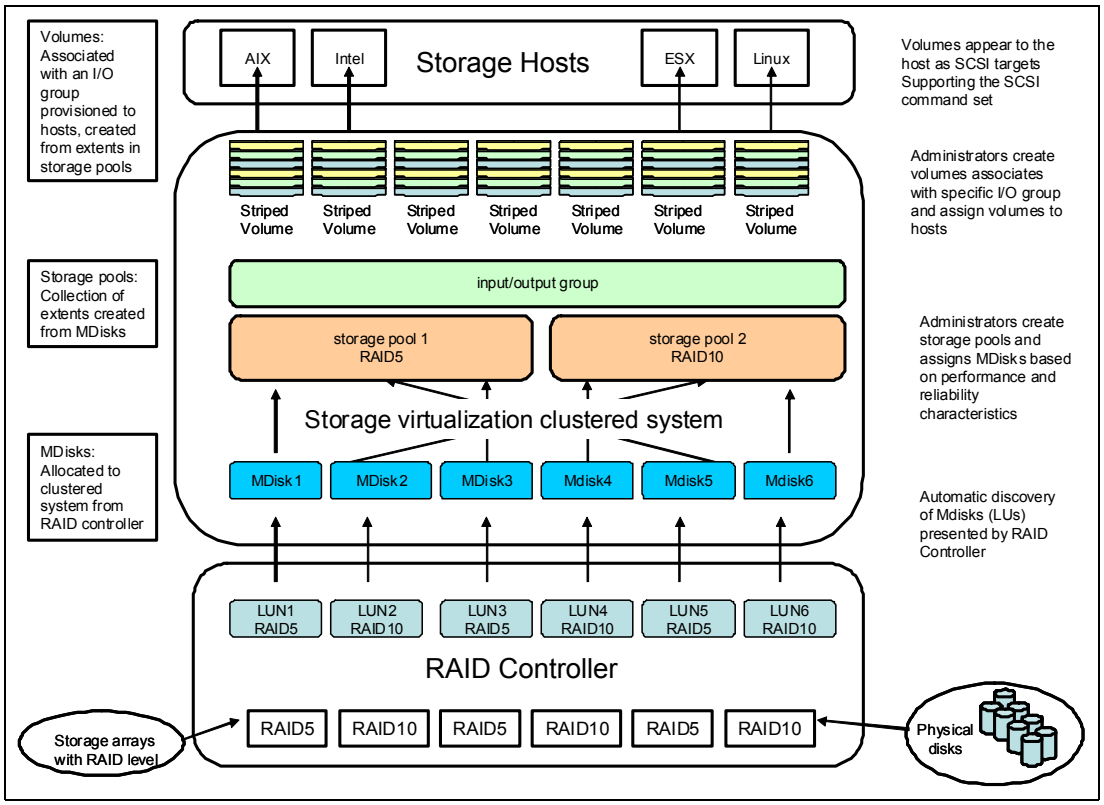


Figure 11-3 Relationships between virtualization components

Consistency group

Consistency groups address the situation where the objective is to preserve data consistency across multiple volumes involved in Copy Services relationships. Because the applications have related data that span multiple volumes, a requirement for preserving the integrity of data being accessed is to ensure that *dependent writes* are executed in the application's intended sequence.

Multipathing device driver

The *multipathing device driver* supported by the virtualization systems is the IBM Subsystem Device Driver (SDD). SDD groups all available paths to a volume and presents it to the operating system. SDD performs all the path handling and selects the active I/O paths.

Always consult the IBM support website for the latest device drivers and the support matrix:

- IBM System Storage SAN Volume Controller: `http://www.ibm.com/storage/support/2145`
- IBM Storwize V7000: `http://www.ibm.com/storage/support/storwize/v7000`

11.1.3 Benefits of the IBM storage virtualization

Storage virtualization delivers a single view of the storage attached to the SAN. Administrators can manage, add, and migrate data on physical disks non-disruptively even between various storage subsystems. The SAN is zoned in such way that the application servers can not see the back-end storage, preventing any possible conflict between storage virtualization systems and the application servers both trying to manage the same back-end storage.

The clustered systems provide storage virtualization by creating a pool of managed disks from attached back-end or internal disk storage subsystems. These managed disks are then mapped to a set of volumes for use by various host systems:

- Performance requirements, availability requirements, and other service level agreements can be addressed by using various storage pools and advanced functions of storage virtualization.
- Dependencies, which exist in a SAN environment with heterogeneous storage and server systems are reduced.

We summarize the key benefits of the storage virtualization here:

Concurrent Migration This function is probably the single most important one that the virtualization systems add to any storage environment. MDisks can be migrated transparently to the host application, which allows data to be moved non-disruptively from a back-end storage subsystem in preparation for a scheduled maintenance activity. It also gives the system administrator the flexibility to balance I/O workload on the fly or easily migrate across to a storage pool with more appropriate performance/resilience characteristics as needs demand.

Single Management Interface The storage virtualization systems provide a single common management interface for provisioning heterogeneous storage.

Copy Services In a storage virtualization environment, there is no need to purchase separate Copy Services licenses from each storage vendor. Furthermore, the source and target of a copy relationship can be on separate storage subsystems.

Increased Utilization Having the flexibility to pool storage across the SAN results in much improved storage capacity utilization. Spare capacity on underlying physical disks can be reallocated non-disruptively from an application server point of view irrespective of the server operating system or platform type. Logical disks can be created from any of the physical disks being managed by the virtualization device (that is, vendor independent).

Performance A well designed storage virtualization system configuration can significantly improve overall performance by reducing hot spots, grouping multiple physical resources, and by making use of the additional cache inside the nodes.

Connectivity Each vendor storage subsystem traditionally requires a vendor's device driver installed on the host to access the subsystem. Where there are many subsystems in the environment, regardless of whether any one host is accessing more than one vendors storage subsystems, then managing the range of device drivers is unnecessarily complex. The IBM approach means that only one device driver, the IBM System Storage Subsystem Device Driver (SDD), is required to access any virtualized storage on the SAN regardless of the vendor storage subsystem.

iSCSI Connectivity Starting from IBM SAN Volume Controller code Version 5.1, there is support for iSCSI attached hosts, which provides both cost saving and flexibility as storage hosts are no longer limited to using FC Host Bus Adapters (HBA) to access the SAN storage.

Scalability It is possible to expand the SVC up to four I/O groups (8 nodes) as performance and capacity requirements increase. Similarly, IBM Storwize V7000 can be dynamically allocated more internal storage by adding new storage expansions.

11.1.4 Key points for using DS5000 with storage virtualization systems

It is important to understand that both SVC and IBM Storwize V7000 do not provide any RAID redundancy protection for the MDisks that are presented to it from back-end storage subsystems. Therefore, it is essential that the back-end storage subsystem is sufficiently resilient to provide the level of protection expected in an enterprise class environment as well as meeting any performance level criteria. In this area, the DS5000 Storage Server becomes a perfect compliment for the storage virtualization systems. Supporting an extensive selection of drive types and capacities together with a choice of RAID level protection and no single points of failure all combine to make the DS5000 Storage Servers an extremely flexible partner for storage virtualization systems.

Tip: IBM Storwize V7000 contains internal storage that can be configured into arrays using standard RAID types. Optionally, the SAN Volume Controller 2145-CF8 and the SAN Volume Controller 2145-CG8 have a number of drives attached to it. These drives are used to create a Redundant Array of Independent Disks (RAID), which are presented as managed disks (MDisks) in the system.

Storage virtualization systems are very flexible in their use. They can be used to manage all of your disk storage requirements or just part of it. In other words, when using storage virtualization with the DS5000, you can still use the DS5000 Storage Manager (SM) functions to allocate part of the storage to certain hosts. The *DS5000 partitioning* feature must be used to separate groups of logical units that are directly attached to hosts or groups of hosts from the logical units that are accessed by the SVC or IBM Storwize V7000.

Storage virtualization systems offer a large scalable cache and can reduce the requirement for additional partitions. The virtualization system consumes only one storage partition for each storage server that connects to it. If you use the virtualization system for all your hosts, then a storage partition upgrade on the DS5000 might not be required. In a configuration with the virtualization system, the DS5000 needs to speak only a single language (the SVC) rather than multiple languages when it is attached in a heterogeneous enterprise. To the DS5000 Storage Server, the virtualization system is the only host that it sees.

In addition, it improves capacity utilization. Spare capacity on underlying physical disks can be reallocated non disruptively from an application server point of view irrespective of the server operating system or platform type.

11.2 IBM System Storage SAN Volume Controller

IBM System Storage SAN Volume Controller is a proven, high-end storage virtualization solution based on SVC Storage Engine running on dedicated SVC software platform. SVC Storage Engines are built on IBM System x server technology and are always deployed in redundant pairs, which are designed to deliver high availability and resiliency.

Latest versions of IBM System Storage SAN Volume Controller use a completely new graphical user interface modeled on the IBM XIV Storage System, which has been very well received by customers. The user interface is designed to be very easy to use and includes many built-in IBM suggestions to help simplify storage provisioning and enable new users to get started quickly.

11.2.1 IBM System Storage SAN Volume Controller hardware

IBM System Storage SAN Volume Controller is composed of hardware processing units called nodes. Each SVC node is an individual server in a SAN Volume Controller clustered system on which the SAN Volume Controller software runs. Figure 11-4 shows the latest SVC model 2145-CG8.

SVC hardware consists of the following components:

- SVC nodes (known as SVC Storage Engine)
- Uninterruptible Power Supply (UPS)
- Serial communication cable for SVC node to communicate with its UPS

Figure 11-4 2 nodes of SVC system 2145-CG8

Each node contains following components:

- System board based on IBM System x3550 M3
- Server Intel Xeon 5600 2.5 GHz quad-core processor
- 24 GB of cache
- Four 8 Gbps FC ports
- Two 1 Gbps and optionally two 10 Gbps iSCSI ports
- Support for solid-state drives (up to four per SVC node)
- Two power supplies

Tip: The node power supply units (PSUs) must be connected to UPS. Although the use of redundant AC power switch is optional, we strongly advise it in order to maintain resiliency in case of the single power circuit failure.

SVC Node 2145-CG8 front panel

The front panel of the SVC node consists of controls and indicators (Figure 11-4). They are used for power functions and navigation and to indicate information, such as system activity, service and configuration options, controller failures, and node identification. Initial configuration of IBM System Storage SAN Volume Controller system is done using front-panel display and navigation buttons.

SVC Node 2145-CG8 rear panel

IBM System Storage SAN Volume Controller connects to back-end storage using the four 8 Gbps FC ports. These ports are also used for communication with host systems. Two 1Gbps and optional 10 Gbps iSCSI ports are only used for host communication. Figure 11-5 shows a rear panel of SVC with the ports labeled.

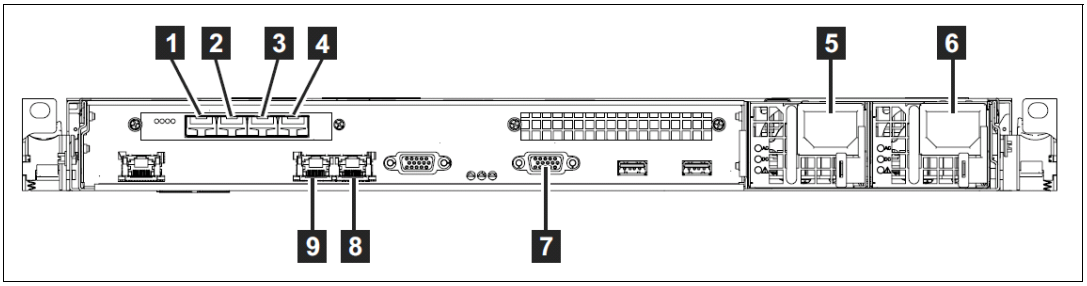


Figure 11-5 SVC Fibre Channel (1-4), Ethernet (9-8) and serial (7) ports, including redundant power supply units (5,6)

11.2.2 IBM System Storage SAN Volume Controller software

IBM System Storage SAN Volume controller is delivered with preinstalled software on SVC Storage Engines. As soon as the SVC is attached to the SAN environment, you can make changes to the configuration quickly and easily as needed.

Starting with IBM SAN Volume Controller code Version 6.1, SAN Volume Controller Console (SVCC, *master console*) is no longer used to manage the SVC system. You can still use the SSPC element manager feature to run the graphical management window, but SVCC is not supported and will not work with IBM SAN Volume Controller code Version 6.1. Instead, a new *SVC Console* graphical user interface runs directly on the SVC clustered system and can be accessed from anywhere on the network using a web browser. Figure 11-6 shows the login screen of the SVC Console.

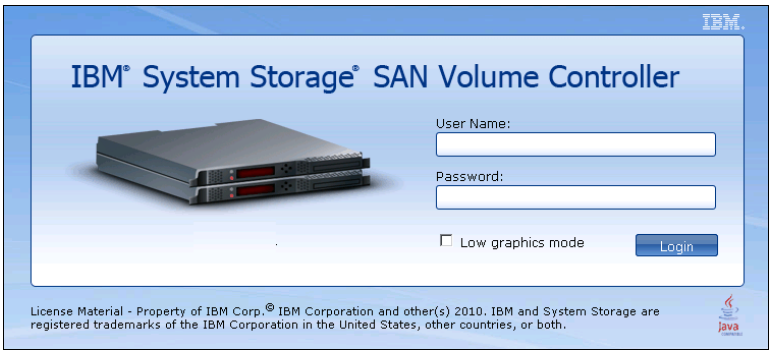


Figure 11-6 SVC user interface running on the clustered system

IBM System Storage SAN Volume controller is composed of nodes. Each of these nodes has SAN Volume Controller code installed on it and is able to run the clustered system independently. At the time of the initial system configuration, one of the nodes is arbitrarily selected to function as *configuration node*. Only one node can be a configuration node at any given time. Configuration node is responsible for managing configuration activity of the clustered system. If the configuration node fails, the system chooses a new configuration node. This action is called a configuration node failover. The new configuration node takes over the systems IP addresses. Thus you can access the clustered system through the same IP addresses although the original configuration node has failed. During the failover, there is a short period when you cannot use the command-line tools or management GUI.

Upgrades: SVC master console users on IBM SAN Volume Controller code Version 5.1 and prior releases will need to migrate from their present user interface to the new user interface when upgrading to IBM SAN Volume Controller code Version V6.1. See the SVC documentation for further instructions.

For information on latest IBM System Storage SAN Volume Controller code and other related information visit the following website:

http://www-01.ibm.com/support/docview.wss?rs=591&uid=ssg1S4000979

11.2.3 IBM System Storage SAN Volume Controller maximum configuration

Table 11-1 summarizes the maximum configuration numbers supported for SVC at the time of writing. Obviously, the numbers in Table 11-1 can change with subsequent hardware and software releases. For the latest configuration information, always see the IBM website:

http://www.ibm.com/storage/support/2145

Table 11-1 SVC maximum configuration

Property	Maximum number	Comments
System (cluster) properties		
Nodes per system	8	Arranged as four I/O groups
Nodes per fabric	64	Maximum number of SVC nodes that can be present on the same Fibre Channel fabric
I/O groups per system	4	Each containing two nodes
Fabrics per system	4	The number of counterpart SANs which are supported
Inter-cluster partnerships per system	3	A system can partnered with up to three remote systems. No more than four systems can be in the same connected se
Node properties		
Logins per node Fibre Channel port	512	HBAs, disk controller ports, node ports within the same system and node ports from remote systems
iSCSI sessions per node	256	512 in IP failover mode (when partner node is unavailable)
Managed Disk properties		
Managed disks (MDisks) per system	4096	The maximum number of logical units which can be managed by a cluster. This number includes external MDisks which have not been configured into storage pools (managed disk groups)
Managed disks per storage pool (managed disk group)	128	

Property	Maximum number	Comments
Storage pools per system	128	
Managed disk extent size	8192 MB	
Capacity for an individual internal managed disk (array)		No limit is imposed beyond the maximum number of drives per array limits
Capacity for an individual external managed disk	1 PB	
Total storage capacity manageable per system	32 PB	Maximum requires an extent size of 8192 MB to be used
Other properties		
Volumes (VDisks) per system	8192	Maximum requires an 8-node cluster; refer to the volumes per I/O group limit next
Volumes per I/O group	2048	
Fully allocated volume capacity	256 TB	Maximum size for an individual fully-allocated volume
Host mappings per system	20000	
Host objects (IDs) per system	1024	A host object can contain both Fibre Channel ports and iSCSI names
Host objects (IDs) per I/O group	256	
FlashCopy mappings per system	4096	
FlashCopy consistency groups per system	127	
Internal drives per 2145-CF8 node	4	
Internal drives per 2145-CG8 node	4	

11.2.4 IBM System Storage SAN Volume Controller licensing

SVC is licensed by the capacity that is being managed. With this type of licensing, you select the number of terabytes available for your license for base virtualization, the FlashCopy feature, and the Metro Mirror and Global Mirror features.

Capacity upgrades can be carried out at anytime during the life of the SVC by purchasing a licence for the additional capacity required.

For more detailed information, you can consult your IBM sales representative.

11.2.5 IBM System Storage SAN Volume Controller publications

SVC supports a wide variety of disk storage and host operating system platforms. For the latest information, see the IBM website:

http://www-03.ibm.com/systems/storage/software/virtualization/svc/interop.html

Now we have reviewed the basic SVC concepts and components. For details and information about SVC implementation and best practices, see the following publications:

- *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933
- *IBM System Storage Solutions Handbook*, SG24-5250
- *SAN Volume Controller Best Practices and Performance Guidelines*, SG24-7521
- *Implementing the SVC in an OEM Environment*, SG24-7275
- *SAN Volume Controller V4.3.0 Advanced Copy Services*, SG24-7574

You must ensure that the firmware level of the DS5000 Storage Server can be used with the SAN Volume Controller clustered system.

For specific firmware levels and the latest supported hardware, as well as maximum number of LUNs per partition that are supported by the firmware level, see the following website:

www.ibm.com/storage/support/2145

11.3 IBM Storwize V7000

The IBM Storwize V7000 midrange storage system provides a modular storage solution that includes the capability to virtualize external SAN-attached storage and its own internal storage. The Storwize V7000 solution is built upon the IBM SAN Volume Controller (SVC) technology base and uses technology from the IBM System Storage DS8000 family.

Storwize V7000 provides a number of configuration options that are aimed at simplifying the implementation process. It also provides automated wizards, called *Directed Maintenance Procedures (DMP)*, to assist in resolving any events that might occur. Storwize V7000 is a clustered, scalable midrange storage system, as well as an external virtualization device.

Figure 11-7 shows the front view of the 2076-112 and 2076-212 enclosures.

Figure 11-7 IBM Storwize V7000 front view for 2076-112 and 2076-212 enclosures

Figure 11-8 shows the front view of the 2076-124 and 224 enclosures.

Figure 11-8 IBM Storwize V7000 front view for 2076-124 and 2076-224 enclosures

11.3.1 IBM Storwize V7000 features

These IBM Storwize V7000 features support enterprise class scalability and connectivity:

- Maximum of 240 internal drives in up to 9 expansion enclosures
- Maximum of 2048 host volumes with up to 512 volumes per host
- Maximum capacity of 256 TB per volume
- Maximum of 256 Fibre Channel attached hosts
- Maximum of 64 iSCSI attached hosts
- Maximum of 256 combined (iSCSI and Fibre Channel hosts)
- Maximum of 4096 MDisks
- Maximum of 1 PB capacity per MDisk
- Maximum virtualization capacity of 32 PB

See 11.3.4, “IBM Storwize V7000 maximum configuration” on page 487 for more information.

11.3.2 IBM Storwize V7000 hardware

The Storwize V7000 platform consists of enclosures and drives. Each enclosure contains two canisters that, although they can be replaced independently, are seen as part of the enclosure. Here are the main components of Storwize V7000:

- One control enclosure containing two node canisters, two PSUs, and 24 x 2.5” or 12 x 3.5” disk drives
- Optionally up to nine expansion enclosures containing two PSUs and 24 x 2.5” or 12 x 3.5” disk drives

Tip: The support for clustering two IBM Storwize V7000 has been added in IBM SAN Volume Controller code Version 6.2.

Key features of Storwize V7000 hardware are:

- 2U rack-mountable chassis
- Twenty four 2.5” drive bays (model x24) or twelve 3.5” drive bays (model x12)
- Up to 24 TB of physical storage per enclosure using 2 TB near-line SAS disk drive modules or up to 14 TB physical storage per enclosure using 600 GB SAS disk drive modules
- SAS disk drives, near-line SAS disk drives and SSDs
- Redundant dual-active intelligent RAID controllers
- 16 GB cache memory per control enclosure (8 GB per internal RAID controller) as a base feature

- For each control enclosure: Eight 8 Gbps Fibre Channel host ports (four 8 Gbps FC ports per RAID controller), four 1 Gbps and optionally four 10 Gbps iSCSI host ports (two 1 Gbps and optionally two 10 Gbps iSCSI host ports per RAID controller)
- RAID controller supports attachment of up to nine storage expansion units with configurations up to 240 TB physical storage capacities (480 TB in clustered systems)
- Dual power supplies and cooling components

IBM Storwize V7000 control enclosure

The control enclosure is a hardware unit that includes the chassis with a midplane for connection of node canisters, drives and power supplies with batteries. Figure 11-9 shows the rear view of IBM Storwize V7000 control enclosure with 2 node canisters.

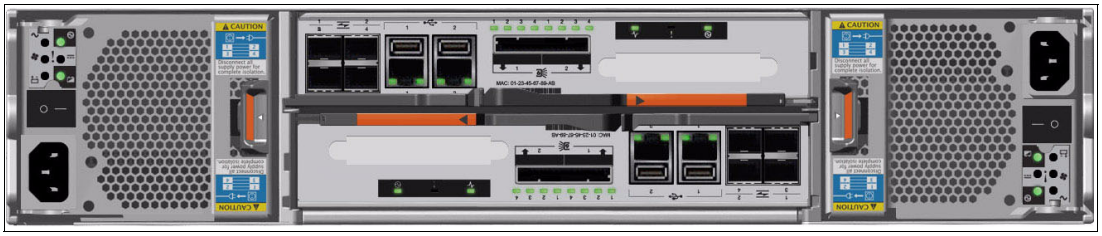


Figure 11-9 IBM Storwize V7000 controller rear view

There are four Fibre Channel ports on the left side of the canister. They are in a block of four in two rows of two connectors. The ports are numbered 1 to 4 from left to right, top to bottom. The ports operate at 2, 4, or 8 Gbps. Use of the ports is optional. There are two green LEDs associated with each port: the speed LED and link activity LED. Each of the nodes can be optionally expanded with iSCSI Host Interface Cards to provide for host connectivity up to 10 Gbps (Figure 11-10).

There are two 10/100/1000 Mbps Ethernet ports side by side on the canister. They are numbered 1 on the left and 2 on the right. Using port 1 is required; port 2 is optional. There are two LEDs associated with each Ethernet port.

There are two power supply slots, on the extreme left and extreme right, each taking up the full 2EIA height.

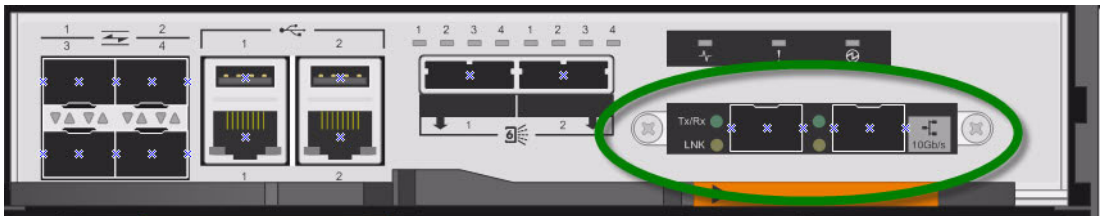


Figure 11-10 10 Gbps iSCSI Host Interface cards

IBM Storwize V7000 expansion enclosures

Enclosure is a hardware unit that includes the electronics to provide serial-attached SCSI (SAS) connections to the internal drives in the enclosure and SAS expansion ports for attachment of additional expansion enclosures.

The optional expansion enclosure contains two expansion canisters, disk drives, and two power supplies. There are two models of the control enclosure with one model providing 12 disk slots (Figure 11-7 on page 482) and the other providing 24 disk slots (Figure 11-8 on page 483).

Figure 11-11 shows rear view of the expansion enclosure.

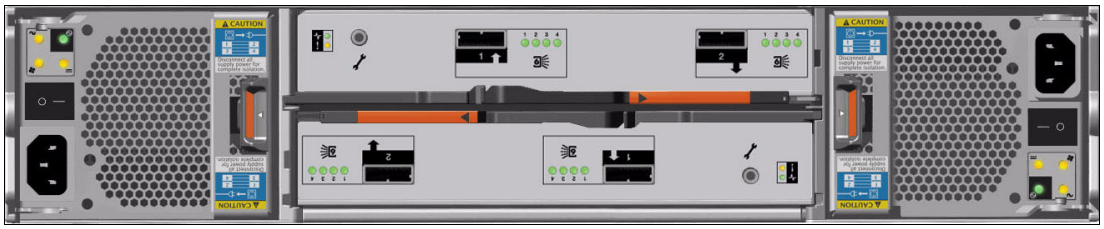


Figure 11-11 Rear view of the Expansion enclosure

11.3.3 IBM Storwize V7000 software

IBM Storwize V7000 code is based on that of the IBM System Storage SAN Volume Controller. Similarly to SVC, IBM Storwize V7000 is controlled by the clustered system, or a node pair, where both nodes hold the same version of internal code for failover purposes. IBM Storwize V7000 uses an easy to use initial setup that is contained within a USB key. The USB key is delivered with each storage system and contains the initialization application called **`InitTool.exe`**. A system management IP address, the subnet mask, and the network gateway address are required. The initialization application creates a configuration file on the USB key.

Initial setup of Storwize V7000

Storwize V7000 starts the initial setup as soon as you plug in the USB key with the newly created file in the storage system. For more details on first-time setup of the Storwize V7000, see *Implementing the IBM Storwize V7000 V6.3*, SG24-7938.

The Storwize V7000 uses the same management interface as SVC, called the SVC Console, as described in 11.2.2, “IBM System Storage SAN Volume Controller software” on page 479.

Figure 11-12 shows the welcome window of Storwize V7000 clustered system.

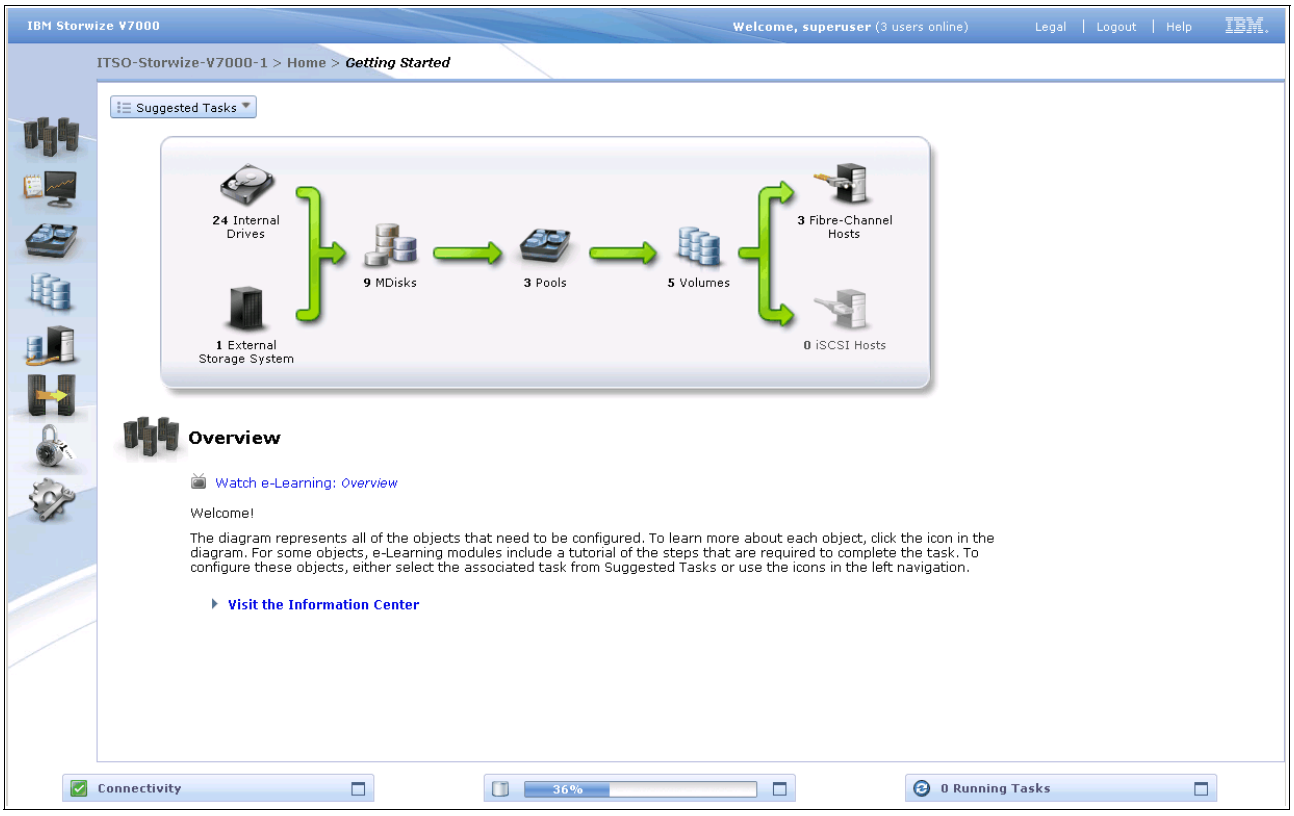


Figure 11-12 Management interface of IBM Storwize V7000

Another method of managing the system is the command-line interface (CLI). The CLI is a flexible tool for system management that uses the SSH protocol. A public / private SSH key pair is required for SSH access. The clustered system can be managed using the CLI, as shown in Example 11-1.

Example 11-1 System management using the CLI

```
IBM_2076:ITSO-Storwize-V7000-1:admin>svcinfo lsuser
id name       password ssh_key remote usergrp_id usergrp_name
0  superuser  yes      no      no     0          SecurityAdmin
1  admin      no       yes     no     0          SecurityAdmin
2  scottmiles yes      no      no     0          SecurityAdmin
3  peterblack yes      no      no     0          SecurityAdmin
```

Common software base for SVC and IBM Storwize V7000

Although IBM Storwize V7000 and IBM System Storage SAN Volume Controller each provide different hardware platforms with different market positioning, they both share the advantages of the common software base built on SAN Volume controller code Version 6.1. Figure 11-13 on page 487 illustrates the attributes of SVC and IBM Storwize V7000 software stack.

Tip: New SAN Volume controller code Version 6.3 is announced to be available for customers during writing of this book. It brings additional features and benefits like lower bandwidth support for Global Mirror, support for remote mirroring between SVC and Storwize V7000, and enhanced interoperability.

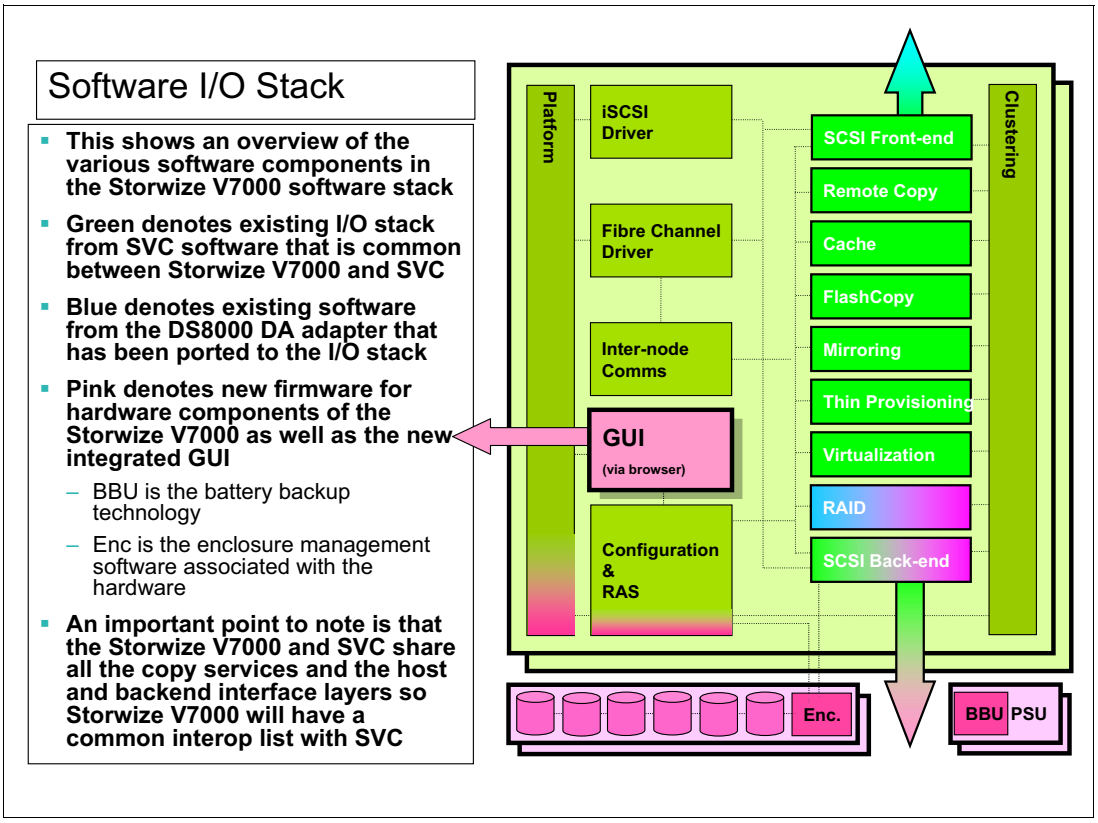


Figure 11-13 Software I/O Stack of IBM Storwize V7000

For information about the latest Storwize V7000 code and other related information, see the following website:

http://www-01.ibm.com/support/docview.wss?rs=591&uid=ssg1S4000978

11.3.4 IBM Storwize V7000 maximum configuration

Table 11-2 summarizes the maximum configuration numbers supported for IBM Storwize V7000 at the time of writing. Obviously, the numbers in Table 11-2 can change with subsequent hardware and software releases. For the latest configuration information, always see the IBM website:

http://www.ibm.com/storage/support/storwize/v7000

Table 11-2 IBM Storwize V7000 maximum configuration

Property	Maximum number	Comments
System (cluster) properties		
Control enclosures per system	2	Each control enclosure contains two node canisters
Nodes per system	4	Arranged as two I/O groups
Nodes per fabric	64	Maximum number of SVC nodes that can be present on the same Fibre Channel fabric
I/O groups per system	2	Each containing two nodes

Property	Maximum number	Comments
Fabrics per system	4	The number of counterpart SANs which are supported
Inter-cluster partnerships per system	3	A system can pretender with up to three remote systems. No more than four systems can be in the same connected set.
Node properties		
Logins per node Fibre Channel port	512	HBAs, disk controller ports, node ports within the same system and node ports from remote systems
iSCSI sessions per node	256	512 in IP failover mode (when partner node is unavailable)
Managed Disk properties		
Managed disks (MDisks) per system	4096	The maximum number of logical units which can be managed by a cluster. This number includes external MDisks which have not been configured into storage pools (managed disk groups)
Managed disks per storage pool (managed disk group)	128	
Storage pools per system	128	
Managed disk extent size	8192 MB	
Capacity for an individual internal managed disk (array)		No limit is imposed beyond the maximum number of drives per array limits
Capacity for an individual external managed disk	1 PB	
Total storage capacity manageable per system	32 PB	Maximum requires an extent size of 8192 MB to be used
Other properties		
Volumes (VDisks) per system	4096	Maximum requires a system containing two control enclosures
Volumes per I/O group	1024	
Fully allocated volume capacity	256 TB	Maximum size for an individual fully-allocated volume
Host mappings per system	20000	
Host objects (IDs) per system	512	A host object might contain both Fibre Channel ports and iSCSI names
Host objects (IDs) per I/O group	256	

Property	Maximum number	Comments
FlashCopy mappings per system	4096	
FlashCopy consistency groups per system	127	
Expansion enclosures per control enclosure	9	
Drives per I/O group	240	
Drives per system	480	Maximum requires a system containing two control enclosures, each with the maximum number of expansion enclosures
RAID arrays per system	128	

11.3.5 IBM Storwize V7000 licensing

Storwize V7000 might require the following licenses:

- Enclosure
- External Virtualization
- Remote Copy (Advanced Copy Services: Metro/Global Mirror)

Tip: If the Storwize V7000 is used as a general migration tool, then the appropriate External Virtualization licenses must be ordered. The only exception is if you want to migrate existing data from external storage to IBM Storwize V7000 internal storage; you can temporarily configure your External Storage license for use within 45 days. For a more-than-45-day migration requirement from external storage to IBM Storwize V7000 internal storage, the appropriate External Virtualization license must be ordered.

11.3.6 IBM Storwize V7000 publications

IBM Storwize V7000 supports a wide variety of disk storage and host operating system platforms. For the latest information, see the IBM website:

http://www-03.ibm.com/systems/storage/disk/storwize_v7000/interop.html

Now we reviewed the basic IBM Storwize V7000 concepts and components. For details and information about IBM Storwize V7000 implementation and best practices, see the following publications:

- *Implementing the IBM Storwize V7000 V6.3*, SG24-7938
- *IBM System Storage Solutions Handbook*, SG24-5250
- *SAN Volume Controller Best Practices and Performance Guidelines*, SG24-7521
- *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933
- *Introduction to Storage Area Networks*, SG24-5470
- *SAN Volume Controller V4.3.0 Advanced Copy Services*, SG24-7574

You must ensure that the firmware level of the DS5000 Storage Server can be used with the IBM Storwize V7000.

For specific firmware levels and the latest supported hardware, as well as maximum number of LUNs per partition that are supported by the firmware level, see the following website:

http://www.ibm.com/storage/support/storwize/v7000

Tip: At the time of writing this book, IBM announces the next major step in the evolution of the product, and the SVC code base: the IBM Storwize V7000 Unified. This integrates the IBM Common NAS software, which is used in the enterprise level SONAS product, with the block based Storwize V7000. While the Storwize V7000 is natively block based, the addition of the new File Modules as interfaces to the V7000 means that users can consolidate both block and file workloads onto one system. It is intended to complement the existing IBM NAS products, the SONAS for enterprise installations, and the IBM N series, which is still the suggested product for pure file workloads.

11.4 Virtualization systems Copy Services

In this section, we describe the Copy Services functions provided by the SVC and Storwize V7000 clustered systems, including FlashCopy and Remote Copy (Metro Mirror and Global Mirror). Copy Services functions are useful for making data copies for backup, application test, recovery, and so on. Virtualization systems make it easy to apply these functions to your environment using its intuitive management interface.

The Copy Services offered by the virtualization systems are FlashCopy, Metro Mirror, and Global Mirror.

If you plan to use SVC or IBM Storwize V7000 for all of your Copy Services, then purchasing the additional premium features, such and FlashCopy, VolumeCopy, or Enhanced Remote Mirroring for the DS5000, might not be necessary.

The SVC and IBM Storwize V7000 Copy Services functions provide various capabilities, for example:

- Support is provided for consistency groups for FlashCopy, Metro Mirror, and Global Mirror.
- Consistency groups can span across underlying storage subsystems.
- FlashCopy source volumes that reside on one disk subsystem can write to target volumes on another disk subsystem.
- Metro Mirror and Global Mirror source volumes can be copied to target volumes on a dissimilar storage subsystems.

11.4.1 SVC and IBM Storwize V7000 FlashCopy

FlashCopy provides the capability to perform an instantaneous point-in-time (PiT) copy of one or more volumes, which is a copy of the volume at that PiT. After being created, it no longer requires the source to be active or available.

FlashCopy works by defining a FlashCopy mapping consisting of a source volume and a target volume. Multiple FlashCopy mappings can be defined, and PiT consistency can be observed across multiple FlashCopy mappings using consistency groups.

Tip: As is the case with the DS5000, the first step before invoking this function is to make sure that all of the application data is written to disk. In other words, determine that the application data is consistent. Such result can be achieved, for example, by quiescing a database and flushing all data buffers to disk.

When FlashCopy is started, it makes a copy of a source volume to a target volume, and the original contents of the target volume is overwritten. The target volume presents the contents of the source volume as they existed at the single point in time (PiT) the FlashCopy was started.

When a FlashCopy operation is started, the source and target volumes are instantaneously available. The reason is that, when started, bitmaps are created to govern and redirect I/O to the source or target volume, respectively, depending on where the requested block is present, while the blocks are copied in the background from the source to the target volume.

Both the source and target volumes are available for read and write operations, although the background copy process has not yet completed copying across the data from the source to target volumes. See Figure 11-14.

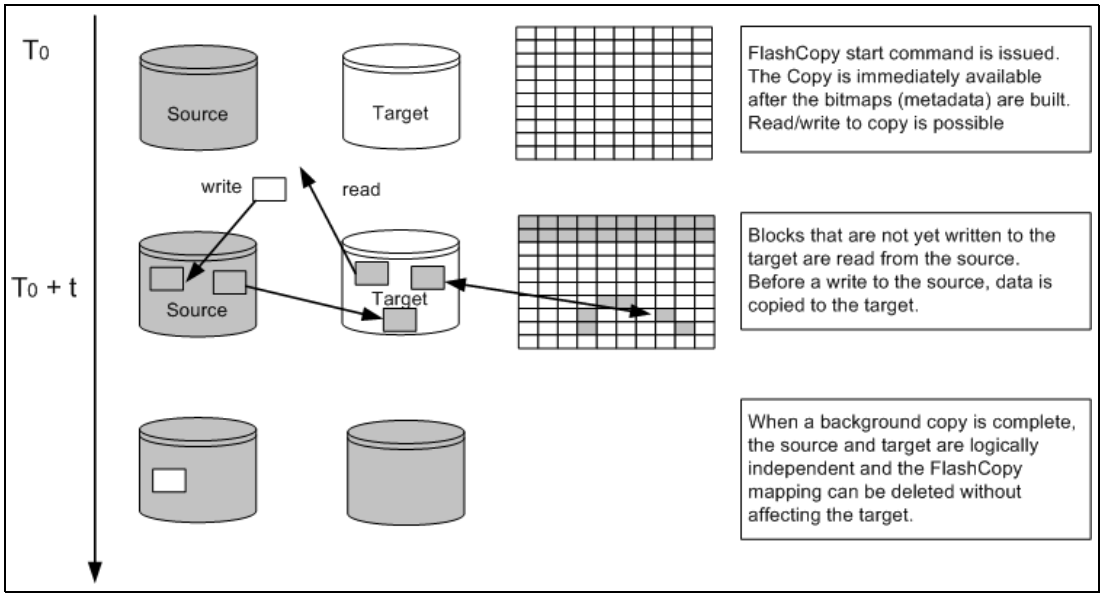


Figure 11-14 Implementation of FlashCopy

Virtualization systems can provide several features that increase the flexibility and usefulness of FlashCopy including:

- Full/Incremental FlashCopy
- FlashCopy Consistency Groups
- Multiple Target FlashCopy
- Cascaded FlashCopy
- Reverse FlashCopy
- FlashCopy Nocopy with Thin Provisioning

For more details about these capabilities, see *Implementing the IBM Storwize V7000 V6.3*, SG24-7938 and *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

Compared to the native DS5000 FlashCopy, the SVC and Storwize V7000 FlashCopy is more like a combination of the DS5000 FlashCopy and VolumeCopy functions, whereas the SVC and Storwize V7000 Thin-Provisioned FlashCopy is very similar to the DS5000 FlashCopy.

11.4.2 Metro Mirror

Metro Mirror is a type of Remote Copy that creates a synchronous copy of data from a master volume to an auxiliary volume. With synchronous copies, host applications write to the master volume but do not receive confirmation that the write operation has completed until the data is written to the auxiliary volume. This action ensures that both the volumes have identical data when the copy completes. After the initial copy completes, the Metro Mirror function maintains a fully synchronized copy of the source data at the target site at all times.

The general application of Metro Mirror is to maintain two real-time synchronized copies of a data set. Often, the two copies are geographically dispersed on two different virtualization clustered systems, although it is possible to use Metro Mirror in a single clustered system (within an I/O group). If the primary copy fails, the secondary copy can then be enabled for I/O operation.

Metro Mirror works by defining a Metro Mirror relationship between volumes of equal size. When creating the Metro Mirror relationship, one volume must be defined as the master, and the other as the auxiliary. In most common applications of Metro Mirror, the master volume contains the production copy of the data, and is used by the host application, whereas the auxiliary volume contains a mirrored copy of the data and is used for failover in disaster recovery scenarios.

The contents of the auxiliary volume that existed when the relationship was created is destroyed.

To provide management (and consistency) across a number of Metro Mirror relationships, consistency groups are supported (as with FlashCopy).

Storage virtualization systems provide both intracluster and intercluster Metro Mirror, as described next.

Intracluster Metro Mirror

Intracluster Metro Mirror can be applied within any single I/O group. Metro Mirror across I/O groups in the same virtualization clustered system is not supported.

Tip: Because intracluster Metro Mirror will consume more resources for a specific clustered system, compared to an intercluster Metro Mirror relationship, use intercluster Metro Mirror whenever possible.

Intercluster Metro Mirror

Intercluster Metro Mirror operations require a pair of virtualization clustered systems that are separated by a number of moderately high bandwidth links. The two clustered systems must be defined in a remote copy partnership, which must be defined on both clustered systems to establish a fully functional Metro Mirror partnership.

Using standard single mode connections, the supported distance between two virtualization clustered systems in a Metro Mirror partnership is 10 km, although greater distances can be achieved by using extenders.

A typical application of this function is to set up a dual-site solution using two clustered systems where the first site is considered the primary production site, and the second site is considered the failover site, which is activated when a failure of the first site is detected.

Metro Mirror is a fully synchronous remote copy technique, which ensures that updates are committed at both primary and secondary volumes before the application is given completion to an update. As shown in Figure 11-15, a write to the master volume is mirrored to the cache for the auxiliary volume before an acknowledge of the write is sent back to the host issuing the write.

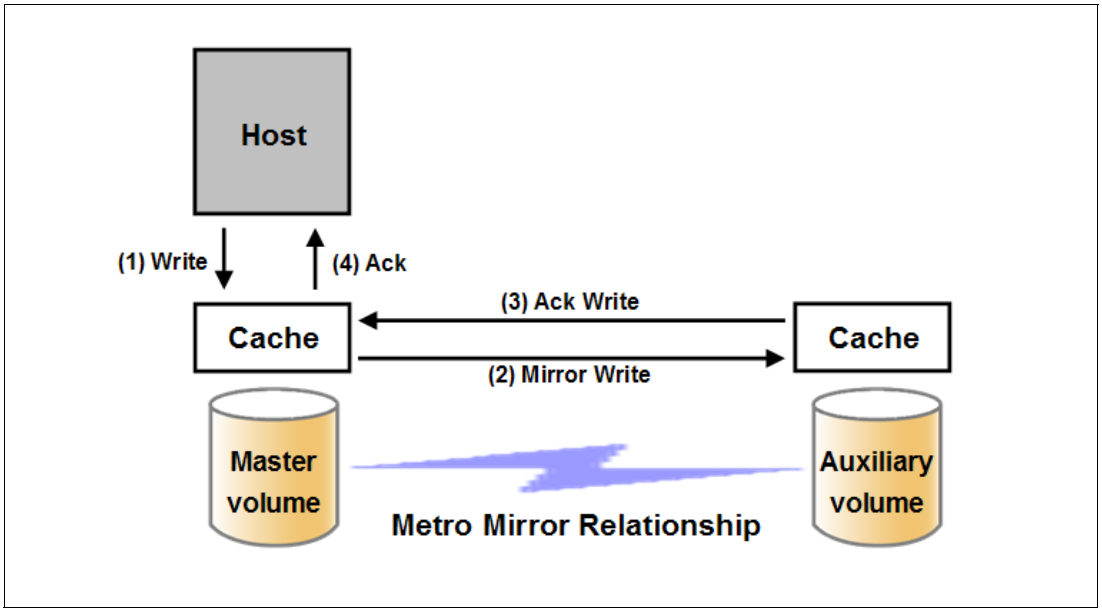


Figure 11-15 Metro Mirroring synchronous write sequence

While the Metro Mirror relationship is active, the secondary copy (volume) is not accessible for host application write I/O at any time. The clustered system allows read-only access to the secondary volume when it contains a consistent image. To enable access to the secondary volume for host operations, the Metro Mirror relationship must first be terminated.

11.4.3 Global Mirror

The Global Mirror provides an asynchronous copy, which means that the secondary volume is not an exact match of the primary volume at every point in time. The Global Mirror function provides the same function as Metro Mirror Remote Copy without requiring the hosts to wait for the full round-trip delay of the long distance link.

Global Mirror works by defining a Global Mirror relationship between two volumes of equal size and maintains the data consistency in an asynchronous manner. Global Mirror copy relationships are mostly intended for intercluster relationships over long distances.

When creating the Global Mirror relationship, one volume is defined as the master, and the other as the auxiliary. The relationship between the two copies is asymmetric. While the Global Mirror relationship is active, the secondary copy (volume) is not accessible for host application write I/O at any time. The clustered system allows read-only access to the secondary volume when it contains a consistent image.

When a host writes to a source volume, the data is copied to the source volume cache. The application is given an I/O completion while the data is sent to the target volume cache. At that stage, the update is not necessarily committed at the secondary site yet, which provides the capability of performing remote copy over distances exceeding the limitations of synchronous Remote Copy.

Figure 11-16 illustrates that a write operation to the master volume is acknowledged back to the host issuing the write before it is mirrored to the cache for the auxiliary volume.

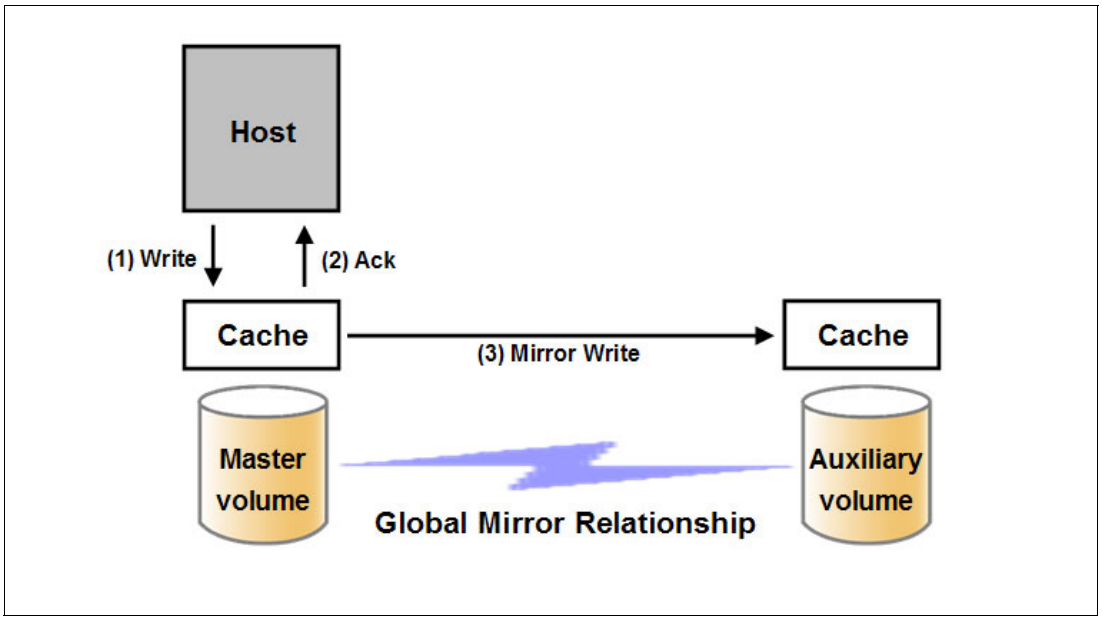


Figure 11-16 Global Mirror asynchronous write sequence

To provide management (and consistency) across a number of Global Mirror relationships, consistency groups are supported.

The clustered systems provide both intracluster and intercluster Global Mirror, which are described next.

Intracluster Global Mirror

Although Global Mirror is available for intracluster, it has no functional value for production use. Intracluster Metro Mirror provides the same capability for less overhead. However, leaving this functionality in place simplifies testing and allows experimentation and testing (for example, to validate server failover on a single test cluster).

Tip: SVC and Storwize V7000 systems running IBM SAN Volume Controller code Version 6.1 or higher do not support the use of intracluster Global Mirror functionality.

Intercluster Global Mirror

Intercluster Global Mirror operations require a pair of virtualization clustered systems that are commonly separated by a number of moderately high bandwidth links. The two clustered systems must each be defined in an Remote Copy partnership to establish a fully functional Global Mirror relationship.

11.4.4 Differences between DS5000 and SVC/Storwize V7000 Copy Services

In this section, we briefly describe the differences between the DS5000 Copy Services, and those implemented on SVC and Storwize V7000. Whenever a DS5000 is managed through a virtualization clustered system, use the clustered system Copy Services with the managed volumes, rather than the DS5000 Copy Services.

FlashCopy

On the DS5000 a target FlashCopy logical drive is not a physical copy of the source logical drive. It is only the logical equivalent of a complete physical copy because only the changed blocks are copied to the target (copy on write). Virtualization systems have a special type of FlashCopy called Thin-provisioned FlashCopy, similar to the DS5000 FlashCopy in terms of functionality. These kinds of FlashCopy are well-suited for applications that read data from the FlashCopy target logical drives or volumes, such as backup and data mining applications.

On the virtualization system, the normal FlashCopy creates a point-in-time image of the Fully Allocated volume, and in this case the target volume always has the same size as a source volume, not only in what it presents to a host, but as well in how much space it allocates from the storage pool. After being created, it no longer requires the source volume to be active or available. This image, after being created, is available to be mounted on other host servers or for recovery purposes.

To get the equivalent function of the normal SVC/Storwize V7000 FlashCopy on the DS5000 Storage Servers, you need to combine FlashCopy and VolumeCopy functions (that is, make a VolumeCopy of the FlashCopy target).

Metro Mirror and Global Mirror

Both the DS5000 Copy Services and the SVC/Storwize V7000 Copy Services offer a Metro Mirror and a Global Mirror. These Copy Services are similar in function between the two technologies, even though the underlying technology that creates and maintains these copies varies.

From a DS5000 standpoint, an important difference when using SVC/Storwize V7000 Copy Services over the DS5000 Copy Services is in host port requirement. The DS5000 Mirroring requires a dedicated host port from each of the DS5000 controllers, which means that ideally you need a DS5000 model with at least four host ports per controller (such as the DS5020 or the DS5100 Storage Servers) if you still need to have dual redundant host SAN fabrics when Remote Mirroring is implemented.

The SVC/Storwize V7000 does not dedicate ports for its mirroring services (nor does it require dedicated ports on the DS5000 Storage Server). Rather, a zoning configuration dedicates paths between the clustered systems.

Virtualization systems mirroring also offers the following advantages:

- SVC/Storwize V7000 maintain a control link on top of the FC link that is used for the Global/Metro Mirror I/O. No IP connection is needed to manage systems inside Remote Copy partnership from one centralized administration point.
- SVC/Storwize V7000 nodes have a large cache that can be further expanded.
- Virtualizations systems make it possible to mirror between various IBM storage subsystem models and also between storage subsystems from other vendors, which gives a possibility for data migration between various types of storage. You can find a complete list of supported storage subsystems at the following links:
  - IBM System Storage SAN Volume Controller:
    http://www-01.ibm.com/support/docview.wss?rs=591&uid=ssg1S4000979
  - IBM Storwize V7000:
    http://www-01.ibm.com/support/docview.wss?rs=591&uid=ssg1S4000978
- Virtualization systems implement a configuration model that maintains the mirror configuration and state through major events such as failover, recovery, and resynchronization to minimize user configuration action through these events.

- Virtualization systems implement flexible resynchronization support, enabling it to re-synchronize volume pairs that have suffered write I/O to both source and target volume and to resynchronize only those regions that are known to have changed.

Consistency groups

Virtualization systems can use consistency groups for all Copy Services including FlashCopy, Metro Mirror and Global Mirror. The use of consistency groups is required so that the data is written in the same sequence as intended by the application. The DS5000 can use consistency groups for Global Mirror relationships only.

Premium features

With the DS5000, you must purchase premium features such as FlashCopy, VolumeCopy, Enhanced Remote Mirror Copy Services, and additional storage partitioning to connect various hosts.

SVC or Storwize V7000 consume only one storage partition of DS5000. All of the hosts that are managed by the virtualization system are masked behind that single storage partition. Of course, if you only partially manage the DS5000 through SVC or Storwize V7000, you still need the additional storage partitions on the DS5000 for the various hosts that need access to the logical drives not managed through SVC virtualization systems.

11.5 Virtualization systems considerations

It is important to know the following restrictions when considering the use of virtualization systems:

- Virtualization system considerations:
  - All nodes in a virtualization system must be located close to one another (the maximum is 100 meters), within the same room or adjacent rooms for ease of service and maintenance. A clustered system can be connected (by the SAN fabric switches) to application hosts, storage subsystems, or other clustered systems, by shortwave optical fiber connections.
  - The maximum distance between the system and host or the system and the storage subsystem is 300 m for shortwave optical connections and 10 km for longwave optical connections. Longer distances are supported between systems that use the intercluster Metro Mirror or Global Mirror feature.
  - With shortwave connections, distances can be up to 150 m (8 Gbps), 380 m (4 Gbps), 500 m (2 Gbps) using OM3 fibre cables between the clustered system and the host, and between the clustered system and the disk subsystem.
- Node considerations:
  - Clustered systems nodes such as the CF8 and CG8 (SVC) and Storwize V7000 always contain one host bus adapter (HBA) with four Fibre Channel (FC) ports which can operate at 2, 4, or 8 Gbps link speeds.

    Tip: The DS5000 Storage Server model can have 8 Gbps host port connectivity through its 8 Gbps Host Interface Cards (HIC).

  - An SVC node will not function unless it is behind the appropriate UPS unit. That is one of the design considerations with the SVC. Even though the room might already be supplied with UPS-protected power, a dedicated UPS unit for each node is required.

- Network considerations:
  - All nodes in a clustered system must be on the same IP subnet so that the nodes can assume the same cluster IP address or service IP address in case of IP failover.
- Fabric considerations:
  - The nodes FC ports must be connected to the Fibre Channel fabric. If you are going to provide iSCSI connectivity to your hosts, you must connect the nodes Ethernet ports to Ethernet switches. Direct connections between node and host, or node and storage subsystem, are not supported.
  - The Fibre Channel switch must be zoned to permit the hosts to see the clustered nodes, and the clustered nodes to see the storage subsystems. The nodes within a clustered system must be zoned in such a way, that allows them to see each other, the disk subsystems, and the front-end host HBAs.
  - When a local and a remote fabrics are connected together for Metro Mirror purposes, the ISL hop count between a local node and a remote node cannot exceed seven hops.
  - No ISL hops are permitted between the nodes within the same I/O group.
  - No ISL hops are permitted between the nodes in various I/O groups inside same clustered system.
  - One ISL hop between the nodes and the DS5000 Storage Server controllers is permitted. Ideally, all storage controllers must be connected to the same Fibre Channel switches as the nodes.
  - In larger configurations, it is common to have ISLs between host systems and the clustered system nodes.

11.5.1 Preferred node

The preferred node is responsible for I/Os for the volume and coordinates sending the I/Os to the alternate node. While both systems will exhibit similar CPU utilization, the preferred node is a little busier. A preferred node is always responsible for the destaging of writes for volumes that it owns.

If a node fails, the other node in the I/O group takes over the I/O responsibilities of the failed node. Data loss during a node failure is prevented by mirroring the I/O write data cache between the two nodes in an I/O group. If a node has failed in an I/O group, the cache goes into write-through mode. Therefore, any write operations for the volume that are assigned to this I/O group are not cached but are sent directly to the disk storage subsystem. If both nodes in an I/O group go offline, the volumes that are assigned to the I/O group cannot be accessed.

11.5.2 Expanding volumes

It is possible to expand a volume in the clustered system, even if it is mapped to a host. Certain operating systems, such as Windows Server 2008, can handle the volumes being expanded even if the host has applications running.

However, a volume that is defined to be in a FlashCopy, Metro Mirror, or Global Mirror mapping on the clustered system cannot get expanded unless the mapping is removed, which means that the FlashCopy, Metro Mirror, or Global Mirror on that volume needs to be terminated before it is possible to expand it.

11.5.3 Multipathing

Each node presents a data volume to the SAN through four paths. Because in normal operation two nodes are used to provide redundant paths to the same storage, this means that a host with two HBAs can see eight paths to each volume presented by the SVC or Storwize V7000. Use zoning to limit the pathing from a minimum of two paths to the maximum available of eight paths, depending on the kind of high availability and performance you want to have in your configuration.

The multipathing device driver supported and delivered by SVC and Storwize V7000 is IBM Subsystem Device Driver (SDD) and two specific packages for AIX (SDDPCM) and Windows Server 2003/2008 (SDDDSM).

Subsystem Device Driver Device Specific Module (SDDDSM) provides Multipath I/O (MPIO) support based on the MPIO technology of Microsoft.

Subsystem Device Driver Path Control Module (SDDPCM) is a loadable path control module for supported storage devices to supply path management functions and error recovery algorithms. When the supported storage devices are configured as MPIO devices, SDDPCM is loaded as part of the AIX MPIO FCP (Fibre Channel Protocol) device driver during the configuration.

Depending on the operating system and the system architecture, SVC and Storwize V7000 can work with various multipath device drivers.

For a complete and updated list of supported Multipath device drivers, see these websites:

- IBM System Storage SAN Volume Controller:
  http://www-01.ibm.com/support/docview.wss?rs=591&uid=ssg1S4000979
- IBM Storwize V7000:
  http://www-01.ibm.com/support/docview.wss?rs=591&uid=ssg1S4000978

Note that the prior levels of host software packages that were specified are also tested for IBM SAN Volume Controller code Version 6.2 and allow for flexibility in maintaining the host software levels with respect to the SAN Volume Controller code Version. In other words, it is possible to upgrade the virtualization clustered system before upgrading the host software levels or after upgrading the software levels, depending on your maintenance schedule.

The number of paths from the nodes in a I/O group to a host must not exceed eight, even if it is not the maximum paths number handled by SDD. It is a supported configuration to have eight paths to each volume, but this design provides no performance benefit (indeed, under certain circumstances, it can even reduce performance), and it does not improve reliability or availability by any significant degree.

Tip: Even though a maximum of eight paths are supported with multipathing software, limiting the number of paths to four solves many issues with high port fanouts, fabric state changes, and host memory management, as well as improving performance.

In case of four host HBA ports, and because eight paths are not an optimum number, you need to configure your host definitions (and SAN zoning) as though the single host is two separate hosts. Then you need to map each of the two pseudo-hosts to volumes properly assigned to various preferred nodes in order to balance the I/O.

11.5.4 SAN aliases for SVC and IBM Storwize V7000: Guidelines

Modern SAN switches have three types of zoning available: port zoning, worldwide node name (WWNN) zoning, and worldwide port name (WWPN) zoning.

Tip: The preferred method is to use WWPN zoning *only*.

If available on your particular type of SAN switch, use zoning aliases when creating the SAN zones. Zoning aliases make your zoning easier to configure and understand, and result n fewer possibilities for errors.

One approach is to include multiple members in one alias, because zoning aliases can normally contain multiple members (just like zones). Create the following aliases:

- Create one alias that holds all the SVC/Storwize V7000 node ports on each fabric. Both have an extremely predictable WWPN structure because it always starts with 50:05:07:68 and ends with two octets that distinguish for you which node is which. The first digit of the third octet from the end is the port number on the node.

  In Example 11-2, notice that in the fabric switch (FabricA) we have connected port 1 and port 3 of each of the two nodes of a SVC cluster composed of only one I/O group (I/O group 0).

  Example 11-2 SVC cluster alias

```
SVC_Cluster_FabricA:
50:05:07:68:01:10:37:e5
50:05:07:68:01:30:37:e5
50:05:07:68:01:10:37:dc
50:05:07:68:01:30:37:dc
```

- Create one alias for each of the two controllers for a DS5000 Storage Server. See Example 11-3.

  Example 11-3 Storage aliases

```
DS5300_CtrlA_FabricA:
20:04:00:a0:b8:17:44:32
20:04:00:a0:b8:17:44:33

DS5300_CtrlB_FabricA:
20:05:00:a0:b8:17:44:32
20:05:00:a0:b8:17:44:33
```

- Create one alias for each I/O group port pair (that is, it needs to contain the first node in the I/O group, port 1, and the second node in the I/O group, port 1), because the best practices that we have described specify that each host HBA port is only supposed to see a single port on each of the two nodes in a SVC/Storwize V7000 system composed of a single I/O group, in order to keep the maximum or four paths. In case of multiple I/O groups within the SVC, assign the hosts to separate I/O groups, maintaining the rule of the four paths per host and considering a comparison among the workload generated by each host, in order to have a fair traffic balance among the nodes. See Example 11-4.

Example 11-4 I/O group port pair aliases

```
SVC_IOGroup0_Port1_FabricA:
50:05:07:68:01:10:37:e5
50:05:07:68:01:10:37:dc

SVC_IOGroup0_Port3_FabricA:
50:05:07:68:01:30:37:e5
50:05:07:68:01:30:37:dc
```

- We can then create an alias for each host HBA port, even if it is not strictly necessary, because each host is only going to appear in a single zone. See Example 11-5.

Example 11-5 Host alias

```
SVC_Host_HBA1_FabricA:
21:00:00:e0:8b:05:28:f9
```

Figure 11-17 shows a zoning configuration in which we have used the aliases in the previous examples.

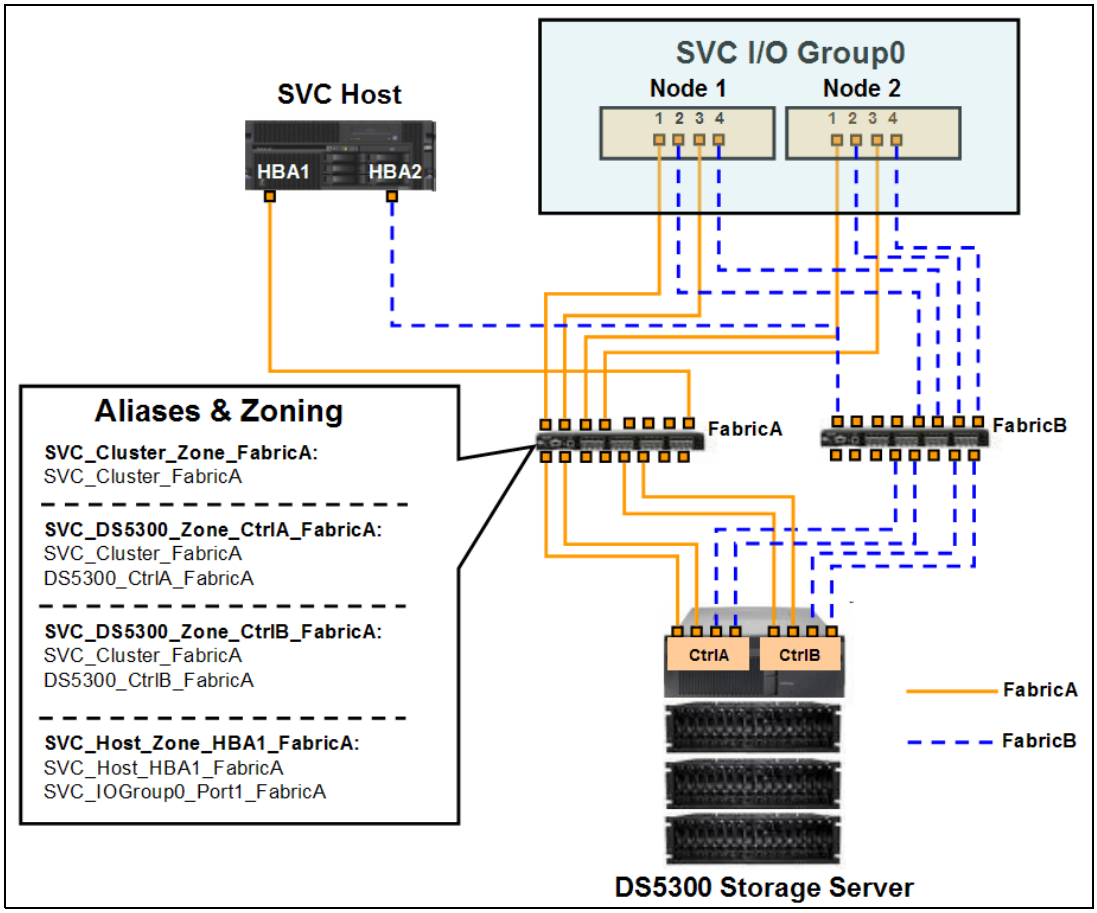


Figure 11-17 Aliases and zoning configuration example

11.5.5 SAN zoning rules

We use a four node SVC clustered system in SAN environment as shown in Figure 11-18 to demonstrate the zoning rules.

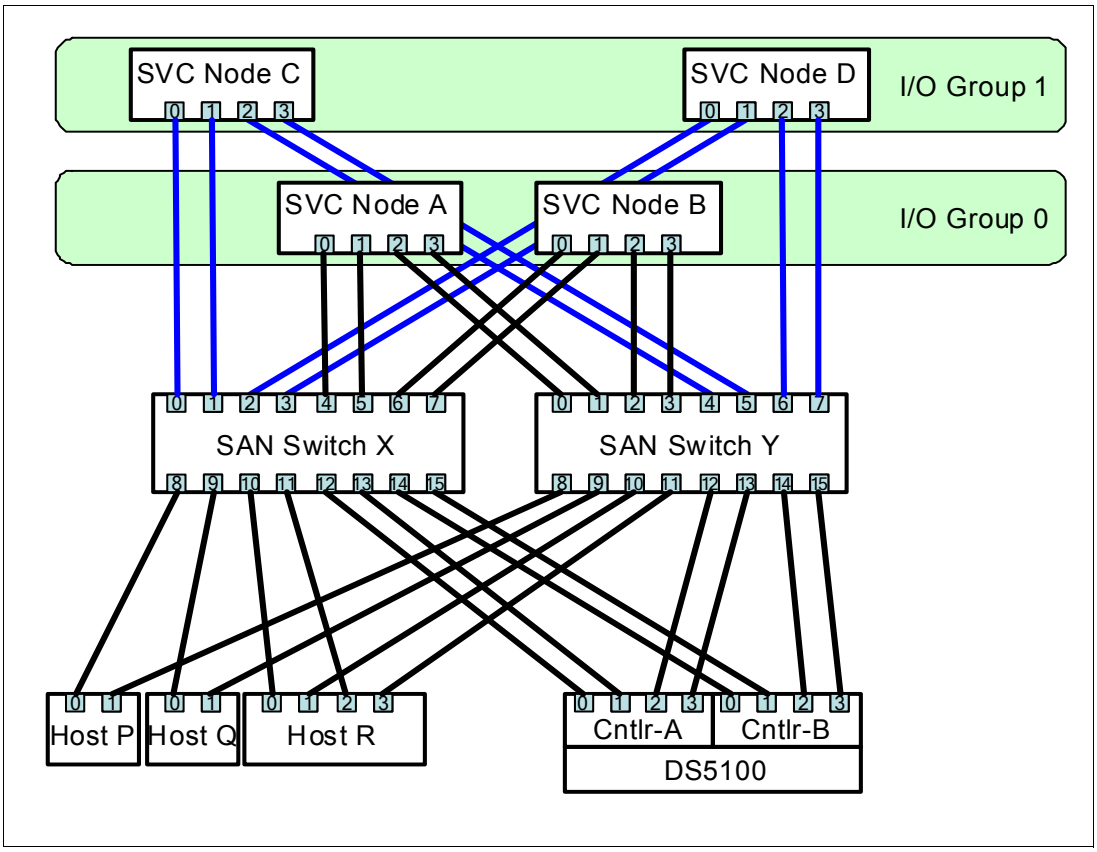


Figure 11-18 SAN Topology

Zoning on the SAN fabric(s) is used to ensure that all nodes in a clustered system can communicate with each other as well as preventing the hosts from accessing the back-end storage LUNs. Having cabled and created our SVC clustered system, we now need to proceed with defining our SAN fabric zones. We assume that aliases are created following the best practices guidelines described in the previous section. The configuration rules state that there must be at least three separate zones on each fabric with connections to the SVC ports.

Cluster zone

A *cluster zone* is required for inter-node communications. Every SVC node must be zoned to see every other node in the SVC clustered system. In our example, Figure 11-19, we have four SVC nodes and two independent SAN fabrics. Eight SVC ports are connected to Switch X and the remaining eight ports connect to Switch Y. Therefore, our cluster zone must include just the eight SVC ports visible on that fabric.

When configuring zones for communication between nodes in the same clustered system, the minimum configuration requires that all Fibre Channel ports on a node detect at least one Fibre Channel port on each other node in the same clustered system.

Tip: Up to 4 independent SAN fabrics are supported with IBM System Storage SAN Volume Controller and IBM Storwize V7000.

Figure 11-19 highlights all the ports in red that must be included in the single cluster zone on Switch X. Identical zones must be created on Switch Y.

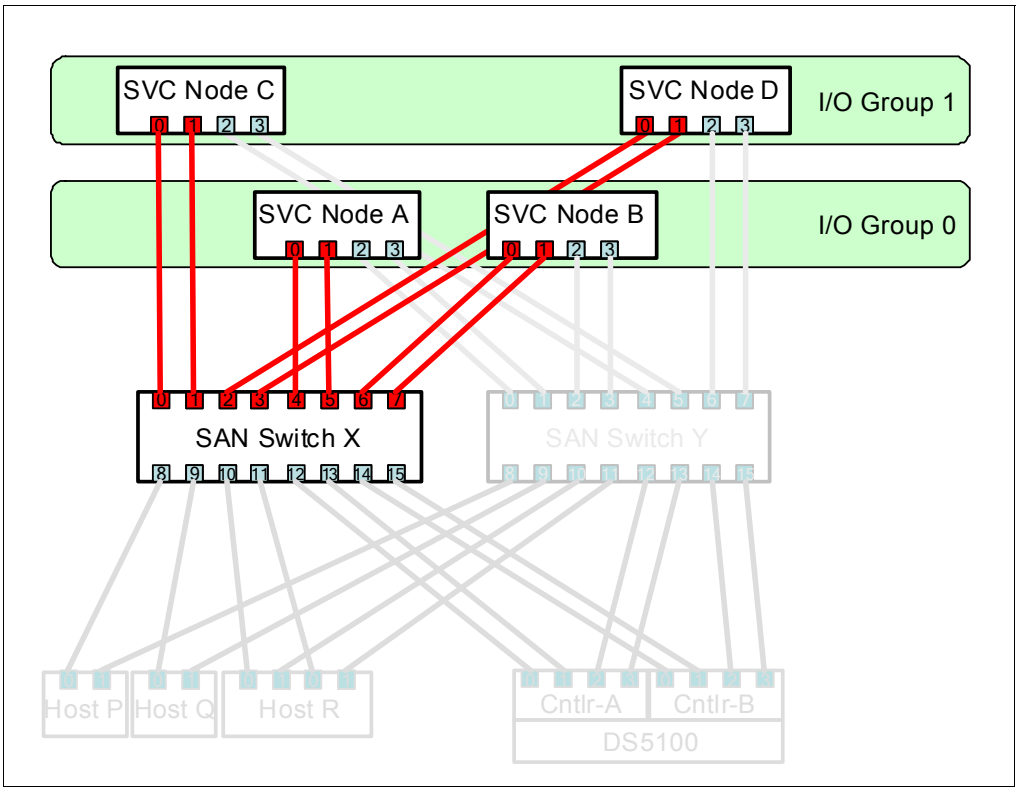


Figure 11-19 The SVC Cluster Zone

Host zone

A volume is always presented through all eight FC ports in the two clustered system nodes that make up the I/O group which owns it. Therefore, in a dual fabric configuration, the number of paths to a volume is four multiplied by the number of host ports if there was no zoning to limit the available paths. Each host with two HBA ports has eight paths and the host with four HBA ports has 16 paths to a disk device.

The number of volume paths from the nodes to a host must not exceed eight. Configurations in which this number is exceeded are not supported.

This rule exists to limit the number of paths that must be resolved by the multipathing device driver. It is best practice to restrict the number of volume paths from a host to four. To achieve this, zone the SANs so that each host bus adapter port is zoned with one port for each node in an I/O group. Zone each HBA port on the same host to another set of node ports to maximize performance and redundancy.

If all hosts are running the same operating system and contain identical HBAs, then it is acceptable to create a single zone containing HBAs from all hosts together with a single port from each I/O group. However, HBAs from miscellaneous vendors or running various operating systems must not be included within the same zone.

It is always preferable to reduce the zoning granularity. Ideally, we must aim to define each zone to include just a single initiator (HBA port), which isolates the host from any SAN fabric related problems on neighboring hosts.

In our example configuration, hosts P and Q each have two HBAs, whereas host R has four HBAs.

Figure 11-20 shows how we can use zoning to restrict the available paths from each host through SAN Switch X. For clarity and best practice, we create an alias for each pair of FC ports within each I/O group, which then makes it easy to define separate host zones without exceeding the path limits. Identical zones must be created on Switch Y.

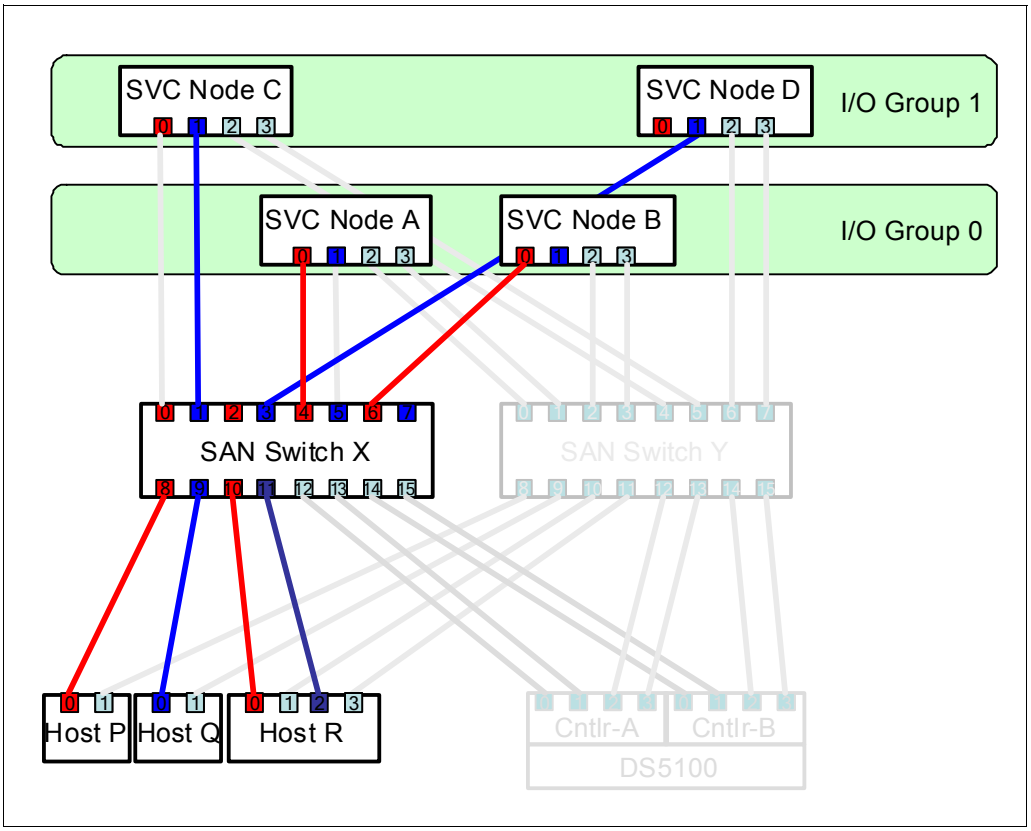


Figure 11-20 SVC Host Zones

By implementing the red and blue zones in our example configuration, we have evenly distributed the host access across all SVC ports whilst keeping within the maximum path limits. Now each volume has just 4 paths to each host.

In our example, we assume that host P only requires access to I/O group 0 volumes, host Q only requires access to I/O group 1 and host R accesses volumes in both I/O groups.

Storage zone

Switch zones that contain storage subsystem ports must not have more than 40 ports. A configuration that exceeds 40 ports is not supported. The DS5000 Storage Server must be configured with at least two ports per controller for a total of four ports per DS5000.

Figure 11-21 shows the ports present within the storage zone on Switch X in our example configuration. It includes all SVC node ports together with a pair of DS5100 host ports from each controller. An identical zone must be created on Switch Y.

The DS5100 has eight host ports per controller. It is acceptable to include additional host ports within the storage zone if increased back-end bandwidth is required. However, all nodes in a clustered system must be able to detect the same ports on each back-end storage subsystem. Operation in a mode where two nodes detect another set of ports on the same storage subsystem is degraded, and the system logs errors that request a repair action, which can occur if inappropriate zoning is applied to the fabric or if inappropriate Storage Partition mapping is used on the DS5000 Storage Server.

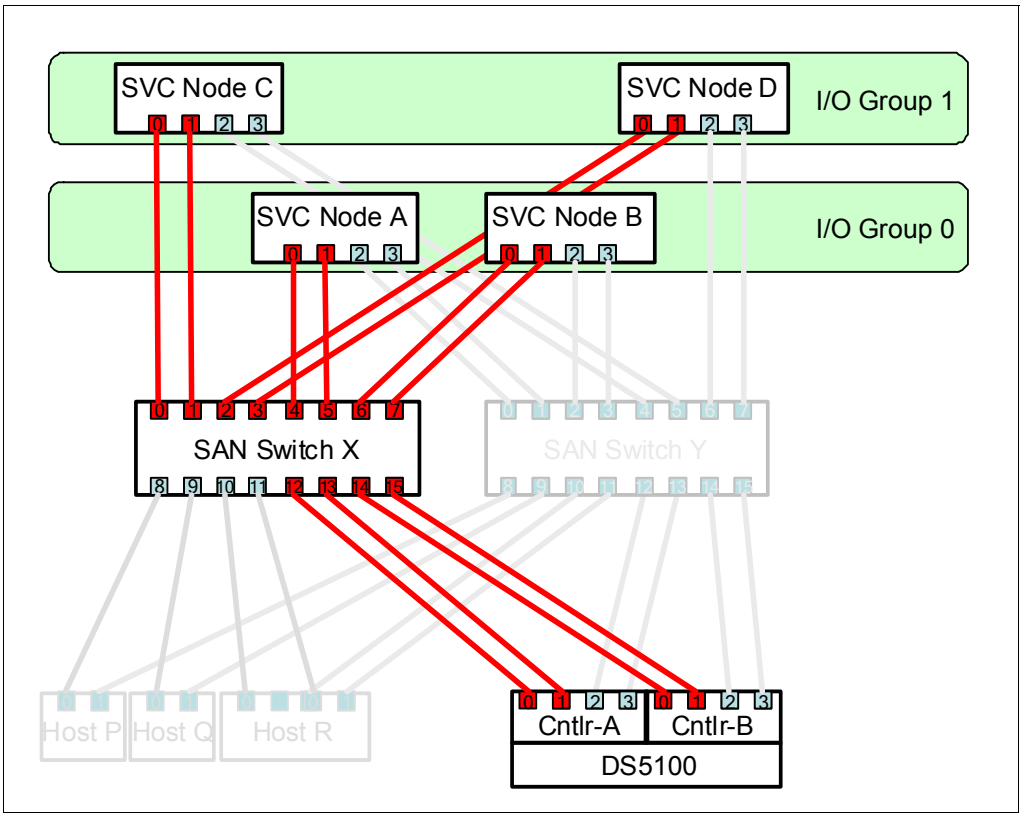


Figure 11-21 SVC Storage Zone

Tip: The diagram shows a single storage zone consisting of ports from both controllers, and it is a supported configuration. However, it is a best practice that these two controllers are not in the same zone if you have attached them to the same SAN. Also, create an alias for each controller.

See Example 11-6 for a sample configuration.

Example 11-6 Alias and storage zoning best practices

```
SVC_Cluster_SAN_Switch_X:
50:05:07:68:01:10:37:e5
50:05:07:68:01:20:37:e5
50:05:07:68:01:10:37:dc
50:05:07:68:01:20:37:dc
50:05:07:68:01:10:1d:1c
50:05:07:68:01:20:1d:1c
50:05:07:68:01:10:27:e2
50:05:07:68:01:20:27:e2
```

```
DS5k_Ctrl_A_SAN_Switch_X:
20:04:00:a0:b8:17:44:32
20:04:00:a0:b8:17:44:33

DS5k_Ctrl_B_SAN_Switch_X:
20:05:00:a0:b8:17:44:32
20:05:00:a0:b8:17:44:33

SVC_DS5k_Zone_Ctrl_A_SAN_Switch_X:
SVC_Cluster_SAN_Switch_X
DS5k_Ctrl_A_SAN_Switch_X

SVC_DS5k_Zone_Ctrl_B_SAN_Switch_X:
SVC_Cluster_SAN_Switch_X
DS5k_Ctrl_B_SAN_Switch_X
```

Inter-cluster zone

An inter-cluster zone is only required when Metro Mirroring or Global Mirror is used between virtualization clustered systems. It includes ports of clustered systems from both sites.

11.6 Storage virtualization systems with DS5000 best practices

There are several best practices and guidelines to follow when a DS5000 is used through virtualization systems so that it performs well with most applications.

The key is to plan how the DS5000 disks (logical drives or LUNs) will be allocated and used by clustered system because LUNs (which are the actual disks seen by clustered system as MDisks) need to be created first on the DS5000 Storage Server.

11.6.1 Disk allocation process

Here we list a simplified process of allocating disk from the storage subsystem array to the SVC or Storwize V7000:

1. The LUN is created from the array using Storage Manager.
2. The LUN is presented to clustered system by assigning the LUN to the SVC or Storwize V7000 Host.
3. The LUN is discovered by the management interface and then gets created as an MDisk.
4. The MDisk is assigned to a storage pool. The storage pool determines the extent size.
5. The host definition is created on clustered system and assigned to I/O group.
6. The extents on MDisks inside storage pool are then used to create volume that is assigned to an I/O group (must be in same I/O group as host).
7. The volume is mapped to the host.

LUN creation

The storage arrays created on the DS5000 when defining LUNs that will be assigned to the virtualization systems must be configured with only one LUN per array. The LUN must be sized to utilize all of the available storage space on the array. In the DS5000, you must have an equal number of arrays and LUNs, equally spread on the two controllers. After having assigned spare disks in the DS5000, define your arrays with RAID protection, and create as few LUNs as you can in the array. If possible, make the LUNs the same size, so you can utilize the full capacity when striping volumes on the virtualization system.

The decision for array RAID level and which DS5000 physical disks it contains is made when creating LUNs on the DS5000 before it is defined as an MDisk and added to a storage pool. It is essential to know at this point which particular host and the nature of the host applications that will access the volumes.

In other words, key decisions for reliability, availability, and performance are still made at the DS5000 level. Therefore you still need to follow the guidelines given specifically for the DS5000 in previous chapters of this book.

Here we summarize the most important guidelines:

- For data protection, ensure that the array is created with enclosure loss protection so that if an enclosure fails on the DS5000, the array will still remain functional.
- For database transaction logs, RAID 10 gives maximum protection and performance. Database transaction logs require sequential writes to the log files. These must be well protected and must deliver high performance. For database data files, RAID 5 and RAID 6 offers a balance between protection and performance.

  For best practice, the transaction logs and data files are best to be in a dedicated array:

  - Create an array with one LUN of the appropriate size for the logs. Because this DS5000 LUN will then be defined as an MDisk and assigned to a storage pool, make sure that the LUN size is appropriate for the extent size assigned to this storage pool.
  - Map the whole MDisk (all extents) and only extents from that MDisk to the volume. Doing so ensures that the host has dedicated use of that physical disk (DS5000 LUN), which also ensures that no other volumes will be mapped to that MDisk and possibly cause disk thrashing.
  - For applications with small, highly random I/Os, you can stripe volumes across multiple MDisks inside the same storage pool for performance improvement. However, it is only true with applications that generate small high random I/Os. In a case of large blocksize I/Os that are mainly sequential data, using multiple MDdisks to stripe the extents across does not give a performance benefit.
- For file Server environment, RAID 5 offers a good level of performance and protection.

  If their requirements are not too demanding, several host applications can use the same MDisk (that is, they use separate volumes but created from extents of the same MDIsk). In other words, in this case several hosts share the same DS5000 LUN through SVC or Storwize V7000. Still, you must monitor the DS5000 LUN (array) to ensure that excessive disk thrashing is not occurring. If thrashing occurs then the volume can be migrated to another storage pool.

Also, ensure that only LUNs from one storage subsystem, such as a specific DS5000, are present in one storage pool, mainly for availability reasons, because the failure of one storage subsystem will make the storage pool go offline, and thereby all volumes using MDisks belonging to the storage pool will go offline.

It is important that LUNs are properly balanced across the controllers prior to performing MDisk discovery. Failing to properly balance LUNs across storage subsystem controllers in advance can result in a suboptimal pathing configuration to the back-end disks, which can cause a performance degradation. Ensure that storage subsystems have all controllers online and that all LUNs have been distributed to their preferred controller (local affinity) prior to performing MDisk discovery. Pathing can always be rebalanced later, however, often not until after lengthy problem isolation has taken place.

Controller affinity and preferred path

In this context, affinity refers to the controller in a dual-controller subsystem (such as the DS5000 Storage Server) that has been assigned access to the back-end storage for a specific LUN under nominal conditions (that is to say, both controllers are active). Preferred path refers to the host side connections that are physically connected to the controller that has the assigned affinity for the corresponding LUN being accessed.

For the DS5000, preferred path is important, because Fibre Channel cards are integrated into the controller. This architecture allows “dynamic” multipathing and “active/standby” pathing through Fibre Channel cards that are attached to the same controller (the SVC or Storwize V7000 do not support dynamic multipathing) and an alternate set of paths, which are configured to the other controller that will be used if the corresponding controller fails. The DS5000 differs from other storage subsystems such as the DS8000, because it has the capability to transfer ownership of LUNs at the LUN level as opposed to the controller level.

The virtualization clustered system attempts to follow IBM DS5000 series specified preferred ownership. You can specify on administrative level which controller (A or B) is used as the preferred path to perform I/O operations to a given LUN. If the clustered system can see the ports of the preferred controller and no error conditions exist, then the clustered system accesses that LUN through one of the ports on that controller. Under error conditions, the preferred ownership is ignored.

ADT for DS5000

The DS5000 has a feature called Auto Logical Drive Transfer (ADT). This feature allows logical drive level failover as opposed to controller level failover. When you enable this option, the DS5000 moves LUN ownership between controllers according to the path used by the host. For the SVC and Storwize V7000, the ADT feature is enabled by default when you select the “IBM TS SAN VCE” host type when you configure the DS5000.

Tip: It is important that you select the “IBM TS SAN VCE” host type when configuring the DS5000 for SVC or Storwize V7000 attachment in order to allow them to properly manage the back-end paths. If the host type is incorrect, virtualization systems will report a 1625 (“incorrect controller configuration”) error.

MDisks

It is important that LUNs are properly balanced across storage controllers prior to performing MDisk discovery. Failing to properly balance LUNs across storage controllers in advance can result in a suboptimal pathing configuration to the back-end disks, which can cause a performance degradation. Ensure that storage subsystems have all controllers online and that all LUNs have been distributed to their preferred controller prior to performing MDisk discovery. Pathing can always be rebalanced later, however, often not until after lengthy problem isolation has taken place.

If you discover that the LUNs are not evenly distributed across the dual controllers in a DS5000, you can dynamically change the LUN affinity. However, the virtualization system will move them back to the original controller, and the DS5000 will generate an error indicating that the LUN is no longer on its preferred controller. To correct this situation, you need to run the SVC command **`svctask detectmdisk`** or use the SVC console option “Detect MDisks.” Virtualization system will query the DS5000 again and access the LUNs through the new preferred controller configuration.

IBM SAN Volume Controller code Version 6.1 provides for better load distribution across paths within storage pools. In previous code levels, the path to MDisk assignment was made in a round-robin fashion across all MDisks configured to the cluster.

With that method, no attention is paid to how MDisks within storage pools are distributed across paths and therefore it is possible and even likely to have certain paths be more heavily loaded than others. This condition became even more likely to occur as the number of MDisks contained in the storage pool reduced. IBM SAN Volume Controller code Version 6.1 contains logic that considers MDisks within storage pools and more effectively distributes their active paths based on the storage controller ports available. The **`svctask detectmdisk`** commands need to be run following the creation or modification (add or remove MDisk) of storage pools for paths to be redistributed.

To ensure sufficient bandwidth to the storage controller and an even balance across storage controller ports, the number of MDisks per storage pool is to be a multiple of the number of storage ports available. For example, if a storage pool has 8 storage controller ports available to it, then it is to contain either 8, 16, 24, 32, or 40 MDisks. Exceeding 40 MDisks per storage pool is not advisable

Volumes and extent size

Volumes are allocated in whole numbers of extents so the volume size must be created as multiples to the extent size, so as not to waste storage at the end of each MDisk. For example if the extent size is 16 MB then the volume must be created in multiples of 16 MB.

When creating a new volume, the first MDisk from which to allocate an extent is chosen in a pseudo-random way rather than simply choosing the next disk in a round-robin fashion. The pseudo-random algorithm avoids the situation whereby the “striping effect” inherent in a round-robin algorithm places the first extent for a large number of volumes on the same MDisk. Placing the first extent of a number of volumes on the same MDisk can lead to poor performance for workloads that place a large I/O load on the first extent of each volume, or that create multiple sequential streams. Figure 11-22 shows the mapping of extents between Mdisks and volumes.

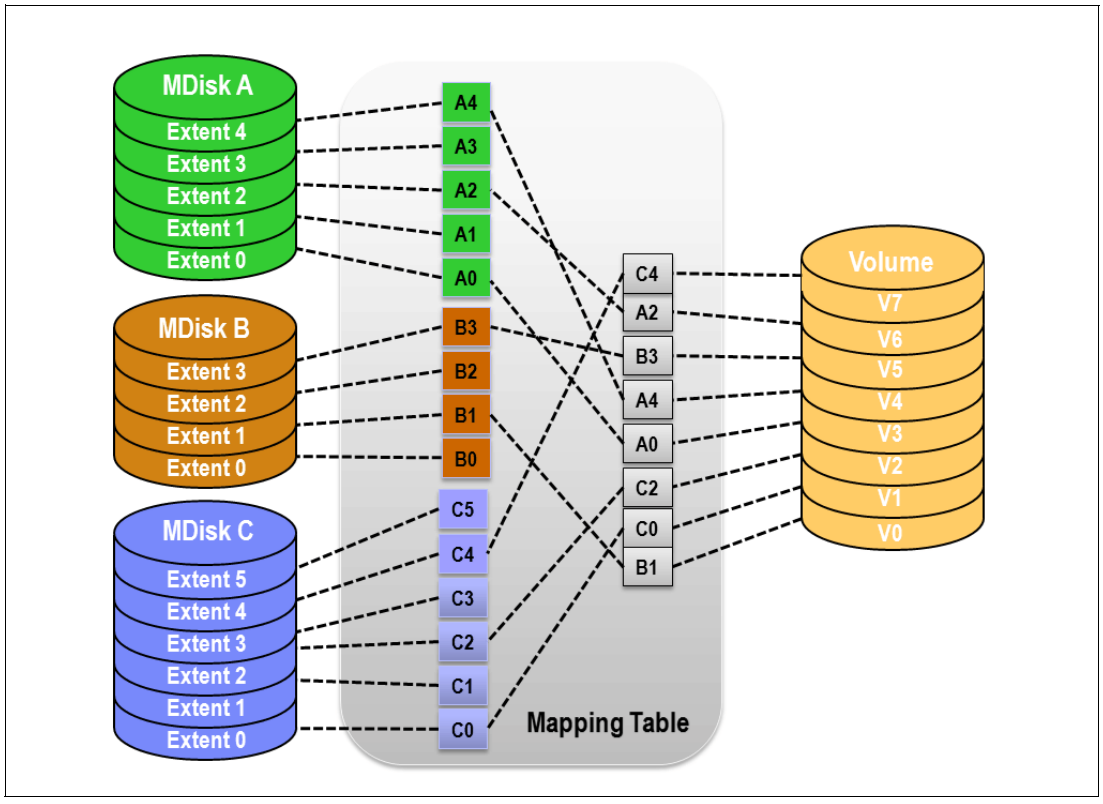


Figure 11-22 Simple view of block virtualization

DS5000 array and segment size

The DS5000 identifies the SVC or Storwize V7000 as a host and does not need to be configured for specific applications. The *stripe width* is a key parameter and in this context it refers to the number of disks that makes up a Redundant Array of Independent Disks (RAID) array.

With RAID 5 arrays, determining the number of physical drives to put into an array always presents a compromise. Striping across a larger number of drives can improve performance for transaction-based workloads. However, striping can also have a negative effect on sequential workloads. A common mistake that people make when selecting array stripe width is the tendency to focus only on the capability of a single array to perform various workloads. However, you must also consider in this decision the aggregate throughput requirements of the entire storage subsystem. A large number of physical disks in an array can create a workload imbalance between the controllers, because only one controller of the DS5000 actively accesses a specific array. On the other hand, be cautious when using large disk drives so that you do not have too few spindles to handle the load.

When selecting stripe width, you must consider its effect on RAID array rebuild time and availability. A larger number of disks in an array increases the rebuild time for disk failures, which can have a negative effect on performance. Additionally, more disks in an array increase the probability of having a second drive fail within the same array prior to the rebuild completion of an initial drive failure, which is an inherent exposure to the RAID 5 architecture.

Tip: For the DS5000, a stripe width of 4+P and 8+P is best suited for RAID 5 protection, and a stripe width of 8+P+Q is best suited for RAID 6.

The segment size is the maximum amount of data that the controller can write at once on one physical disk in a logical drive before writing to the next physical disk.

With hosts directly attached to the storage subsytem without SVC or Storwize V7000, considerations are often made to align device data partitions to physical drive boundaries within the storage subsystem. For the virtualization clustered systems, aligning device data partitions to physical drive boundaries within the storage subsystem is less critical based on the caching that the virtualization systems provide, and based on the fact that there is less variation in its I/O profile, which is used to access back-end disks. The dynamic cache stage size with back-end destage ranging 32 - 256 KB is implemented with IBM SAN Volume Controller code Version 6.1 and higher.

For the virtualization system, the opportunity for full stride writes occurs with large sequential workloads, and in that case, the larger the segment size is, the better. However, larger segment sizes can adversely affect random I/O, because they can cause disk fragmentation. The virtualization system and controller cache do a good job of hiding the RAID 5 write penalty for random I/O, and therefore, larger segment sizes can be accommodated. The primary consideration for selecting segment size is to ensure that a single host I/O will fit within a single segment to prevent accessing multiple physical drives. Testing has shown that the best compromise for handling all workloads is to use a segment size of 256 KB.

Overall, in an SVC and Storwize V7000 configuration, larger DS5000 segment sizes will give you better performance even in a random IO environment due to the use of extents that generally provide a sequential I/O pattern on the DS5000 even with random I/O requests from the host.

Tip: Use 256 KB segment size as the best compromise for mixed type workloads.

The smaller the MDisk / array / LUN, then the greater the spread of the extents across the MDisks in the storage pools, which reduces disk thrashing as the number of volumes increases. When a striped volume is created, extents are taken one at a time from each MDisk in the group, so with more MDisks in the group, the array contention is kept to a minimum. When adding more disks to a subsystem, consider adding the new MDisks to existing storage pools rather then creating additional small storage pools. Scripts are available to restripe volume extents evenly across all MDisks in the storage pools if required. Go to the website http://www.ibm.com/alphaworks and search for svctools.

Tip: To evenly spread the performance, arrays (LUNs) must be all the same size within a storage pools.

Cache block size

The size of the cache memory allocation unit can be either 4 K, 8 K, 16 K, or 32 K. The DS5000 series uses a 4 KB cache block size by default; however, it can be changed to 8 KB, 16 KB or 32 KB. Depending on the IBM DS5000 model, use a host type of IBM TS SAN VCE to establish the correct cache block size for the SVC or Storwize V7000 system. Either set this as the system default host type or, if partitioning is enabled, associate each virtualization clustered system port with the host type mentioned before.

Disk thrashing

Storage virtualization clustered systems make it simple to carve up storage and allocate it to the hosts, which also can cause problems if multiple hosts use the same physical disks and cause them to be very heavily utilized. Multiple hosts using the same array is not a problem in itself, it is the disk requirements of each host that might cause the problems. For example, you might not want a database to share the same disks as the email server or a backup server. The disk requirements might vary widely. Careful planning must be undertaken to ensure that the hosts that share the same array do not cause a problem due to heavy I/O operations.

Dual SAN fabric

To meet the business requirements for high availability, build a dual fabric network (that is, two independent fabrics that do not connect to each other). Resources such as DS5000 Storage Server have two or more host ports per controller. They are used to connect both controllers of the storage subsystem to each fabric, which means that controller A in the DS5000 is connected to counterpart SAN A, and controller B in the DS5000 is connected to counterpart SAN B, which improves data bandwidth performance and provides redundancy for high availability. However, this does increase the cost, as switches and ports are duplicated

Distance limitations

Ensure that the nodes of the clustered system do not exceed any of the distance limitations. Ideally, you will want all nodes in the same room. If you want to have SVC or Storwize V7000 in separate locations that are in separate buildings or farther away, then create a separate clustered system. For disaster recovery and business continuity purposes, using the appropriate Copy Services between the two clustered systems will provide the level of protection required.

Similarly, if you have more than one data center, then you want a single clustered system in each data center and replicate any data between them.

11.6.2 DS5000 tuning summary

In Table 11-3 we summarize the best values for the main SVC/Storwize V7000 and DS5000 tunable parameters.

Table 11-3 Tuning summary

Models	Attribute	Value
SVC/Storwize V7000	Extent Size (MB)	256
SVC/Storwize V7000	Managed Mode	Striped
DS5000	Segment Size (KB)	256
DS5000	Cache Block Size (KB)	8
DS5000	Cache Flush Control	50/50
DS5000	Read Ahead	1(Enabled)
DS5000	RAID5	4+P, 8+P
DS5000	RAID6	8+P+Q

11.7 DS5000 configuration with SVC and IBM Storwize V7000

In this section, we describe the configuration tasks required on the DS5000 Storage Server, including rules for the configuration.

11.7.1 Setting DS5000 so both controllers have the same WWNN

The virtualization system recognizes that the DS5000 controllers belong to the same storage subsystem unit if they both have the same World Wide Node Name (WWNN). This WWNN setting can be changed through the script editor on the DS5000.

There are a number of ways to determine whether it is set correctly for virtualization system. The World Wide Port Name (WWPN) and World Wide Node Name of all devices logged in to the fabric can be checked from the SAN switch management interface. Confirm that the WWPN of all DS5000 host ports are unique but the WWNN are identical for all ports belonging to a single storage unit.

The same information can be obtained from the Controller section when viewing the Storage Subsystem Profile from the Storage Manager GUI, which will list the WWPN and WWNN information for each host port:

```
World-wide port identifier:  20:27:00:80:e5:17:b5:bc
World-wide node identifier:  20:06:00:80:e5:17:b5:bc
```

Alternatively, to confirm the status of the WWNN setting from, type in the statements in Figure 11-23 on the upper pane of the script editor. Press ENTER after each statement.

To run a script, select **Tools** → **Execute Script** from the Storage Manager Enterprise Management Window.

```
show "Showing the current state of WWNN NVSRAM setting";
show controller[a] nvsrambyte[0x34];
show controller[b] nvsrambyte[0x34];
```

Figure 11-23 Script to check state of WWNN setting

Then select **Tools** → **Verify and Execute** from the menu options in the script editor. If there are no typing errors, then the output of the commands must be displayed in the bottom pane as shown in Figure 11-24.

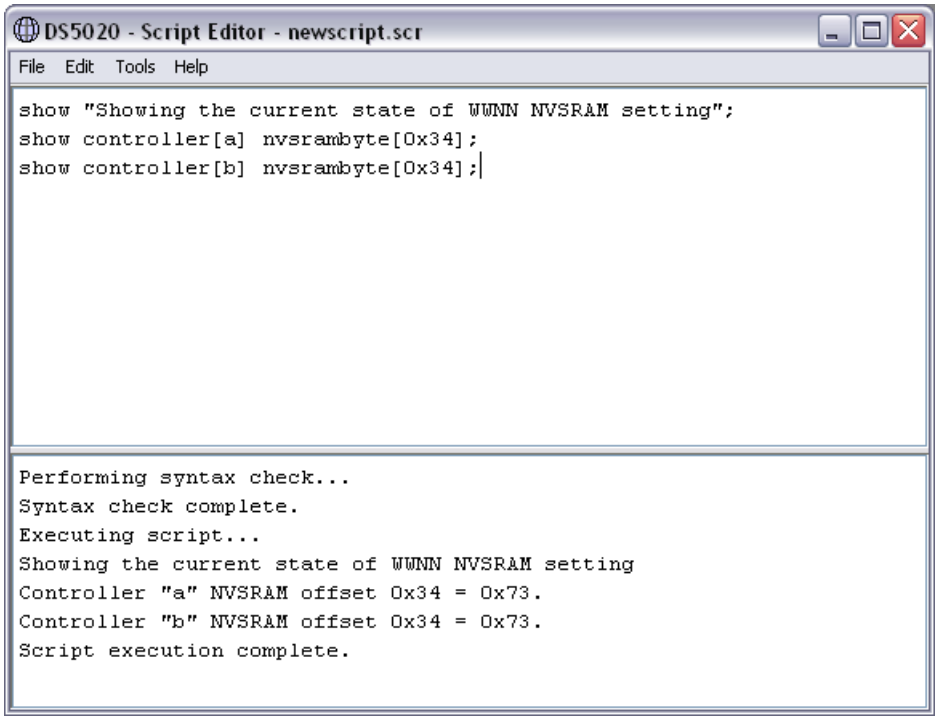


Figure 11-24 Storage Manager script editor

Unfortunately, this does not provide a simple YES/NO answer.

The returned output is the hexadecimal value of the NVSRAM byte at offset 0x34 for each controller. The binary value of bit #1 (where bit #0 is the Least Significant Bit) determines whether the WWNNs are identical or not. See Table 11-4.

Table 11-4 Node names on each controller

Bit #1	Description
0	Various node names (WWNN) on each controller
1	Identical node names (WWNN) on each controller

In our example, the script returned a value of 0x73 for each controller. The second digit value of 3 confirms that bit #1 is set and hence the World Wide Node Names are identical on both controllers, which is the correct setting for the storage virtualization system, therefore no further action is required.

If the controllers are configured with various WWNNs, then the script shown in Figure 11-24 must be executed. It will finish by resetting both controllers for the changes to take effect.

Caution: This procedure is intended for initial configuration of the DS5000. The script must *not* be run in a live production environment because all hosts accessing the storage subsystem will be affected by the changes.

```
show "Setting to enable same WWN for each controller ...";
set controller[a] nvsrambyte[0x34] = 0x06,0x02;
set controller[b] nvsrambyte[0x34] = 0x06,0x02;
show controller[a] nvsrambyte[0x34];
show controller[b] nvsrambyte[0x34];

show "Resetting controller A";
reset Controller [a];
show "Resetting controller B";
reset Controller [b];
```

Figure 11-25 Script to set controllers to have same WWNN

Similar scripts are also bundled with the Storage Manager client download file:

```
- SameWWN.script: Setup RAID controllers to have the same World Wide
     Names. The World Wide Names (node) will be the same for each
     controller pair. The NVSRAM default sets the RAID controllers
     to have the same World Wide Names.

   - DifferentWWN.script:  Setup RAID controllers to have different World
     Wide Names. The World Wide Names (node) will be different for each
     controller pair. The NVSRAM default sets the RAID controllers
     to have the same World Wide Names.
```

11.7.2 Host definition in Storage Manager

The WWPNs of all of the Fibre Channel ports in a virtualization system must be defined within a single unique Host Group. Therefore, a single Storage Partition license is required for each virtualization system. The only exception to this rule is when all logical drives on the DS5000 unit are mapped to a single virtualization system. In that case, all LUNs can be mapped within the Default Group.

In Storage Manager mappings, the host type must be set to:

```
IBM TS SAN VCE SAN Volume Contr
```

This host type has Automatic Drive Transfer (ADT) set by default.

To set the Default Host Type, click **Storage Subsystem** → **Change** → **Default Host Type**.

To set the Host Type for each of the clustered system host port, you can specify the host type of that port or modify existing ports in the Mappings View.

It will also be useful for easy reference in the future to assign meaningful alias names in Storage Manager for all the SVC host groups, hosts, and host ports:

```
HOST GROUP:    LAB_A_SVC

Host:          LAB_A_SVC_Node1
Host Port:     SVC_LABA1_P1   50:05:07:68:01:40:63:11
Host Port:     SVC_LABA1_P2   50:05:07:68:01:30:63:11
Host Port:     SVC_LABA1_P3   50:05:07:68:01:10:63:11
Host Port:     SVC_LABA1_P4   50:05:07:68:01:20:63:11

Host:          LAB_A_SVC_Node2
Host Port:     SVC_LABA2_P1   50:05:07:68:01:40:41:28
Host Port:     SVC_LABA2_P2   50:05:07:68:01:30:41:28
Host Port:     SVC_LABA2_P3   50:05:07:68:01:10:41:28
Host Port:     SVC_LABA2_P4   50:05:07:68:01:20:41:28
```

Both the virtualization system node number, and physical port number can be extracted from the systems port WWPN.

There are a number of ways to access the Host Definition panel in the Storage Manager GUI, such as **Mappings** → **Define** → **Host**.

This presents an option to select the host ports from a pull-down list of known unassociated host port identifiers. The following two subsections describe how to identify the ports that are presented in the list.

Determining the WWNNs of the SVC and Storwize V7000 nodes

Issue the following SAN Volume Controller CLI command to determine the WWNN of each node:

```
svcinfo lsnode -delim :
```

Here is an example of the output:

```
id:name:UPS_serial_number:WWNN:status:IO_group_id:
IO_group_name:config_node:UPS_unique_id:hardware
1:group1node1:1OL3ASH:500507680100[6349]:online:0:io_grp0:yes:202378101C0D18D8:CG8
2:group1node2:1OL3ANF:500507680100[2809]:online:0:io_grp0:no:202378101C0D1796:CG8
3:group2node1:1OL3ASH:500507680100[1426]:online:1:io_grp1:no:202378101C0D18D8:CG8
4:group2node2:1OL3ANF:500507680100[82F4]:online:1:io_grp1:no:202378101C0D1796:CG8
```

The fourth field displays the WWNN for the node, which can be used to identify the node ID and name. The last two bytes (bytes #7 & 8) will match the same two bytes of the node port WWPN.

Determining the node physical port number from the WWPN

Figure 11-26 shows the physical position of the four Fibre Channel ports on the rear view of a SVC node. In this case we show a 2145-CG8 unit, although similar left-to-right numbering applies to all previous SVC models. See Figure 11-9 on page 484 for port layout on IBM Storwize V7000.

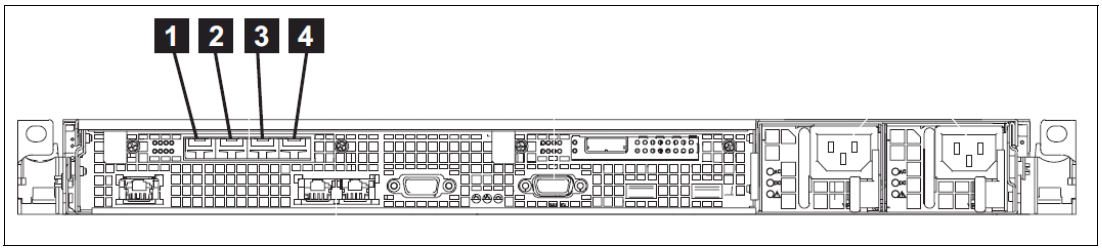


Figure 11-26 Physical Fibre Channel port numbers on the 2145-CG8 SVC node

The first digit of byte #6 of the SVC port WWPN is used to identify the physical port:

50:05:07:68:01:40:41:28

Table 11-5 can then be used to identify the physical port numbers from the list of unassociated port identifiers in Storage Manager.

Table 11-5 Identifying physical port numbers

Physical port #	WWPN byte #6 SVC	WWPN byte #6 Storwize V7000
1	4	1
2	3	2
3	1	3
4	2	4

11.7.3 Arrays and logical drives

You must follow normal logical drive creation guidelines when defining storage for use with virtualization system on the DS5000 Storage Server. Distribute the logical drives evenly between the two controllers.

According to SVC and Storwize V7000 best practices, wait for the logical drive initialization to complete before mapping it to the virtualization system because it can result in excessive discovery times. Waiting might not be necessary on the DS5000 Storage Server because the Immediate Availability Format (IAF) feature makes the LUNs accessible while initialization is in progress.

11.7.4 Logical drive mapping

All logical drives must be mapped to the single host group representing the entire virtualization clustered system. It is not permitted to map LUNs to certain nodes or ports in the clustered system while excluding others.

The Access LUN allows in-band management of an IBM System Storage DS5000. It must only be mapped to hosts capable of running the Storage Manager Client and Agent. The storage virtualization system will ignore the Access LUN if it is mapped to it. Nonetheless, it is good practice to remove the Access LUN from the SVC/Storwize V7000 host group mappings.

Important: The Access LUN must never be mapped as LUN #0.

The IBM SAN Volume Controller code does not support any iSCSI storage devices. All DS5000 logical drives must be mapped to the SVC or Storwize V7000 by Fibre Channel SAN ports.

11.8 Managing SVC and IBM Storwize V7000 objects

In this section, we describe various procedures relevant for monitoring and managing DS5000 Storage Server devices on the SVC and Storwize V7000.

11.8.1 Adding a new DS5000 to a virtualization system configuration

DS5000 Storage Servers can be added to an existing SVC or Storwize V7000 configuration at any time and without disruption. This procedure can be performed using either the SVC Console graphical user interface or CLI commands.

First, see “Setting DS5000 so both controllers have the same WWNN” on page 511 to ensure that the controllers are set up correctly for use with the SVC or Storwize V7000.

Then, if required, create the arrays, logical drives, host group definition, and mappings following the basic guidelines in 11.7, “DS5000 configuration with SVC and IBM Storwize V7000” on page 511.

When the new DS5000 Storage Server is added to the Fibre Channel SAN and included in the same switch Storage Zone as a SVC or Storwize V7000 clustered system, the clustered system automatically discovers it. It then integrates the unit to determine the storage which is presented to the SVC or Storwize V7000 nodes. The logical drives that are mapped to them are displayed as unmanaged MDisks.

Tip: The automatic discovery that is performed by SVC and Storwize V7000 does not write anything to an unmanaged MDisk. You must instruct the virtualization clustered system to add an MDisk to a storage pool or use an MDisk to create an image mode volume

Procedure using SVC Console graphical user interface

Follow these steps to add a new DS5000 Storage Server to the virtualization clustered system:

1. To discover new storage subsystem, select **Physical Storage** in the Welcome menu and then select **External**. The **External** panel shown in Figure 11-27 opens.

   For more detailed information about a specific controller, click one Storage System in the left column (highlighted in the figure)

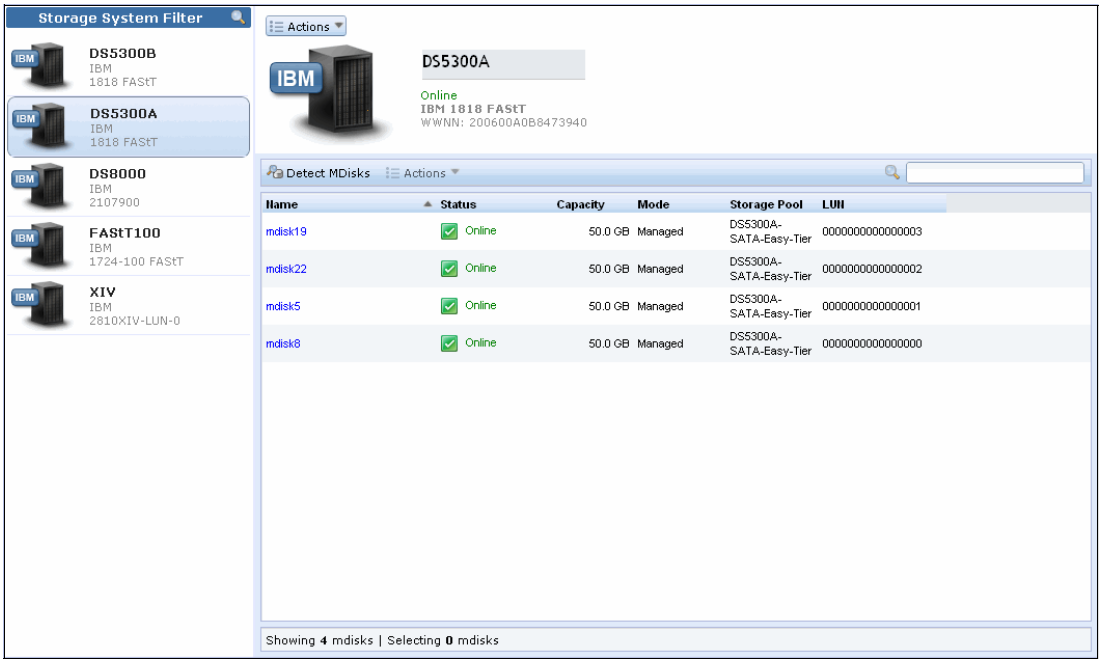


Figure 11-27 Disk controller systems

2. To rename the newly discovered storage subsystem, in the left panel, select the controller that you want to rename as shown in Figure 11-27. Simply click it and you will be able to edit the name, as shown in Figure 11-28. Type the new name that you want to assign to the controller, and press Enter as shown in Figure 11-28.

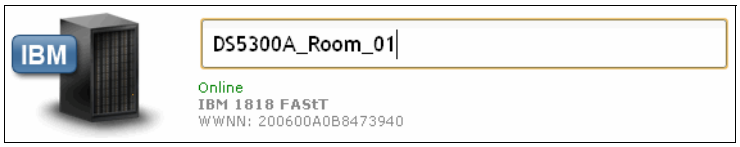


Figure 11-28 Changing the name for Storage System

A task is launched to change the name of this Storage System. When it is completed, you can close this window. The new name of your controller is displayed on the Disk Controller Systems panel.

3. You can discover managed disks (MDisk) from the External panel. Select a controller in the left panel and click **Detect MDisks** button to discover MDisks from this controller, as shown in Figure 11-29.

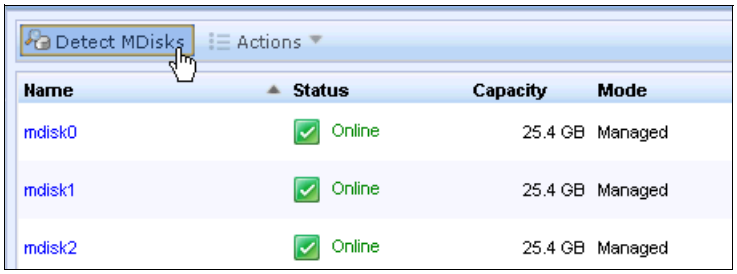


Figure 11-29 Detect MDisks action

The Discover devices task runs and when the task is completed, click **Close** and see the new MDisks available.

4. To create storage pools and volumes using new storage subsystem back-end storage, see the steps in 11.10.4, “Using the LUN in SVC” on page 532.

Procedure using the CLI

Follow these steps to add a new DS5000 Storage Server to the virtualization clustered system:

1. Issue the following CLI command to ensure that the cluster has detected the new storage (MDisks):

   ```
   svctask detectmdisk
   ```

2. Determine the storage controller name to validate that it is the correct controller. The new DS5000 Storage Server will have automatically been assigned a default name.
3. If you are unsure which DS5000 Storage Server is presenting the MDisks, then:
   a. issue the following command to list the controllers:

      ```
      svcinfo lscontroller
      ```

   b. Find the new controller in the list. The new controller has the highest numbered default name.
4. Record the name of the new controller.
5. Issue the following command to change the controller name to something that you can easily use to identify it:

   ```
   svctask chcontroller -name newname oldname
   ```

Where *newname* is the name that you want to change the controller to and *oldname* is the name that you are changing.

6. Issue the following command to list the unmanaged MDisks:

   ```
   svcinfo lsmdisk -filtervalue mode=unmanaged:controller_name=new_name
   ```

These MDisks must correspond with the logical drives mapped from the new DS5000 Storage Server.

7. Record the field controller LUN number. The controller LUN number corresponds with the LUN number that we assigned in Storage Manager when mapping the logical drives to the clustered system.
8. Create a new storage pool and add only logical drives belonging to the new DS5000 subsystem to this storage pool. A separate storage pool must be created for each group of new logical drives with separate performance or resilience characteristics. Give each storage pool a descriptive name, such as SATA_8+P, FC15K_4+P, FC10K_RAID10.

   ```
   svctask mkmdiskgrp -ext 16 -name mdisk_grp_name -mdisk {colon separated list of
   new mdisks returned in step 6}
   ```

This creates a new storage pool with an extent size of 16 MB.

Tip: If you use command-line interface with IBM SAN Volume Controller code version 6.2, the command prefixes (svcinfo, svctask) are optional.

11.8.2 Removing a storage subsystem

Normally, you can use one of the procedures described in 11.9, “Migration” on page 522 to migrate data before the storage subsystem is removed from the SVC or Storwize V7000 system. This procedure assumes that the DS5000 Storage Server has been removed or the logical drives are no longer accessible. The managed disks (MDisks) that represent those logical drives might still exist in the virtualization system. However, SVC or Storwize V7000 cannot access these MDisks because the logical drives that these MDisks represent have been unconfigured or removed from the storage subsystem. You must remove these MDisks.

Perform the following steps to remove MDisks:

1. Run the **`svctask includemdisk`** CLI command on all the affected MDisks.
2. Run the **`svctask rmmdisk`** CLI command on all affected MDisks. Doing it puts the MDisks into the unmanaged mode.
3. Run the **`svctask detectmdisk`** CLI command. The clustered system detects that the MDisks no longer exist in the storage subsystem.

All of the MDisks that represent unconfigured logical drives are removed from the clustered system.

11.8.3 Monitoring the MDisk Status

Use the SVC Console or CLI commands to check the status of the MDisks. On the SVC console, from the SVC Welcome panel, click **Physical Storage** → **MDisks**. The MDisks panel opens as shown in Figure 11-30.

To retrieve more detailed information about a specific MDisk, perform the following steps:

- In the MDisks panel right-click an MDisk.
- Click **Properties**.
- For the selected MDisk, an overview is displayed showing its various parameters and dependent volumes.

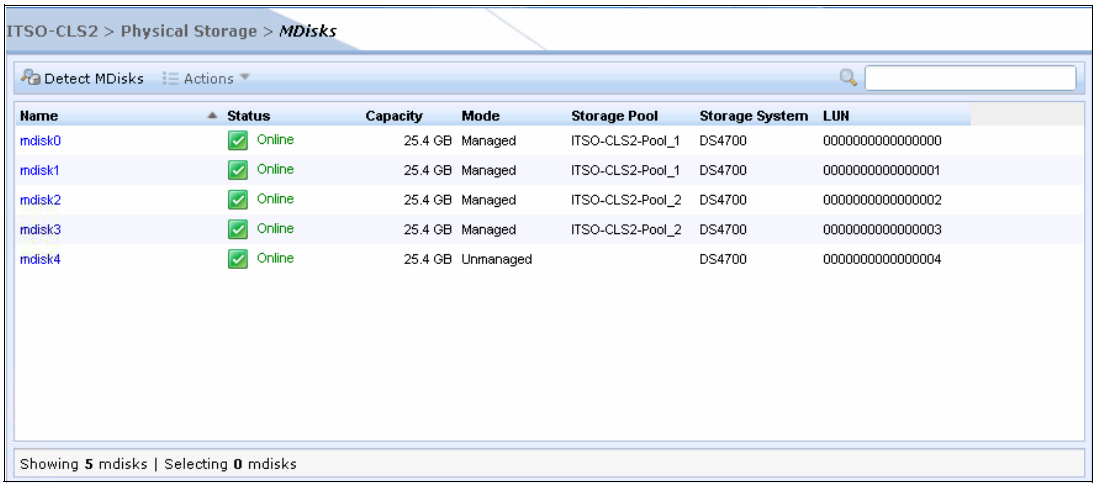


Figure 11-30 Viewing Managed Disks panel

Alternatively, you can use the **`svctask lsmdisk`** CLI command to generate a similar list.

The MDisk can be in one of the following states:

Online The MDisk can be accessed by all online nodes. That is, all the nodes that are currently working members of the clustered system can access this MDisk. The MDisk is online when the following conditions are met:
All timeout error recovery procedures complete and report the disk as online.
Logical unit number (LUN) inventory of the target ports correctly reported the MDisk.
Discovery of this LUN completed successfully.
All of the MDisk target ports report this LUN as available with no fault conditions.

Degraded Paths The MDisk is not accessible to one or more nodes in the cluster. Degraded path status is most likely the result of incorrect configuration of either the disk controller or the Fibre Channel fabric. However, hardware failures in the disk controller, Fibre Channel fabric, or node can also be a contributing factor to this state.

Complete the following actions to recover from this state:
1. Verify that fabric configuration rules for storage systems are correct.
2. Ensure that you have configured the storage system properly.
3. Correct any alerts in the Event Log.

Degraded Ports The MDisk has one or more 1220 events in the Event Log. The 1220 code indicates that the remote Fibre Channel port has been excluded from the MDisk. This event might cause reduced performance on the storage controller and usually indicates a hardware problem with the storage controller. To fix this problem you must resolve any hardware problems on the storage controller and fix the 1220 events in the Event Log. To resolve these events in the log, select an alert with four-digit error code and click **Run Fix Procedure** in the **Actions** menu**.** The “Directed Maintenance Procedure.” windows opens. You must follow the wizard and its steps to fix the event.

Excluded The MDisk has been excluded from use by the cluster after repeated access errors. Run the Directed Maintenance Procedures to determine the problem.

Offline The MDisk cannot be accessed by any of the online nodes. That is, all of the nodes that are currently working members of the clustered system cannot access this MDisk. This state can be caused by a failure in the SAN, storage subsystem, or one or more physical disks connected to the storage subsystem. The MDisk is reported as offline if all paths to the disk fail.

11.8.4 Event reporting and notification

Events detected by the virtualization system are saved in an Event Log. As soon as an entry is made in this Event Log, the condition is analyzed and classified to help you diagnose problems. If any service activity is required, the user is notified of the event.

Viewing the Event Log

You can view the Event Log by using the SVC command line interface (CLI) or the SVC Console.

Viewing the full contents of the Event Log using the CLI

To view the contents of the log, follow these steps:

1. Dump the contents of the Event Log to a file.
   a. Issue the **`svctask dumperrlog`** command to create a dump file that contains the current Event Log data.
   b. Issue the **`svcinfo lserrlogdumps`** command to determine the name of the dump file that you have just created.
   c. Issue the **`secure copy command (scp)`** to copy the dump file to the local workstation.
2. View the file with a text viewer after the dump file is extracted from the clustered system.

Viewing the Event Log using the SVC console

The Event Log panel displays two types of events, namely messages and alerts, and it indicates the cause of any log entry.

To access this panel, from the SVC Welcome panel, select **Troubleshooting** → **Event Log**. Figure 11-31 shows the Event Log from SVC Console.

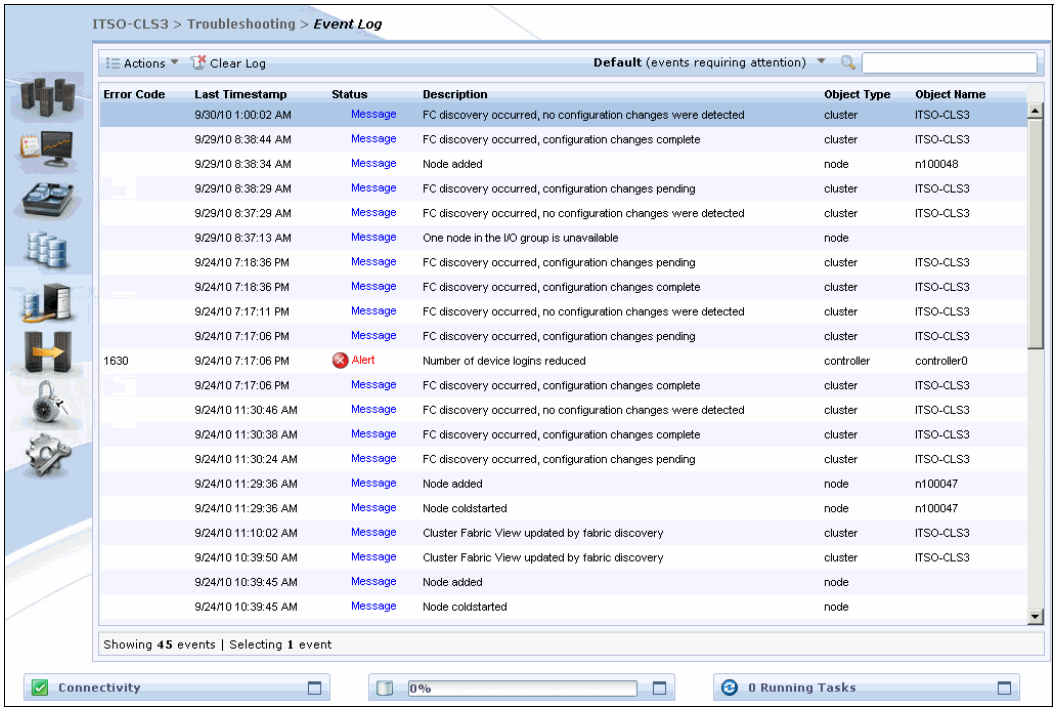


Figure 11-31 Event Log panel

Certain alerts have a four-digit error code and a fix procedure that helps you fix the problem. Other alerts also require action, but do not have a fix procedure. Messages are fixed when you acknowledge reading them. You can use different types of filtering to ease up the search inside this view.

11.9 Migration

In this section, we describe a migration scenario where we are replacing a pre-virtualized storage subsystem (a DS4000 in this case) with a DS5000 Storage Server using SVC clustered system.

11.9.1 Migration overview and concepts

In this section, we describe various migration concepts.

MDisk modes: It is important to understand these three basic MDisk modes:

Unmanaged MDisk An MDisk is reported as unmanaged when it is not a member of any storage pool. An unmanaged MDisk is not associated with any volumes and has no metadata stored on it. The virtualization system will not write to an MDisk that is in unmanaged mode except when it attempts to change the mode of the MDisk to one of the other modes.

Image Mode MDisk Image Mode provides a direct block-for-block translation from the MDisk to the volume with no virtualization. Image Mode volume have a minimum size of one block (512 bytes) and always occupy at least one extent. An Image Mode MDisk is associated with exactly one volume.

Managed Mode MDisk Managed Mode Mdisks contribute extents to the pool of extents available in the storage pool.

The virtualization clustered systems cannot differentiate unmanaged-mode MDisks that contain existing data from unmanaged-mode MDisks that are blank. Therefore, it is vital that you control the introduction of these MDisks to the system by adding them one at a time. For example, map a single LUN from your RAID controller to the clustered system and refresh the view of MDisks. The newly detected MDisk is displayed.

11.9.2 Migration procedure

Figure 11-32 shows a DS4000 unit with three attached hosts. Each host has a number of logical drives mapped to it through a separate partition. All data on the DS4000 must be preserved.

In addition, a new DS5000 Storage Server has been introduced to the SAN. A storage zone has been created on the SAN fabrics to allow the SVC to access the DS5000 logical drives. Logical drives have been created on the DS5000 unit and mapped to the SVC.

Note that the DS5000 only requires a single partition for the SVC clustered system.

In this example, we assume that all the logical drives mapped to these 3 hosts have similar performance and resilience requirements, such as all 4+P RAID 5 FC 10K. Therefore, we can migrate them all into the same storage pool. However, sometimes it is preferred to create separate storage pools for each host operating system.

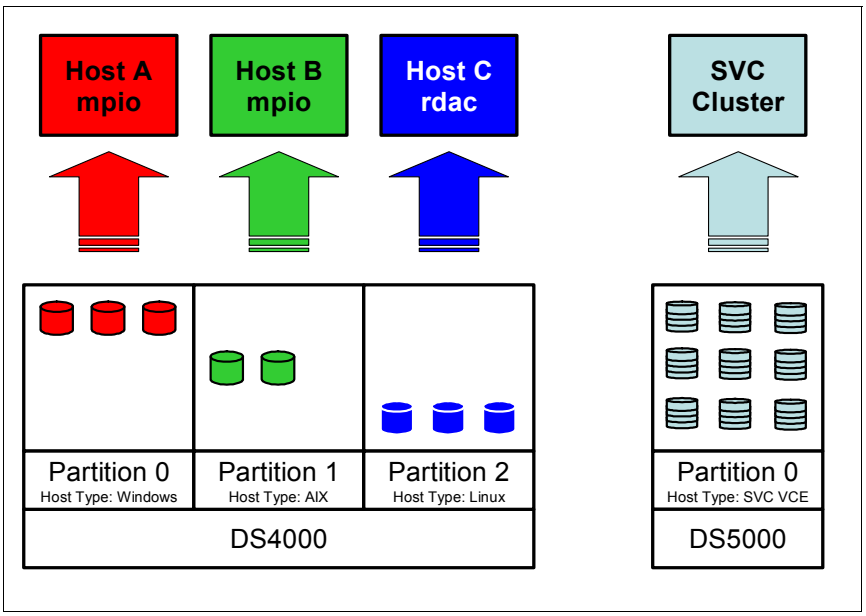


Figure 11-32 Migration - Initial Configuration

The host C LUNs will be the first to be migrated to the DS5000 by SVC. Here, we summarize the basic steps involved in the migration. For further details, see the SVC reference documentation.

1. Create an MDisk Group for the new DS5000 LUNs:
   a. All new volumes must be automatically discovered.
   b. Use the CLI command **`svctask detectmdisk`** to manually scan for new devices.
   c. Use the CLI command **`svcinfo lsdiscoverystatus`** to check status of the discovery.
   d. Use the CLI command **`svcinfo lsmdiskcandidate`** to show unmanaged MDisks.
   e. Use the CLI command **`svctask mkmdiskgrp`** to add the new MDisks to a storage pool:

      `svctask mkmdiskgrp -name` *`ds5kgrp`* `-ext 512 -mdisk` *`mdisk1:mdisk2:mdisk3:mdisk4`*
2. Stop all applications on host C.
3. Remove the RDAC multipath driver on host C.
4. Install the SDD multipath driver on host C.

5. Change SAN zoning so that host C is only presented the SVC system ports.
6. Define Host C to the SVC:
   a. Use the CLI command **svcinfo lshbaportcandidate** to list unassigned host WWPNs.
   b. Use the CLI command **svctask mkhost** to define the host.
   c. Use the CLI command **svctask lshost** to verify that host was defined successfully.
7. Shut down Host C.
8. Change Host Type definitions for partition 2 on the DS4000 to IBM TS SVC VCE.
9. Detect new MDisks:
   a. All new volumes must be automatically discovered.
   b. Use the CLI command **svctask detectmdisk** to manually scan for new devices.
   c. Use the CLI command **svcinfo lsdiscoverystatus** to check status of the discovery.
   d. Use the CLI command **svcinfo lscontroller** to list all storage systems.
   e. Use the CLI command **svcinfo lsmdiskcandidate** to show unmanaged MDisks.
10. Create an empty storage pool for the imported image volumes from the DS4000:
    a. Use the CLI command **svctask mkmdiskgrp** to create the empty storage pool.
    b. Use the CLI command **svcinfo lsmdiskgrp** to list storage pools.
11. Create image mode volume from data volumes on the DS4000:
    a. Use the CLI command **svcinfo lsmdiskcandidate** to show unmanaged MDisks.
    b. Use the CLI command **svctask mkvdisk** to create the image mode volume:

       svctask mkvdisk -mdiskgrp *imggrp* -iogrp 0 -vtype image -name *DS401* -mdisk *mdisk4*

    c. Use the CLI command **svcinfo lsvdisk *vdiskname*** to check new volume properties.
12. Map the Image volume to the Host:
    a. Use the CLI command **svctask mkvdiskhostmap** to map the new volume to the host.
    b. Virtualized images of the LUNs in partition 2 on the DS4000 Storage System are now accessible on host C.
13. Boot Host C.
14. You can start all applications on Host C to minimize downtime.

15. Start migrating data from DS4000 to DS5000 (see Figure 11-33):

    a. Use the CLI command **svctask migratevdisk** to migrate from the image mode source MDisks on the DS4000 onto the managed mode target on the DS5000:

       `svctask migratevdisk -vdisk DS401 -mdiskgrp ds5kgrp`

    b. Use the CLI command **svcinfo lsmigrate** to monitor the progress of the migration.

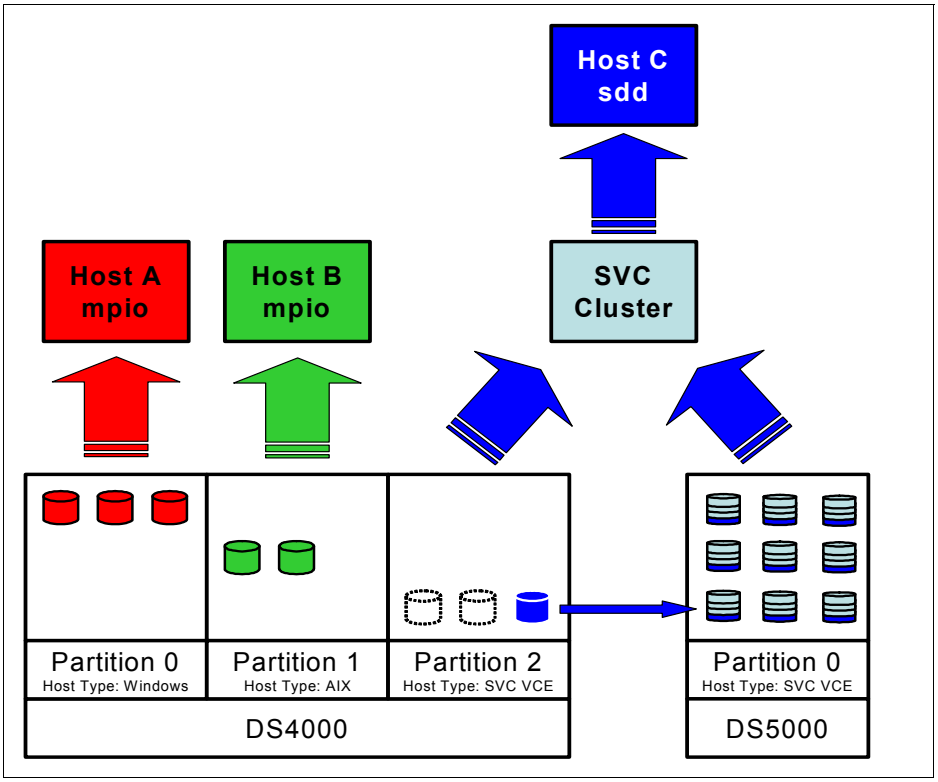


Figure 11-33 Migration - migration in progress

The host outage required for implementing these changes is minimal. When the host is rebooted in step 13 on page 524, it will detect all the original volumes mapped from partition 2 on the DS4000 as Image mode volumes through the SVC clustered system. The migration procedure in step 15 on page 525 is completely transparent to the host.

Do not manually add an unmanaged-mode MDisk that contains existing data to a storage pool. If you do, the data is lost. When you use the command to convert an image mode volume from an unmanaged-mode MDisk, you will select the storage pool where it must be added.

Likewise, in step 11 on page 524, it is essential to specify the volume type as “image.” If you add the existing volumes to a storage pool as a normal striped MDisks, then all data on those volumes is lost. Because we intend to virtualize previously existing disks with data intact, we must create image mode volumes instead.

Repeat the same procedure for all remaining hosts with storage on the DS4000 (Figure 11-34).

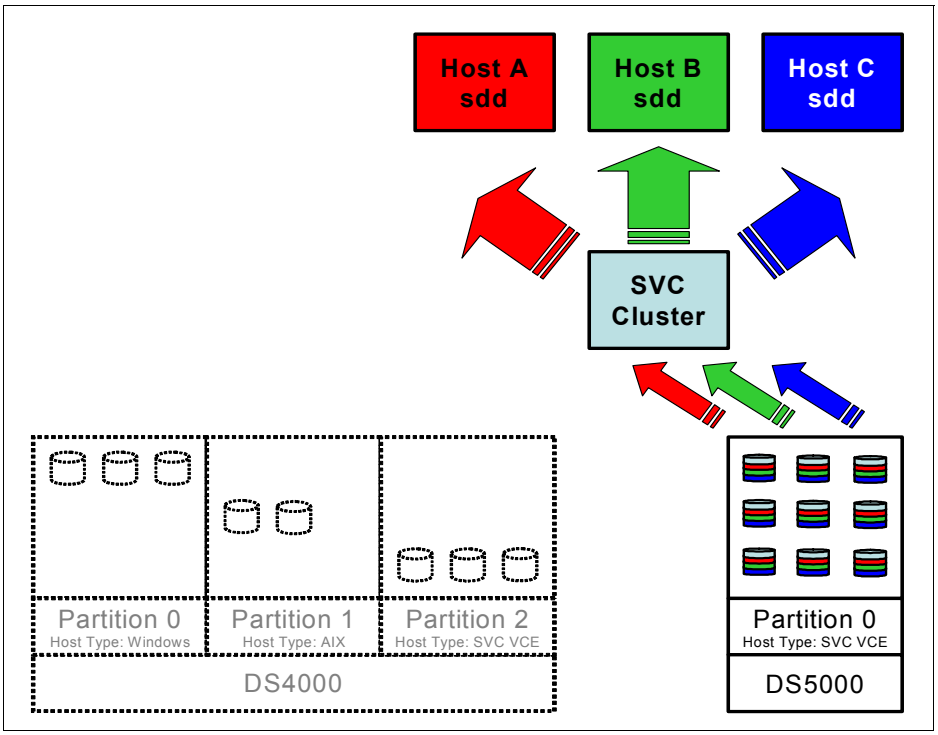


Figure 11-34 Migration - All LUNs moved to DS5000 by SVC

You can remove the original partitions and create a new partition on the DS4000 in order to present the unused capacity as managed disks on the SVC. Then you can migrate certain members of the original data volumes back onto it if required.

Although we used a migration from a DS4000 Storage Server to a DS5000 as an example, the same procedure applies if we were migrating from other storage platforms.

Tip: The steps in the SVC migration example described above are exactly the same for IBM Storwize V7000.

11.10 SVC with DS5000 configuration example

In our example, we have a DS5000 Storage Server for which certain LUNs are used by a host connected to SVC and certain LUNs are directly mapped to other hosts. Figure 11-35 shows the architecture and the aliases used following the naming convention mentioned in 11.5.4, “SAN aliases for SVC and IBM Storwize V7000: Guidelines” on page 499.

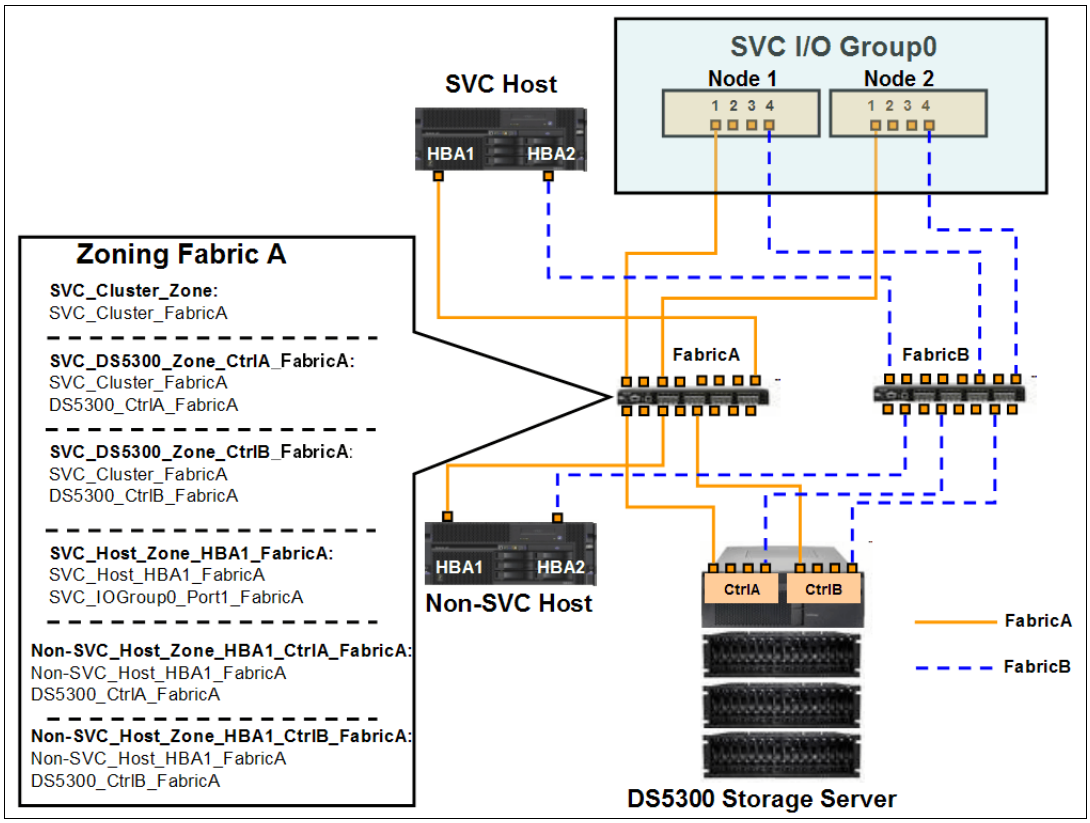


Figure 11-35 DS5000 Storage Server with SVC and non -SVC host

In this example the SVC is a two-node clustered system. Each node has one HBAs with four ports, but only ports 1 and 4 of each node are currently being used.

A host group is created in Storage Manager. The name of the host group must reflect something meaningful to your environment to signify that the group is an SVC group. In the example shown in Figure 11-39 on page 532, we named the host group *SVCGroup*.

Next, a host is created. The host name must conform to your naming convention. In the example shown in Figure 11-39 on page 532, we called the host *SVCHost*. This host definition is the only one required in Storage Manager for all the hosts or servers that will access the storage subsystem using SVC. The ports assigned are the ports of the SVC nodes.

The LUNs created with the Storage Manager client are then assigned to the host *SVCHost*. They can now be discovered using the SVC Console or CLI, and assigned as SVC managed disks.

11.10.1 Zoning for a non-SVC host

Zoning rules for hosts that will not use the SVC remain unchanged. They must still follow the best practice where, for each host, multiple zones are created such that each host HBA port is paired with a specific controller on the DS5000 Storage Server. For example, for a host with two HBA ports, the zoning is as follows:

- Host zone 1: HBA_1 is in a zone with DS5000 controller A.
- Host zone 2: HBA_2 is in a zone with DS5000 controller A.
- Host zone 3: HBA_1 is in a zone with DS5000 controller B.
- Host zone 4: HBA_2 is in a zone with DS5000 controller B.

In Figure 11-18 on page 501 we show only the zoning for the Fabric A, therefore two zones are created:

- **Non-SVC_Host_Zone_HBA1_CtrlA_FabricA** → contains two aliases
  - Non-SVC_Host_HBA1_FabricA
  - DS5300_CtrlA_FabricA
- **Non-SVC_Host_Zone_HBA1_CtrlB_FabricA** → contains two aliases
  - Non-SVC_Host_HBA1_FabricA
  - DS5300_CtrlB_FabricA

Obviously the counterpart FabricB is going to get the other two zones:

- **Non-SVC_Host_Zone_HBA2_CtrlA_FabricB** → contains two aliases
  - Non-SVC_Host_HBA2_FabricB
  - DS5300_CtrlA_FabricB
- **Non-SVC_Host_Zone_HBA2_CtrlB_FabricB** → contains two aliases
  - Non-SVC_Host_HBA2_FabricB
  - DS5300_CtrlB_FabricB

11.10.2 Zoning for SVC and hosts that will use the SVC

All SVC nodes in the SVC clustered system are connected to the same SAN, and present volumes to the hosts. These volumes are created from managed disks presented by the storage subsystem.

In our example, the zoning for SVC must be such that each node (node 1 and node 2) can address the DS5000 Storage Server.

Each SVC node has one HBA with four Fiber Channel ports. We have only used two ports from the HBA present in each SVC node (ports 1 and 4).

Storage zoning needs to be done for the SVC to utilize the DS5000 Storage Server (see Figure 11-18 on page 501):

For the Fabric A, we have the following zones (one per controller):

- **SVC_DS5300_Zone_CtrlB _FabricA** → contains two aliases:
  - SVC_Cluster_FabricA
  - DS5300_CtrlB_FabricA

- **SVC_DS5300_Zone_CtrlA_FabricA** → contains two aliases:
  - SVC_Cluster_FabricA
  - DS5300_CtrlA_FabricA

For the Fabric B, we create the following zones:

- **SVC_DS5300_Zone_CtrlB_FabricB** → contains two aliases:
  - SVC_Cluster_FabricB
  - DS5300_CtrlB_FabricB
- **SVC_DS5300_Zone_CtrlA_FabricB** → contains two aliases:
  - SVC_Cluster_FabricB
  - DS5300_CtrlA_FabricB

Note that the two aliases **SVC_Cluster_FabricA** and **SVC_Cluster_FabricB** contains respectively the ports 1 of each node and the ports 4 of each node.

With this zoning, each port of the SVC nodes connected to the fabric can address each DS5000 controller. It also offers many paths from the nodes to the storage for redundancy.

Host zoning needs to be done for every host that will use SVC to manage the disks.

- Host zone 1: HBA_1 is in a zone with SVC node 1 port 1 and SVC node 2 port 1.
- Host zone 2: HBA_2 is in a zone with SVC node 1 port 4 and SVC node 2 port 4.

With this zoning, each host HBA can see a port on each node that is connected to the fabric. It also offers many paths to the storage that the SVC will manage.

In Figure 11-18 on page 501 we show only the zoning for the Fabric A, therefore we have:

- **SVC_Host_Zone_HBA1** → contains two aliases
  - SVC_Host_HBA1_FabricA
  - SVC_IOGroup0_Port1_FabricA

On the other side, the counterpart FabricB is going to get the following zone:

- **SVC_Host_Zone_HBA2** → contains two aliases
  - SVC_Host_HBA2_FabricA
  - SVC_IOGroup0_Port4_FabricA

11.10.3 Configuring the DS5000 Storage Server

The storage arrays must be created with only one LUN per array. The LUN must be sized to utilize all of the available storage space on the array. These single LUNs will then be used to create volumes, which will be made available to the hosts. Choose the RAID level required for the LUN. This RAID level is dependent upon the nature of the application that will use this LUN. See section 11.6, “Storage virtualization systems with DS5000 best practices” on page 505 for suggestions.

When creating the host type, ensure that you select the particular host type for SVC of IBM TotalStorage SAN Volume Controller Engine (TS SAN VCE). Remember that the LUNs can be used by all types of hosts and applications attached to SVC.

In our example, we created three SVC LUNs that will be managed by the SVC. These LUNs start with the letters SVC for clarity.

- The database LUN is for SQL database files.
  - It has been created with RAID 5 and is 73 GB in size (Figure 11-36).
  - It has been created using 750 GB 7.2K drives.
  - It was created with a larger segment size of 128 K.
  - The read ahead multiplier is set to enabled (1).

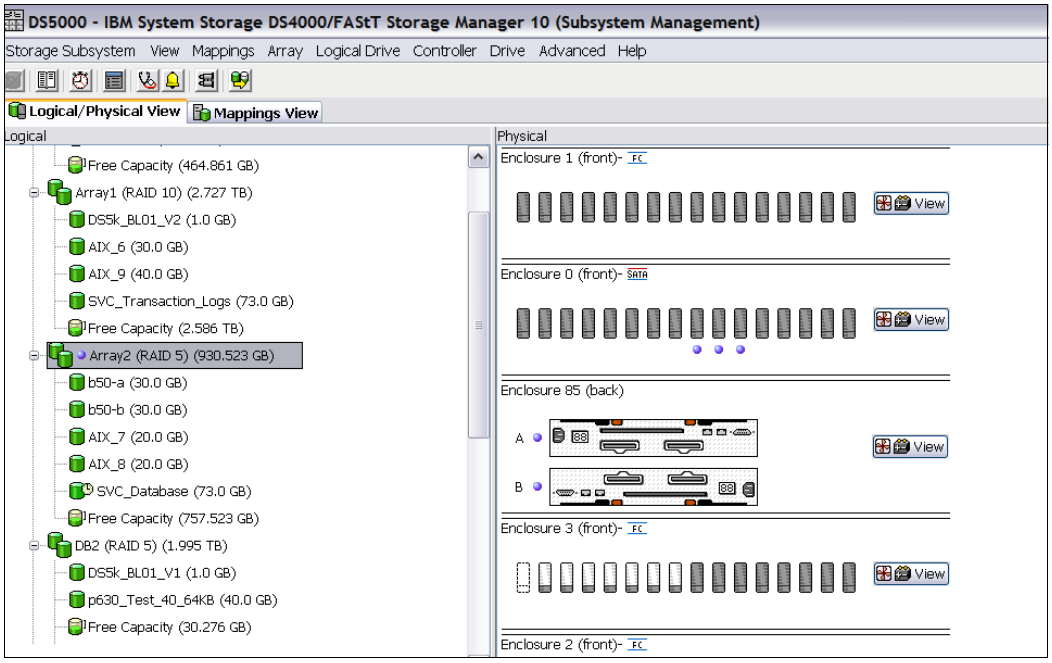


Figure 11-36 Storage Manager with LUNs created

- The second LUN is the transaction logs for that SQL database (Figure 11-37).
  - It was created with RAID 10.
  - It has been created with a larger segment size of 128 KB.
  - It has been created using 750 GB 7.2 K drives.
  - The read ahead multiplier is set to disabled (0).

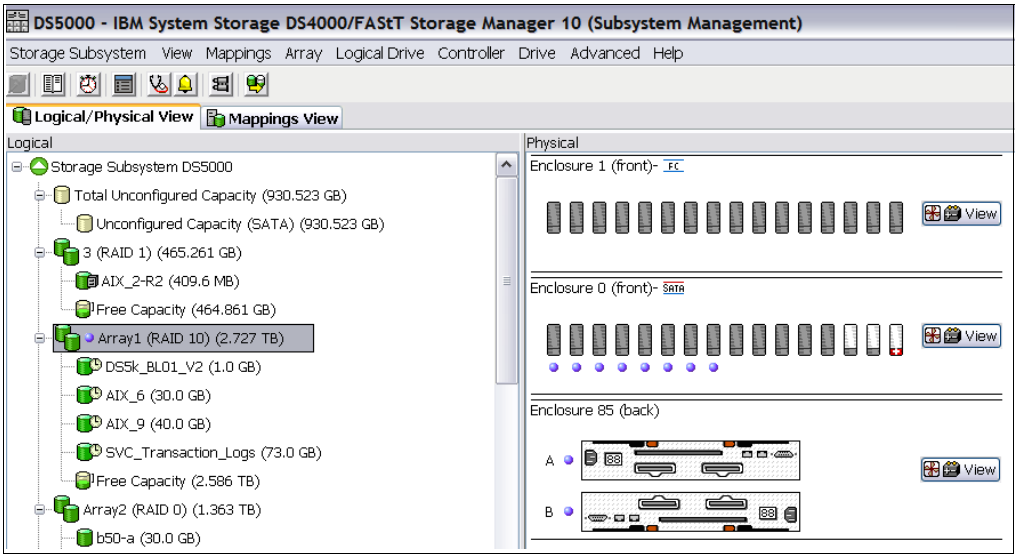


Figure 11-37 Transaction Log LUN

- The third LUN is the file and print LUN (Figure 11-38).

  As the file and print LUN is not as read or write intensive as the database or transaction log LUNs, it can be shared between multiple hosts. Most of the files are very small, and the environment is more of a read than write.

  - It has been created with a segment size of 64 KB.
  - It has been created using 7.2 K drives.
  - The read ahead multiplier is set to enabled (1).

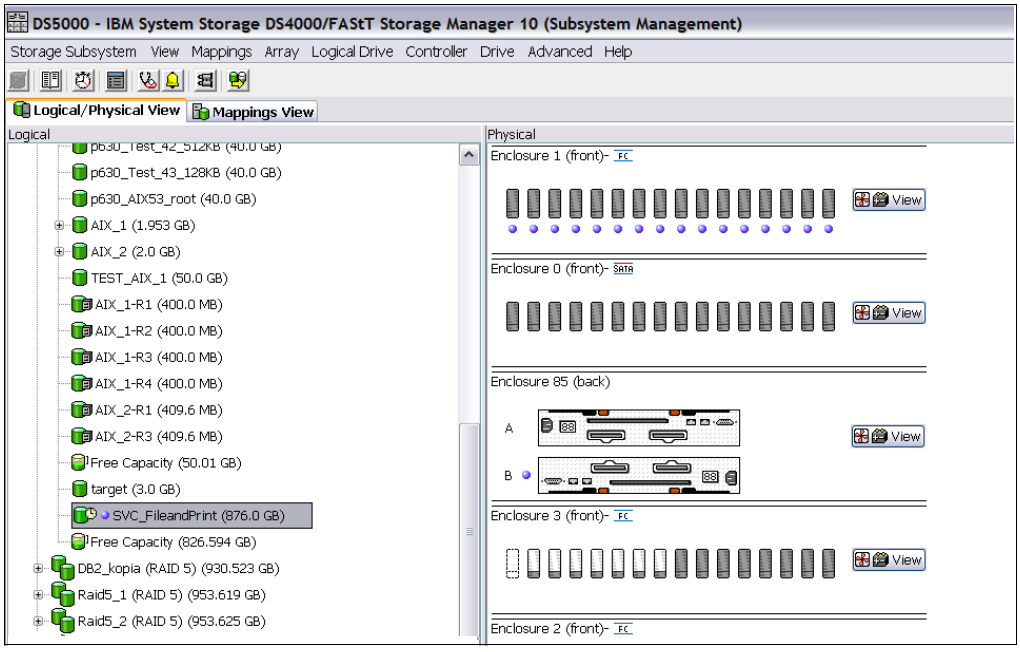


Figure 11-38 File and Print LUN

The SVC nodes must be able to access all of the LUNs that SVC will manage. Therefore, the LUNs need to be mapped to all of the available SVC Fibre Channel host ports. The storage is mapped at the host SVCHost. The host type must be set to IBM TS SAN VCE (Figure 11-39).

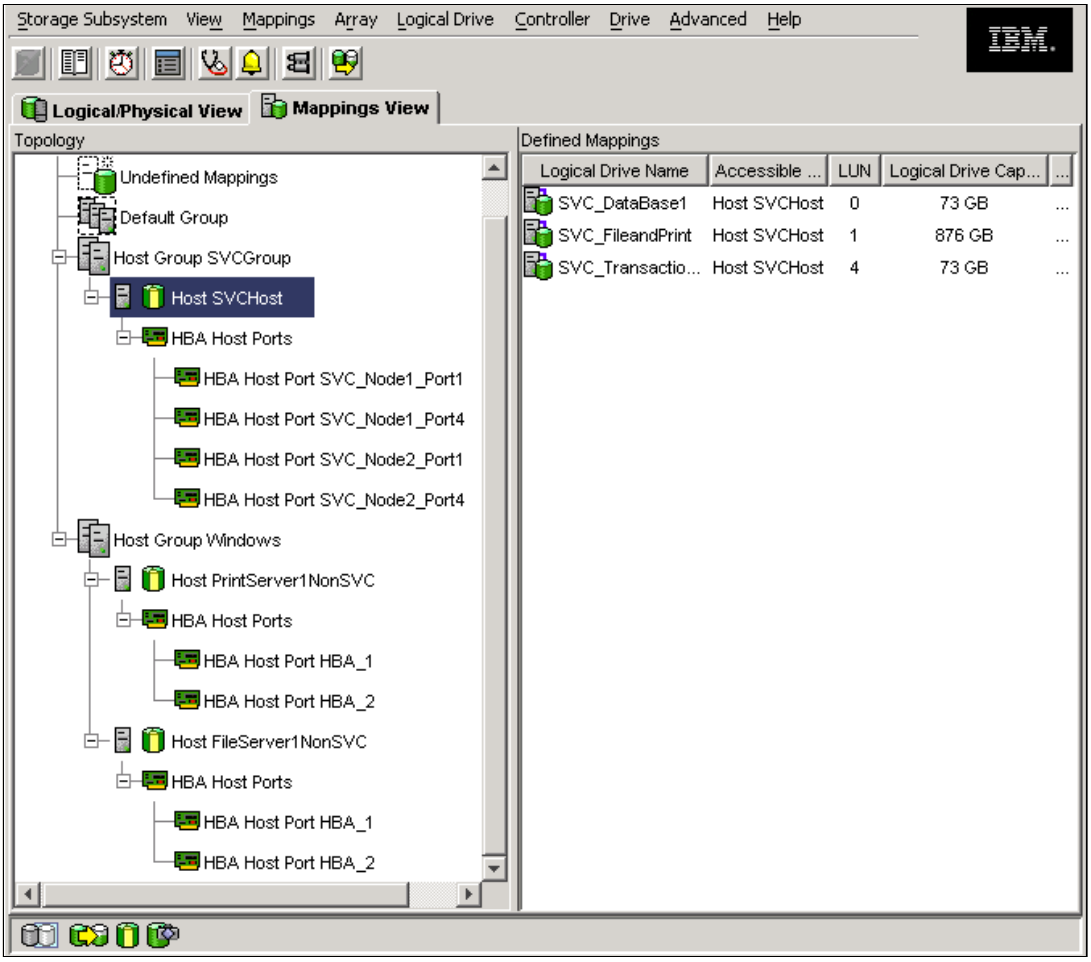


Figure 11-39 LUNs managed by the SVC are mapped to the host SVCHost

11.10.4 Using the LUN in SVC

The LUN can now be discovered by the SVC and defined as an MDisk. The MDisk is then assigned to a storage pool. Extents from the storage pool are used to define a volume that can in turn be mapped to an SVC host.

Follow these steps:

1. Log in at the SVC console with appropriate privileges.

   To create new storage pool, from the SVC Welcome panel, click **Physical Storage**, and then click **Pools**.

   The Pools panel opens. On this page, click **New Pool**, as shown in Figure 11-40.

   The wizard **Create Storage Pools** opens.

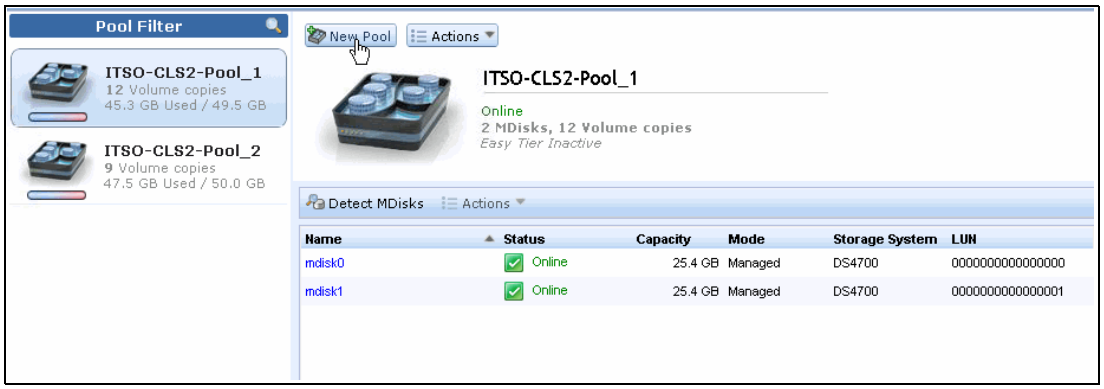


Figure 11-40 Selecting the option to create the storage pool

   On this first page, complete the wizard by providing the pool name and choosing the extent size (you need to expand Advanced settings box)

   Click **Next**.

   On the next page you are able to detect the new MDisks, this will be described next as a separate step.

   Click Finish to create the new *empty* storage pool.

2. If you created an empty storage pool or you simply assign additional MDisks to your SVC environment later, you can add MDisks to existing storage pools by performing the following steps:

   Select the unmanaged MDisk that you want to add to a storage pool

   Click **Add to Pool** in the **Actions** menu (Figure)

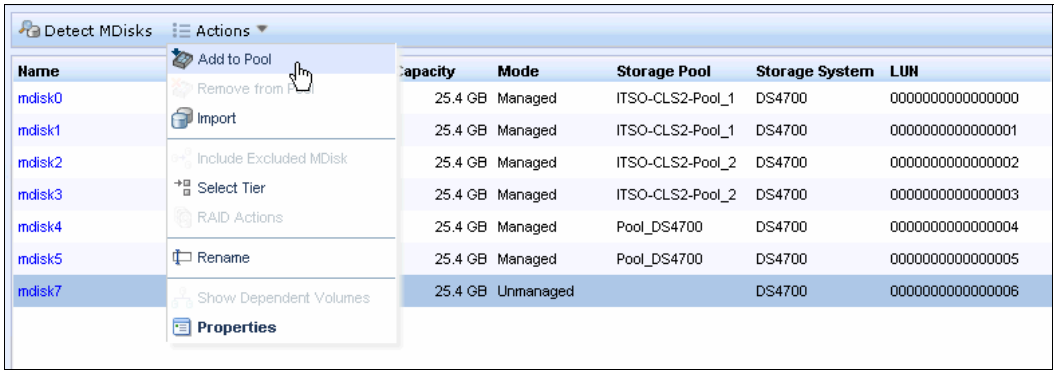


Figure 11-41 Actions: Add to Pool

Tip: You can also access the Add to Pool action by right-clicking an unmanaged MDisk.

From the Add MDisk to Pool window, select in which pool you want to integrate this MDisk and then click **Add to Pool**, as shown in Figure 11-42.

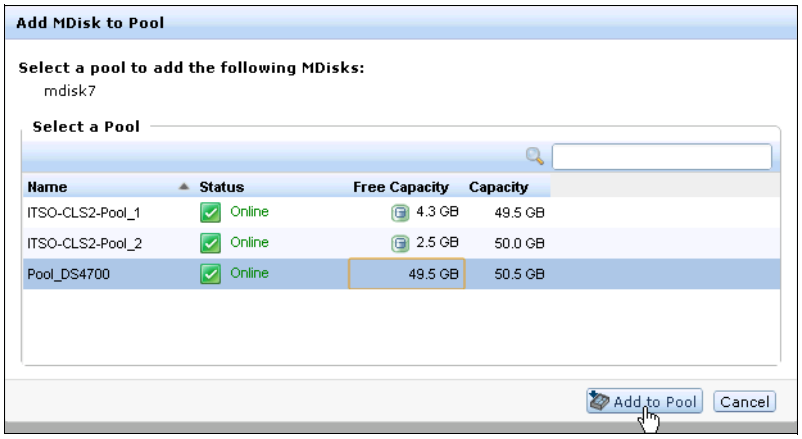


Figure 11-42 Adding an MDisk to existing storage pool

In the storage pools panel, the new storage pool free capacity is displayed

3. To create a new volume, perform the following steps (Figure 11-43):

   Go to the All Volumes panel from the SVC Welcome panel, and click **Volumes** → **All Volumes**.

   Click **New Volume**

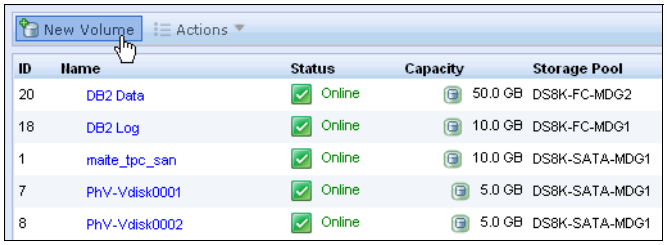


Figure 11-43 New volume action

Select one of the following presets, as shown in Figure 11-44.

- **Generic**: Create volumes that use a set amount of capacity from the selected storage pool.
- **Thin Provision**: Create volumes whose capacity is large, but which only use the capacity that is written by the host application from the pool.
- **Mirror**: Create volumes with two physical copies that provide data protect. Each copy can belong to a different storage pool to protect data from storage failures.
- **Thin Mirror**: Create volumes with two physical copies to protect data from failures while using only the capacity that is written by the host application

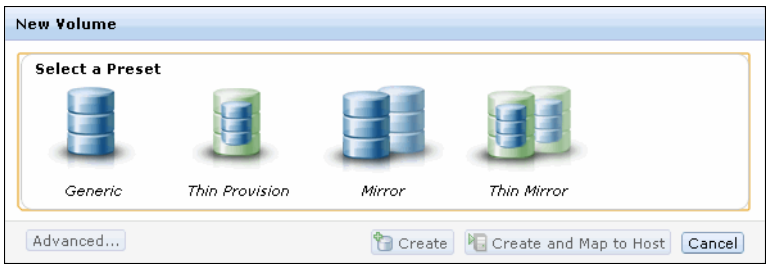


Figure 11-44 New Volume: Select a Preset

After selecting a preset, in our example **Generic**, you must select the storage pool on which the data will be striped (Figure 11-45).

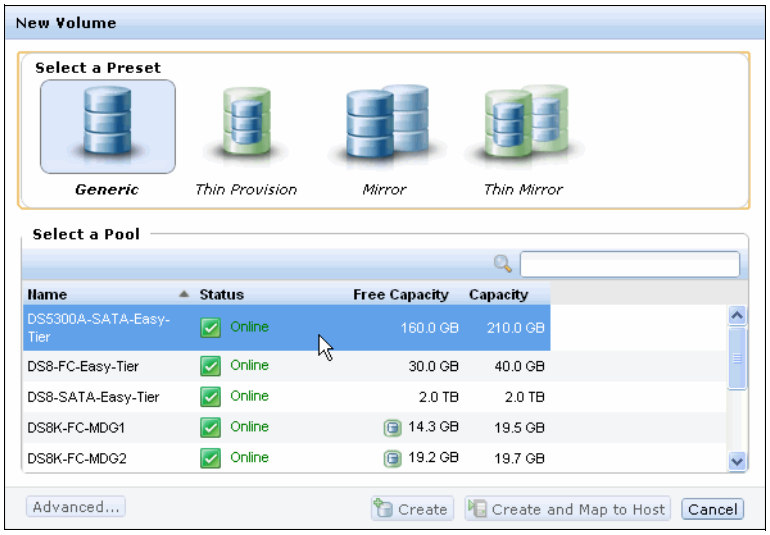


Figure 11-45 Select the storage pool

After the storage pool has been selected, the window will be updated automatically and you will need to select a volume name and size as shown in Figure 11-46.

- Enter a name if you want to create a single volume, or a naming prefix if you want to create multiple volumes
- Enter the size of the volume that you want to create and select the capacity measurement (bytes, KB, MB, GB or TB) from the list.

An updated summary automatically appears in the bottom of the window to give you an idea of the space that will be used and that is remaining in the pool.

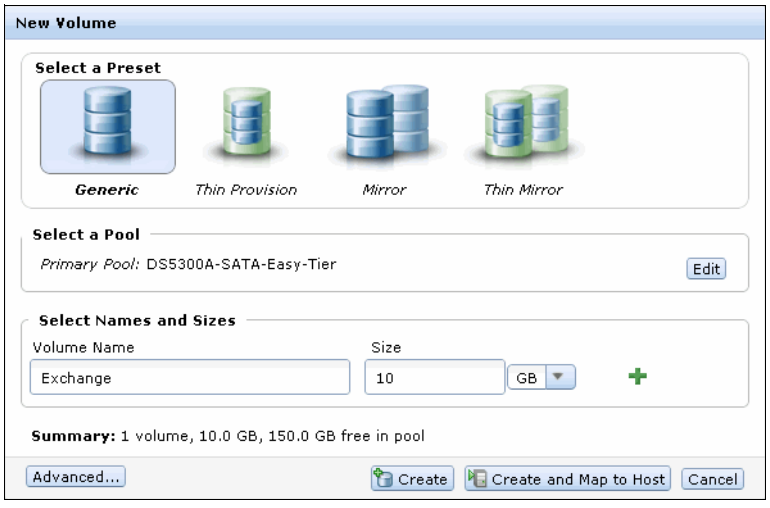


Figure 11-46 New volume: Select Name and Size

Various optional actions are available from this window:

- You can modify the storage pool by clicking **Edit**. In this case, you can select another storage pool.
- You can create additional volumes by clicking the + button. This action can be repeated as many times as necessary. You can remove them by clicking the button.

You can activate and customize advanced features such as thin-provisioning or mirroring, depending on the preset you selected. To access these settings, click **Advanced.**

After all the advanced settings have been set, click **OK** to return to the main menu

Then, you have the choice to only create the volume using the **Create** button, or to create and map it using the **Create and Map to Host** button**.**

Before mapping the volume to a host, you need to create the host definition first.

4. To create a new host on SVC, perform following steps:

   There are two types of connections to hosts, Fibre Channel (FC) and iSCSI. In this section, we detail Fibre Channel (FC) method.

   To create a new host that uses the FC connection type, perform the following steps:

   Go to the All Hosts panel from the SVC Welcome panel and then click **Hosts** → **All Hosts**

   Click **New Host** as shown in Figure 11-47.

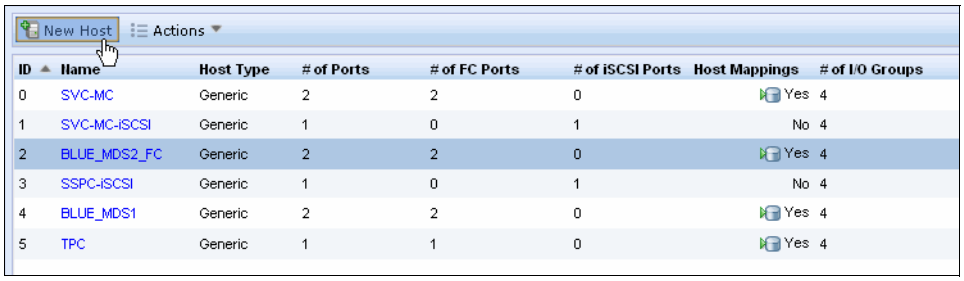


Figure 11-47 New Host action

Select **Fibre-Channel Host** from the two types of connection available (Figure 11-48).

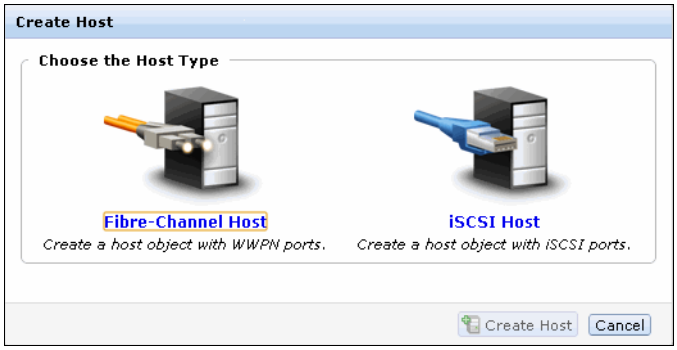


Figure 11-48 Create Host window

In the Creating Hosts window (Figure 11-49), type a name for your host (Host Name).

Fibre-Channel Ports Section: Use the drop-down list to select the WWPNs that correspond to your HBA or HBAs and click **Add Port to List** in the Fibre-Channel Ports window. To add additional ports, repeat this action.

If your WWPNs are not being displayed, click **Rescan** to rediscover new WWPNs available since the last scan.

Tip: In certain cases your WWPNs still might not be displayed, even though you are sure that your adapter is functioning (for example, you see the WWPN in the switch name server) and your zones are correctly set up. To rectify this, type the WWPN of your HBA or HBAs into the drop-down list and click **Add Port to List**. It will be displayed as `unverified`.

Advanced Settings Section: If you need to modify the I/O Group, the Port Mask or the Host Type, you must select **Advanced** to access these Advanced Settings:

- Select one or more I/O groups from which the host can access volumes. By default, all I/O Groups are selected.
- You can use a port mask to control the node target ports that a host can access. The port mask applies to the logins from the host initiator port that is associated with the host object.
- Select the Host Type. The default type is Generic. Use generic for all hosts, unless you use Hewlett-Packard UNIX (HP-UX) or Sun. For these, select HP_UX (to have more than eight LUNs supported for HP_UX machines) or TPGS for Sun hosts using MPxIO.

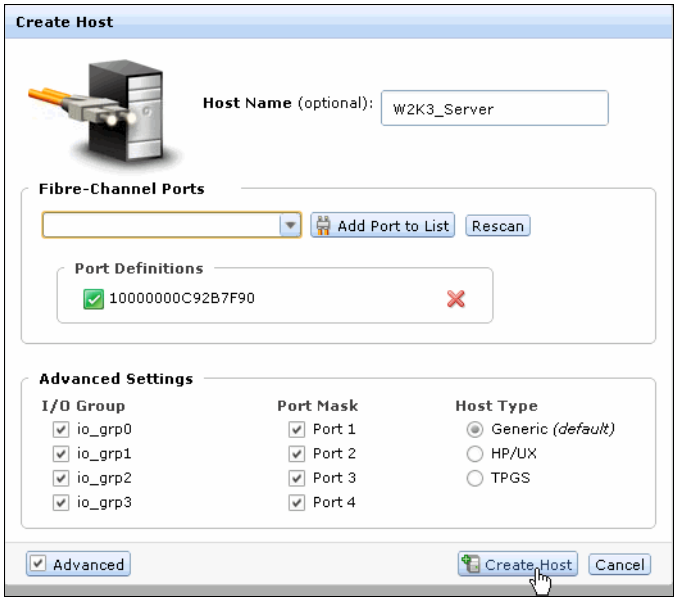


Figure 11-49 Creating a new Fibre Channel connected host

Click the **Create Host button** as shown in Figure 11-49. This action brings you back to the **All Hosts** panel on Figure 11-50 where you can see the newly added FC host.

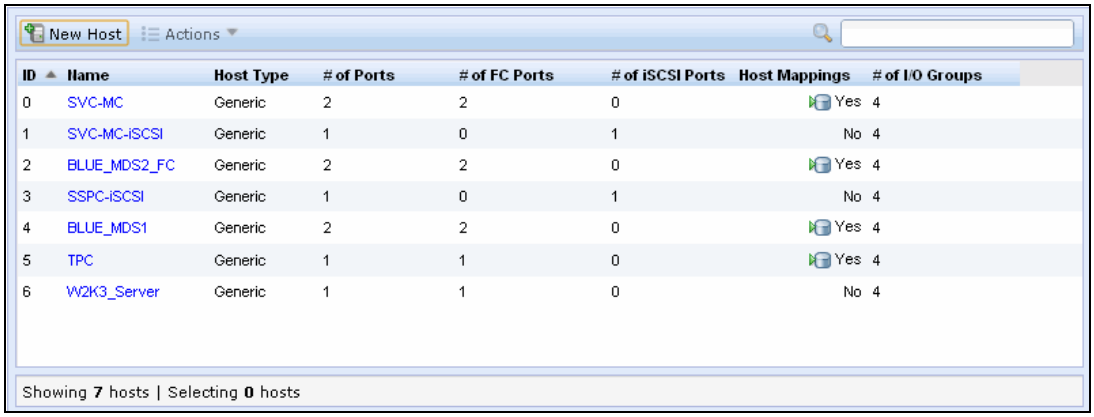


Figure 11-50 Create host results

5. To map a new volume to the SVC host follow this procedure:

   Go to the All Volumes panel from the SVC Welcome panel and click **Volumes** → **All Volumes**.

   Select the volume in the table.

   Click **Map to Host** in the **Actions** menu

   Tip: You can also right-click a volume and select **Map to Host** from the list.

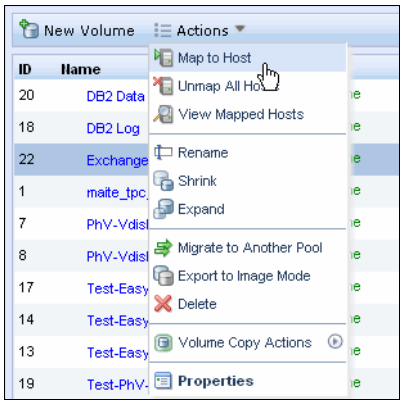


Figure 11-51 Map to Host action

On the Modify Mappings window, select the host on which you want to map this volume using the drop-down button and then click **Next** (Figure 11-52).

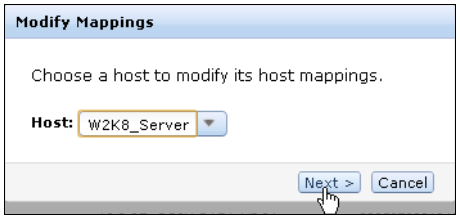


Figure 11-52 Select the host to which you want to map your volume

On the Modify Mappings window, verify the mapping. If you want to modify it, select the volume or volumes that you want to map to a host and move each of them to the right table using the right arrow as shown in Figure 11-53 on page 539. If you need to remove them, use the left arrow.

In the right table, you can edit the SCSI ID. Select a mapping that is highlighted in yellow, which indicates that the mapping is new, and click **Edit SCSI ID.**

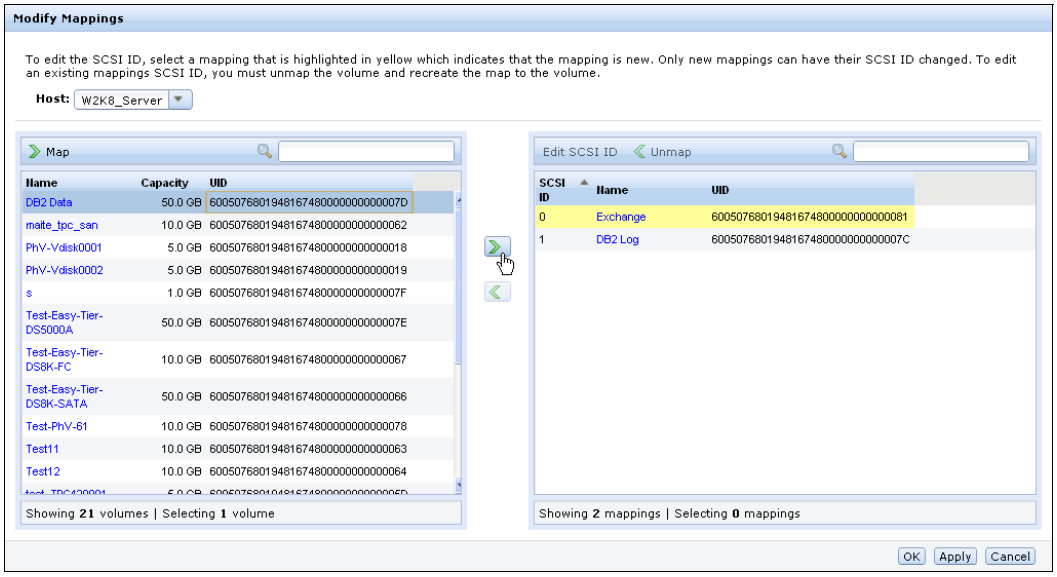


Figure 11-53 Modify Mappings window: Adding volumes to a host

After all the volumes you want to map to this host have been added, click **OK**. You will return to the main All Volumes panel.

The volumes are mapped to the host. Perform host-specific actions to discover and add the new volume to your host.

Tip: The steps in the SVC configuration example described above are exactly the same for IBM Storwize V7000.

12

DS5000 with AIX, PowerVM, and PowerHA

In this chapter, we present configuration information relevant to the DS5000 Storage Server attached to IBM POWER servers. We also review special considerations for PowerHA formerly known as High Availability Cluster Multiprocessing (HACMP) configurations in AIX.

AIX 5L and AIX6 are award winning operating systems, delivering superior scalability, reliability, and manageability. AIX runs across the entire range of IBM POWER from entry-level servers and workstations to powerful supercomputers able to handle the most complex commercial and technical workloads in the world. With the continuos development of POWER systems and their hardware capabilities, the latest version of AIX 7.1 further exploits features and benefits of intensively computing, highly available, or fault tolerant systems.

In addition, AIX has an excellent history of binary compatibility, which provides assurance that your critical applications will continue to run as you upgrade to this newer versions of AIX.

IBM PowerVM® provides industrial-strength virtualization for AIX, IBM i, and Linux environments on IBM POWER processor-based systems. IBM Power Systems servers integrated with PowerVM technology are designed to allow clients to build a dynamic infrastructure that will help them to reduce costs, manage risk and improve service levels. PowerVM also offers a secure and resilient virtualization environment, built on the advanced RAS (reliability, availability and serviceability) features, extreme scalability and leadership performance[1] of the IBM Power Systems platform, based on the outstanding IBM POWER7® processors.

PowerHA is IBM software to build highly available or fault tolerant clusters on a combination of POWER systems. It is supported by a wide range of POWER servers, with the new storage systems, and network types, and it is one of the best-rated, UNIX-based clustering solutions in the industry.

[1] Power Systems benchmark results: http://www.ibm.com/systems/power/hardware/benchmarks/

12.1 Configuring DS5000 in an AIX environment

In this section, we review the prerequisites and configuration step for successful attachment of the DS5000 Storage Server to the AIX host environment and the available enhanced features such a virtualization (PowerVM) or clustering (PowerHA). With the latest POWER product enhancements, that address many advanced clients' requirements, it is necessary to plan and scope the best option available for these requirements.

12.1.1 Host Bus Adapters in an AIX environment for DS5000 attachment

Multiple combinations of systems, AIX versions, and Host Bus Adapters (HBAs) can be used in order to connect a POWER system to the DS5000 storage. For detailed HBA and systems information regarding combinations, see the following website to access the System Storage Interoperation Center (SSIC):

http://www.ibm.com/systems/support/storage/config/ssic

The SSIC provides the interoperability matrix, based on selected HW and SW options. The SSIC is a selection driven tool that offers various options to satisfy your query (Table 12-1).

Table 12-1 SSIC Options for AIX interoperability matrix required information

Option type	Select options
Product Family	IBM System Storage Midrange Disk
Product Model	DS5020, DS5100, DS5300
Product Version	DS5xxx (select version)
Host Platform	IBM Power Systems
Operation Systems	AIX (select version and TL level)
Connection Protocol	Fibre Channel, FCoE
HBA Model	(Select HBA model)
SAN Vendor	ANY, Cisco, Brocade, McData (Select Vendor)
SAN or Networking Model	Select SAN Switch model
Clustering	IBM PowerHA (select version)
Multipathing	IBM MPIO, IBM SDD PCM (select version)

To be able to complete the SSIC interoperability matrix, you must have all the requested details available. On new POWER6 and POWER7 systems, we strongly advise to use the latest firmware versions. For the existing systems, your planning must include ascertaining the prerequisites and upgrade paths to achieve a fully supported system.

For further interoperabilty information, see the following website:

http://www.ibm.com/systems/storage/product/interop.html

12.1.2 Independent Software Vendors

To achieve full compatibility of IBM POWER systems and various levels of AIX operating system with applications and SW products provided by IBM partner Independent Software Vendors (ISV), the IBM offers the resource library of best practices and typical deployment scenarios of various products. The ISV Resource Library is available at this website:

http://www.ibm.com/systems/storage/solutions/isv

The solution entries provided in the ISV Resource Library are intended to aid in the identification of high quality solutions with our partner Independent software vendors (ISVs) for leading IBM storage platforms. IBM's commitment to interoperability is visible through its ongoing work with such ISVs as Oracle, Microsoft, SAP, or VMware.

IBM and/or the ISV has tested to a high degree of interoperability but does not warrant functionality or problem resolution of any listed product. Clients should contact corresponding vendor for related pre-sales and post-sales support issues, as IBM Product Support can only assist on specific IBM Products.

12.1.3 Verifying AIX and microcode level

Always make sure that the HBA is at a supported microcode level for the model and firmware version installed on your DS5000 and for the version of Technology Level of AIX (use command **oslevel** as in the Example 12-1).

Example 12-1 Determine version of AIX

```
#oslevel -s
7100-00-01-1037
```

Numerous methods are available to check the current microcode level on the HBA adapter. For example, with adapter fcs0, you can use the following methods:

1. The first method uses the **lscfg** command. It returns Vital Product Data (VPD) information of the adapter. The Z9 or ZA field contains the firmware level. This method also displays the FRU number and Assembly part number, and the World Wide Name WWN. The following output in Example 12-2 from the **lscfg** command is typical:

   Example 12-2 Obtaining VPD using lscfg

```
# lscfg -vpl fcs0
fcs0       U5802.001.00B5146-P1-C2-T1  8Gb PCI Express Dual Port FC Adapter

        Part Number.................10N9824
        Serial Number...............1B00504823
        Manufacturer................001B
        EC Level....................D76482B
        Customer Card ID Number.....577D
        FRU Number..................10N9824
        Device Specific.(ZM)........3
        Network Address.............10000000C99A7F9E
        ROS Level and ID............02781174
        .
        .
        .
        Device Specific.(Z9)........US1.11X4
        Device Specific.(ZA)........U2D1.11X4
        Device Specific.(ZB)........U3K1.11X4
        Device Specific.(ZC)........00000000
        Hardware Location Code......U5802.001.00B5146-P1-C2-T1
```

2. The second method uses the **lsmcode** command. It returns the extension of the firmware image file that is installed. Not all adapters support this method, but the majority of 4 Gbps and all 8 Gbps HBAs do support it (see Example 12-3).

Example 12-3 Gathering HBA firmware level

```
# lsmcode -c -d fcs0
The current microcode level for fcs0 is 111304.
```

3. The third method uses the **fcstat** command, which shows all Fibre Channel statistics details of given HBA. At the beginning of the output, as displayed in Example 12-4, we see the firmware level:

Example 12-4 Fibre Channel statistics report of HBA

```
# fcstat fcs0

FIBRE CHANNEL STATISTICS REPORT: fcs0

Device Type: FC Adapter (adapter/pciex/df1000f114108a0)
Serial Number: 1B00504823
Option ROM Version: 02781174
ZA: U2D1.11X4
World Wide Node Name: 0x20000000C99A7F9E
World Wide Port Name: 0x10000000C99A7F9E

FC-4 TYPES:
  Supported: 0x0000012000000000000000000000000000000000000000000000000000000000
  Active:    0x0000010000000000000000000000000000000000000000000000000000000000

FC-4 TYPES (ULP mappings):
  Supported ULPs:
        Internet Protocol (IP) over Fibre Channel (IETF RFC2625)
        Small Computer System Interface (SCSI) Fibre Channel Protocol (FCP)
  Active ULPs:
        Small Computer System Interface (SCSI) Fibre Channel Protocol (FCP)
Port Speed (supported): 8 GBIT
Port Speed (running):   8 GBIT
Port Type: Fabric
Seconds Since Last Reset: 190818
```

Always confirm that you have the latest supported level of the firmware and drivers. If not, download the latest version and upgrade using IBM Fix Central at this website:

http://www.ibm.com/support/fixcentral/

Fix Central requires machine, adapter, or feature code options before providing the relevant latest adapter firmware and links in order to download. When you have entered necessary details, the description link for that adapter will provide you the full instructions on how to download and install the code including the summary of changes in the firmware version of the driver you are about to install.

Hint: When checking the firmware and adapter update instructions or `readme` files from Fix Central, cross-check the feature code (FC) with the FRU number of the adapter listed in the **lscfg -vpl fcs*x*** command to assure the correct firmware.

12.1.4 Upgrading HBA firmware levels

Upgrading the firmware level consists of downloading the firmware (microcode) from your AIX host system to the adapter. After the update has been installed on AIX, you must apply it to each adapter.

Follow these steps to complete this operation:

1. From the AIX command prompt, enter **diag** and press Enter.
2. Select the **Task Selection** → **Microcode Tasks** option.
3. Highlight the **Download Microcode** option.
4. Press Enter to select all the Fibre Channel adapters to which you want to download firmware. Press F7.

   The Download panel is displayed with one of the selected adapters highlighted. Press Enter to continue.
5. Highlight **/etc/microcode** and press Enter.
6. Follow the instructions that are displayed to download the firmware, one adapter at a time.

12.2 AIX device drivers

For proper functionality and maximum performance of all SAN attached disk drives in your AIX system, make sure, that the appropriate disk device driver is present. In this section, we describe the AIX device drivers available for IBM Midrange Storage Systems of DS5000 family.

12.2.1 AIX MPIO

With Multiple Path I/O (MPIO), a device can be separately detected through one or more physical connections, or paths. A path-control module (PCM) provides the path management functions. An MPIO-capable device driver can control more than one type of target device. A PCM can support one or more specific devices. Therefore, one device driver can be interfaced to multiple PCMs that control the I/O across the paths to each of the target devices.

Figure 12-1 shows the interaction between the components that make up the MPIO solution. The MPIO device driver can also control multiple types of target devices, each requiring a particular PCM.

The AIX PCM consists of the following components:

- PCM RTL configuration module, which is a runtime loadable module that enables the device methods to detect additional PCM Kernel Extension (KE) device-specific or path ODM attributes that the PCM KE requires. The PCM Runtime Library (RTL, sometimes expressed as Run-Time Loadable) is loaded by a device method. One or more routines within the PCM RTL are then accessed to perform specific operations that initialize or modify PCM KE variables.
- PCM KE kernel extension, which supplies path-control management capabilities to any device driver that supports the MPIO interface. The PCM KE depends on device configuration to detect paths and communicate that information to the device driver. Each MPIO-capable device driver adds the paths to a device from its immediate parent or parents. The maintenance and scheduling of I/O across various paths is provided by the PCM KE and is not apparent to the MPIO-capable device driver. The PCM KE can provide more than one routing algorithm, which can be selected by the user. The PCM KE also

helps collect information that can be used to determine and select the best path for any I/O request. The PCM KE can select the best path based on a variety of criteria, including load balancing, connection speed and connection failure.

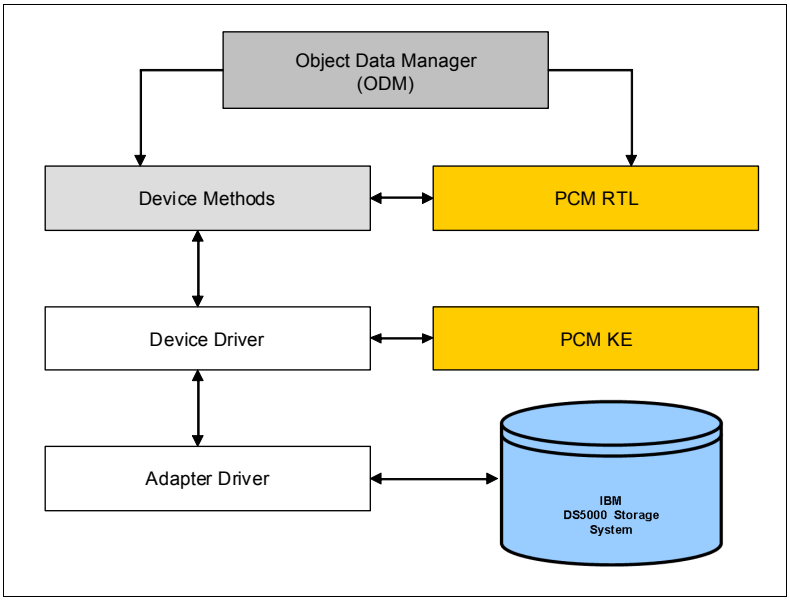


Figure 12-1 The MPIO conceptual diagram

Before a device can take advantage of MPIO, the device's driver, methods, and predefined attributes in the Object Data Manager (ODM) must be modified to support detection, configuration, and management of multiple paths. The parallel SCSI and Fibre Channel device drivers and their device methods have been modified to support MPIO devices. Also, the predefined attributes for certain devices in the ODM have been modified for MPIO.

The AIX PCM supports a set of devices defined in the `devices.common.IBM.mpio.rte` file set, including DS5000 logical drives. For other types of devices, the device vendor must provide attributes predefined in the ODM, a PCM, and any other supporting code necessary to recognize the device as MPIO-capable. The standard AIX MPIO is therefore capable of managing paths using the **`mkpath`**, **`chpath`**, **`rmpath`**, and **`lspath`** commands for DS5000 logical drives.

12.2.2 SDDPCM

Attention: Starting with the AIX version 7.1, SDDPCM no longer supports IBM Midrange disk subsystems; a native MPIO driver needs to be used instead. In addition, support for AIX fcp_array is being phased out. The AIX fcp_array users must migrate to the AIX MPIO multipath driver at the earliest time window.

In a multipath configuration environment SDDPCM provides support for a host system attachment to storage devices. It provides enhanced data availability, dynamic input/output (I/O) load balancing across multiple paths, and automatic path failover protection.

SDDPCM functionality

SDDPCM is a loadable path control module for supported storage devices to supply path management functions and error recovery algorithms. When the supported storage devices are configured as Multipath I/O (MPIO) devices, SDDPCM is loaded as part of the AIX MPIO FCP (Fibre Channel Protocol) device driver during the configuration. The AIX MPIO-capable device driver with the supported storage devices SDDPCM module enhances the data availability and I/O load balancing.

SDDPCM benefits

AIX MPIO-capable device drivers will automatically discover, configure, and make available every storage device path (Figure 12-2). SDDPCM manages the paths to provide the following benefits:

- High availability and load balancing of storage I/O
- Automatic path-failover protection
- Concurrent download of supported storage devices licensed machine code
- Prevention of a single-point-failure

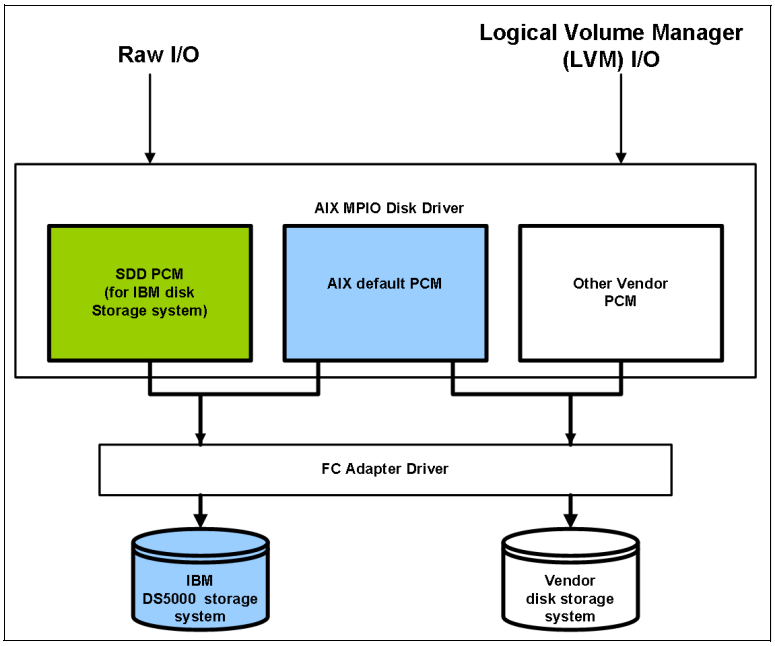


Figure 12-2 SDDPCM position in the AIX protocol stack

As seen in Figure 12-2, where SDDPCM is positioned in the AIX protocol stack, its path selection routine is invoked to select an appropriate path for each I/O operation.

DS5000 is a dual array controller disk subsystem where each logical drive is assigned to one controller that is considered the owner, or the active controller, of a particular logical drive. The other controller is considered as an alternate, or passive, controller. Thus, SDDPCM distinguishes the following paths to the DS5000 logical drive:

- Paths to the ownership (active) controller
- Paths to the alternate (passive) controller

With this type of active/passive dual-controller subsystem device, I/O can be sent only to the ownership controller. When the SDDPCM selects paths for I/O, it selects paths that are connected only to the ownership controller. If there is no path on the ownership controller that can be used, SDDPCM changes the logical drive controller ownership to the alternate controller, switches the paths that were passive to active, and then selects these active paths for I/O.

Attention: SDD and SDDPCM are different, exclusive software packages on a server and must not be mixed. You cannot install both software packages on a server for supported storage devices. Only SDDPCM is supported on DS5000 where the logical volumes presented to AIX are MPIO capable. SDD supports storage devices which are configured as non-MPIO-capable devices (that means multiple logical device instances are created for a physical LUN), such as DS8000 LUNs.

SDDPCM is functional on DS5100, DS5300, and DS5020 as well as certain older models in the DS4000 range. As this support is being phased out, we suggest to migrate your AIX system to utilize MPIO instead as soon as possible.

As mentioned, SDDPCM supports a maximum of 1200 configured devices and a maximum of 16 paths per device, meaning that the maximum number of host adapter ports that are supported is 16. However, with the round robin or load balance path selection algorithms, configuring more than four paths per device might impact the I/O performance.

Number of paths: Use the minimum number of paths necessary to achieve sufficient redundancy in the SAN environment. The preferred number of paths per device is four.

12.2.3 RDAC drivers on AIX

The Redundant Disk Array Controller (RDAC) has limited support on older versions of DS4000 and AIX, it is not supported on DS5000 or AIX version 6.1 (or VIO 1.5) and later with the exception of HBA FC5758. Multipath IO (MPIO) and Subsystem Device Driver Patch Control Module (SDDPCM) has replaced RDAC and will be supported on existing and future versions of AIX and DS5000. Therefore, RDAC should not be used for any new installations.

12.3 Installing the AIX device drivers

This section briefly guide you how to install the AIX MPIO and SDDPCM device drivers required for connectivity to the DS5000 Storage System. Throughout the following text, we expect readers to have the minimal administrative knowledge of UNIX systems and basic understanding of AIX Logical Volume Manager (LVM).

12.3.1 AIX MPIO

This device driver is packaged with the AIX set of installation media and will automatically be installed with AIX if the POWER system has HBAs physically present on the machine or LPAR when it was installed (or assigned as virtual adapter from VIOS). If the HBAs are available in the system afterwards or dynamically associated with the LPAR, then there is a possibility that all License Program Products (LPPs) and prerequisites for the MPIO (and the HBA device drivers) are not installed in full extent.

In this case the AIX installation media might be required when the command **cfgmgr** is run. The command will automatically configure the HBAs to AIX and, with the installation media install all the required HBA and MPIO LPPs. For example, to use the media on device /dev/cd0, insert installation media CD1 in the CDROM and enter the command:

```
cfgmgr -i /dev/cd0
```

You might be prompted to insert another CD media if needed to complete the configuration, after which you can re-apply the Technology Level (TL) package again, containing updates to the base level file sets just installed from the CD media. To do it, follow the standard procedure of applying TL maintenance to AIX.

12.3.2 SDDPCM

You can install SDDPCM on AIX by using the **installp** command or by using the System Management Interface Tool (SMIT) with the shortcut command, **smitty installp**. To obtain the LPPs and instructions on how to download and install, see the following website:

http://www.ibm.com/support/docview.wss?uid=ssg1S4000201

It is always a good practice to use the latest versions available for SDDPCM with DS5000 Storage Manager v10.7x or higher. However, consult the support matrix for the current suggested version of SSDPCM, available at this website:

http://www.ibm.com/systems/support/storage/ssic/interoperability.wss

SDDPCM (at the time of writing, SDDPCM v2.6.0.2) is supported on the following AIX levels with APARs specified:

AIX5.3	AIX53 TL06: APAR IZ36257
	AIX53 TL07: APAR IZ36258
	AIX53 TL08: APAR IZ36259
	AIX53 TL09: APAR IZ36203
	AIX53 TL10: APAR IZ36343
	AIX53 TL11: APAR IZ36538
AIX6.1	AIX61 TL00: APAR IZ36255
	AIX61 TL01: APAR IZ36256
	AIX61 TL02: APAR IZ36221
	AIX61 TL03: APAR IZ37573
	AIX61 TL04: APAR IZ37960

Important: For AIX7.1 and later, use MPIO instead.

Here is a prerequisite for this level of SDDPCM:

```
devices.fcp.disk.ibm.mpio.rte (version 1.0.0.15)
```

12.4 Attachment to the AIX host

You need to plan the physical or logical attachment of the DS5000 to the AIX host carefully to avoid a single point of failure. The POWER machine with AIX can be attached directly to the DS5000 Storage System utilizing two independent Host Bust Adapters (HBA), but typically it is connected through a SAN fabric with multiple SAN switches or SAN directors. This enables sharing of the host port connectivity on DS5000 with other fabric-attached servers in your SAN.

The AIX host must be able to communicate with both DS5000 controllers. Figure 12-3 shows a dual HBA system where each HBA is zoned to a DS5000 controller. This configuration is preferred where the AIX initiators are isolated from each other. In our example the SAN fabric can be deemed as a single point of failure, but the good practice it to connect HBAs and DS5000 controllers to separate SAN switches in the independent fabrics.

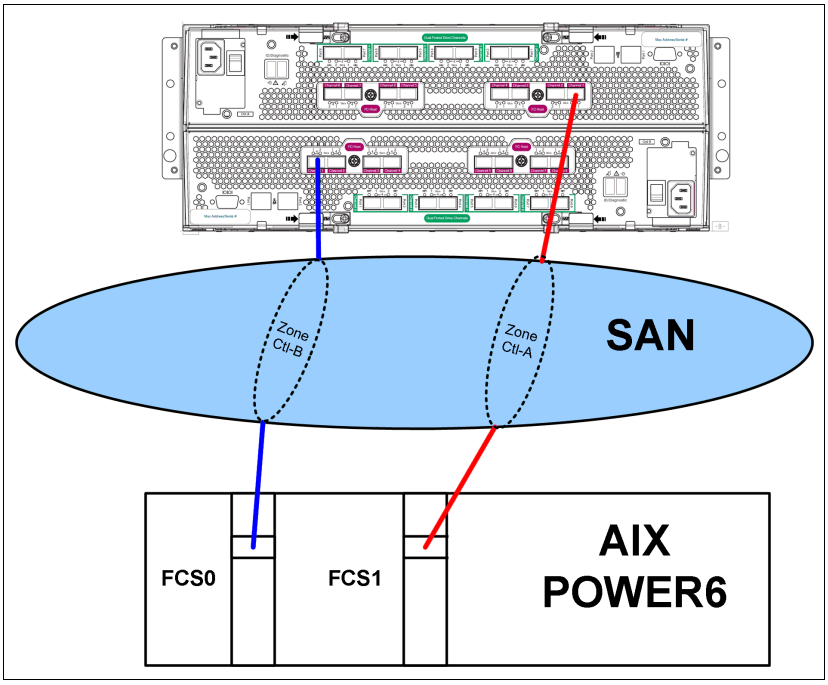


Figure 12-3 Single SAN fabric represents the Single Point of Failure

When zoning is configured on the SAN switch as shown in Figure 12-3 on page 550, the HBA worldwide names (WWN) appear on AIX host. To double-check on the DS5000, navigate to the main Storage Manager (SM) window, select **Mappings** → **View Unassociated Host Port Identifiers**. You need to see the same WWNs as you get them for each HBA from the AIX command prompt using command:

```
lscfg -vl fcsx | grep "Network Address"
```

Where "*x*" stands for the associated HBA. For instance, two WWNs of different adapters are shown in Example 12-5.

Example 12-5 Determining the WWN of HBAs in AIX

```
# lscfg -vpl fcs0 | grep "Network Address"
        Network Address.............10000000C955E581
# lscfg -vpl fcs1 | grep "Network Address"
        Network Address.............10000000C955E566
```

Tip: You need to run **`cfgmgr`** from the AIX command prompt (as *root* user) when the SAN zoning is complete or changed, so the DS5000 receives the WWN details from the SAN fabric. It confirms the activation of new changes in SAN zoning.

The WWNs can then be verified as in our example in Figure 12-4. Our host has already been defined and the WWNs labelled and associated with a host name.

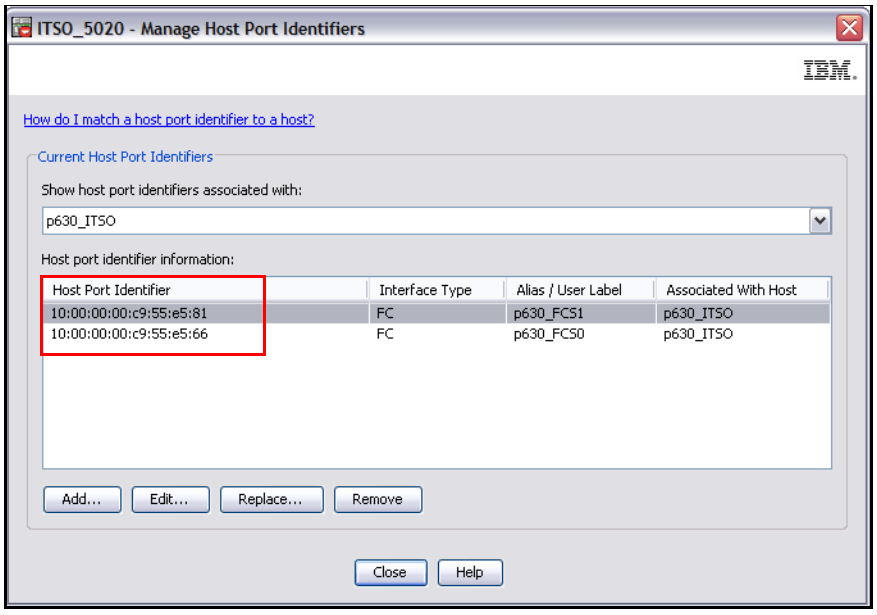


Figure 12-4 Host WWNs as seen from DS5020

12.4.1 Storage partitioning for AIX

The definition of storage partitions allows controlled access to the logical drives on the DS5000 storage subsystem to only those hosts that are defined in that particular Storage Partition. Storage partitioning is defined by specifying the world Wide Names (WWN) of the host ports to the DS5000 and creating hosts names on the DS5000 Storage Manager. The hosts can then be added to a Host Group.

Storage partitioning Host Groups on the DS5000 are used to define AIX clusters for PowerHA and also for shared disk resources when using dual Virtual Input/Output (VIO) servers in a POWER virtualized environment. They allow mapping of logical drives to all the HA cluster nodes or VIO servers which are defined in the appropriate Host Group.

When you define the host ports, you specify the operating system of the attached hosts. Each operating system expects settings that varies slightly, and handles SCSI commands accordingly. In the AIX host group all defined hosts must have a host type of “**AIX**” including VIO servers. Failure to do so will result in undetermined errors at each host in the Host Group.

Details on how to define and configure hosts and Host Groups on DS5000 are in *IBM System Storage DS5000 Series Hardware Guide*, SG24-8023.

Storage partition mapping with one HBA on one AIX server

This configuration has only one HBA in one AIX server (Figure 12-5). This configuration is *supported but not considered secure*, because SAN connections using single HBA become a single point of failure for the server when accessing external storage. However, even the single HBA requires zoning to the both controllers A and B (see Figure 12-8 on page 554).

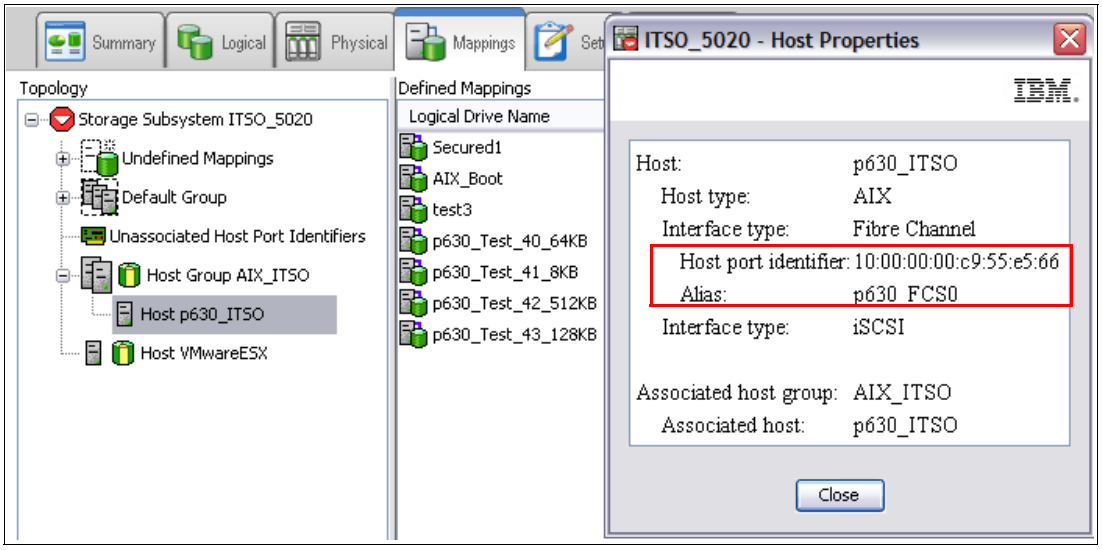


Figure 12-5 Single HBA on AIX host - single point of failure

Storage partition mapping with two HBAs on one AIX server

This configuration, which is the most common, is shown in Figure 12-6. For an example of the cabling and zoning of the HBA adapters to a DS5000 controller, see Figure 12-3 on page 550.

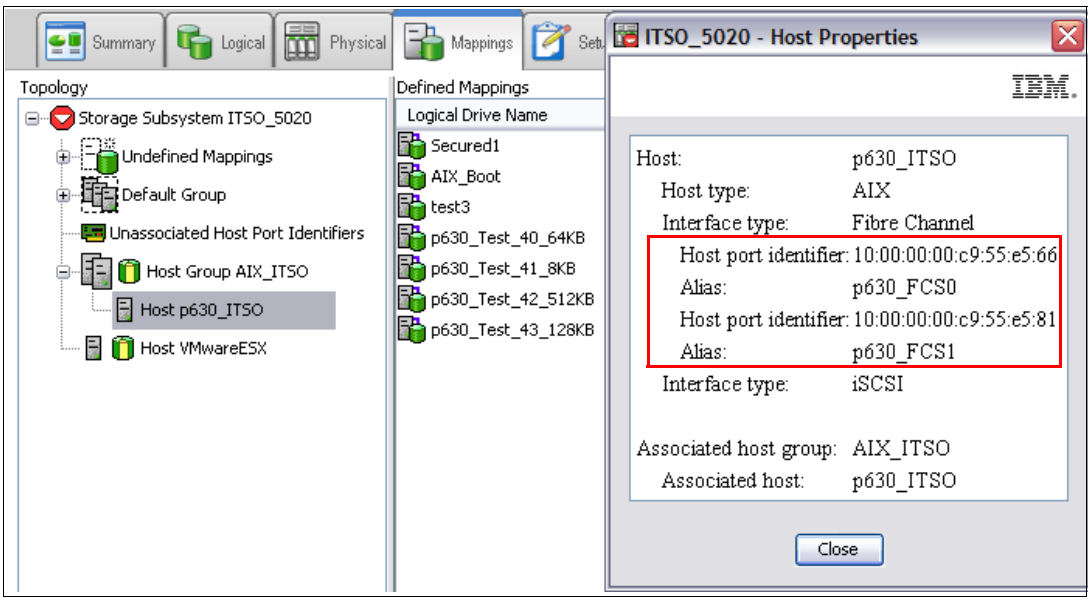


Figure 12-6 Two HBAs on a single AIX host

Two storage partitions mapping with four HBAs on two AIX servers

Figure 12-7 and Figure 12-8 represent this configuration, which is a common configuration used in PowerHA with two cluster nodes. More AIX servers can be added to the Storage Group if required in the PowerHA cluster. This configuration is also typical in a PowerVM environment where dual VIO servers are used.

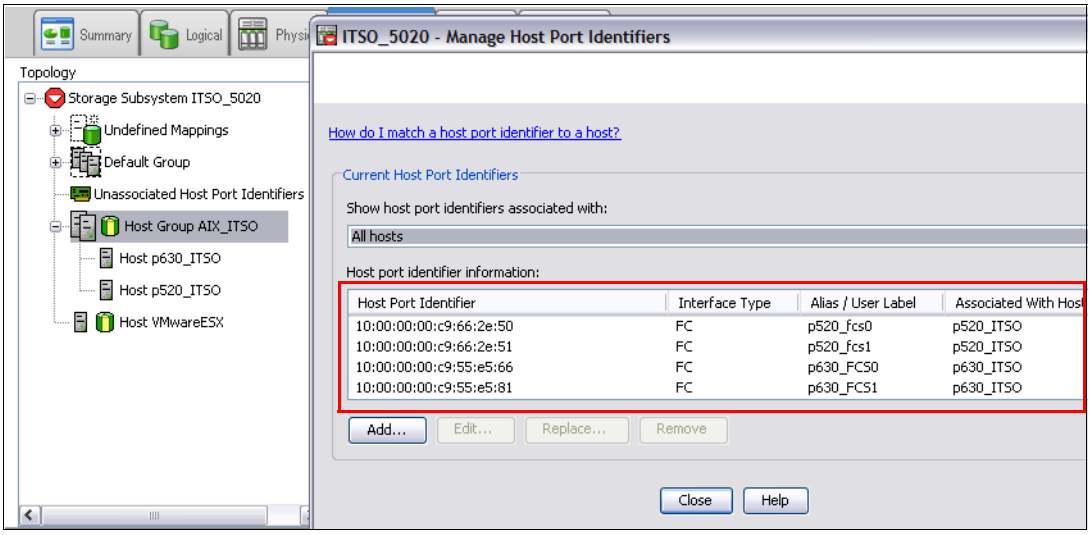


Figure 12-7 Two storage partitions - Two HBAs for each storage partition

Mapping to a default group (not supported)

The Default Host Group in DS5000 is used for hosts that are not participating in specific mappings. For AIX hosts, a Host Group (or groups) must be created where AIX defined hosts can be allocated, which is a best practice in order to keep various host types segregated from each other, AIX hosts must be in a separate Host Group.

One storage partition mapping with four HBAs on one AIX server

Mapping with four HBA on one server with only one Storage Partition is supported only when MPIO device driver is used. The SDDPCM does not offer this feature. The AIX host with four HBAs must be defined on DS5000 as two separate hosts each in a separate host group. The logical drives on DS5000 can then be mapped to either one Host Group or the other.

12.4.2 HBA configurations

In this section, we review the guidelines for HBA configurations.

Configuration with one HBA on host and two controllers on DS5000

Although, the dual HBA controllers are considered as good practice, one HBA zoned to two controllers on DS5000 is *supported*. Single HBA configurations require both controllers in the DS5000 to be connected to the host. In a switched environment, the both controllers must be connected to the switch within the same SAN zone as the HBA.

In this case, we have one HBA on the AIX server included in a single zone for access to controller A and controller B respectively. In Storage Manager, you can define one storage partition with one host HBA. See Figure 12-5 on page 552.

This configuration is *supported, but not considered suitable*. We prefer SAN administrators to create dedicated zone for each Fibre Channel HBA, as it complies with SAN best practices and standards. There is only one initiator in each zone in Figure 12-8, hence it simplifies the fabric and SAN operation and management.

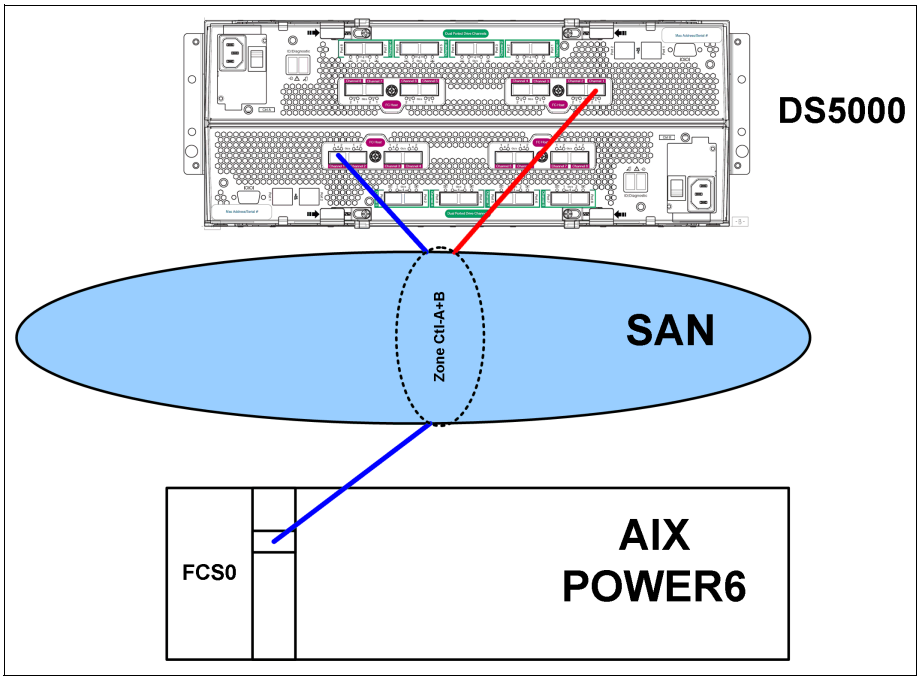


Figure 12-8 AIX system with one HBA

Configuration with two HBAs on AIX host to DS5000

Two HBAs connected to both controllers on DS5000 with appropriate zoning are *supported* (Figure 12-6 on page 552). This configuration shows a zone that includes the first HBA to the DS5000 controller A and another zone that accesses second HBA to DS5000 controller B.

In the Storage Manager, create one Storage Partition with a host and two HBAs (Figure 12-9).

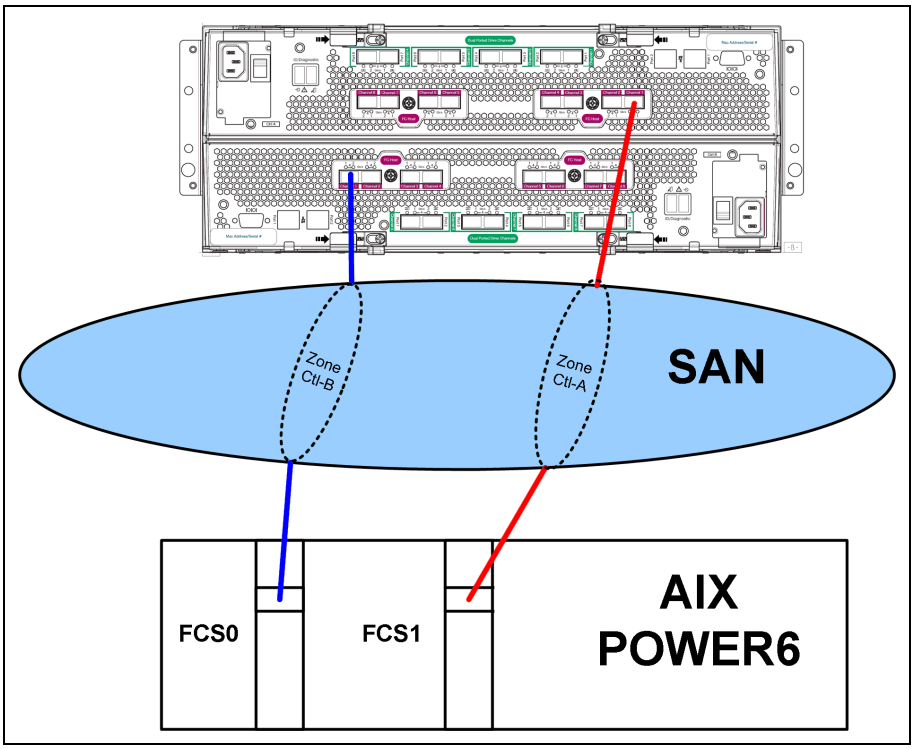


Figure 12-9 Two HBAs, configured to the two DS5000 controllers

Configuration with four HBAs on host and four host ports on DS5000

A configuration with four HBAs and four host ports on DS5000, two on each DS5000 controller is supported with appropriate zoning as shown in Figure 12-10. It is very important to define two hosts on DS5000 mapping, each defined host residing in a separate host group, which results in a logical drive being allocated to either one Host Group or the other so is only accessed by one set of HBAs.

In Figure 12-10, dual HBAs have been used to eliminate consequences of an adapter failure, therefore the connectivity is not completely lost to the host group. HBA `fcs0` and `fcs2` reside in one DS5000 Host Group, while `fcs1` and `fcs3` are in the other. If `fcs0` and `fcs1` would be in the same Host Group, then the physical adapter is a single point of failure and connectivity is lost to that Host Group when the adapter fails.

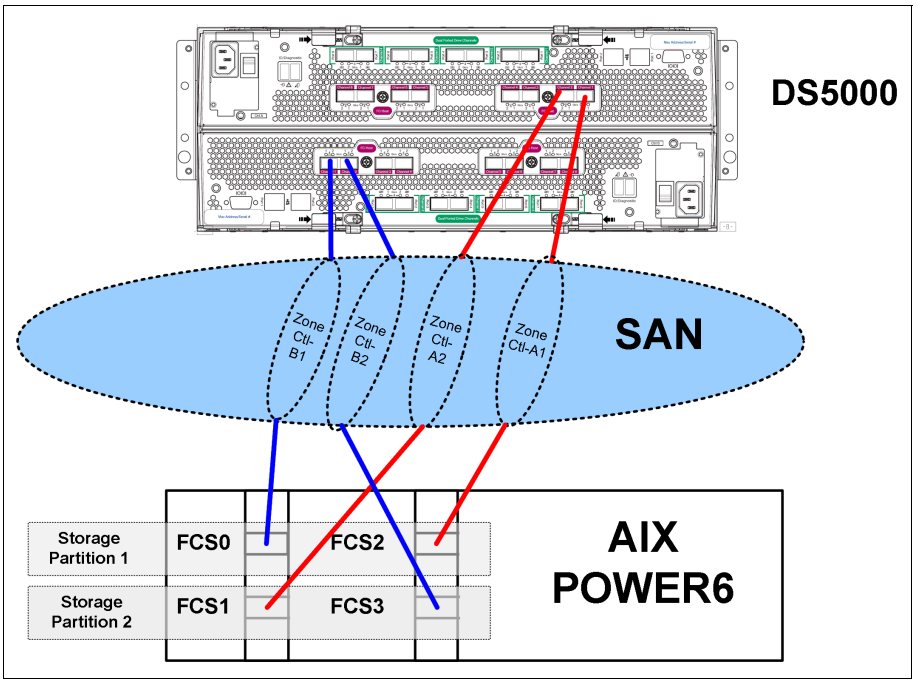


Figure 12-10 Four HBAs, two storage partitions

12.4.3 Unsupported HBA configurations

Many unsupported HBA configurations on AIX systems can be summarized as follows:

- A configuration with single HBA and only one DS5000 controller is unsupported because the AIX server requires connectivity to both DS5000 controllers to maintain logical drive connectivity in the event of controller failover. See Figure 12-8 on page 554 for the correct configuration.
- One HBA and zoning to more than two controllers ports on DS5000 are not supported.
- More than two HBAs on AIX host *without* the use of multiple Host Groups defined on DS5000 (see Figure 12-10 for an example) are not supported.

12.5 Multiple device drivers in the system

In this section, we look at both device drivers used with AIX and their coexistence of both on a single AIX server. In addition, certain servers can also be attached to other external storage subsystem in the same SAN fabric, with devices that are not compatible with the AIX MPIO device driver. These would require either SDD or SDDPCM device driver.

MPIO and SDD (or SDDPCM): AIX MPIO and SDD (used by IBM DS6000™ and DS8000) are supported on the same AIX host but on separate HBAs and using separate zones (Figure 12-11). This configuration is commonly used when the same AIX server accesses DS5000 storage as well as DS8000 or SVC managed storage.

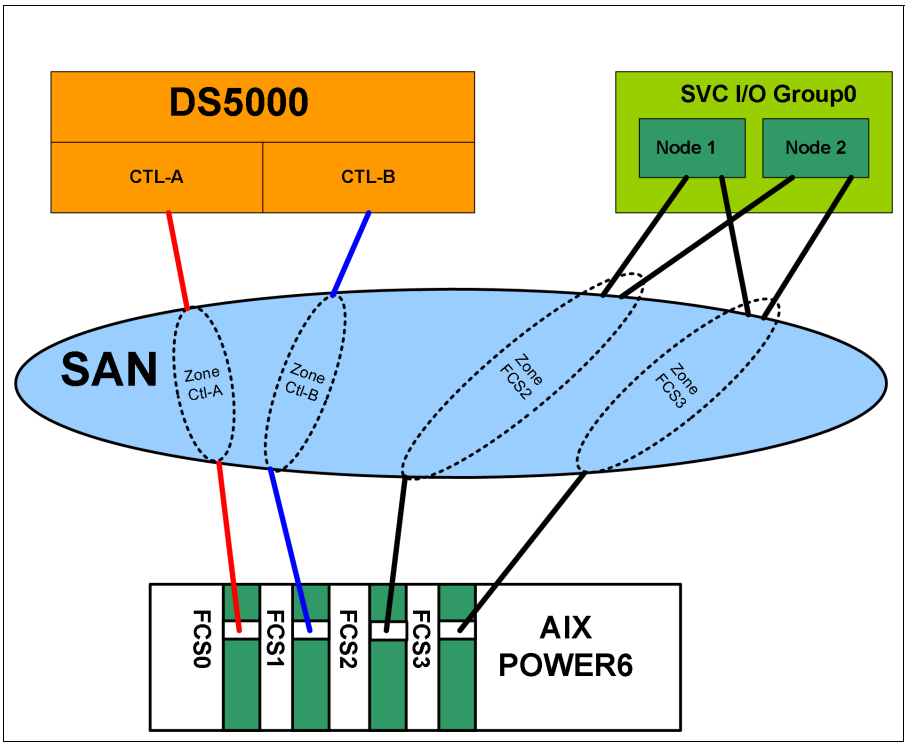


Figure 12-11 AIX MPIO and SDD coexistence using separate HBAs and zoning

Figure 12-11 shows that it is possible to configure the SDD driver to attach to ESS, DS8000, DS6000 or SVC, and the MPIO driver to be used by the DS5000. However, it requires independent pair of HBAs for access to the DS5000 and another pair of HBAs for access to the ESS or similar on the AIX machine. Separate zones are needed for each HBA.

Tip: You cannot install both SDD and SDDPCM software packages on a server for supported storage devices.

12.6 HBA and device settings

There is a variety of Fibre Channel adapters and settings of their parameters that impact performance, throughput, load balancing. The appropriate configuration can vary, based upon the environment that the AIX server is deployed in.

12.6.1 HBA configuration

In this section, we describe various considerations regarding HBA settings and performance in AIX environments.

HBA settings that affect performance

Review the following three parameters for the HBA settings that can affect performance. The type of workload and I/O rate is a major consideration when making any changes to these parameters of used HBA adapter:

- num_cmd_elems:

  `num_cmd_elems` controls the maximum number of commands to queue to the adapter. The default value is 200. You can set it to a higher value in I/O intensive environments. The maximum of num_cmd_elems for a 2 Gb HBA is 2048. However, there is a cost in real memory (DMA region) for a `num_cmd_elem` with a high value. In the throughput based environment, you want to decrease the queue depth setting to a smaller value, such as 16. In a mixed application environment, you do not want to lower the `num_cmd_elem` setting, as other logical drives might need this higher value for optimal performance. In a pure high throughput workload, this value will have no effect. There is no real specified value.

  Use performance measurement tools as described in Chapter 8, “Storage Manager Performance Monitor” on page 343 and observe the results for various values of num_cmd_elems.

- lg_term_dma:

  `lg_term_dma` controls the size of a DMA address region requested by the Fibre Channel driver at startup. Doing it does not set aside real memory, rather it sets aside PCI DMA address space. Each device that is opened uses a portion of this DMA address region. The region controlled by lg_term_dma is not used to satisfy I/O requests.

  The `lg_term_dma` can be set to 0x1000000 (16 MB). The default is 0x200000 (2MB). You must be able to safely reduce this to the default. The first symptom of `lg_term_dma` exhaustion is that disk open requests begin to fail with ENOMEM message in the error log (**errpt** command). In that case you probably cannot vary on certain volume groups. Too small a value is not likely to cause any runtime performance issue. Reducing the value will free up physical memory and DMA address space for other uses.

- max_xfer_size:

  This value is the maximum I/O size that the adapter will support. The default maximum transfer size is 0x100000 which is 16 MB of a memory area used by the adapter. Consider changing this value to 0x200000 or larger, for other allowable values of `max_xfer_size`, the memory area is 128 MB in size.

  Increasing this value increases the DMA memory area used for data transfer. You must resist the urge to just set these attributes to the maximums. There is a limit to the amount of DMA region available per slot and PCI bus. Setting these values too high might result in certain adapters failing to configure because other adapters on the bus have already exhausted the resources. On the other hand, if too little space is set aside here, I/Os might be delayed in the FC adapter driver waiting for previous I/Os to finish. You will generally see errors in **errpt** if this happens.

Review and modify HBA settings

You can review and change the HBA settings by using the following commands, we consider fcs0 is used for storage traffic, as in an Example 12-6:

- To view all three described attribute values of fcs0 adapter, enter the command:

  ```
  lsattr -Rl fcs0 -a <attribute>
  ```

  Example 12-6 Sample output of possible HBA attributes

  ```
  # lsattr -Rl fcs0 -a max_xfer_size
  0x100000
  0x200000
  0x400000
  0x800000
  0x1000000
  #
  #lsattr -Rl fcs0 -a lg_term_dma
  0x100000
  0x200000
  0x400000
  0x800000
  0x1000000
  0x2000000
  0x4000000
  0x8000000
  #
  # lsattr -Rl fcs0 -a num_cmd_elems
  20...2048 (+1)
  ```

- The following command changes the maximum transfer size (`max_xfer_size`) and the maximum number of commands to queue (`num_cmd_elems`) of an HBA (fcs0) upon the next system reboot:

  ```
  chdev -l fcs0 -P -a max_xfer_size=<value> -a num_cmd_elems=<value> -P
  ```

  This will not take effect until the next system reboot as indicated by “-P” parameter.

- To avoid a system reboot, make sure all activity is stopped on the adapter, and issue the following commands:

  ```
  rmdev -l fcs0 -R
  chdev -l fcs0 -a max_xfer_size=<value> -a num_cmd_elems=<value>
  ```

  Then recreate all child devices with the **`cfgmgr`** command.

12.6.2 Device settings

In conjunction with the HBA settings, there is a number of disk device settings that can be changed depending on the environment that is being used and what device driver is managing the hdisk. These devices include both disk (hdisk) and the Fibre Channel device driver

Disk device settings

To display disk device settings for a particular hdisk, enter the following command:

```
lsattr -El hdiskx
```

The typical output for hdisk5 is shown in Example 12-7.

Example 12-7 Device attributes of hdisk5

```
# lsattr -El hdisk5
PCM             PCM/friend/sddpcm                                 PCM                                          True
PR_key_value    none                                              Reserve Key                                  True
algorithm       load_balance                                      Algorithm                                    True
clr_q           no                                                Device CLEARS its Queue on error             True
dist_err_pcnt   1                                                 Distributed Error Percentage                 True
dist_tw_width   10                                                Distributed Error Sample Time                True
hcheck_interval 100                                               Health Check Interval                        True
hcheck_mode     nonactive                                         Health Check Mode                            True
location                                                          Location Label                               True
lun_id          0x3000000000000                                   Logical Unit Number ID                       False
lun_reset_spt   yes                                               Support SCSI LUN reset                       True
max_transfer    0x40000                                           Maximum TRANSFER Size                        True
node_name       0x500507680100b06b                                FC Node Name                                 False
pvid            00c9da3e77a2569100000000000000000                 Physical volume identifier                   False
q_err           yes                                               Use QERR bit                                 True
q_type          simple                                            Queuing TYPE                                 True
qfull_dly       2                                                 delay in seconds for SCSI TASK SET FULL      True
queue_depth     20                                                Queue DEPTH                                  True
reserve_policy  no_reserve                                        Reserve Policy                               True
retry_timeout   120                                               Retry Timeout                                True
rw_timeout      60                                                READ/WRITE time out value                    True
scbsy_dly       20                                                delay in seconds for SCSI BUSY               True
scsi_id         0x10600                                           SCSI ID                                      False
start_timeout   180                                               START unit time out value                    True
unique_id       332136005076801810575C80000000000000F04214503IBMfcp Device Unique Identification              False
ww_name         0x500507680140b06b                                FC World Wide Name                           False
```

The following key device parameters must be set according to the DS5000 and AIX environment for which the hdisk is being used:

- **reserve_policy:**

  There are four reserve policies that can be set:

  no_reserve: There is no reserve being made on MPIO devices. A device without reservation can be accessed by any initiators at any time. I/O can be sent from all the paths of the MPIO device. PowerHA supports **no_reserve** policy with Enhanced Concurrent Mode volume group and is the most suitable setting for VIO servers in a PowerVM environment. *This policy is the default reserve policy of SDDPCM.*

  single_path: This policy is the SCSI-2 reservation policy. If you set this reserve policy for MPIO devices, only the **fail_over** path selection algorithm can be selected for the devices. With this reservation policy, an MPIO device has all paths being opened; however, only one path made a SCSI-2 reservation on the device. I/O can only be sent through this path. When this path is broken, reserve is released, another path is selected, and SCSI-2 reserve is reissued by this new path. All input and output is now routed to this new path. *This policy is the default reserve policy of AIX MPIO*.

  PR_exclusive: If you set an MPIO device with this persistent reserve policy, a persistent reservation is made on this device with a persistent reserve (PR) key. Any initiators who register with the same PR key can access this device. Normally, you must pick a unique PR key for a server. Separate servers must have unique PR keys. I/O is routed to all paths of the MPIO device, because all paths of an MPIO device are registered with the same PR key. In a nonconcurrence clustering environment, such as PowerHA (HACMP), it is the reserve policy that you must select.

  PR_shared: A persistent reservation is made on this device with a persistent reserve (PR) key. However, any initiators that implemented persistent registration can access this MPIO device, even if the initiators are registered with other PR keys. In a concurrent clustering environment, such as PowerHA, it is the reserve policy that you must select for sharing resources among multiple servers.

- **algorithm:**

  There are two algorithm settings for MPIO and three with SDDPCM, as follows:

  failover: Sends all I/O down a single path. If the path is determined to be faulty, an alternate path is selected for sending all I/O. This algorithm keeps track of all the enabled paths in an ordered list. If the path being used to send I/O becomes marked failed or disabled, the next enabled path in the list is selected. The sequence within the list is determined by the path priority attribute.

  round_robin: Distributes the I/O across all enabled paths. The path priority is determined by the path priority attribute value. If a path becomes marked failed or disabled, it is no longer used for sending I/O. The priority of the remaining paths is then recalculated to determine the percentage of I/O that must be sent down each path. If all paths have the same value, then the I/O is then equally distributed across all enabled paths.

  load_balance: (SDDPCM only) This algorithm is based on the existing load balancing algorithm, but it also uses the incorporated I/O statistics on each target port to which the host is connected. You can use this algorithm for SAN configurations where I/O throughput is unbalanced on the storage targets.

- **hcheck_mode:**

  Healthchecking supports the following modes of operation:

  enabled: When this value is selected, the healthcheck command will be sent to paths that are opened with a normal path mode.

  failed: When this value is selected, the healthcheck command is sent to paths that are in failed state.

  nonactive: When this value is selected, the healthcheck command will be sent to paths that have no active I/O, including paths that are opened or in failed state, which is the default setting for MPIO devices.

Fibre Channel controller settings

Review the parameters in the Fibre Channel SCSI I/O controller, using the **lsattr** command, as shown in Example 12-8.

Example 12-8 Attributes of FC SCSI I/O Controller

```
# lsattr -El fscsi0
attach       switch            How this adapter is CONNECTED         False
dyntrk       no                Dynamic Tracking of FC Devices        True
fc_err_recov delayed_fail      FC Fabric Event Error RECOVERY Policy True
scsi_id      0x70800           Adapter SCSI ID                       False
sw_fc_class  3                 FC Class for Fabric                   True
```

The adjustable parameters which must be reviewed and changed depending on the AIX and DS5000 environment being used are:

- **dyntrk:** The default for this setting is “no”. When set to “yes” it will enable dynamic tracking of FC devices, the FC adapter driver detects when the Fibre Channel N_Port ID of a device changes. The FC adapter driver then reroutes traffic destined for that device to the new address while the devices are still online. Events that can cause an N_Port ID to change include moving a cable between a switch and storage device from one switch port to another, connecting two separate switches using an inter-switch link (ISL), and possibly rebooting a switch.

- **fc_err_recov:** This parameter controllers the fast I/O failure and is useful in situations where multipathing software is used. Setting the `fc_err_recov` attribute to `fast_fail` can decrease the I/O fail times because of link loss between the storage device and switch. This technique supports faster failover to alternate paths when FC adapter driver detects a link event, such as a lost link between a storage device and a switch. The FC adapter driver waits a short period of time, approximately 15 seconds, so that the fabric can stabilize, this occurs when the default setting of `delayed_fail` is set. At that point, if the FC adapter driver detects that the device is not on the fabric, it begins failing all I/Os at the adapter driver. Any new I/O or future retries of the failed I/Os are failed immediately by the adapter until the adapter driver detects that the device has rejoined the fabric. In single-path configurations, especially configurations with a single path to a paging device, the `delayed_fail` default setting is preferred, one such scenario is AIX server direct attachment to the DS5000 controller host port.

For both parameters, support is dependent on firmware levels being at their latest for certain HBAs (FC 6227, FC 6228, FC 6239).

The following command shows how to change the FC I/O controller:

```
chdev -l fscsi0 -a fc_err_recov=fast_fail -a dyntrk=yes
```

Changing these HBA and device attribute parameters can require stopping I/O to the devices. The previous command requires you to issue the following command beforehand in order to make the change:

```
rmdev -l fscsi0 -R
```

After the **chdev** command was successful, you need to issue **cfgmgr** command to make all devices available again, which significantly impacts any running applications on AIX, therefore making any changes to HBAs or devices requires a certain amount of pre-planning. Changes can be made to the ODM only using the “-P” option with any chdev command, the change is then be applied after the system has been re booted and can be scheduled with other maintenance tasks.

Reviewing these parameters must be undertaken after AIX has been installed and before any logical drives have been mapped by DS5000.

Hence the order of the changes must be:

1. Host Bus Adapter (fcs*n*)
2. Fibre Channel Controller (fscsi*n*)
3. Logical Drive (hdisk*n*)

As subsequent logical drives are mapped by the DS5000 to the running AIX system, changes can be made to the hdisks prior to adding them to a volume group or allocating to LPAR by a VIO server.

12.7 PowerVM with DS5000 attachment

IBM Power Systems combined with PowerVM technology are designed to help consolidate and simplify your IT environment. Key capabilities include these:

- Improve server utilization and sharing I/O resources to reduce total cost of ownership and make efficient use of IT assets.
- Improve business responsiveness and operational speed by dynamically re-allocating resources to applications as needed — to better match changing business needs or handle unexpected changes on demand.

- Simplify IT infrastructure management by making workloads independent on hardware resources, thereby enabling you to make business-driven policies to deliver resources based upon time, cost, and service-level requirements.

PowerVM is the industry-leading virtualization solution for AIX, IBM i, and Linux environments on IBM POWER technology. PowerVM offers a secure virtualization environment, built on the advanced RAS features and leadership performance of the Power Systems platform. It consists of leading technologies such as Power Hypervisor, IBM Micro-Partitioning®, Dynamic Logical Partitioning, Shared Processor Pools, Shared Storage Pools, Integrated Virtualization Manager, PowerVM Lx86, Live Partition Mobility, IBM Active Memory™ Sharing, N_Port ID Virtualization, and Suspend/Resume. PowerVM is a combination of hardware enablement and value-added software.

There are three versions of PowerVM, suited for various deployments and implementation scenarios:

- **PowerVM Express Edition** is designed for customers looking for an introduction to more advanced virtualization features at a highly affordable price.
- **PowerVM Standard Edition** provides advanced virtualization functionality for AIX, IBM i, and Linux operating systems. PowerVM Standard Edition is supported on all POWER processor-based servers and includes features designed to allow businesses to increase system utilization.
- **PowerVM Enterprise Edition** includes all the features of PowerVM Standard Edition plus two new industry-leading capabilities called Active Memory Sharing and Live Partition Mobility. It provides the most complete virtualization for AIX, IBM i, and Linux operating systems in the industry.

12.7.1 Functions and features

Only PowerVM Standard Edition can be upgraded to the Enterprise Editions. There is no upgrade path available from PowerVm Express Edition. Table 12-2 outlines the functions and features of each edition.

Table 12-2 Features of PowerVM

PowerVM Edition	Express	Standard	Enterprise
PowerVM Hypervisor	Yes	Yes	Yes
Dynamic Logical Partitioning	Yes	Yes	Yes
Management	VMControl, IVM	VMControl, IVM, HMC	VMControl, IVM, HMC
Maximum Partitions	3 per server	254 per server	254 per server
Virtual I/O Server	Yes	Yes (dual)	Yes (dual)
Integrated Virtualization Manager	Yes	Yes	Yes
PowerVM Lx86	Yes	Yes	Yes
Suspend/Resume	No	Yes	Yes
N_Port_ID Virtualization	Yes	Yes	Yes
Multiple Shared Processor Pool	No	Yes	Yes
Shared Storage Pools	No	Yes	Yes

PowerVM Edition	Express	Standard	Enterprise
Thin Provisioning	No	Yes	Yes
Active Memory Sharing	No	No	Yes
Live Partition Mobility	No	No	Yes

12.7.2 Dual VIO Server and DS5000

In this example we look at the requirements to attach dual VIO servers from a POWER6 system to DS5000. A VIO server is an appliance server with which you can associate physical resources and that allows you to share these resources amongst a group of logical partitions (LPARs) on the POWER system. The Virtual I/O Server can use both virtualized storage and network adapters, making use of the virtual SCSI and virtual Ethernet facilities.

The VIO server has a specifically adapted version of AIX operating system where functions of AIX are familiar to the AIX administrator after they have exited the VIO shell with the **oem_setup_env** command. In our test environment we use this commonly known shell.

The two VIO servers act as a cluster as each of them might need to access all storage resources in the event of the other VIO server becomes unavailable. These resources are then distributed to the logical partition (LPAR) in the POWER6 virtualized environment. In our example both VIO servers are using AIX MPIO device driver.

Our lab setup is shown in Figure 12-12. The diagram represents VIO server with dual HBAs and correct zoning in the SAN fabric - one HBA on each VIO server accesses dedicated DS5000 controller.

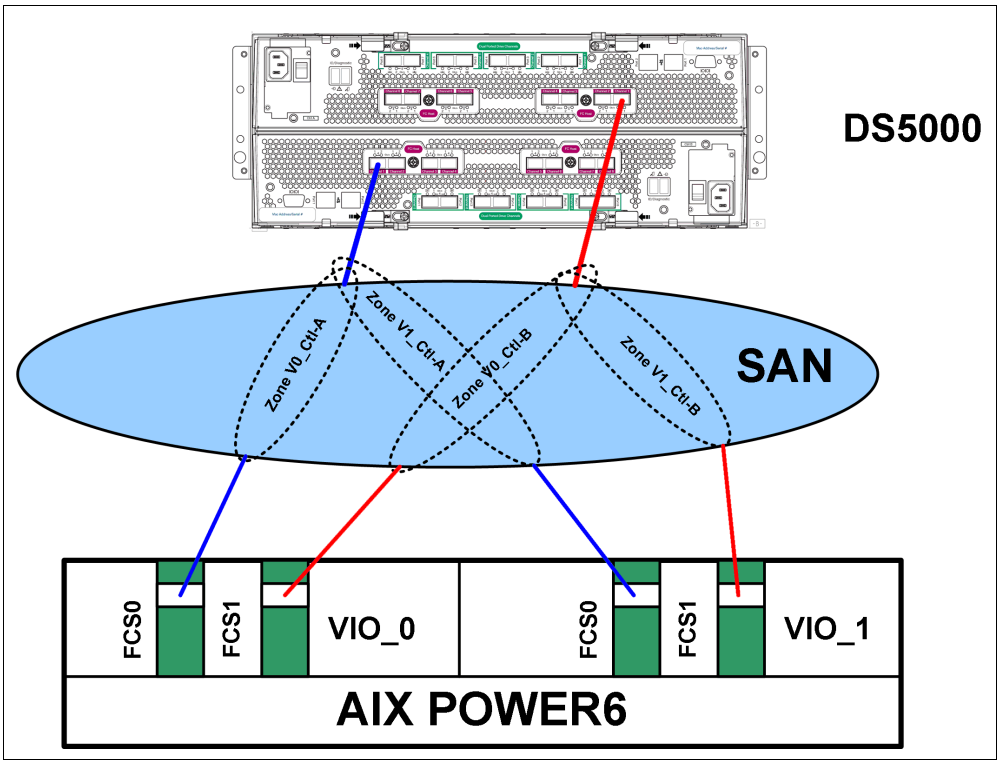


Figure 12-12 VIO servers zoned to DS5000 controllers

With the zoning completed, the Host Group on the DS5000 can be created and hosts defined. Figure 12-13 shows this action completed with four logical drives VIOR1-4, mapped to the group.

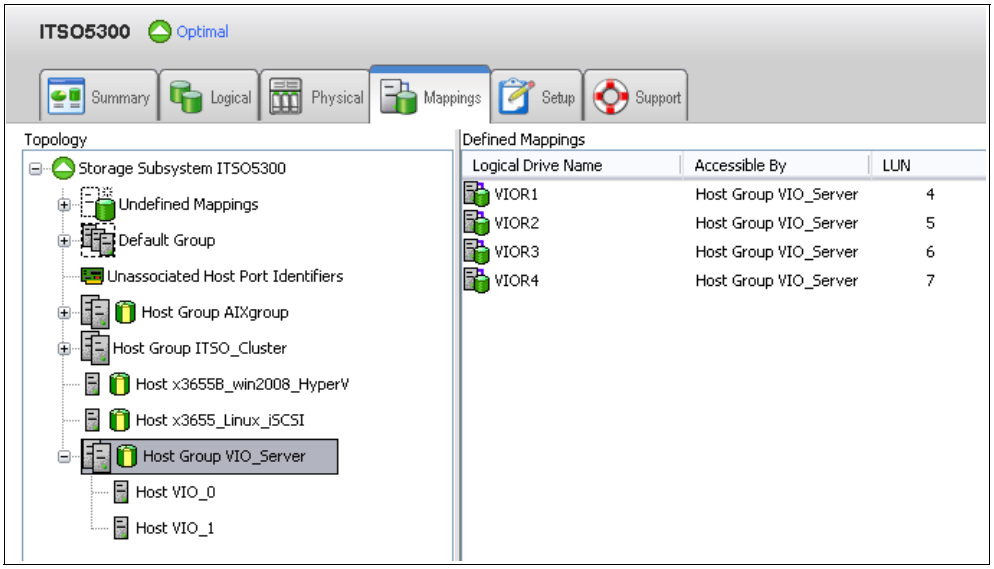


Figure 12-13 VIO Servers defined in the same Host Group

When zoning and mapping are done, the VIO servers must be able to configure each logical drive as specific `hdiskn` using **`cfgdev (cfgmgr)`** command. Before the logical drives are mapped to the VIO servers, the HBA and FC controller settings can be configured on each VIO server, as there is no I/O yet on the adapters or any child devices present, therefore these changes have least impact on resource availability for the VIO server.

For those less experienced with the VIO shell, we suggest to switch into more familiar AIX root shell using command **`oem_setup_env`** (as we did in our test environment).

HBA setting `max_xfer_size` is the only parameter considered to be changed for both adapters on each VIO server (Example 12-9).

Example 12-9 Change max_xfer_size

```
# chdev -l fcs0 -a max_xfer_size=0x200000
# chdev -l fcs1 -a max_xfer_size=0x200000
```

The FC I/O controller settings are then changed to the suggested settings for VIO servers and clusters on DS5000 (Example 12-10).

Example 12-10 Change FC controller settings

```
# chdev -l fscsi0 -a dyntrk=yes -a fc_err_recov=fast_fail
# chdev -l fscsi1 -a dyntrk=yes -a fc_err_recov=fast_fail
```

The **`cfgmgr`** command now configures and makes available disk drives (hdisks) on particular VIO server, where those four logical drives are assigned to. These hdisks will not be used in volume groups but for allocation by each VIO server to a specific LPAR in the PowerVM environment. The layout of disks allocated to VIO server should look as shown in Example 12-11.

Example 12-11 hdisks mapped to the VIO server

```
Frame id 0:
    Storage Subsystem worldwide name: 60ab8004777d800004a956964
    Controller count: 2
    Partition count: 1
    Partition 0:
    Storage Subsystem Name = 'ITSO5300'
        hdisk      LUN #   Ownership User Label
        hdisk9         4   B (preferred) VIOR1
        hdisk10        5   A (preferred) VIOR2
        hdisk11        6   B (preferred) VIOR3
        hdisk12        7   A (preferred) VIOR4
```

Before being assigned to the particular LPAR, the attributes need to be changed to reflect the requirements of a dual VIO server environment utilizing AIX MPIO device driver. Especially reservation policy needs to be changed. See the Example 12-12 for details.

Example 12-12 Changing disk attributes on VIO servers

```
# chdev -l hdisk9 -a reserve_policy=no_reserve -a algorithm=round_robin
hdisk9 changed
# chdev -l hdisk10 -a reserve_policy=no_reserve -a algorithm=round_robin
hdisk10 changed
# chdev -l hdisk11 -a reserve_policy=no_reserve -a algorithm=round_robin
hdisk11 changed
# chdev -l hdisk12 -a reserve_policy=no_reserve -a algorithm=round_robin
hdisk12 changed
```

These logical drives are now ready to be assigned by each VIO server to the designated LPAR in the PowerVM environment. For further details about PowerVM, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

12.8 Dynamic functions of DS5000

The DS5000 offers the following dynamic functions for mixed workloads:

- Dynamic Segment (stripe) Size migration (DSS)
- Dynamic RAID Level Migration (DRM)
- Dynamic Capacity Expansion (DCE)
- Dynamic Volume Expansion (DVE)
- Dynamic (ERM) Mode Switching

In the following text of the chapter, we briefly describe how AIX platforms on IBM POWER systems can benefit from these dynamic functions.

12.8.1 The dynamic functions in AIX environments

The dynamic functions are supported on AIX operating environments with certain limitations (for detailed information about supported platforms and operating systems, see the DS5000 compatibility matrix). Here we present the details of Dynamic Volume Expansion with AIX host, as the other dynamic functions basically run on background of the AIX systems using DS Storage Manager, and without interruption to the applications.

Tip: DVE for AIX 5.3 requires PTF U499974 installed before expanding any file systems. The AIX 6.1 and 7.1 already incorporate this feature in base installation and natively support Dynamic Volume Expansion.

With AIX, there is no real Dynamic Volume Expansion in terms of filesystem size, because when you increase the size of the logical drive in the DS5000 Storage Manager, in order to use the additional disk space, it is necessary to modify the volume group containing that drive. A short downtime is required to perform that operation, which includes the following steps:

1. Stop the application.
2. Unmount all file systems on the involved volume group.
3. Change the characteristics of the volume group with **`chvg -g vgname`**.
4. Re-mount file systems.
5. Restart the application.

In the following sections, we provide the practical example of the procedure.

12.8.2 Example: Increasing DS5000 logical volume size in AIX step by step

This example assumes the following configuration:

- The logical drive to be increased is test3, which is hdisk4 with existing capacity of 60 GB.
- In AIX Logical Volume Manager (LVM), we have defined one volume group (ds5kvg3) with one logical volume (test4_lv) on the hdisk4 with `filesystem /test4` mounted.

We show how to increase the size of the logical volume (test3), by increasing the size of hdisk4 without creating a new disk.

Using the DS5000 Storage Manager, we increase the size of the DS5000 logical drive by 10 GB (see Figure 12-14).

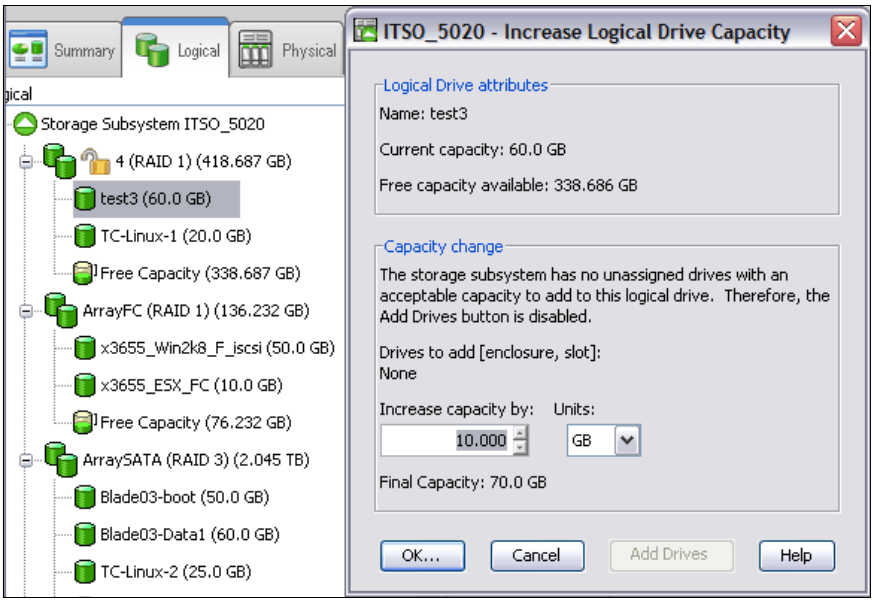


Figure 12-14 Increase the size of logical drive - test3 by 10GB

You can verify the size (in MegaBytes) of the corresponding `hdisk` in AIX, by typing the **`lspv`** and **`bootinfo`** commands as shown in Example 12-13 prior to the increase. Each of these two commands uses slightly different calculation of disk size (bootinfo gathers information from ODM, lspv multiples the Physical Partition size with the number of PPs), therefore both results are also slightly dissimilar. When recalculated to GigaBytes (divided by 1024), the difference disappears.

Example 12-13 Verification of disk size in AIX

```
# bootinfo -s hdisk4
61440
#
# lspv hdisk4
PHYSICAL VOLUME:    hdisk4                  VOLUME GROUP:     ds5kvg3
PV IDENTIFIER:      0007041a0cf29ecf VGIDENTIFIER 0007041a00004c00000001240cf29fcd
PV STATE:           active
STALE PARTITIONS:   0                       ALLOCATABLE:      yes
PP SIZE:            64 megabyte(s)          LOGICAL VOLUMES:  1
TOTAL PPs:          959 (61376 megabytes)   VG DESCRIPTORS:   2
FREE PPs:           659 (42176 megabytes)   HOT SPARE:        no
USED PPs:           300 (19200 megabytes)   MAX REQUEST:      256 kilobytes
FREE DISTRIBUTION:  192..192..00..83..192
USED DISTRIBUTION:  00..00..191..109..00
MIRROR POOL:        None
```

The output of the command gives the size of disk in megabytes, 61440 (60 GB).

In the next step, stop the application, unmount all file systems on the particular volume group and set it offline:

```
varyoffvg ds5kvg3
```

Tip: We suggest to wait for DS5000 increase capacity process on the array to complete, as shown in Figure 12-15, before continuing with the **`chvg`** command in AIX.

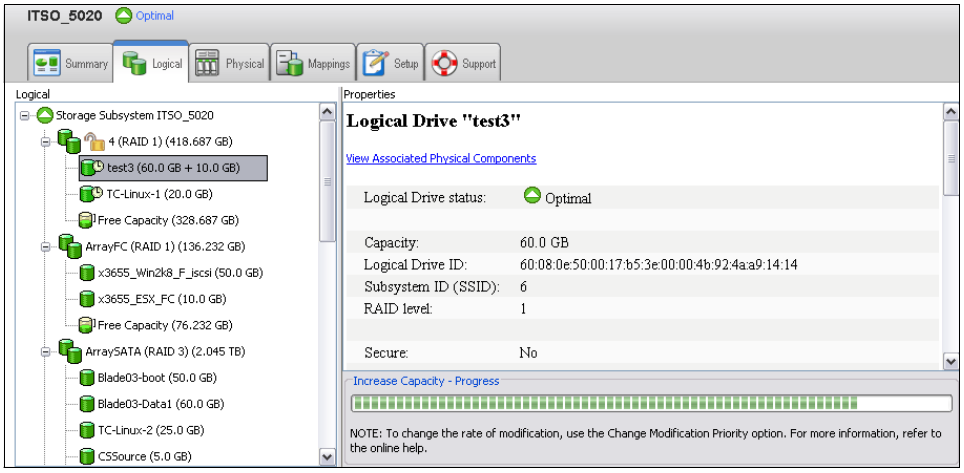


Figure 12-15 Wait for increase capacity process to complete

Modify the volume group to use the new space, with the following commands:

```
varyonvg ds5kvg3
chvg -g ds5kvg3
```

Example 12-14 shows the same disk hdisk4 with its new size of 71680 MB (70 GB).

Example 12-14 Increased capacity of allocated volume

```
# chvg -g ds5kvg3
0516-1164 chvg: Volume group ds5kvg3 changed.  With given characteristics ds5kvg3
        can include upto 16 physical volumes with 2032 physical partitions each.
#
# bootinfo -s hdisk4
71680
#
# mount /test4
#
# lspv hdisk4
PHYSICAL VOLUME:    hdisk4                   VOLUME GROUP:     ds5kvg3
PV IDENTIFIER:      0007041a0cf29ecf VGIDENTIFIER 0007041a00004c00000001240cf29fcd
PV STATE:           active
STALE PARTITIONS:   0                        ALLOCATABLE:      yes
PP SIZE:            64 megabyte(s)           LOGICAL VOLUMES:  1
TOTAL PPs:          1119 (71616 megabytes)   VG DESCRIPTORS:   2
FREE PPs:           819 (52416 megabytes)    HOT SPARE:        no
USED PPs:           300 (19200 megabytes)    MAX REQUEST:      256 kilobytes
FREE DISTRIBUTION:  224..160..00..211..224
USED DISTRIBUTION:  00..64..223..13..00
MIRROR POOL:        None
```

Finally, mount all remaining file systems and restart the application.

You can also verify the available new space for the volume group with the command **`lsvg DS5kvg3`**. Your newly extended space in the volume group is ready to be used and the size of filesystem can be increased using command **`chfs`** and its parameters, in our example `size=+11G`.

12.9 PowerHA and DS5000

Clustering servers is the linking of two or more computers or nodes into a single, unified resource. High-availability clusters are designed to provide continuous access to the business critical data and applications through the component redundancy and application failover.

PowerHA (in previous versions known as Highly Available Cluster Multi Processing - HACMP) is designed to automatically detect system or network failures and eliminate a single point-of-failure by managing failover to a recovery processor with a minimal loss of end-user time. The current release of PowerHA can detect and react to software failures severe enough to cause a system crash and network or adapter failures. The Enhanced Scalability capabilities of PowerHA offer additional availability benefits through the use of the Reliable Scalable Cluster Technology (RSCT) function of AIX or Cluster Aware AIX (CAA) in PowerHA 7.1 and higher.

PowerHA makes use of the redundant hardware configured in the cluster to keep an application running, and restart it on a backup cluster-node if necessary. Using PowerHA virtually eliminates planned outages, because users, applications, and data can be moved to back up the system during the scheduled system maintenance. Such advanced features as Cluster Single Point of Control and Dynamic Reconfiguration allow the automatic addition of users, files, hardware, and security functions without interruption to the mission-critical jobs.

PowerHA clusters can be configured to meet complex and various applications' availability and potential recovery needs. Configurations can include mutual takeover or idle standby recovery processes. With an PowerHA mutual takeover configuration, applications and their workloads are assigned to specific servers, thus maximizing application throughput and effectively utilizing investments in hardware and software. In an idle standby configuration, an extra node is added to the cluster to back up any of the other nodes in the cluster.

In an PowerHA environment, each server in a cluster is a node. Up to 32 POWER AIX servers can participate in a single PowerHA cluster. Each node has access to shared disk resources that are accessed by other nodes. When there is a failure, PowerHA transfers ownership of shared disks and other resources based on how you define the relationship among nodes in a cluster. This process is known as resource group (RG) *failover* or resource group *fallback* (on Microsoft Clusters also known as *failback*).

Ultimately, the goal of any IT solution in a critical environment is to provide continuous service and data protection. The high availability is just one building block to achieve the continuous operation goal. The high availability is based on the availability of the hardware, software (operating system and its components), application, and network components.

To build a highly available solution, the following components are needed:

- Redundant servers
- Redundant network paths
- Redundant network adapters
- Redundant network paths
- Redundant storage (data) paths
- Redundant (mirrored/RAID) storage
- Monitoring
- Failure detection
- Failure diagnosis
- Automated failover
- Automated reintegration

The main objective of the PowerHA is eliminate Single Points of Failure (SPOFs) as detailed in Table 12-3.

Table 12-3 Eliminate SPOFs

Cluster object	Eliminated as a single point of failure by:
Node (servers)	Multiple nodes
Power supply	Multiple circuits or power supplies
Network adapter	Redundant network adapters
Network	Multiple networks connected to each nodes, redundant network paths with independent hardware between each node and the clients.
TCP/IP subsystem	A non- IP networks to back up TCP/IP
I/O Adapter	Redundant I/O Adapters
Controllers	Use of redundant controllers
Sites	Use of more than one site for disaster recovery
Storage	Redundant hardware, enclosures, disk mirroring / RAID technology, redundant data paths
Resource Groups	Use of resource groups to control all resources required by an application

Each of the items listed in Table 12-3 on page 570 in the Cluster Object column is a physical or logical component that, if it fails, will result in the application being unavailable for serving clients.

The PowerHA software provides the framework and a set of tools for integrating applications in a highly available system. Applications to be integrated in a PowerHA cluster require a fair amount of customization, not at the application level, but rather at the PowerHA and AIX platform level. PowerHA is a flexible platform that allows integration of generic applications running on AIX platform, providing for high available systems at a reasonable cost.

12.9.1 HACMP/ES and ESCRM

Scalability, support of large clusters, and therefore, large configurations of nodes and potentially disks leads to a requirement to manage “clusters” of nodes. To address management issues and take advantage of new disk attachment technologies, PowerHA Enhanced Scalable (HACMP/ES) was released. It was originally only available for the RS6000/SP where tools were already in place with PSSP to manage larger clusters.

The additional variety of HACMP/ES is HACMP/XD that allows applications and resource groups to failover to the remote sites over Extended Distance (that is why /XD in the name). After the required installation packages and licenses are applied to the AIX, you see the additional option in the **`smitty hacmp`** menu as shown in Figure 12-16 (follow the sequence: **System Management (C-SPOC)** → **HACMP Resource Group and Application Management** → **Move a Resource Group to Another Node/Site**).

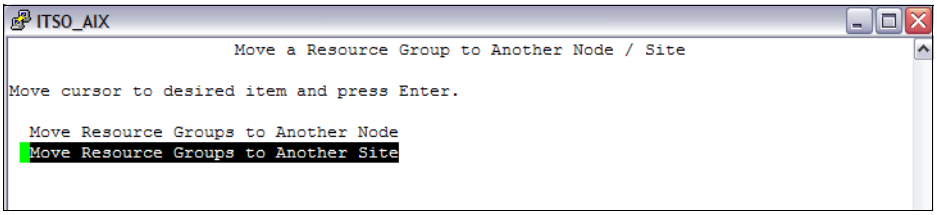


Figure 12-16 Options of HACMP resource group movement

Enhanced Scalability Concurrent Resource Manager (ESCRM) optionally adds concurrent shared-access management for the supported RAID disk subsystems. Concurrent access is provided at the raw disk level. The application must support a mechanism to control access to the shared data, such as locking. The ESCRM includes the HACMP/ES components and the HACMP distributed lock manager.

12.9.2 Cluster Aware AIX and Non-Cluster Aware AIX with PowerHA

In this section we describe these different types of clustering capabilities.

Cluster Aware AIX

Cluster Aware AIX (CAA) introduces fundamental clustering capabilities into the base operating system AIX. Such capabilities include the creation and definition of the set of nodes that comprise the cluster. CAA provides the tools and monitoring capabilities for the detection of node and interface health.

Filesets: CAA is provided by the non-PowerHA file sets `bos.cluster.rte`, `bos.ahafs`, and `bos.cluster.solid`. These filesets are on the AIX Install Media or in the TL6 of AIX 6.1.

CAA provides a set of tools and APIs to enable clustering on the AIX operating system. CAA does not provide the application monitoring and resource failover capabilities that PowerHA provides. PowerHA uses the CAA capabilities. Other applications and software programs can use the APIs and command-line interfaces (CLIs) that CAA provides to make their applications and services “Cluster Aware” on the AIX operating system.

The following products and parties can benefit from CAA technology:

- RSCT (3.1. and later)
- PowerHA (7.1 and later)
- VIOS (2.2 and later)
- Third-party ISVs, service providers, and software products

When CAA is active in your cluster, you notice the daemon services running as shown in Example 12-15.

Example 12-15 CAA daemons and services

```
# lssrc -g caa
Subsystem              Group    PID         Status
clcomd                 caa      4849670     active
cld                    caa      7012500     active
solid                  caa      11010276    active
clconfd                caa      7340038     active
solidhac               caa      10027064    active
```

CAA includes the following services:

clcomd The **clcomd** daemon is the cluster communications daemon, which has changed in PowerHA 7.1. In previous versions of PowerHA, it was called **clcomdES**. The location of the rhosts file that PowerHA uses has also changed. The rhosts file used by the **clcomd** service is in the /etc/cluster/rhosts directory. The old **clcomdES** rhosts file in the /usr/es/sbin/cluster/etc directory is not used.

cld The **cld** daemon runs on each node and determines whether the local node must be the primary or the secondary IBM **solidDB**® database server.

solid The **solid** subsystem provides the database engine, and **solidhac** is used for high availability of the IBM solidDB database. Both run on the primary and the secondary database servers. In a two-node cluster, the primary database is mounted on node 1 (/clrepos_private1), and the secondary database is mounted on node 2 (/clrepos_private2). These nodes have the **solid** and **solidhac** subsystems running.

clconfd The **clconfd** subsystem runs on each node of the cluster and wakes up every 10 minutes to synchronize any necessary cluster changes.

Non-Cluster Aware AIX with PowerHA

Instead of utilizing Cluster Aware AIX (CAA) services and daemons, the main communication goes from PowerHA to Group Services (`grpsvcs`), then to Topology Services (`topsvcs`), and back to PowerHA. The communication path from PowerHA to Resource Monitoring and Control (RMC) is used for PowerHA Process Application Monitors. In another case where PowerHA uses RMC, a resource group is configured with the Dynamic Node Priority policy.

Tip: Non-CAA mode is still used when you have a PowerHA version 6.1 or earlier, even if you are running AIX 7.1.

Group Services now use the subsystem name `cthags`, replacing `grpsvcs`. Group Services are now started with a different control script (`cthags`) and in turn from a different subsystem name `cthags`. Example 12-16 is a detailed list of all available PowerHA services, even where CAA is being used, as indicated by their active status, in contrast with Group Services shown as inactive.

Example 12-16 PowerHA-related services and daemons

```
# lssrc -a | egrep "rsct|ha|svcs|caa|cluster" | grep -v _rm
cld              caa         4980920    active
clcomd           caa         4915400    active
clconfd          caa         5243070    active
cthags           cthags      4456672    active
ctrmc            rsct        5767356    active
clstrmgrES       cluster     10813688   active
solidhac         caa         10420288   active
solid            caa         5832836    active
clevmgrdES       cluster     5177370    active
clinfoES         cluster     11337972   active
ctcas            rsct                   inoperative
topsvcs          topsvcs                inoperative
grpsvcs          grpsvcs                inoperative
grpglsm          grpsvcs                inoperative
emsvcs           emsvcs                 inoperative
emaixos          emsvcs                 inoperative
```

For more information about the Cluster Aware AIX, related changes in PowerHA 7.1, and the IBM AIX 7.1, see the *IBM AIX Version 7.1 Differences Guide*, SG24-7910.

12.9.3 Supported environments

Important: Before installing DS5000 in a PowerHA environment, always read the AIX `readme` file, the DS5000 *readme* for the specific Storage Manager version and model, and the PowerHA configuration and compatibility matrix information.

For up-to-date information about the supported environments for DS5000, see the IBM System Storage Interoperability Center site, where in the field *Clustering* choose your version of IBM HACMP or PowerHA:

`http://www.ibm.com/systems/support/storage/ssic/interoperability.wss`

For PowerHA, see the website:

`http://www.ibm.com/systems/power/software/availability/aix/resources/`

12.9.4 General rules

The benefit of a PowerHA environment is to eliminate the single points of failure. Figure 12-17 shows the concept of a two-node PowerHA cluster (it is not a limitation; you can have more nodes) attached to a DS5000 Storage Server through a fully redundant, independent SAN fabrics. This type of configuration eliminates a Fibre Channel (FC) adapter, switch, or cable from being a single point of failure (PowerHA itself protects against a node failure).

If there is only one application (one Resource Group) running on the cluster, we understand the cluster as *Active-Passive* PowerHA; if we install two independent applications (Resource Groups), and each of them running on dedicated cluster node, we talk about *Active-Active* cluster (sometimes referred as *Mutual* cluster). The typical deployment of Active-Active cluster is the database system (DB2, Oracle, IBM Informix®, mySQL, and so on) installed on one of the cluster nodes, and the application managing the data in that database is running on opposite cluster node (IBM WebSphere®, web server, user interfaces, and so on).

Using only one, single FC switch is possible (with the appropriate zoning), but is considered a single point of failure. If the FC switch fails, you cannot access the DS5000 volumes from none of the PowerHA cluster nodes. So, with the only a single FC switch, PowerHA becomes ineffective in the event of a switch failure. In our example we show the configuration for a fully redundant production environment. Each PowerHA cluster node must also contain at least two Fibre Channel host adapters to eliminate the adapter as a single point of failure. Note that each adapter in a particular cluster node is connected to a separate switch (cross cabling).

Zoning on FC switches is necessary in order to cater for the cluster requirements. Every adapter in the AIX system can access only one controller (they are AIX-specific zoning restrictions, not PowerHA specific).

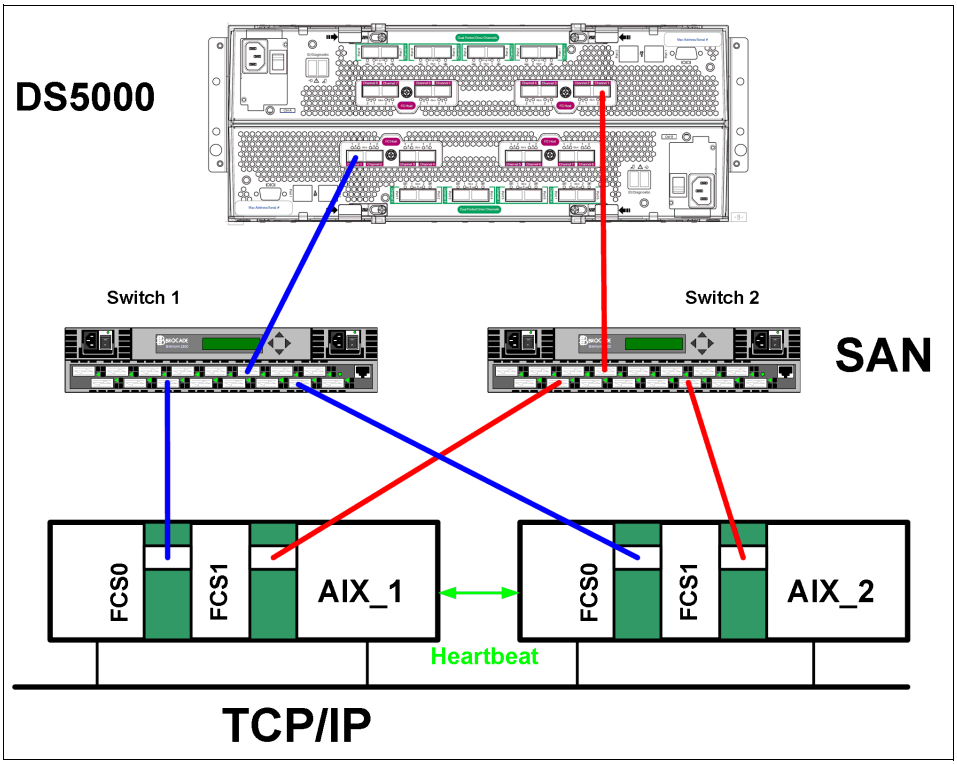


Figure 12-17 PowerHA cluster and DS5000

12.9.5 Limitations and restrictions of PowerHA

When installing a DS5000 in an PowerHA environment, there are certain restrictions and guidelines to take into account, which we list here. It does not mean that any other configuration fails, but it can lead into the unpredictable results, making it hard to manage and troubleshoot.

Note the following general limitations and restrictions for PowerHA:

- Switched fabric connections between the host nodes and the DS5000 storage subsystem are preferred. However, direct attachment from the host nodes to the DS5000 storage subsystem in an PowerHA environment is now supported when all of the following restrictions and limitations are met:
  - Only dual-controller DS5000 storage subsystem versions are supported for direct attachment in a high-availability (HA) configuration.
  - The AIX operating system must be Version 5.3 or later.
  - The PowerHA clustering software must be Version 5.1 or later, however, it is always best to use the latest version.
  - All host nodes that are directly attached to the DS5000 storage subsystem must be part of the same PowerHA cluster.
  - All logical drives that are members the DS5000 storage subsystem host group are part of one or more enhanced concurrent-mode volume groups.
  - The volume group (`varyonvg`) is in the active state only on the host node that owns the PowerHA non-concurrent resource group (which contains the enhanced concurrent mode volume group or groups). For all other host nodes in the PowerHA cluster, the enhanced concurrent-mode volume group (`varyoffvg`) is in the passive state.

- Direct operations on the logical drives in the enhanced concurrent-mode volume groups cannot be performed from any host nodes in the PowerHA cluster if the operations bypass the Logical Volume Manager (LVM) layer of the AIX operating system. For example, you cannot use a DD command while logged in as the root user.
- Each host node in the PowerHA cluster must have two Fibre Channel connections to the DS5000 storage subsystem. One direct Fibre Channel connection must be to controller A in the DS5000 storage subsystem, and the other direct Fibre Channel connection must be to controller B in the DS5000 Storage System.
- You can directly attach a maximum of two host nodes in an PowerHA cluster to a DS5000 storage subsystem. Each host node must have two direct Fibre Channel connections to the storage subsystem.

- PowerHA Cluster-Single Point of Control (C-SPOC) cannot be used to add a DS5000 disk to AIX through the **`Add a Disk to the Cluster`** facility.
- PowerHA C-SPOC does not support enhanced concurrent mode volume groups.
- PowerHA Versions 5.x are supported on the IBM Power* 595 (9119-FHA). clustered configurations.
- PowerHA is now supported in Heterogeneous server environments. For more information regarding a particular operating system environment, see the specific *Installation and Support Guide*.
- PowerHA clusters can support 2-32 servers for DS5000 partition. In this environment, be sure to read and understand the AIX device drivers queue depth settings, as documented in “Changing ODM attribute settings in AIX” on page 203.
- Non-clustered AIX hosts can be connected to the same DS5000 that is attached to an PowerHA cluster, but must be configured on separate DS5000 host partitions (Host Group).
- Single HBA configurations are allowed, but each single HBA configuration requires that both controllers in the DS5000 be connected to a switch within the same SAN zone as the HBA. although single HBA configurations are supported, using a single HBA configuration is not considered suitable for PowerHA environments, due to the fact that it introduces a single point of failure in the storage I/O path.
- PowerHA from version V5.4 onwards supports Linux, extending many of its robust capabilities and heritage to the Linux environment. Linux support includes the base capabilities for reliable monitoring and failure detection as available for AIX.

For further PowerHA information, see the following website:

`http://www.ibm.com/systems/power/software/availability/aix/resources/`

12.9.6 Planning considerations

When planning a high availability cluster, you must consider the sizing of the nodes, storage, network, and so on, to provide the necessary resources for the applications to run properly, even in a takeover situation. Important part of the planning phase is a disk and network heartbeats’ resources consideration.

Sizing: Choosing the nodes in the cluster

Before you start the implementation of the cluster, you must know how many nodes are required, and the type of the nodes that must be used. It is important in terms of the resources required by the applications.

Sizing of the nodes must consider the following aspects:

- Processors (number of processors and speed)
- Amount of memory in each node
- Disk storage (internal)
- Number of communication and disk adapters in each node
- Node reliability

The number of nodes in the cluster depends on the number of applications to be made highly available, and also on the degree of availability desired. Having more than one spare node for each application in the cluster increases the overall availability of the applications.

PowerHA V5.x and higher supports a variety of nodes, ranging from desktop systems to high-end servers. Logical Partitions (LPARs) are supported as well.

The cluster resource sharing is based on the applications requirements. Nodes that perform tasks that are not directly related to the applications to be made highly available and do not need to share resources with the application nodes must be configured in separate clusters for easier implementation and administration.

All nodes need to provide sufficient resources (processors, memory, and adapters) to sustain execution of all the designated applications in a fail-over situation (to take over the resources from a failing node).

Where possible, use cluster nodes with a similar hardware configuration, especially when implementing clusters with applications in mutual takeover or concurrent configurations. Doing it makes it easier to distribute resources and to perform administrative operations (software maintenance and so on).

Sizing: Storage considerations

Applications to be made highly available require a shared storage space for application data. The shared storage space is used either for concurrent access, or for making the data available to the application on the takeover node (in a failover situation).

The storage to be used in a cluster must provide shared access from all designated nodes for each application. The technologies currently supported for PowerHA shared storage are SCSI, SAS, and Fibre Channel including SSD and FDE (SED) drives — as is the case with the DS5000.

The storage configuration must be defined according to application requirements as non-shared (“private”) or shared storage. The private storage can reside on internal disks and is not involved in any takeover activity.

Shared storage must provide mechanisms for controlled access, considering the following reasons:

- Data placed in shared storage must be accessible from whichever node the application might be running at a point in time. In certain cases, the application is running on only one node at a time (non-concurrent), but in certain cases, concurrent access to the data must be provided. On DS5000 the Host Groups are used for the PowerHA cluster
- In a non-concurrent environment, if the shared data is updated by the wrong node, it can result in data corruption.
- In a concurrent environment, the application must provide its own data access mechanism, because the storage controlled access mechanisms are by-passed by the platform concurrent software (AIX/PowerHA).

12.9.7 Cluster disks setup

The following sections outlines the important information about cluster disks setup, and in particular describes cabling, AIX configuration, microcode loading, and configuration of DS5000 disks.

Figure 12-18 shows a simple two-node PowerHA cluster. The basic cabling configuration ensures redundancy and allows for possible future expansion. It also supports remote mirroring because it leaves two controller ports on the DS5000 available.

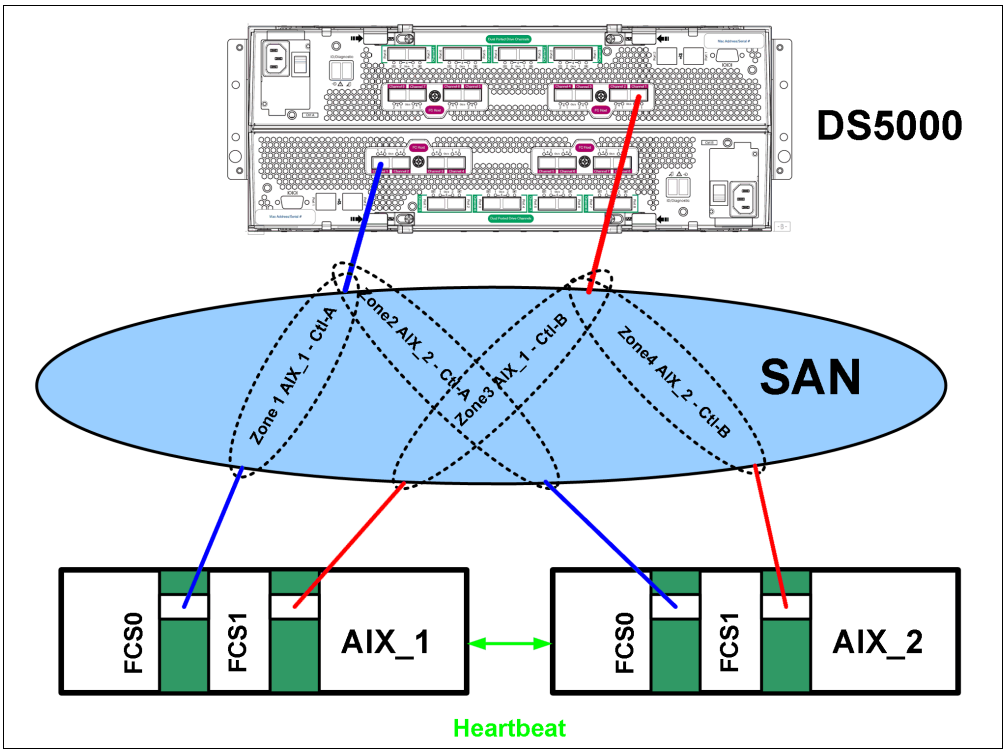


Figure 12-18 HACMP - Preferred cabling and zoning

Log on to each of the AIX nodes in the cluster and verify that you have a working configuration. You must get an output similar to what is illustrated next for the various commands:

- List of adapters on each cluster node (Example 12-17).

Example 12-17 Available Fiber Channel adapters

```
AIX_1# lsdev -Cc adapter | grep fcs
fcs0    Available 1Z-08    FC Adapter
fcs1    Available 1D-08    FC Adapter

AIX_2# lsdev -Cc adapter | grep fcs
fcs0    Available 1Z-08    FC Adapter
fcs1    Available 1D-08    FC Adapter
```

- Using Storage Manager, define a Host Group for the cluster and include the various hosts (nodes) and host ports as illustrated in Figure 12-19.

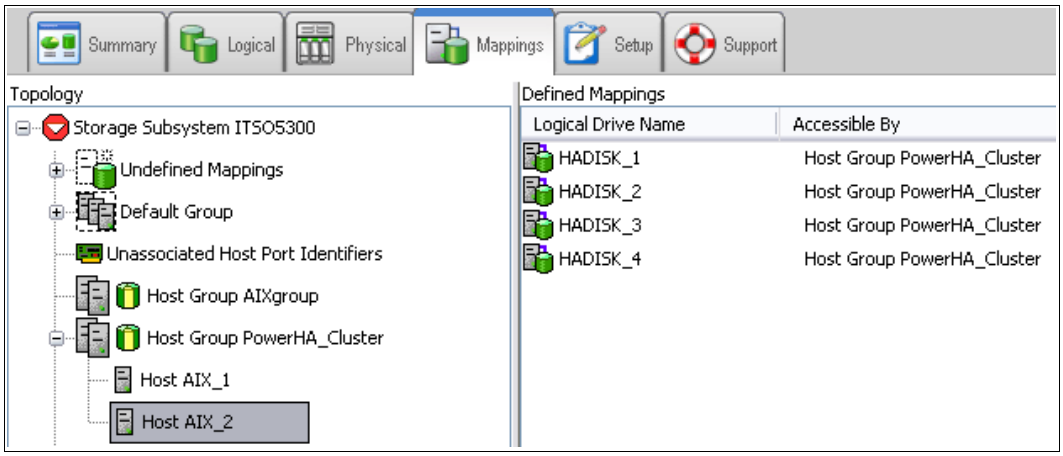


Figure 12-19 Cluster - Host Group and Mappings

- Verify disks on cluster node 1 (AIX_1 in Example 12-18).

Example 12-18 Allocated disks to AIX_1

```
# lsdev -Cc disk
hdisk9  Available 1n-08-02     IBM MPIO DS5000 Array Disk
hdisk10 Available 1n-08-02     IBM MPIO DS5000 Array Disk
hdisk11 Available 1n-08-02     IBM MPIO DS5000 Array Disk
hdisk12 Available 1n-08-02     IBM MPIO DS5000 Array Disk
# lspv
hdisk9          0007041a08050518                    None
hdisk10         0007041a080506c5                    None
hdisk11         0007041a0805086f                    None
hdisk12         0007041a080509be                    None
```

- Verify disks on cluster node 2 (AIX_2 in Example 12-19).

Example 12-19 Allocated disks to AIX_2

```
# lsdev -Cc disk
hdisk9  Available 1n-08-02     IBM MPIO DS5000 Array Disk
hdisk10 Available 1n-08-02     IBM MPIO DS5000 Array Disk
hdisk11 Available 1n-08-02     IBM MPIO DS5000 Array Disk
hdisk12 Available 1n-08-02     IBM MPIO DS5000 Array Disk
# lspv
hdisk9          0007041a08050518                    None
hdisk10         0007041a080506c5                    None
hdisk11         0007041a0805086f                    None
hdisk12         0007041a080509be                    None
```

Both examples confirm that AIX Multi Path I/O driver (MPIO) is used and correctly loaded.

12.9.8 Shared LVM component configuration

In this section, we describe how to define the Logical Volume Manager (LVM) components shared by cluster nodes in an PowerHA for AIX cluster environment.

Creating the volume groups, logical volumes, and file systems shared by the nodes in an PowerHA cluster requires that you perform steps on all nodes in the cluster. In general, you define the components on one node (source node) and then import the volume group on the other nodes in the cluster (destination nodes), which ensures that the ODM definitions of the shared components are the same on all nodes in the cluster.

Non-concurrent access environments typically use journaled file systems to manage data, whereas concurrent access environments use raw logical volumes. In this chapter, we provide other instructions for defining shared LVM components in non-concurrent access and concurrent access environments.

Creating a shared volume group

In the following sections we provide information about creating a non-concurrent volume group as well as a volume group for concurrent access.

Creating a non-concurrent volume group

Use the **`smit mkvg`** fast path to create a shared volume group. Use the default field values unless your site has other requirements. See Table 12-4 for the **`smit mkvg`** options.

Table 12-4 smit mkvg options

Options	Description
VOLUME GROUP name	The name of the shared volume group must be unique within the cluster.
Physical partition SIZE in megabytes	Accept the default.
PHYSICAL VOLUME NAMES	Specify the names of the physical volumes you want included in the volume group.
Activate volume group AUTOMATICALLY at system restart?	Set to no so that the volume group can be activated as appropriate by the cluster event scripts.
Volume Group MAJOR NUMBER	If you are not using NFS, you can use the default (which is the next available number in the valid range). If you are using NFS, you must make sure to use the same major number on all nodes. Use the **`lvlstmajor`** command on each node to determine a free major number common to all nodes.
Create VG concurrent capable?	Set this field to no (leave default)
Auto-varyon concurrent mode?	Accept the default.

Creating a volume group for concurrent access

The procedure used to create a concurrent access volume group varies, depending on which type of device you are using. In our case we will assume DS5000 disks.

To use a concurrent access volume group, defined on a DS5000 disk subsystem, you must create it as a concurrent-capable volume group. A concurrent-capable volume group can be activated (varied on) in either non-concurrent mode or concurrent access mode.

To define logical volumes on a concurrent-capable volume group, it must be varied on in non-concurrent mode.

Use **smit mkvg** with the options shown in Table 12-5 to build the volume group.

Table 12-5 Options for volume group

Options	Description
VOLUME GROUP name	The name of the shared volume group must be unique within the cluster.
Physical partition SIZE in megabytes	Accept the default.
PHYSICAL VOLUME NAMES	Specify the names of the physical volumes you want included in the volume group.
Activate volume group AUTOMATICALLY at system restart?	Set this field to no so that the volume group can be activated, as appropriate, by the cluster event scripts.
Volume Group MAJOR NUMBER	While it is only really required when you are using NFS, it is always good practice in an HACMP cluster to have a shared volume group have the same major number on all the nodes that serve it. Use the **lvlstmajor** command on each node to determine a free major number common to all nodes.
Create VG concurrent capable?	Set this field to yes so that the volume group can be activated in concurrent access mode by the HACMP for AIX event scripts
Auto-varyon concurrent mode?	Set this field to no so that the volume group can be activated, as appropriate, by the cluster event scripts.

Creating shared logical volumes and file systems

Use the **smit crjfs** fast path to create the shared file system on the source node. When you create a journaled file system, AIX creates the corresponding logical volume. Therefore, you do not need to define a logical volume. However, you would need to rename both the logical volume and the log logical volume later, for the file system and volume group (Table 12-6).

Table 12-6 Creating journaled file systems from smitty

Options	Description
Mount AUTOMATICALLY at system restart?	Make sure this field is set to no.
Start Disk Accounting	Make sure this field is set to no.

Renaming a jfslog and logical volumes on the source node

AIX assigns a logical volume name to each logical volume it creates. Examples of logical volume names are `/dev/lv00` and `/dev/lv01`. Within an PowerHA cluster, the name of any shared logical volume must be unique. Also, the journaled file system log (`jfslog`) is a logical volume that requires a unique name in the cluster.

To make sure that logical volumes have unique names, rename the logical volume associated with the file system and the corresponding `jfslog` logical volume. Use a naming scheme that indicates the logical volume is associated with a certain file system. For example, **lvsharefs** can name a logical volume for the `/sharefs` file system.

Follow these steps to rename the logical volumes:

1. Use the **lsvg -l volume_group_name** command to determine the name of the logical volume and the log logical volume (`jfslog`) associated with the shared volume groups. In the resulting display, look for the logical volume name that has type *jfs*, which is the logical volume. Then look for the logical volume name that has type `jfslog`, which is the log logical volume.

2. Use the **smit chlv** fast path to rename the logical volume and the log logical volume.
3. After renaming the jfslog or a logical volume, check the /etc/filesystems file to make sure the dev and log attributes reflect the change. Check the log attribute for each file system in the volume group, and make sure that it has the new jfslog name. Check the dev attribute for the logical volume that you renamed, and make sure that it has the new logical volume name.

Importing to other nodes

The following sections cover varying off a volume group on the source node, importing it onto the destination node, changing its startup status, and varying it off on the destination nodes.

Varying off a volume group on the source node

Use the **varyoffvg** command to deactivate the shared volume group. You vary off the volume group so that it can be properly imported onto a destination node and activated as appropriate by the cluster event scripts. Enter the following command:

```
varyoffvg volume_group_name
```

Make sure that all the file systems of the volume group have been unmounted; otherwise, the **varyoffvg** command does not work.

Importing a volume group onto the destination node

To import a volume group onto destination nodes you can use the SMIT interface or the TaskGuide utility. The TaskGuide uses a graphical interface to guide you through the steps of adding nodes to an existing volume group. Importing the volume group onto the destination nodes synchronizes the ODM definition of the volume group on each node on which it is imported.

You can use the **smit importvg** fast path to import the volume group (Table 12-7).

Table 12-7 smit importvg options

Options	Description
VOLUME GROUP name	Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the source node.
PHYSICAL VOLUME name	Enter the name of a physical volume that resides in the volume group. Note that a disk might have another logical name on other nodes. Make sure that you use the disk name as it is defined on the destination node.
Volume Group MAJOR NUMBER	If you are not using NFS, you can use the default (which is the next available number in the valid range). If you are using NFS, you must make sure to use the same major number on all nodes. Use the **lvlstmajor** command on each node to determine a free major number common to all nodes.

Changing a volume group startup status

By default, a volume group that has just been imported is configured to automatically become active at system restart. In an PowerHA for AIX environment, a volume group must be varied on as appropriate by the cluster event scripts. Therefore, after importing a volume group, use the SMIT Change a Volume Group window to reconfigure the volume group so that it is not activated automatically at system restart.

Use the **smit chvg** fast path to change the characteristics of a volume group (Table 12-8).

Table 12-8 Options of "smitty chvg"

Options	Description
Activate volume group automatically at system restart?	Set this field to *no*.
A QUORUM of disks required to keep the volume group online?	If you are using DS5000 with RAID protection, set this field to *no*.

Varying off the volume group on the destination nodes

Use the **varyoffvg** command to deactivate the shared volume group so that it can be imported onto another destination node or activated as appropriate by the cluster event scripts. Enter:

```
# varyoffvg volume_group_name
```

12.9.9 Fast disk takeover

By utilizing the Enhanced Concurrent Mode volume groups, in non-concurrent resource groups, it almost eliminates disk takeover time by removing the need for disk reservations, and breaking these reserves. The volume groups are varied online in *Active* mode on only the owning node, and all other fall over candidate nodes have it varied on in *Passive* mode. RSCT is utilized for communications to coordinate activity between the nodes so that only one node has it varied on actively.

Time for disk takeover now is fairly consistent at around 10 seconds. Now with the multiple resource groups and hundreds to thousands of disks, it might be a little more — but not that significantly more.

Note that whereas the VGs are varied on concurrently to both nodes, **lsvg -o** will only show the VG as active on the node accessing the disk; however, running **lspv** will show that the VG disks are active on both nodes. Also note, that **lsvg vgname** will tell you if the VG is in the active or passive state.

12.9.10 Forced varyon of volume groups

For situations in which you are using LVM mirroring, and want to survive failure of half the disks, there is a new attribute in the Resource Group **smit** panel to force varyon of the VG, provided that a complete copy of the LVs are available. Previously, this was accomplished by setting the HACMP_MIRROR_VARYON environment variable to yes, or by user-written cluster pre/post/recovery event scripts.

The new attribute in the Resource Group Smit panel is used instead:

```
Volume Groups Use forced varyon of volume groups, if necessary [false]
```

To take advantage of this feature, set this attribute to *true*, as the default is false.

12.9.11 Disk heartbeat

Heartbeat provides the ability to use existing shared disks, regardless of disk type, to provide serial network type connectivity, that can replace needs of using integrated serial ports or 8-port asynchronous adapters and the additional cables needed for it.

This feature utilizes a special reservation area, previously used by SSA Concurrent volume groups. Because the Enhanced Concurrent mode does not utilize this space, therefore it makes it available for heartbeat. The advantage of this features is that the disk chosen for the serial heartbeat is automatically part of a data volume group and no further configuration is needed.

The disk heartbeat function was introduced in code version 2.2.1.30 of RSCT. Certain APARs can be used to bring that version up to 2.2.1.31. If you have that level installed, as well as HACMP 5.1 and higher, you can use disk heartbeating. The relevant file to look at for confirmation is `/usr/sbin/rsct/bin/hats_diskhb_nim`.

Starting in HACMP 5.1 with AIX 5.1, the enhanced concurrent mode of volume groups can be used only in concurrent (or “online on all available nodes”) type of resource groups. From AIX 5.2 onwards, disk heartbeats can exist on an enhanced concurrent VG that resides in a non-concurrent resource group.

To use disk heartbeats, none of the cluster nodes can issue a SCSI reservation command for the disk; both nodes must be able to read and write to that disk temporary data for disk heartbeat concurrently. It is accomplished for example by definition of the disk in an enhanced concurrent volume group.

Configuration of disk heartbeat device

This example consists of a two-node cluster (nodes AIX_1 and AIX_2) with shared DS5000 disk devices. If more than two nodes exist in your cluster, you need *n* number of non-IP heartbeat networks, where *n* represents the number of nodes in the cluster (for instance, a three node cluster requires three non-IP heartbeat networks), those create a heartbeat ring (Figure 12-20).

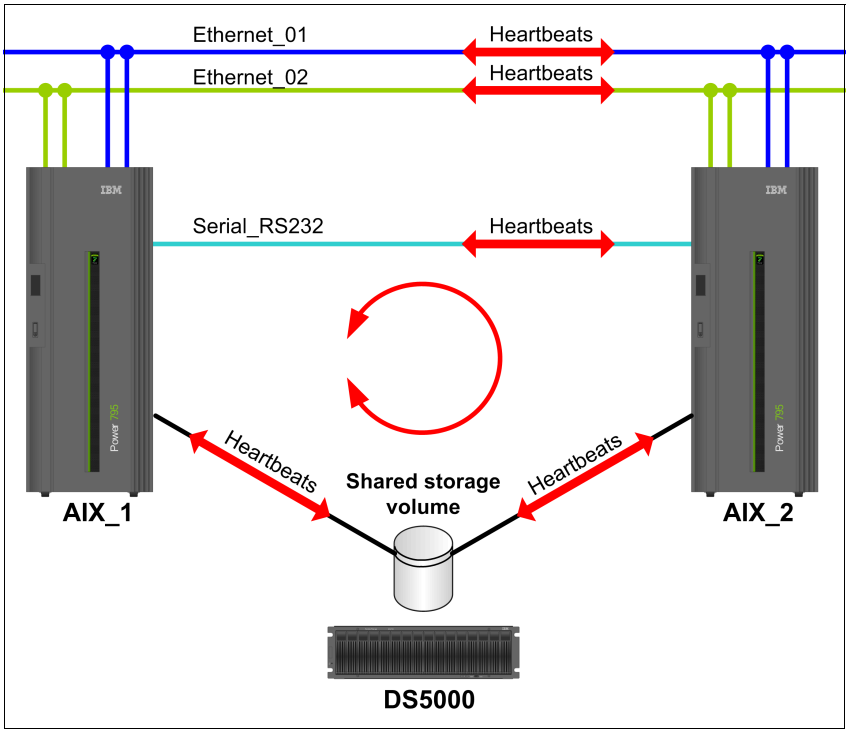


Figure 12-20 Disk heartbeat ring

Prerequisites

We assumed that the shared storage devices are already made available and configured to AIX, and that the proper levels of RSCT or CAA, and HACMP are already installed.

Tip: To enable enhanced-concurrent volume groups, make sure that `bos.clvm.enh` is installed, in addition to the standard PowerHA installation using **`installp`** command.

Configuring disk heartbeat

As mentioned previously, disk heartbeat utilizes enhanced-concurrent volume groups. When starting with a new configuration of disks, you usually create enhanced concurrent volume groups by Cluster Single Point of Control (C-SPOC).

To be able to use C-SPOC successfully, it is required that an IP-based network already exists, and that the storage devices have their PVIDs in both systems' ODMs; that can be verified by running **`lspv`** on each system. If a common PVID does not exist, it is necessary to run **`chdev -l <hdisk#> -a pv=yes`** on both system, to allow C-SPOC match up the devices as known shared storage devices.

In our test scenario, the following **`smitty cl_admin`** menu was used:

HACMP Concurrent Logical Volume Management → **Concurrent Volume Groups** → **Create a Concurrent Volume Group**.

Choose the appropriate nodes, and then choose the appropriate shared storage devices based on pvids (hdisk*n*). Select the name for shared VG, desired PP size, make sure that Enhanced Concurrent Mode is set to true and press Enter, which will create the shared enhanced concurrent VG needed for our disk heartbeat.

Hint: When done, it is good practise to verify by **`lspv`** that heartbeat device and VG are shown correctly with desired attributes.

On cluster node AIX_1:

```
# lspv
   hdisk7 000a7f5af78e0cf4 hacmp_hb_vg
```

On cluster node AIX_2:

```
# lspv
   hdisk3 000a7f5af78e0cf4 hacmp_hb_vg
```

Creating disk heartbeat devices and network

There are two ways to create these items. Because we have already created the enhanced concurrent VG, we can use the discovery method (1) and let PowerHA find it for us. Or we can do it manually by the Pre-defined devices method (2).

Following is an example of each method.

1. Creating by the Discover Method:

   Type **smitty hacmp** and select **Extended Configuration → Discover HACMP-related Information from Configured Nodes**.

   This runs the discovery automatically and creates a `clip_config` file that contains all relevant information as discovered. After being completed, navigate back to the Extended Configuration menu and choose **Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Add Discovered Communication Interface and Devices → Communication Devices**. Choose the appropriate devices (for example, `hdisk3` and `hdisk7`).

   a. Select **Point-to-Point Pair of Discovered Communication Devices to Add**.
   b. Move the cursor to the desired item and press F7. Use arrow keys to scroll.
   c. *One or more* items can be selected.
   d. Press Enter *after* making all selections.

   ```
   # Node         Device   Pvid
   > nodeAIX_1    hdisk7   000a7f5af78
   > nodeAIX_2    hdisk3   000a7f5af78
   ```

2. Creating by the Pre-Defined Devices Method:

   When using this method, it is necessary to create a `diskhb` network first, then assign the disk-node pair devices to the network. Create the `diskhb` network by entering `smitty hacmp` and selecting **Extended Configuration → Extended Topology Configuration → Configure HACMP Networks → Add a Network to the HACMP cluster**. Select **diskhb**.

   Enter the desired network name (for example, `disknet1`) and press Enter. Type **smitty hacmp** and select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Add Pre-Defined Communication Interfaces and Devices**.

   a. Communication Devices → Choose your diskhb Network Name

   b. Add a communication device.

   c. Type or select values in the entry fields.

   d. Press Enter *after* making all desired changes:

   ```
   Device Name                          [AIX_1_hboverdisk]
   Network Type                         diskhb
   Network Name                         disknet1
   Device Path                          [/dev/hdisk7]
   Node Name                            [AIX_1]
   ```

   For Device Name, use a unique name that will show up in your topology.

   For the Device Path, you might want to put in a /dev/<device name>. Then choose the corresponding node (AIX_1) for this device and device name. Then press Enter.

   You need to repeat the same process for the other node (AIX_2) and the other device (hdisk3), which will complete the configuration of both devices for the `diskhb` network.

Testing disk heartbeat connectivity

After the device and network definitions have been created, test the system and make sure communications are working properly. (If the volume group is varied on in normal mode on one of the nodes, the test will probably not work, so make sure it is varied off).

To test the validity of a disk heartbeat connection, use the following command:

```
/usr/sbin/rsct/bin/dhb_read
```

The usage of dhb_read is as follows:

```
dhb_read -p devicename    //dump diskhb sector contents
dhb_read -p devicename -r //receive data over diskhb network
dhb_read -p devicename -t //transmit data over diskhb network
```

To verify that `disknet1` can communicate from AIX_2 to node AIX_1 and that it is able to send heartbeat packets, you can run the following commands:

- On node AIX_1, type:

  ```
  dhb_read -p hdisk7-r
  ```

- On node AIX_2, execute:

  ```
  dhb_read -p hdisk3 -t
  ```

If the heartbeat link from AIX_2 to AIX_1 is operational, the both nodes display:

```
Link operating normally.
```

You can run this again and swap transmitters and receivers of heartbeat packets. To make the network active, it is necessary to sync up the cluster. Because the volume group has not been added to the resource group yet, we will sync up once instead of twice.

Adding shared disk as a shared resource

In most cases you have your `diskhb` device on a shared data volume group. It is necessary to add that VG into your resource group and synchronize the cluster.

1. Use the command **`smitty hacmp`** and select **Extended Configuration → Extended Resource Configuration → Extended Resource Group Configuration → Change/Show Resources and Attributes for a Resource Group**.
2. Press Enter and choose the appropriate resource group.
3. Enter the new VG (`enhconcvg`) into the volume group list and press Enter.
4. Return to the top of the Extended Configuration menu and synchronize the cluster.

Monitoring disk heartbeat

While the cluster is up and running, you can monitor the activity of the disk heartbeats using the command **`lssrc -ls topsvcs`**, which produces output as shown in Example 12-20:

Example 12-20 Disk heartbeat verification

```
Subsystem         Group            PID     Status
 topsvcs          topsvcs          201184  active

Network Name   Indx Defd  Mbrs  St   Adapter ID      Group ID
diskhb_1       [ 3] 2     2     S    255.255.10.1    255.255.10.3
diskhb_1       [ 3] rhdisk110        0x867c98a8      0x867c9909
HB Interval = 2.000 secs. Sensitivity = 4 missed beats
Missed HBs: Total: 204 Current group: 204
Packets sent    : 1159137 ICMP 0 Errors: 0 No mbuf: 0
Packets received: 1221456 ICMP 0 Dropped: 0
NIM's PID: 360646
```

Be aware that there is a grace period for heartbeats to start processing, which is normally around 60 seconds. So if you run this command immediately after starting the cluster, you might not see any results until the heartbeat processing grace period has elapsed.

Performance concerns with disk heartbeat

The most modern disks take somewhere around 15 milliseconds to service an I/O request, which means that they cannot do much more than 60 seeks per second. The sectors used for disk heartbeating are part of the VGDA, which is at the outer edge of the disk, and might not be near the application data.

This means that every time a disk heartbeat is triggered, a seeking sequence has to complete. Disk heartbeating typically (with the default parameters) requires four seeks per second. The **filemon** tool can be used to monitor the seek activity on each heartbeat disk.

In case when the disk, that already indicates a high seek activities, must be used for heartbeating, it might be necessary to change the heartbeat timing parameters to prevent long write application delays from being seen as a failure.

12.9.12 More information

The comprehensive information about cluster configuration using PowerHA on IBM POWER systems, including introduction, planing, and implementation, can be found in *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845, and *Exploiting IBM PowerHA SystemMirror Enterprise Edition*, SG24-7841.

Additionally, the online resources are accessible under:

http://www.ibm.com/systems/power/software/availability/

A

GPFS

In this Appendix, we look at the basic concepts behind the GPFS file system and show how the DS5000 Storage Server can be configured in this environment.

GPFS concepts

The IBM General Parallel File System (GPFS) is a powerful file system that provides the following capabilities:

- Global namespace
- Shared file system access among GPFS clusters
- Simultaneous file access from multiple nodes
- High recoverability and data availability due to replication
- Ability to make certain changes while a file system is mounted
- Simplified administration that is similar to existing UNIX systems

GPFS provides file system services to parallel and serial applications. It allows parallel applications simultaneous access to the same files, or other files, from any node that has the GPFS file system mounted, as well as managing a high level of control over all file system operations.

GPFS is particularly appropriate in an environment where the aggregate peak need for data bandwidth exceeds the capability of a distributed file system server.

GPFS allows users shared file access within a single GPFS cluster and across multiple GPFS clusters. Shared file system access among GPFS clusters GPFS allows users shared access to files in either the cluster where the file system was created or other GPFS clusters. Each site in the network is managed as a separate cluster, while allowing shared file system access.

A GPFS cluster can consist of the following types of nodes:

- Linux nodes
- AIX nodes
- Windows nodes
- combination of Linux, AIX and Windows nodes

GPFS includes a Network Shared Disk component which provides a method for cluster-wide disk naming and access. All disks used by GPFS must first be given a globally-accessible NSD name.

Performance advantages with GPFS file system

Using GPFS to store and retrieve your files can improve system performance as follows:

- Allowing multiple processes or applications on all nodes in the cluster simultaneous access to the same file using standard file system calls.
- Increasing aggregate bandwidth of your file system by spreading reads and writes across multiple disks.
- Balancing the load evenly across all disks to maximize their combined throughput. One disk is no more active than another.
- Supporting very large file and file system sizes.
- Allowing concurrent reads and writes from multiple nodes.
- Allowing for distributed token (lock) management. Distributing token management reduces system delays associated with a lockable object waiting to obtaining a token.
- Allowing for the specification of other networks for GPFS daemon communication and for GPFS administration command usage within your cluster.

Achieving high throughput to a single, large file requires striping data across multiple disks and multiple disk controllers. Rather than relying on striping in a separate volume manager layer, GPFS implements striping in the file system. Managing its own striping affords GPFS the control it needs to achieve fault tolerance and to balance load across adapters, storage controllers, and disks. Large files in GPFS are divided into equal sized blocks, and consecutive blocks are placed on various disks in a round-robin fashion.

To exploit disk parallelism when reading a large file from a single-threaded application, whenever it can recognize a pattern, GPFS prefetches data into its buffer pool, issuing I/O requests in parallel to as many disks as necessary to achieve the bandwidth of which the switching fabric is capable. GPFS recognizes sequential, reverse sequential, and various forms of striped access patterns.

Data availability advantages with GPFS

GPFS failover support allows you to organize your hardware into failure groups. A failure group is a set of disks that share a common point of failure that might cause them all to become simultaneously unavailable. When used in conjunction with the replication feature of GPFS, the creation of multiple failure groups provides for increased file availability if a group of disks fails. GPFS maintains each instance of replicated data and metadata on disks in various failure groups. If a set of disks becomes unavailable, GPFS fails over to the replicated copies in another failure group.

Disks can be added and deleted while the file system is mounted. Equally, nodes can be added or deleted without having to stop and restart the GPFS daemon on all nodes.

GPFS configuration

GPFS can be configured in a variety of ways. Here we just explore a configuration where all NSDs are SAN-attached to all nodes in the cluster and the nodes in the cluster are either all Linux (Figure A-1) or all AIX. For documentation listing the latest hardware that GPFS has been tested with, see the following website:

http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfs_faqs/gpfsclustersfaq.html

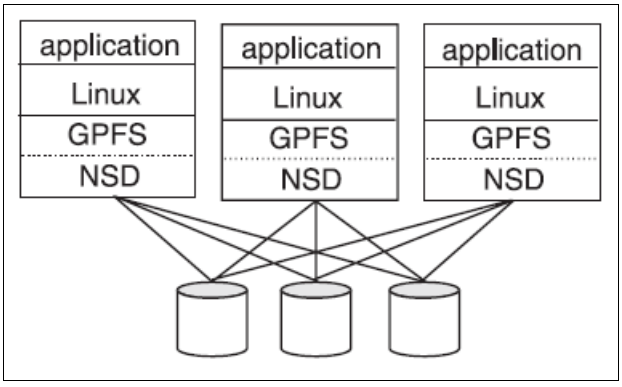


Figure A-1 Linux only GPFS cluster with SAN attached disks

DS5000 configuration limitations with GPFS

The following limitations apply to PSSP and GPFS configurations:

- Direct connection is not allowed between the host node and a DS5000 storage subsystem.
- Only switched fabric connection is allowed.
- RVSD clusters can support up to two IBM Virtual Shared Disk and RVSD servers for each DS5000 partition.
- Single node quorum is not supported in a dual-node GPFS cluster with DS5000 disks in the configuration.
- Heterogeneous configurations are not supported.

DS5000 settings for GPFS environment

In GPFS file systems, the following DS5000 cache settings are supported:

- Read cache enabled or disabled
- Write cache enabled or disabled
- Cache mirroring enabled or disabled (depending upon the write cache mirroring setting)

The performance benefits of read or write caching depend on the application. Because the cache settings can be easily adjusted from Storage Manager (SM), preferably carry out your performance tests during implementation. Here is a sample configuration you can use as a starting point for NSD servers or other GPFS nodes that are directly attached to a SAN over a Fibre Channel network:

1. Use the following cache settings:

   Read cache = enabled
   Read ahead multiplier = 0
   Write cache = disabled
   Write cache mirroring = disabled
   Write block size = 16 K

2. When the storage server disks are configured for RAID5, certain configuration settings can affect GPFS performance:

   – GPFS block size
   – Maximum I/O size of host Fibre Channel (FC) host bus adapter (HBA) device driver
   – Storage server RAID5 stripe size

   For optimal performance, GPFS block size must be a multiple of the maximum I/O size of the FC HBA device driver. In addition, the maximum I/O size of the FC HBA device driver must be a multiple of the RAID5 stripe size.

3. The following guidelines can help avoid the performance penalty of read-modify-write at the storage server for GPFS writes. Here are a few examples of the settings:

   – 8+P RAID5:

     • GPFS block size = 512K
     • Storage Server RAID5 segment size = 64 K (RAID5 stripe size=512 K)
     • Maximum IO size of FC HBA device driver = 512K

   – 4+P RAID5:

     • GPFS block size = 256 K
     • Storage Server RAID5 segment size = 64 K (RAID5 stripe size = 256 K)
     • Maximum IO size of FC HBA device driver = 256 K

For the example settings using 8+P and 4+P RAID5, the RAID5 parity can be calculated from the data written and will avoid reading from disk to calculate the RAID5 parity. The maximum IO size of the FC HBA device driver can be verified using **iostat** or the Storage Server performance monitor. In certain cases, the device driver might need to be patched to increase the default maximum IO size.

4. The GPFS parameter maxMBpS can limit the maximum throughput of an NSD server or a single GPFS node that is directly attached to the SAN with a FC HBA. Increase the maxMBpS from the default value of 150 to 200 (200 MBps). The maxMBpS parameter is changed by issuing the mmchconfig command. After this change is made, restart GPFS on the nodes and test both read and write performance of both a single node in addition to a large number of nodes.
5. The layout and distribution of the disks across the drive-side channels on the DS5000 Storage Server can also be an important factor when attempting to maximize throughput and IOPs.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM System Storage DS5000 Series Implementation and Best Practices Guide*, SG24-8024
- *IBM System Storage DS Storage Manager Copy Services Guide*, SG24-7822
- *IBM Midrange System Storage Hardware Guide*, SG24-7676
- *IBM Midrange System Storage Implementation and Best Practices Guide*, SG24-6363
- *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation*, SG24-7544
- *IBM System Storage Copy Services and IBM i: A Guide to Planning and Implementation*, SG24-7103
- *IBM System Storage DS3000: Introduction and Implementation Guide*, SG24-7065
- *IBM System Storage DS4000 and Storage Manager V10.30*, SG24-7010
- *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116
- *Implementing an IBM/Cisco SAN*, SG24-7545
- *IBM System Storage DS3500 Introduction and Implementation Guide*, SG24-7914
- *VMware Implementation with IBM System Storage DS5000*, REDP-4609
- *IBM System Storage EXP5060 Storage Expansion Enclosure Planning Guide*, REDP-4679
- *SAN Boot Implementation and Best Practices Guide for IBM System Storage*, SG24-7958

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *IBM System Storage DS3000, DS4000, and DS5000 Command Line Interface and Script Commands Programming Guide,* GC52-1275
- *IBM System Storage DS4000 Concepts Guide,* GC26-7734
- *IBM System Storage DS4000/DS5000 EXP810 Storage Expansion Enclosure Installation, User's and Maintenance Guide,* GC26-7798
- *IBM System Storage DS4000/DS5000 Fibre Channel and Serial ATA Intermix Premium Feature Installation Overview,* GC53-1137
- *IBM System Storage DS4000/DS5000 Hard Drive and Storage Expansion Enclosure Installation and Migration Guide,* GC53-1139
- *IBM System Storage DS4800 Storage Subsystem Installation, User's, and Maintenance Guide,* GC26-7845
- *IBM System Storage DS4800 Storage Subsystem Quick Start Guide,* GC27-2148
- *IBM System Storage DS5000 EXP5000 Storage Expansion Enclosure Installation, User's, and Maintenance Guide,* GC53-1141
- *IBM System Storage DS5100, DS5300, and EXP5000 Quick Start Guide,* GC53-1134
- *IBM System Storage DS5100 and DS5300 Storage Subsystems Installation, User's, and Maintenance Guide,* GC53-1140
- *IBM System Storage DS Storage Manager Version 10 Installation and Host Support Guide,* GC53-1135
- *IBM System Storage DS Storage Manager Version 10.50 Copy Services User's Guide,* GC53-1136
- *IBM System Storage DS Storage Manager Version 10.60 Copy Services User's Guide,* GC53-1136

Online resources

These websites are also relevant as further information sources:

- System Storage Interoperation Center (SSIC): `http://www-03.ibm.com/systems/support/storage/config/ssic/displayesssearchwithoutjs.wss?start_over=yes`
- Support for IBM Disk systems: `http://www.ibm.com/systems/support/storage/disk`

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM System Storage DS5000 Series Implementation and Best Practices Guide

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

IBM System Storage DS5000 Series Implementation and Best Practices Guide

Host configuration and performance tuning tips

Performance measurement using TPC for Disk

Storage virtualization with IBM SVC and Storwize v7000

This IBM Redbooks publication represents a compilation of best practices for deploying and configuring the IBM System Storage DS5000 Series family of products. This book is intended for IBM technical professionals, Business Partners, and customers responsible for the planning, deployment, and maintenance of the IBM System Storage DS5000 Series family of products. We realize that setting up DS5000 Storage Servers can be a complex task. There is no single configuration that will be satisfactory for every application or situation.

First, we provide a conceptual framework for understanding the hardware in a Storage Area Network. Then, we offer our guidelines, hints, and tips for the physical installation, cabling, and zoning, using the Storage Manager setup tasks. Next, we provide a quick guide to help you install and configure the DS5000 using best practices.

After that, we turn our attention to the performance and tuning of various components and features, including numerous guidelines. We look at performance implications for various application products such as IBM DB2, Oracle, IBM Tivoli Storage Manager, Microsoft SQL server, and in particular, Microsoft Exchange server.

Then we review the various tools available to simulate workloads and to measure, collect, and analyze performance data. We also consider the IBM AIX environment, including IBM High Availability Cluster Multiprocessing (HACMP) and IBM General Parallel File System (GPFS). This edition of the book also includes guidelines for managing and using the DS5000 with the IBM System Storage SAN Volume Controller (SVC) and IBM Storwize V7000.

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

SG24-8024-00 ISBN 0738437476