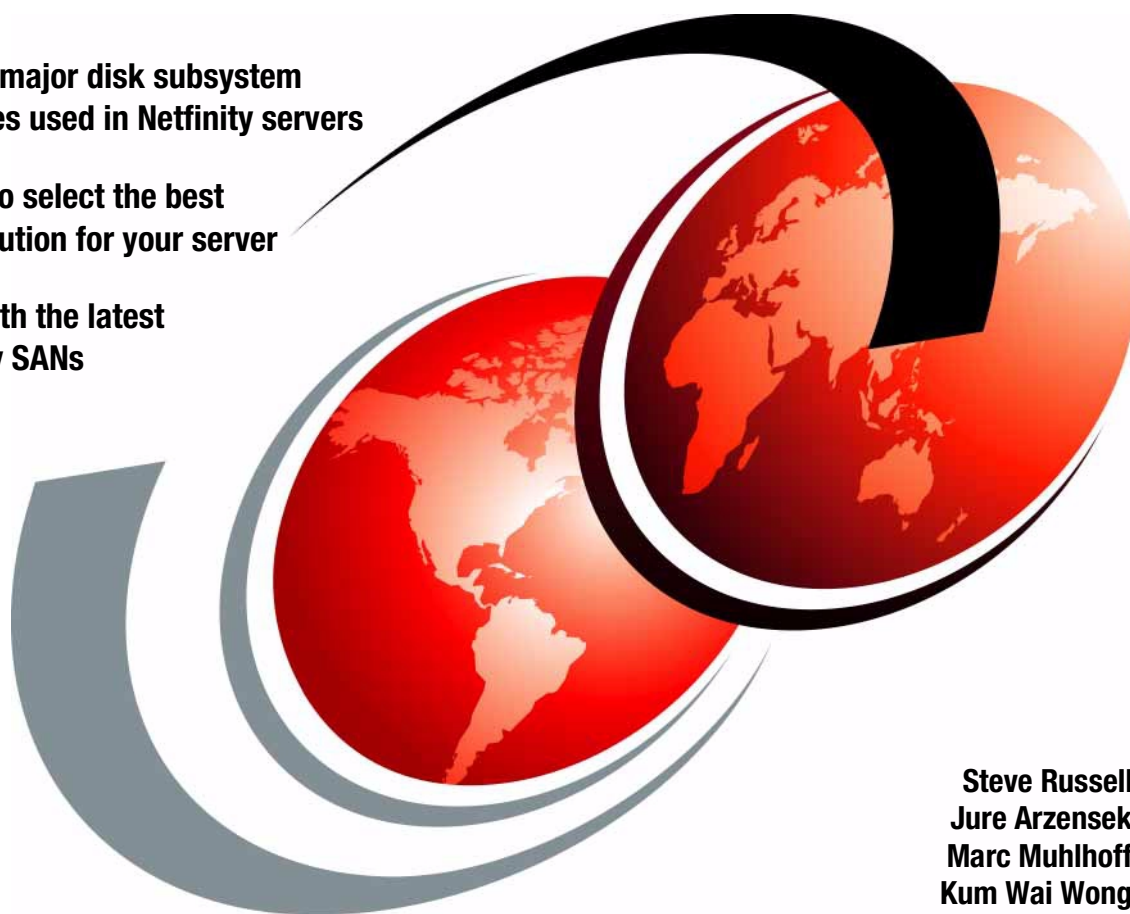


# Netfinity Server Disk Subsystems

Covers the major disk subsystem technologies used in Netfinity servers

Helps you to select the best storage solution for your server

Updated with the latest on Netfinity SANs



Steve Russell  
Jure Arzensek  
Marc Muhlhoff  
Kum Wai Wong

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

**Redbooks**





International Technical Support Organization

**Netfinity Server Disk Subsystems**

June 2000

**Take Note!**

Before using this information and the product it supports, be sure to read the general information in Appendix C, "Special notices" on page 329.

**Fourth Edition (June 2000)**

This edition applies to the following Netfinity storage products:

- The ServeRAID family of SCSI-based RAID controllers
- The Netfinity Fibre Channel family of products
- The SSA family of adapters

Comments may be addressed to:

IBM Corporation, International Technical Support Organization  
Dept. HZ8 Building 678  
P.O. Box 12195  
Research Triangle Park, NC 27709-2195

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

**© Copyright International Business Machines Corporation 1997 2000. All rights reserved.**

Note to U.S Government Users – Documentation related to restricted rights – Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

---

# Contents

<b>Preface</b> .....	ix
The team that wrote this redbook .....	x
Comments welcome .....	xii
<hr/>	
<b>Part 1. Selecting a disk subsystem</b> .....	1
<b>Chapter 1. Introduction</b> .....	3
1.1 How to use this book .....	3
1.2 Audience .....	4
<b>Chapter 2. Netfinity storage technology</b> .....	7
2.1 Evolution of storage technology .....	7
2.1.1 Enterprise data storage .....	7
2.1.2 Intel processor-based systems .....	8
2.1.3 What drives this evolution? .....	9
2.2 What technologies are available today? .....	10
2.2.1 Small computer system interface (SCSI) .....	11
2.2.2 Fibre Channel .....	11
2.2.3 Serial storage architecture (SSA) .....	12
2.3 Comparing storage solutions .....	12
2.3.1 Distance .....	13
2.3.2 Performance .....	14
2.3.3 Fault tolerance .....	14
2.3.4 Availability .....	15
2.3.5 Scalability .....	16
2.4 Mapping your requirements to technology .....	16
2.4.1 SCSI disk subsystems .....	17
2.4.2 Fibre Channel disk subsystems .....	17
2.4.3 SSA disk subsystems .....	18
<b>Chapter 3. Sample disk configurations</b> .....	19
3.1 Netfinity SCSI disk subsystems .....	19
3.1.1 Standard SCSI configurations .....	19
3.1.2 Fault-tolerant configurations .....	21
3.1.3 A high-capacity SCSI subsystem .....	22
3.2 Typical Netfinity Fibre Channel disk subsystems .....	25
3.2.1 Basic configuration .....	25
3.2.2 Multiple hosts .....	26
3.2.3 Two-node cluster configurations .....	28
3.2.4 A high-capacity Fibre Channel subsystem .....	29
3.3 Netfinity disk and tape pooling .....	32

3.4	Netfinity remote backup, archiving and recovery . . . . .	35
3.5	Netfinity high-availability solutions using Microsoft Cluster Server . . .	37

---

**Part 2. ServeRAID SCSI subsystems . . . . . 43**

<b>Chapter 4. Introduction to ServeRAID . . . . .</b>	<b>45</b>
4.1 Netfinity ServeRAID hardware . . . . .	45
4.2 ServeRAID features and options . . . . .	46
4.2.1 Arrays and logical drives . . . . .	46
4.2.2 RAID levels supported by ServeRAID adapters . . . . .	48
4.2.3 ServeRAID-4H adapter and Ultra3 160/m SCSI . . . . .	58
4.2.4 ServeRAID-4M adapter . . . . .	60
4.2.5 ServeRAID-4L adapter . . . . .	60
4.2.6 ServeRAID-3HB adapter. . . . .	60
4.2.7 ServeRAID-3L adapter . . . . .	61
4.2.8 Older ServeRAID adapters . . . . .	62
4.2.9 LVDS SCSI connectivity . . . . .	63
4.2.10 “Optimal” SCSI speed. . . . .	64
4.2.11 64-bit PCI data path . . . . .	64
4.2.12 ServeRAID adapter cache . . . . .	64
4.2.13 I <sub>2</sub> O enabled . . . . .	66
4.2.14 Active PCI support . . . . .	67
4.2.15 Fault-tolerant adapter pair . . . . .	67
4.2.16 Hot-swap rebuild. . . . .	67
4.2.17 Data scrubbing . . . . .	68
4.2.18 Autosync . . . . .	69
4.2.19 Configuration data stored in multiple locations . . . . .	69
4.2.20 ServeRAID utilities . . . . .	70
4.2.21 Logical drive migration . . . . .	72
4.2.22 Command-line utilities . . . . .	72
4.2.23 FlashCopy . . . . .	73
4.2.24 BIOS and firmware . . . . .	75
4.2.25 Feature comparison . . . . .	76
4.3 External storage enclosures . . . . .	78
4.3.1 Netfinity EXP200 Storage Expansion Unit. . . . .	78
4.3.2 Netfinity EXP300 Storage Expansion Unit. . . . .	79
<b>Chapter 5. Implementing ServeRAID subsystems . . . . .</b>	<b>81</b>
5.1 Utilities . . . . .	81
5.2 Configuration utilities . . . . .	82
5.2.1 ServeRAID Configuration Program . . . . .	82
5.2.2 ServeRAID Mini-Configuration Utility . . . . .	89
5.3 ServeRAID Manager . . . . .	92

5.3.1	Creating arrays and logical drives . . . . .	93
5.3.2	Logical drive migration (LDM) . . . . .	94
5.3.3	Recovering from physical disk drive failure . . . . .	99
5.3.4	Remote system management . . . . .	104
5.4	Active PCI support . . . . .	106
5.4.1	Active PCI software and hardware components . . . . .	107
5.4.2	The hot-plug tools . . . . .	108
5.5	Configuring a fault-tolerant pair . . . . .	114
5.5.1	Failover . . . . .	115
5.5.2	Guidelines and restrictions . . . . .	115
5.5.3	Installation . . . . .	116
5.5.4	Working with the fault-tolerant pair . . . . .	121
5.6	Clustering with ServeRAID . . . . .	124
5.7	Boot-time messages . . . . .	127
5.8	ServeRAID command-line utilities . . . . .	130
5.8.1	IPSSSEND subcommands . . . . .	131
5.9	Managing the subsystem: Netfinity Director and Netfinity Manager . . . . .	132
5.9.1	Netfinity Director integration . . . . .	132
5.9.2	Netfinity Manager integration . . . . .	137
5.10	Performance considerations . . . . .	145
5.10.1	Factors affecting ServeRAID performance . . . . .	146
5.10.2	RAID subsystem planning . . . . .	146
5.10.3	Number of drives . . . . .	148
5.10.4	Drive performance . . . . .	150
5.10.5	Logical drive configuration . . . . .	152
5.10.6	Stripe unit size . . . . .	153
5.10.7	SCSI bus organization . . . . .	156
5.10.8	Write-back cache operation . . . . .	157
5.10.9	Write-back versus write-through cache . . . . .	158
5.10.10	RAID adapter cache size . . . . .	159
5.10.11	Device drivers . . . . .	161
5.10.12	Firmware . . . . .	162
5.10.13	SCSI bus transfer rate . . . . .	164

---

**Part 3. Fibre Channel subsystems . . . . . 165**

<b>Chapter 6. Introduction to Fibre Channel . . . . .</b>	<b>167</b>
6.1 IBM's implementation of Fibre Channel . . . . .	168
6.1.1 Fibre Channel topology . . . . .	168
6.2 Netfinity Fibre Channel hardware . . . . .	171
6.2.1 Fibre Channel cabling and components . . . . .	172
6.2.2 Fibre Channel PCI adapters . . . . .	175
6.2.3 The Fibre Channel Controller Unit 3526 . . . . .	176

6.2.4	The Netfinity FAST500 RAID Controller . . . . .	177
6.2.5	Fibre Channel controller RAID levels . . . . .	179
6.2.6	The IBM Netfinity FAST200 RAID/Storage Unit . . . . .	183
6.2.7	The IBM Netfinity EXP200 Storage Expansion Unit . . . . .	185
6.2.8	The IBM Netfinity EXP300 Storage Expansion Unit . . . . .	185
6.2.9	The IBM FAST EXP500 Storage Expansion Unit . . . . .	185
6.2.10	Cabling requirements for the EXP500 . . . . .	187
6.2.11	The IBM SAN Fibre Channel Switches . . . . .	193
6.2.12	IBM SAN Fibre Channel Managed Hub . . . . .	196
6.2.13	The IBM SAN Data Gateway for SCSI . . . . .	196
<b>Chapter 7. Implementing Fibre Channel disk subsystems . . . . .</b>		<b>199</b>
7.1	Terminology and definitions . . . . .	199
7.2	Introduction to Netfinity Fibre Channel Storage Manager 7 . . . . .	200
7.3	Migrating from SYMlicity Storage Manager . . . . .	203
7.4	Controller management . . . . .	204
7.4.1	Direct management . . . . .	204
7.4.2	Host-agent management . . . . .	206
7.5	Managing the controllers with Netfinity Storage Manager . . . . .	208
7.5.1	Downloading firmware and NVSRAM contents . . . . .	208
7.5.2	Changing NVSRAM settings . . . . .	209
7.6	Managing storage . . . . .	210
7.6.1	Advanced storage administration tasks . . . . .	211
7.6.2	Storage partitioning . . . . .	219

---

**Part 4. SSA subsystems . . . . . 225**

<b>Chapter 8. Introduction to serial storage architecture (SSA) . . . . .</b>		<b>227</b>
8.1	The SSA protocol: IBM's implementation . . . . .	227
8.1.1	SSA topology . . . . .	228
8.1.2	The SSA frame . . . . .	231
8.1.3	SSA performance . . . . .	232
8.2	Netfinity SSA hardware . . . . .	234
8.2.1	SSA cabling . . . . .	234
8.2.2	How to identify SSA adapters and their features . . . . .	238
8.2.3	The IBM Advanced SerialRAID/X Adapter . . . . .	242
8.2.4	The write cache . . . . .	244
8.2.5	The 7133 disk storage enclosures . . . . .	245
8.2.6	The IBM SAN Data Gateway for SSA . . . . .	248
<b>Chapter 9. Implementing SSA disk subsystems . . . . .</b>		<b>251</b>
9.1	Configuration of bypass cards . . . . .	251
9.1.1	Bypass card modes . . . . .	252



9.2	Building SSA loops . . . . .	254
9.2.1	Single-host configurations . . . . .	255
9.2.2	Dual-host configurations . . . . .	257
9.2.3	Cabling summary . . . . .	259
9.3	Disaster recovery configuration . . . . .	260
9.3.1	RAID-10 and RAID-1 explained . . . . .	260
9.3.2	Split-site operation with RAID-10 and RAID-1 . . . . .	261
9.3.3	Operation of hot-spares . . . . .	264
9.3.4	Performance . . . . .	265
9.3.5	Disk placement . . . . .	266
9.3.6	Summary . . . . .	269
9.4	Managing the advanced SerialRAID/X adapter . . . . .	269
9.4.1	Operating system-specific tools . . . . .	270
9.4.2	Remote System Management . . . . .	272
9.4.3	Managing disk resources with RSM . . . . .	274
<hr/>		
<b>Part 5.</b>	<b>Storage area networks (SANs) . . . . .</b>	<b>277</b>
	<b>Chapter 10. Introduction to SAN . . . . .</b>	<b>279</b>
10.1	Traditional storage architectures . . . . .	280
10.1.1	Problems arising with server-attached storage . . . . .	281
10.2	What is a storage area network? . . . . .	283
10.3	SAN challenges . . . . .	289
10.4	SAN benefits . . . . .	290
10.5	SAN evolution . . . . .	291
	<b>Chapter 11. SANs and Netfinity servers . . . . .</b>	<b>293</b>
11.1	Netfinity server SAN challenges . . . . .	293
11.2	Netfinity SAN components . . . . .	294
11.2.1	SAN software for Netfinity servers . . . . .	295
11.2.2	Netfinity servers . . . . .	301
11.2.3	Netfinity SAN interconnects . . . . .	303
11.2.4	Netfinity SAN storage . . . . .	305
11.3	Netfinity SAN solutions . . . . .	305
11.3.1	Storage consolidation . . . . .	305
11.3.2	Netfinity server consolidation . . . . .	307
11.4	Netfinity SAN at present and beyond . . . . .	310
<hr/>		
<b>Part 6.</b>	<b>Appendixes . . . . .</b>	<b>313</b>
	<b>Appendix A. Network operating system support . . . . .</b>	<b>315</b>
A.1	IBM ServeRAID adapters . . . . .	315
A.2	Netfinity Fibre Channel and FASTT RAID Controllers . . . . .	316

A.3 IBM Advanced SerialRAID/X Adapter . . . . .	316
<b>Appendix B. Troubleshooting</b> . . . . .	<b>317</b>
B.1 Troubleshooting ServeRAID solutions . . . . .	317
B.1.1 Windows NT Server . . . . .	317
B.1.2 Novell NetWare . . . . .	318
B.1.3 Linux . . . . .	318
B.1.4 SCO UnixWare . . . . .	319
B.1.5 OS/2 . . . . .	320
B.2 Troubleshooting Netfinity Fibre Channel solutions . . . . .	320
B.2.1 Replacing controllers in a controller unit. . . . .	321
B.2.2 The Major Event Log (MEL) . . . . .	321
B.2.3 Power on/off sequence . . . . .	322
B.3 Troubleshooting SSA disk subsystems. . . . .	323
B.3.1 Using the SSA service aids . . . . .	323
B.3.2 Service Request Number list . . . . .	325
B.3.3 RSM Proxy configuration . . . . .	325
B.3.4 Service port. . . . .	325
B.3.5 7133 bypass card settings . . . . .	326
B.3.6 Documentation . . . . .	326
B.4 e-Gatherer. . . . .	327
<b>Appendix C. Special notices</b> . . . . .	<b>329</b>
<b>Appendix D. Related publications</b> . . . . .	<b>333</b>
D.1 IBM Redbooks . . . . .	333
D.2 IBM Redbooks collections . . . . .	333
D.3 Other resources . . . . .	333
D.4 Referenced Web sites . . . . .	334
<b>How to get IBM Redbooks</b> . . . . .	<b>335</b>
IBM Redbooks fax order form . . . . .	336
<b>Abbreviations and acronyms</b> . . . . .	<b>337</b>
<b>Index</b> . . . . .	<b>339</b>
<b>IBM Redbooks review</b> . . . . .	<b>347</b>

---

## Preface

This IBM Redbook is the definitive guide to IBM Netfinity disk subsystems. Featuring current hardware and software for ServeRAID, Fibre Channel, and SSA subsystems, it is aimed at technical staff within IBM, customers, and business partners who wish to understand the range of available storage options for IBM's Netfinity family of servers. It will provide you with sufficient information to enable you to make informed decisions when selecting disk subsystems for Netfinity servers, and will also prove invaluable to anyone involved in the purchase, support, sale, and use of these leading-edge storage solutions.

We start by providing an overview of the three disk technologies used by Netfinity servers. The subsystems covered are the ServeRAID family of SCSI-based RAID controllers, Netfinity Fibre Channel products, and the SSA products supported by Netfinity servers. All currently available products are discussed. Guidance in selecting appropriate storage subsystems for your applications is also given in the opening section.

The three subsequent parts are devoted to each of the three technologies in turn. Each part describes the available hardware for the technology under discussion. Implementation details for that particular disk subsystem are then explained. This structure allows readers who wish to implement a specific solution quickly to locate the information they need. Anyone requiring a broader view of Netfinity disk technology may also choose to explore the other sections.

With the rapid increase in interest in storage area network (SAN) approaches to data storage, and a growing number of Netfinity SAN products, we have included a new section in this edition to address this topic. After an introduction to SAN concepts, we discuss current Netfinity SAN products and solutions.

---

## The team that wrote this redbook

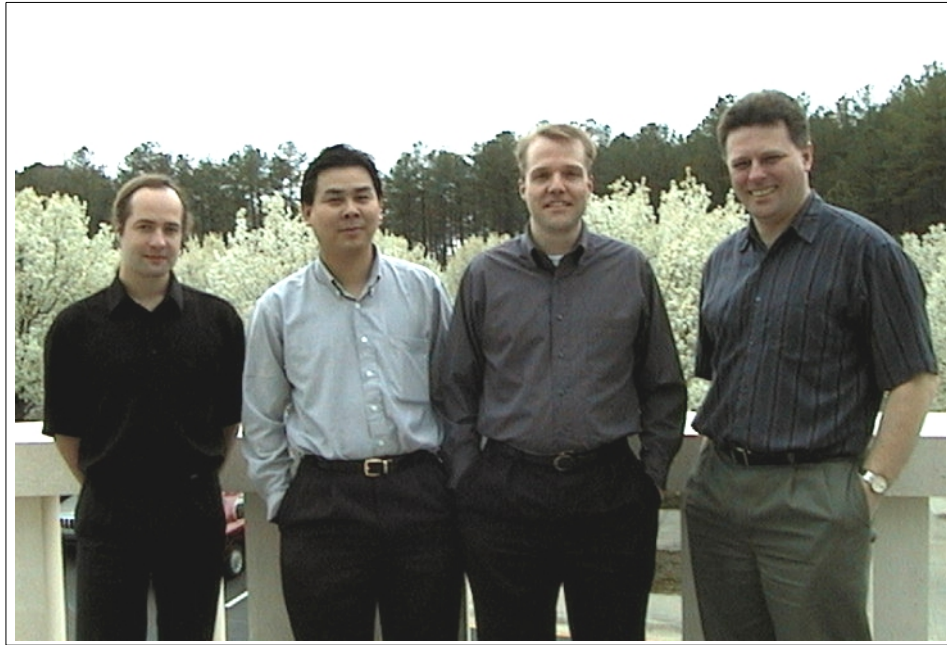
This redbook was produced by a team of specialists from around the world, working at the International Technical Support Organization, Raleigh Center.

**Steve Russell** is a Certified Senior IT Specialist at the International Technical Support Organization, Raleigh Center. Before joining the ITSO in January 1999, Steve had a Technical Marketing role, working in the UK as a member of IBM's Netfinity organization in EMEA. Prior to that, he spent nearly 15 years managing and developing PC-based hardware and software projects at IBM's Hursley laboratory in the UK. He holds a degree in Electrical and Electronic Engineering, and is a member of the Institution of Electrical Engineers and a Chartered Engineer.

**Jure Arzensek** is a PC Institute instructor in Slovenia and has been teaching PC Institute courses in various EMEA countries. He has four years of experience in Netfinity and PC Servers. He holds a degree in Computer Science from the University of Ljubljana. His areas of expertise include Windows NT, Microsoft Cluster Server, and Novell NetWare. He has written extensively on ServeRAID solutions.

**Marc Muhlhoff** is an IBM Netfinity Level 2 Engineer in Greenock, United Kingdom. He has four years of experience working in the technical support of Netfinity servers and operating systems. He specializes in Novell NetWare (CNE 5) and Red Hat Linux (RHCE) and has extensive knowledge in Windows NT support, especially MSCS. His areas of expertise in hardware include IBM ServeRAID, Fibre Channel and SSA solutions on Netfinity. He holds a degree in Physics from the University of Kaiserslautern, Germany.

**Kum Wai Wong** is a Senior MIS Manager in Ernst & Young, Malaysia. Prior to that, he was the Consulting Manager for Technology Enablement in Management Consultancy Services. He has ten years of experience in managing ERP and business intelligence, technology infrastructure and services, strategic IT planning, and project management. His areas of expertise include developing high availability and high performance networking and computing services. He holds a Bachelor of Science degree in Mathematics and Computer Science.



*The team: (left-to-right) Jure, Wong, Marc, and Steve*

This is the fourth edition of this redbook. Our thanks to the authors of earlier editions:

David Watts	IBM ITSO, Raleigh
Andreas Groth	IBM Netfinity Level 2 Support, Greenock
Agustino Kurniawan	IBM Technical Support, Indonesia
Gerard Stolvoort	IBM IT Specialist, Netherlands
Joe Laverty	IBM Greenock
Trevor Simchowitz	IBM South Africa

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

Farrel Benton	IBM Server Product Engineering, Raleigh
Julianne Bielski	IBM Netfinity Software Development, Raleigh
Steve Britner	IBM PC Institute, Raleigh
Dave Gover	IBM SSA Development, Hursley
Michael Halisch	IBM EMEA Help Centre, Greenock
Jeffrey Macfarland	IBM Server Product Engineering, Raleigh
Fabiano Matassa	IBM EMEA Netfinity Pre-sales Team
Gregg McKnight	IBM Netfinity Performance Laboratory
Christopher Morton	IBM SAN Product Engineering, San Jose

Chris Neophytou	IBM Associate Network Specialist, Melbourne
Tom Newsom	IBM Project Manager SCSI Development, Raleigh
Simone Nova	IBM Netfinity University, Greenock
Steve Powell	IBM PCI Instructor, Raleigh
Howard Su	IBM PC Institute, Raleigh
David Worley	LSI Logic Storage Systems Inc.

---

## Comments welcome

### Your comments are important to us!

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

- Fax the evaluation form found in “IBM Redbooks review” on page 347 to the fax number shown on the form.
- Use the online evaluation form found at [ibm.com/redbooks](http://ibm.com/redbooks)
- Send your comments in an Internet note to [redbook@us.ibm.com](mailto:redbook@us.ibm.com)

---

## Part 1. Selecting a disk subsystem

## **2** Netfinity Server Disk Subsystems



---

## Chapter 1. Introduction

Data storage is arguably the most important element of your computing system. As businesses have become more dependent on information technology, as IT permeates into every aspect of a company's operations, the need for fast, reliable data storage has become paramount.

Disk storage for Intel-based servers used to be relatively simple to understand and there was little choice, with most installations using basic SCSI-attached disks. Now, the need for reliability and large capacity have provided the impetus for companies such as IBM to develop sophisticated storage subsystems.

These include multi-channel, SCSI RAID controllers, used to attach disks in external storage enclosures to the server. For some applications, however, SCSI has limitations that need to be overcome. To support disk connection over distances greater than a few meters, such as for disaster recovery solutions, to implement multi-node clusters of servers, or to support very large arrays of disks, alternative technologies must be used.

Netfinity servers offer great flexibility in configuring disk storage, with solutions that include SCSI RAID controllers, Fibre Channel disk arrays, and serial storage architecture (SSA) adapters and disks. With this flexibility comes choice: you now need to select the most appropriate technology for the demands you will place on the system. This book is designed to help you do so.

---

### 1.1 How to use this book

The book is divided into six parts. In Part 1, "Selecting a disk subsystem" on page 1, we discuss the factors that help determine which type of subsystem is most appropriate for typical server implementations. We also provide several sample configurations and information specific to individual operating system environments.

Part 2, "ServeRAID SCSI subsystems" on page 43 examines SCSI-based disk subsystems. We assume the reader is familiar with SCSI principles and do not discuss the basic SCSI adapters included in entry level Netfinity servers, focussing instead on the RAID adapters more often used in business-critical servers. This part, therefore, provides a detailed discussion of the IBM ServeRAID family of adapters, including the latest ServeRAID-4 products, including a description of the capabilities of these adapters and a chapter describing how to implement ServeRAID-based systems.

The latest Fibre Channel products supported by Netfinity are covered in Part 3, “Fibre Channel subsystems” on page 165. Again, we describe the available products and their features. A chapter is also devoted to explaining how to implement Fibre Channel solutions using Netfinity components.

Part 4, “SSA subsystems” on page 225 describes the technology behind serial storage architecture (SSA), a novel high-performance disk subsystem solution that has unique capabilities in the Intel-based server environment. It discusses available hardware and the configurations that it makes possible. Included in this section is information on disaster recovery and managing SSA disk subsystems.

Fibre Channel technology is an important element of the emerging storage area network (SAN) approach to providing data storage, and Part 5, “Storage area networks (SANs)” on page 277 discusses SAN in some depth. It describes the Netfinity SAN products that are available today and gives some glimpses of future possibilities deriving from this exciting and promising technology.

Finally, Part 6, “Appendixes” on page 313 includes useful information about operating system support and troubleshooting that will be invaluable when you want to install or debug a disk subsystem.

---

## 1.2 Audience

We have written this book to address three distinct types of readers. First, we aim to meet the needs of the IT staff within customers and other implementors of Netfinity server-based systems who need to select storage solutions to solve real business problems. If you fall into this category of reader, you will find the first part of the book particularly interesting. In it, we give guidelines on how to match a specific storage solution to your storage requirements.

You will probably find Part 5, “Storage area networks (SANs)” on page 277 of interest also. SANs are becoming increasingly important in the Intel-based server marketplace as these systems take on a growing number of business-critical roles in enterprises of all sizes. Part 5 provides information about IBM’s SAN strategy and how it can be implemented on Netfinity servers.

Our second audience are the sales professionals and others who design and plan storage solutions. Readers in this category will find Parts 2, 3 and 4 of special interest. Each contains information about one of the three disk storage technologies supported by Netfinity servers. A description of the

available products that use the technology under discussion, their features and capabilities, and how to implement disk subsystems using the technology can be found in these sections.

The implementation details referred to are aimed at our third audience, the technical support and implementation staff at customers, business partners and within IBM. For these readers, we have also included Appendix B, “Troubleshooting” on page 317, which provides troubleshooting information for those times when things go awry.

Throughout the book you will find background information on storage technologies and principles, and also tables and figures with data that will be invaluable to anybody working in this fast-developing and exciting field.

## **6** Netfinity Server Disk Subsystems

---

## Chapter 2. Netfinity storage technology

In this chapter, we examine the evolution of data storage based on SCSI, SSA, and Fibre Channel technologies, comparing them in terms of factors such as distance, performance, fault tolerance, availability and scalability. Finally, we will focus on the elements to assist you to map your storage requirements to the appropriate storage technology.

---

### 2.1 Evolution of storage technology

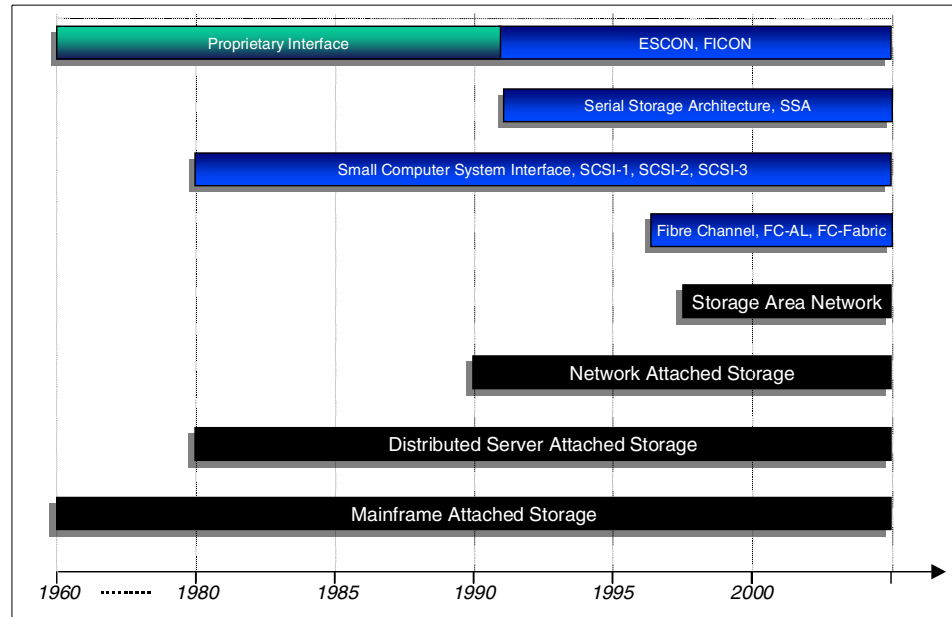


Figure 1. Evolution of storage interface and storage technology

The mainstream adoption of any new technology is an evolutionary process. To understand where on the evolutionary scale of storage technology we are today and where we will be heading in the future, we will take a brief look at the development of storage interfaces and storage technologies, starting with mainframe storage in the 1960s.

#### 2.1.1 Enterprise data storage

The first generation of storage system was based on system-attached storage in the mainframe environment. This is where the storage devices

such as disk and tape are directly connected to the server and thus, data access is dependent on the server platform and file system. The host server controls input and output (I/O) to the storage devices, issuing low level device commands, and listening for device responses. Initially, the storage devices were limited to executing data I/O requests from the server. Subsequently, more function was added such as caching to improve performance, and features such as RAID-1 mirroring to improve data availability.

In system-attached storage, at first the processor was used to control and perform I/O to the disks. To reduce the load on the processor, the channel subsystem was introduced to manage data flow to and from I/O devices. In the System/390 mainframe environment, connections between I/O devices and the processor used bulky bus and tag cables with a relatively low transfer rate of up to 4.5 MBps over a distance of up to 120 m. In 1991, IBM introduced the much faster and less cumbersome ESCON (Enterprise Systems Connection), based on fiber optics. It supports transfer rate of up to 17 MBps over a distance of 43 km.

The ESCON architecture was further enhanced with ESCON directors, which are high-speed switches providing dynamic connection capability between servers and storage devices. With this dynamic connection capability, a single channel can communicate with many control units, and a single control unit can communicate with many channels on one or more host servers. This reduces the number of channels and control units required to support an equivalent configuration that uses direct paths (that is, a configuration without ESCON directors). ESCON channels, coupled with ESCON directors, represent the first SAN architecture managed by ESCON manager.

FICON is an enhancement of ESCON that is based on Fibre Channel technology. FICON architecture is used in System/390 to support heterogeneous SAN environments. FICON channels are capable of transfer rates up to 100 MBps full duplex and extend the channel up to 100 km.

### **2.1.2 Intel processor-based systems**

Moving to the Intel-based PC system environment, after several lower-level interfaces were adopted and then superseded, the ATA interface became the predominant standard supported. The evolution of the ATA interface includes IDE (for Integrated Drive Electronics), EIDE (Enhanced IDE) and UltraATA. These interfaces support a transfer rate of up to 66 MBps, a maximum of two devices per interface, typically two interfaces per system, and a maximum connection distance of 25 cm. It is popular in the PC desktop environment due to its low cost. However, in a server environment, the small number of devices supported and the severe limitation on cable length make these

interfaces unsuitable for most systems. The least expensive alternative to ATA disk subsystems are those based on SCSI technology.

A typical SCSI disk subsystem today can support 15 or more disks, transfer data at 80 MBps or higher, and support connections to disks that are several feet away from the controller. As data storage requirements increased, SCSI adapters were developed to meet the requirements for scalability, high availability and high performance, primarily through the introduction of the redundant array of independent disks (RAID) technology.

To provide increased capacity and distance beyond that possible using parallel SCSI technology, IBM's Netfinity servers introduced disk subsystems based on serial storage architecture (SSA) and Fibre Channel technology. These technologies allow many more disk drives to be attached to a single controller and support high data transfer rates over distances that may be measured in kilometers.

### **2.1.3 What drives this evolution?**

Prior to 1997, client solutions were based mainly on stand-alone servers with internal SCSI disks or SCSI storage expansion enclosures located near the server. RAID technology provides a level of redundancy and protection against disk failure, but cannot help if the server itself fails. Business-critical applications require higher levels of availability than afforded by RAID alone.

Using external storage enclosures, such as one of the Netfinity storage expansion units, allows two-way clustering systems to be implemented using Microsoft Cluster Server (MSCS). This high availability solution allows two servers to access the same external disk subsystems, and maintains availability of server resource in the event of a server failure. In 1999, IBM extended the MSCS solution and introduced n-way clustering products based on Fibre Channel disk subsystems that support attachment of SCSI disk drives but over greater distances than previously possible.

In a centralized or distributed server-attached storage environment, the server may perform one of several roles such as a database server, a file and print server, an application server, a communications server, and so on. As the number of users connecting to this kind of system increases, the performance of the server will degrade. Importantly, if the server becomes unavailable, so will the storage system and clients will not be able to access their server-based data.

As the need for a high performance, cost-effective and centralized storage system continued to grow, the concept of *network-attached storage* (NAS)

was introduced. NAS is essentially a dedicated thin-client file server connected to a local area network using common interconnect technology such as Ethernet. In this case, the storage is externalized from the server. NAS is shared among multiple servers connected to the same network. It is able to process requests for data access from multiple servers - data copy sharing between heterogeneous servers and data sharing between homogeneous servers. NAS is differentiated from SAN because, whereas SAN utilizes a dedicated network between servers and storage, NAS is connected to the public LAN.

The evolution to SAN is a response to the explosion of demand for disk storage. Fibre Channel technology is the key enabling factor in this process of evolution. As mentioned, storage devices in a SAN are not connected to the primary network, but rather use a dedicated network specifically for storage access. As Fibre Channel technology continues to advance, a shared storage repository may attach to multiple host servers.

The SAN is effectively an extended, shared storage bus that can be interconnected using similar technologies to those used in LAN or WAN networks, including routers, switches and gateways. However, addressing connectivity alone is not sufficient to implement a true SAN environment. Also required is an intelligent disk subsystem to communicate between networks; products providing full SAN capability will be developed in the not-too-distant future.

In a nutshell, a SAN is a data-centric environment where storage is an independent, high-performance and high-availability network. It is also highly scalable and centralized. As it is in its own network, network traffic in the primary network is no longer an issue.

SAN technology will become critical to the success of most data-intensive computer environments, driven by applications. To build a flexible and secure storage environment it takes a good understanding of your own storage needs and knowledge of how to leverage current technology to meet those requirements.

---

## **2.2 What technologies are available today?**

In this section, we give a functional overview of the three storage technologies used in Netfinity servers today. These are based on SCSI, Fibre Channel and SSA technologies.



### **2.2.1 Small computer system interface (SCSI)**

The small computer systems interface (SCSI) has been around for almost 20 years now, having evolved greatly from the original interface developed by Shugart in 1980. It has become the standard disk subsystem technology for almost all Intel-based servers available on the market today, including the entire IBM Netfinity server range. Very few disk subsystem technologies are available to rival the performance and features offered by SCSI at an affordable price. SCSI allows you to connect hard disk drives, tape drives, CD-ROM drives and other SCSI devices on the same bus.

Due to its relatively long history and popularity in the Intel processor-based server market, SCSI hardware manufacturers abound and consequently the choice of SCSI hardware is huge. Since SCSI has been adopted, developed, and well-documented by the American National Standards Institute (ANSI), interoperability with existing devices, backward compatibility, and ease of upgrading are key factors in the popularity of SCSI subsystems.

The SCSI bus is a parallel interface and may be used as the final disk connection even in Fibre Channel and SSA subsystems. SCSI devices are directly connected to the bus, which requires adherence to a set of rules determining bus length and proper bus termination to ensure correct operation. It supports a transfer rate of up to 160 MBps using the latest Ultra3 SCSI interface, with a maximum of 16 devices (including the adapter) per interface. Connection distances have now been extended up to 15m using differential SCSI connections.

For more information on SCSI, in particular the Netfinity ServeRAID family of SCSI RAID adapters, refer to Part 2, "ServeRAID SCSI subsystems" on page 43.

### **2.2.2 Fibre Channel**

Fibre Channel (FC) is the next generation in high-performance storage interface technology. It consists of an integrated set of standards that define new protocols for flexible information transfer using several interconnection topologies. These standards have been created by ANSI, with IBM playing a leadership role as it has done before in the development of many new industry-standard technologies.

Fibre Channel combines the standard SCSI command set and protocol with the flexibility and connectivity of networks. Its flexibility and scalability enable it to handle different protocols simultaneously. This allows a Fibre Channel network to serve as a high-speed LAN, supporting network protocols such as TCP/IP and attachment of storage devices.

For more detailed information about Fibre Channel, refer to Part 3, “Fibre Channel subsystems” on page 165.

### **2.2.3 Serial storage architecture (SSA)**

Serial storage architecture (SSA) was introduced by IBM in 1991. IBM created the base SSA technology and helped to found the SSA Industry Association (SSAIA), an independent body to help promote SSA as an open industry standard. Over 35 companies, representing all major segments of the computer industry, now participate in the SSAIA.

Developed originally for use in RS/6000 systems, SSA technology is fast becoming an industry standard, with many manufacturers including SSA hardware offerings in their product lines. The responsible ANSI committee continues to develop the SSA standard as a part of the SCSI-3 family of standards.

IBM has taken SSA a step further, by announcing the availability of SSA technology and hardware for the Intel processor-based server market. IBM is now producing SSA hardware for the Netfinity range, including RAID and cluster-aware adapters, hard disk drives, enclosures and the necessary cabling to enable users to fully exploit the power of SSA technology.

For details on SSA, you may refer to Part 4, “SSA subsystems” on page 225.

---

## **2.3 Comparing storage solutions**

Each type of storage subsystem has particular features and strengths that make it a more appropriate solution for a specific application. In this section we examine the major factors that influence the decision of which subsystem

to implement. Table 1 summarizes these features, and the following paragraphs explain how these influence the capabilities of each technology.

Table 1. Feature comparison among disk technologies

<b>Comparison of SCSI, Fibre Channel, and SSA</b>			
<b>Features</b>	<b>SCSI</b>	<b>FC-AL</b>	<b>SSA</b>
Data transmission	Half duplex	Full duplex	Full duplex
Array support	Parity	Hot-plugging, dual porting	Hot-plugging, dual porting
Data transfer rate	40 MBps for Ultra SCSI, 80 MBps for Ultra2 SCSI, 160 MBps for Ultra3 SCSI	100 MBps, 200 MBps (with dual porting)	80 MBps, 160 MBps (with dual porting)
Number of devices supported	15 per bus	126 per FC-AL, 16 million per FC-fabric	126 per loop
Connection distance	Up to 12 m	Up to 10 km	Up to 10 km
Cable type	Copper	Copper, fiber optic	Copper, fiber optic
Data integrity	Parity, CRC with Ultra3 SCSI	CRC	CRC

### 2.3.1 Distance

Single-ended SCSI technology supports connections over relatively limited distances, and as SCSI bus speeds have increased, there is a commensurate reduction in the supported bus length. Ultra SCSI, for example, allows the bus to extend only to 1.5 meters. To overcome this limitation, the implementation of a differential SCSI bus, introduced with Ultra2 SCSI adapters, allows somewhat longer cable lengths due to the improved noise immunity. For example, Netfinity ServeRAID-3H Ultra2 SCSI adapters are able to support disk drives up to a maximum of 12 m from the servers.

In comparison, the Netfinity Fibre Channel and SSA technologies support connections over much greater distances. Fibre Channel supports distances of up to 25 m using copper cabling. Short-wave optical cable extends this to 500m and long-wave optical cable boosts the distance even more: up to 10 km may separate the server from the disks.

SSA disks may be up to 25 m from the server using copper cable and, by using a Fiber Optical Extender, this may be extended to a distance of up to 10 km. Even at a separation of 10 km, there is no loss in data throughput or reliability with either Fibre Channel or SSA.

### **2.3.2 Performance**

Initially, SCSI technology provided a maximum transfer rate of only 5 MBps, but enhancements have increased this to 80 MBps using Ultra2 SCSI, and 160 MBps using Ultra3 SCSI. However, SCSI technology is also limited by its shared, arbitrated bus, for which connected SCSI devices have to compete in order to gain access and transfer data. Because of this, actual data throughput is somewhat less than the maximum, as performance suffers when multiple devices wish to transfer data across the shared bus.

In terms of performance, Fibre Channel transfers data at a speed of 100 MBps, providing high performance over large distances. This makes Fibre Channel an ideal technology for the sharing of storage resources, and for disaster protection and recovery solutions for business-critical applications. Fibre Channel can deliver data transfer rates of 200 MBps. Future Fibre Channel enhancements are expected to push the performance levels up to 1 GBps.

SSA supports an aggregate bandwidth of 160 MBps using the Advanced SerialRAID/X Adapter - 80 MBps read and 80 MBps write on one initiator on a loop. Because SSA utilizes a non-arbitrated loop for communications, multiple transactions can take place concurrently on different sections of a single loop. Thus the aggregate bandwidth can be higher than the basic loop speed.

### **2.3.3 Fault tolerance**

SCSI technology has no inherent fault-tolerant features. Should an adapter fail, or the bus be damaged, communication with the disks will be disrupted. However, Netfinity servers can use SCSI-based disk subsystems in fault-tolerant configurations such as those made possible by Microsoft Cluster Server (MSCS).

Fibre Channel Arbitrated Loop (FC-AL) is a full-duplex, bidirectional loop. It is designed to permit devices to be removed from the loop without interrupting throughput or sacrificing data integrity. To prevent loop failure in the event of a failed device, the Fibre Channel port to which the device is connected has bypass circuits that can quickly route around the problem so that all other devices on the loop remain accessible.

The Fibre Channel Arbitrated Loop interface also has a sophisticated error-detection scheme. Several bytes of cyclic redundancy check (CRC) information are transmitted along with each packet of user data. The receiving device uses this CRC information to check the integrity of the data received and requests a resend if an error is detected.

SSA is also full-duplex; each device on a loop can be accessed from either direction. If a node or connection should fail, the initiator will automatically reroute frames around the other side of the loop. IBM's 7133 Serial Disk Enclosures also use host port bypass circuits that provide a path around a failed link eliminating the need for the data to be rerouted away from the failed connections to reach its target. Port bypass circuits also provide protection for multiple SSA device failures.

SSA also provides significant link recovery capability. The SSA interface causes data to be retransmitted if errors are detected. Both FC-AL and SSA schemes are more effective than the parity scheme used in parallel SCSI transfers, where error detection is performed on a byte-to-byte basis. The new generation Ultra3 SCSI has CRC added, closing the reliability gap between SCSI and FC-AL or SSA.

#### **2.3.4 Availability**

As more business-critical applications are executed on Intel-based servers, the high availability features of the computing environment becomes very important. Netfinity servers offer several features to increase availability, including RAID technology and redundant components. Moving beyond the levels of availability that can be achieved with a single system requires more sophisticated approaches. Fault tolerance and disaster protection and recovery may be implemented through clustering technology using Fibre Channel, SCSI or SSA.

Fibre Channel and SSA clustering solutions can be particularly attractive as they may span long distances, thus enabling a distance mirroring or backup clustering solution. Should a failure occur in a clustered environment, work is transferred to the backup or mirrored server with minimum or no interruption. For the largest distances, Fibre Channel has the advantage that data throughput is maintained, whereas throughput for SSA reduces as distance increases.

SCSI can provide some of these benefits, but does not have the same flexibility as the other two technologies. This is due to the limitations in connection distance and also because it is impractical to implement clusters with more than two servers using SCSI as the disk subsystem.

### 2.3.5 Scalability

SCSI is based on a shared arbitrated bus in which multiple devices transfer data over the shared bus. As the number of devices increases, competition for access to the bus increases and may create a bottleneck for performance. A single SCSI bus can support a maximum of 16 devices, one of which is the controlling adapter.

Fibre Channel uses an arbitration protocol which means that, at any point in time, the entire Fibre Channel loop is used for at most one transmission in each direction. The arbitration scheme is used to decide which node gets control of the loop. Once a node has control, it will create a point-to-point logical connection with the node with which it wishes to communicate. After completing the transmission, the loop is freed and the devices on the loop have to arbitrate for loop access once again (hence the term FC-AL).

System expansion using Fibre Channel is enabled by the increased number of devices that can be attached per PCI slot in comparison with SCSI. Fibre Channel provides 127 FC-AL IDs, one of which is used by the RAID controller. This allows connection of up to 126 devices in a loop. However, by connecting the devices to Fibre Channel switches, it will create a Fibre Channel fabric which, theoretically, is able to support up to 16 million devices. It may connect up to 126 host systems per connection. In addition, the FC-AL architecture allows connection of multiple hosts, whereas SCSI is limited to only two.

SSA is arbitration-free. Each device has a guaranteed bandwidth available to it. Additionally, SSA allows spatial reuse, which means that multiple transmissions may be executed at the same time between different devices and in different directions, depending on their topological position in the loop.

SSA, theoretically, can provide a connection of up to 127 devices including one being used for the RAID adapter. IBM supports SSA adapters with connection to up to 48 devices per loop or 96 devices per adapter.

---

## 2.4 Mapping your requirements to technology

Elements to consider when selecting a disk subsystem are the applications and the computing environment, connection distance, performance, fault tolerance, availability, and scalability.

In this section, we will look at the considerations for SCSI, Fibre Channel, and SSA disk subsystems.

### 2.4.1 SCSI disk subsystems

SCSI offers excellent performance and moderate capacity at reasonable cost, and is used as the core disk subsystem technology in IBM Netfinity servers. The popularity that SCSI enjoys is the result of it being a technology that is extremely stable and well-proven, having undergone several iterations, the latest of which is specified in the SCSI-3 standard.

Technologies such as LVDS and Ultra2 SCSI provide a maximum transfer rate of 80 MBps with cable lengths of up to 12 m and provide good disk subsystem performance. The latest SCSI technology, Ultra3 SCSI provides a 160 MBps maximum data transfer rate.

SCSI technology has come a long way and is still undergoing development. Future developments are still likely to be backward-compatible, as has been the case so far in SCSI evolution. The Netfinity ServeRAID adapter family, exemplified by the Netfinity ServeRAID-4H Ultra3 SCSI adapter, also offer the improved availability of RAID technology.

SCSI remains an important and evolving technology due to its features, cost-effectiveness and dominant position in the server market.

The primary considerations that would direct you towards selecting a SCSI disk subsystem are as follows:

- You have moderate storage requirements.
- You have only moderate-size databases or none at all.
- You do not wish to invest in new technology at present.
- You are satisfied with existing SCSI disk subsystem performance.
- You do not need to separate your disk enclosures from your servers by great distances.
- Maintenance costs have a higher priority than disk subsystem performance.

### 2.4.2 Fibre Channel disk subsystems

Fibre Channel's immediate benefits lie in its improved performance, distance and device connectivity in comparison with SCSI. Solutions that can be derived from these benefits include remote disaster protection, archiving and recovery, and server and storage consolidation.

Fibre Channel can transfer data at speeds up to 100 MBps, giving you high performance over distances up to 10 km from the server, which offers real benefits to your business-critical applications. Speed and distance, coupled

with its ability to attach up to 126 devices using physically longer and smaller cables than SCSI, make it an attractive alternative in many cases.

While Netfinity Fibre Channel solutions offer high functionality, the cost to implement them are higher than comparable SCSI implementations. Fibre Channel is likely to be most attractive in advanced, large-enterprise environments where installation costs can be justified by the availability and performance requirements of core applications. Applications that need access to large amounts of online data storage are also candidates for Fibre Channel subsystems.

The primary considerations that would direct you towards selecting a Fibre Channel disk subsystem are as follows:

- You have large storage requirements.
- Your application utilizes large databases.
- You wish to invest in technology that can allow you to consolidate servers and storage in the future.
- You need better subsystem performance than that offered by SCSI.
- You need to separate your disk enclosures from your servers by great distances.

### **2.4.3 SSA disk subsystems**

SSA offers many of the same advantages as Fibre Channel. As SSA is used widely in IBM's RS/6000 product family, this can be a consideration in making your selection. Using technology with which your support staff is already familiar can reduce overall maintenance costs.

SSA technology can bring many benefits when used in a Netfinity environment. SSA is excellent for increased scalability to support up to multiple terabytes of storage separated by big distances. SSA can also be useful as a way to provide disaster recovery solutions.

The primary considerations that would direct you towards selecting an SSA disk subsystem are as follows:

- You have large storage requirements.
- Your application utilizes large databases.
- You need better subsystem performance than that offered by SCSI.
- You need to separate your disk enclosures from your servers by great distances but do not need the performance offered by Fibre Channel.
- Your organization already has an investment in SSA technology



---

## Chapter 3. Sample disk configurations

In this chapter, we will look at various scenarios and illustrate some sample disk configurations, based on currently available Netfinity disk subsystems technology. The configurations presented are as follows:

- Several typical Netfinity SCSI disk subsystem configurations, including a large disk storage configuration.
- Several typical Netfinity Fibre Channel disk subsystem configurations, including a large disk storage configuration.
- Netfinity disk and tape pooling.
- Netfinity remote backup and recovery.
- A Netfinity high availability solution using Microsoft Cluster Server.

Additional configuration examples can be found in Part 5, “Storage area networks (SANs)” on page 277.

---

### 3.1 Netfinity SCSI disk subsystems

We now move on to examine some typical configurations that you could use when a ServeRAID-based disk subsystem has been selected.

#### 3.1.1 Standard SCSI configurations

Figure 2 on page 20 shows a solution with three Netfinity servers providing various services to attached clients. Two of the servers may be, for example, basic file and print servers, while the third is shown as a backup server with an attached tape library. The two servers providing disk resources each have ServeRAID adapters installed to protect the data held on the disks in the external enclosures.

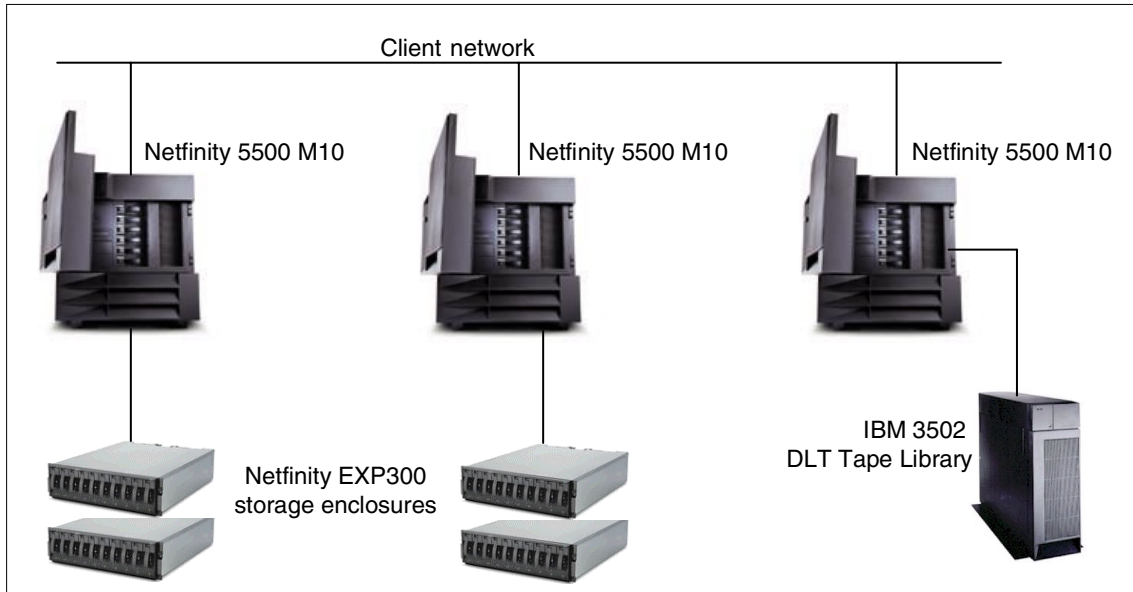


Figure 2. Typical Netfinity SCSI disk subsystem configuration

The specifics of the configuration in Figure 2 are shown in Table 2:

Table 2. Hardware/software requirements for Figure 2

Component	Requirements / recommendations
Software	<ul style="list-style-type: none"> <li>• Microsoft Windows NT 4.0 Server</li> <li>• Microsoft Windows 2000</li> <li>• Novell NetWare 4.2 or 5</li> <li>• Linux</li> <li>• IBM OS/2 Warp Server Advanced 4.0</li> <li>• IntraNetWare 1.0</li> <li>• SCO Open Server 5.0.4</li> <li>• SCO UnixWare 7.0</li> </ul>
Server	<ul style="list-style-type: none"> <li>• Two Netfinity 5500 M10 servers, each configured with two internal 36.4 GB hard drives for RAID-1 hardware mirroring. The internal disks will hold the network operating system and any server application files.</li> <li>• One Netfinity 5500 M10 server to function as a tape backup server. Alternatively, one of the Netfinity 5500 M10 may also be used as the tape backup server if some performance impact during backup processing is acceptable.</li> </ul>

Component	Requirements / recommendations
Interconnects	<ul style="list-style-type: none"> <li>• A Netfinity 10/100 Ethernet PCI Adapter connects each server to the public LAN</li> <li>• Netfinity Ultra3 SCSI cables used for server-to-storage connections</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• Two Netfinity ServeRAID-4H Adapters (one per file/print server)</li> <li>• Four Netfinity EXP300 Storage Expansion Units</li> <li>• 56 (4x14) 36.4 GB Ultra3 SCSI Hard Disk Drives</li> <li>• IBM 3502 DLT Tape Library</li> </ul>

- Netfinity ServeRAID-4H Adapters are installed in the servers to provide RAID protection for both the two internal hard disk drives containing the network operating system and the external storage expansion units containing user data.
- In this example, two external channels are used to connect to the Netfinity EXP300 Storage Expansion Units. Each EXP300 is able to house 14 hard disk drives. If more external storage is required, the ServeRAID-4H controllers support up to four external channels. In addition, the internal drives could be connected to the server's integrated controller (this is itself a RAID controller on the Netfinity 5500 M10; other servers have standard SCSI controllers).
- Servers may also be configured with an additional ServeRAID adapter to create a fault-tolerant pair for improved availability, as shown in our next example.

### 3.1.2 Fault-tolerant configurations

This example illustrates a configuration utilizing the failover feature of the ServeRAID adapter. In Figure 3 on page 22 we have a Netfinity server configured with redundant Netfinity ServeRAID-4H adapters.

If one of the ServeRAID adapters fails, the system will remain operational. Advanced Netfinity servers that support hot-swapping of adapter cards, such as the Netfinity 5500 M10, allow you to replace the failed adapter without taking the server offline.

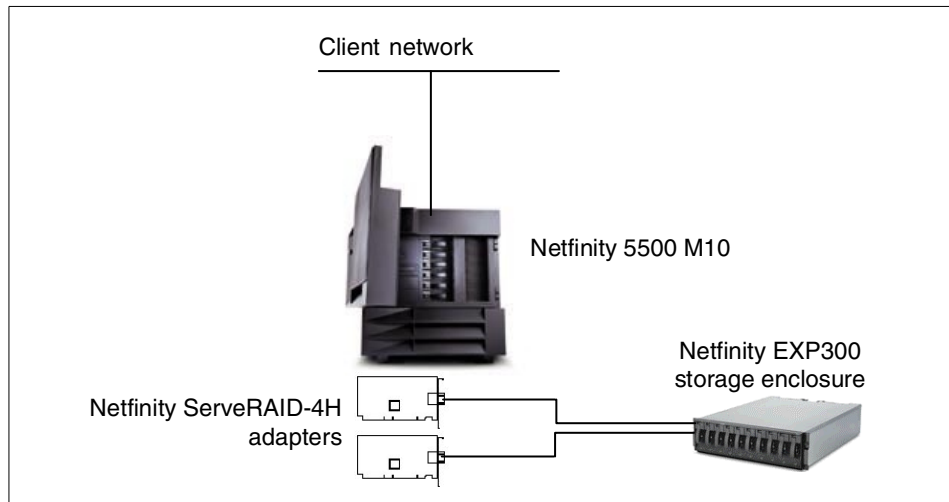


Figure 3. Typical Netfinity SCSI disk subsystem with redundant adapters

#### Netfinity EXP300 Storage Expansion Unit

There is one important limitation that you need to be aware of in this type of configuration. The EXP300 comes with two SCSI ports to allow connection of two SCSI adapters for fault-tolerant and clustered configurations. It supports up to 14 hard disk drives. A single SCSI channel is able to address up to 16 devices.

With two RAID adapters connected to an EXP300, they use one SCSI ID each and the expansion unit backplane also uses one SCSI ID. In this configuration, therefore, only 13 SCSI IDs remain to be allocated to the hard disk drives. The implication of this is that fault-tolerant pairs of ServeRAID adapters cannot be configured with the maximum of 14 hard disk drives. One bay in the EXP300 must be left empty.

### 3.1.3 A high-capacity SCSI subsystem

High-capacity disk subsystems can be implemented using Netfinity servers configured with multiple ServeRAID-4H adapters. In Figure 4, a Netfinity 8500R server is shown in this type of configuration. With support for spanned arrays, the ServeRAID adapter can utilize these disks to create very large disk volumes as seen by the operating system:

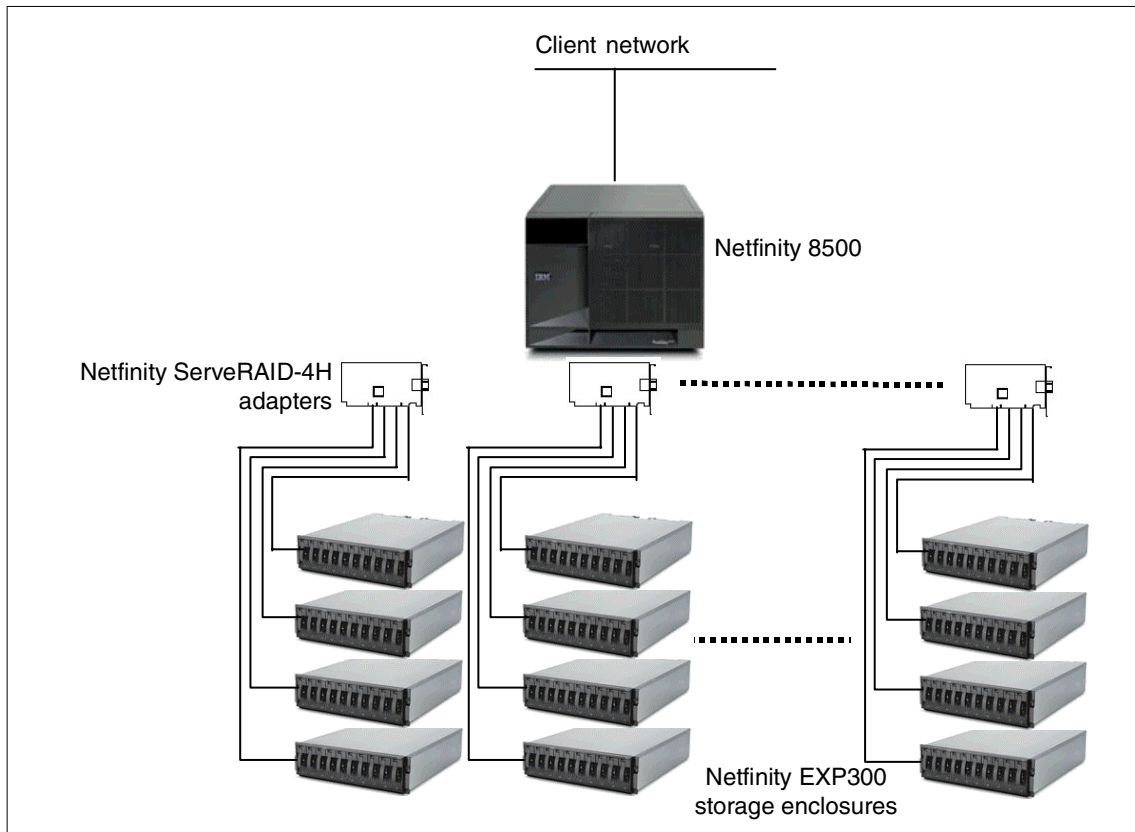


Figure 4. Netfinity large SCSI disk subsystem configuration

The specifics of this configuration are shown in Table 3:

Table 3. Hardware/software requirements for Netfinity large SCSI disk subsystems

Component	Requirements / recommendations
Software	<ul style="list-style-type: none"> <li>• Microsoft Windows NT 4.0 Server</li> <li>• Microsoft Windows 2000</li> <li>• Novell NetWare 4.2 or 5</li> <li>• Linux</li> <li>• IBM OS/2 Warp Server Advanced 4.0</li> <li>• IntraNetWare 1.0</li> <li>• SCO Open Server 5.0.4</li> <li>• SCO UnixWare 7.0</li> </ul>
Server	<ul style="list-style-type: none"> <li>• One Netfinity 8500R server configured with two internal 36.4 GB hard drives for RAID-1 mirroring of the network operating system</li> </ul>

Component	Requirements / recommendations
Interconnects	<ul style="list-style-type: none"> <li>• One Netfinity 10/100 Ethernet PCI Adapter to connect to the public LAN.</li> <li>• Netfinity Ultra3 SCSI cables used for server-to-storage connections.</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• 11 Netfinity ServeRAID-4H adapters</li> <li>• 44 Netfinity EXP300 Storage Expansion Units</li> <li>• 440 36.4 GB Ultra3 SCSI Hard Disk Drives</li> </ul>

- The Netfinity 8500R server has 12 64-bit active PCI slots. One is allocated to the Netfinity 10/100 Ethernet PCI Adapter for connection to a public LAN, leaving 11 slots for Netfinity ServeRAID-4H adapters to connect to the storage expansion units. No disks are installed internal to the server, allowing all RAID controller channels to be used for external disk enclosure attachment. If RAID protection is not required for the operating system (unlikely), internal disks could be driven by the server's integrated SCSI controllers.
- Each Netfinity ServeRAID-4H adapter has four SCSI channels of which two are available as internal channels, while all four are available externally.
- Each channel is connected to a Netfinity EXP300 Storage Expansion Unit which supports up to 14 Ultra3 SCSI 160 MBps 36.4 GB hard disk drives. The full capacity of each expansion storage unit is 509.6 GB without redundancy (RAID-0). In total, the disk capacity supported is 22,422.4 GB or approximately 22 TB.

This configuration is developed based on a single Netfinity 8500R server connected to a large number of SCSI hard disk drives. Typically, this type of configuration is used to store large databases. However, consideration should be given to performance based on demand and usage.

**Multiple ServeRAID-4 adapters**

In a Windows NT or Windows 2000 environment, up to 12 ServeRAID adapters may be installed in a single server. For all other supported operating systems, the maximum supported number is eight.

However, if you need servers with this amount of disk capacity, you may well decide to implement a Fibre Channel solution with its additional benefits (see Part 3, "Fibre Channel subsystems" on page 165).

## 3.2 Typical Netfinity Fibre Channel disk subsystems

The configurations we examine are based on the Netfinity Fibre Array Storage Technology (FAStT) products.

### Supported configurations

Configurations described in this chapter are for example only. The flexibility of Fibre Channel subsystems allows many different configurations to be implemented. You should check the IBM Netfinity Fibre Channel Web site for current information about supported configurations:

<http://www.pc.ibm.com/ww/netfinity/fibrechannel/>

### 3.2.1 Basic configuration

Figure 5 illustrates a basic configuration using a Fibre Channel disk subsystem with redundant Netfinity FAStT Host Adapters connected to the Netfinity FAStT500 RAID Controller with redundant control units. Using this configuration, if one of the adapters, cables or controllers fails, the system will remain operational.

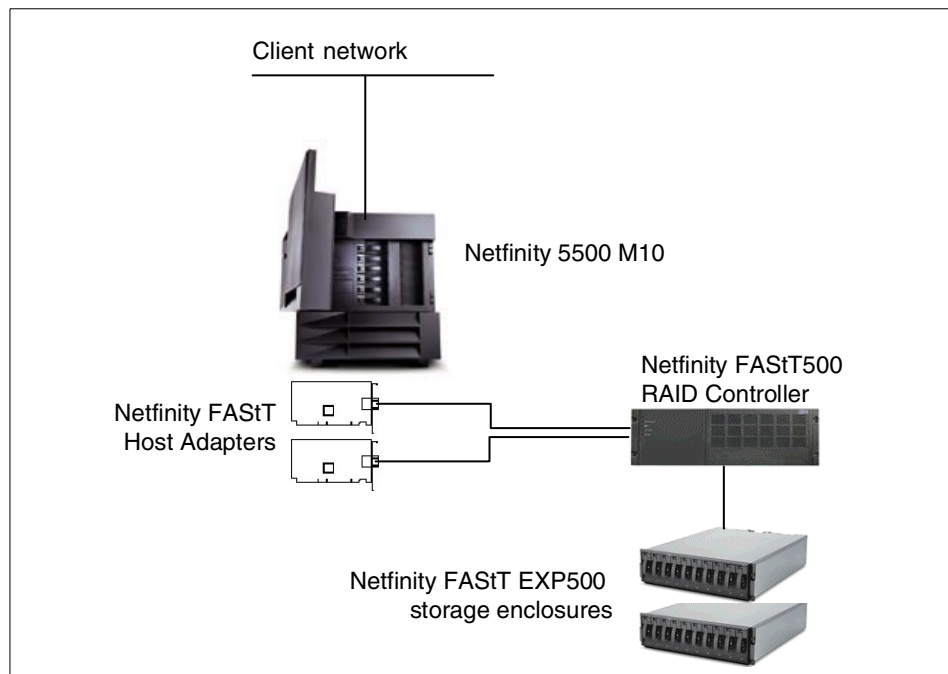


Figure 5. Netfinity FAStT500 configuration with redundant adapters and RAID controllers

The Netfinity FAStT500 RAID Controller in Figure 5 is configured with two control units as standard, to form a redundant pair. These control units can be configured as either an active-active or an active-passive pair. Each control unit supports two Fibre Channel loops, and can attach a theoretical maximum of 126 devices per loop.

The controller enclosure has four pairs of drive interface connectors to support drive loop configurations requiring dual cables. Each pair of connectors is housed in its own drive interface mini-hub. At present only one port per pair used (the other is reserved for future use) and loops are configured for redundancy,

#### **Dual-loop wiring**

It is anticipated that many customers will implement dual-loop wiring to provide increased availability through redundancy. You can find out more about cabling of Fibre Channel subsystems in 6.2.10, “Cabling requirements for the EXP500” on page 187.

Each control unit uses a Fibre Channel ID and the EXP500 units themselves use one Fibre Channel ID each plus one ID for each installed disk. This fact means that, in practice, a loop can be configured with up to 11 Netfinity Fibre Channel EXP500 Storage Expansion Units, each holding 10 hard disk drives, that is, up to 110 disks, using 123 IDs in total ( $11 \times 11 + 2$ ).

### **3.2.2 Multiple hosts**

Our next example of a Netfinity Fibre Channel disk subsystem configuration shows how multiple servers may be connected together to utilize a common disk subsystem in a non-clustered environment (Figure 6):



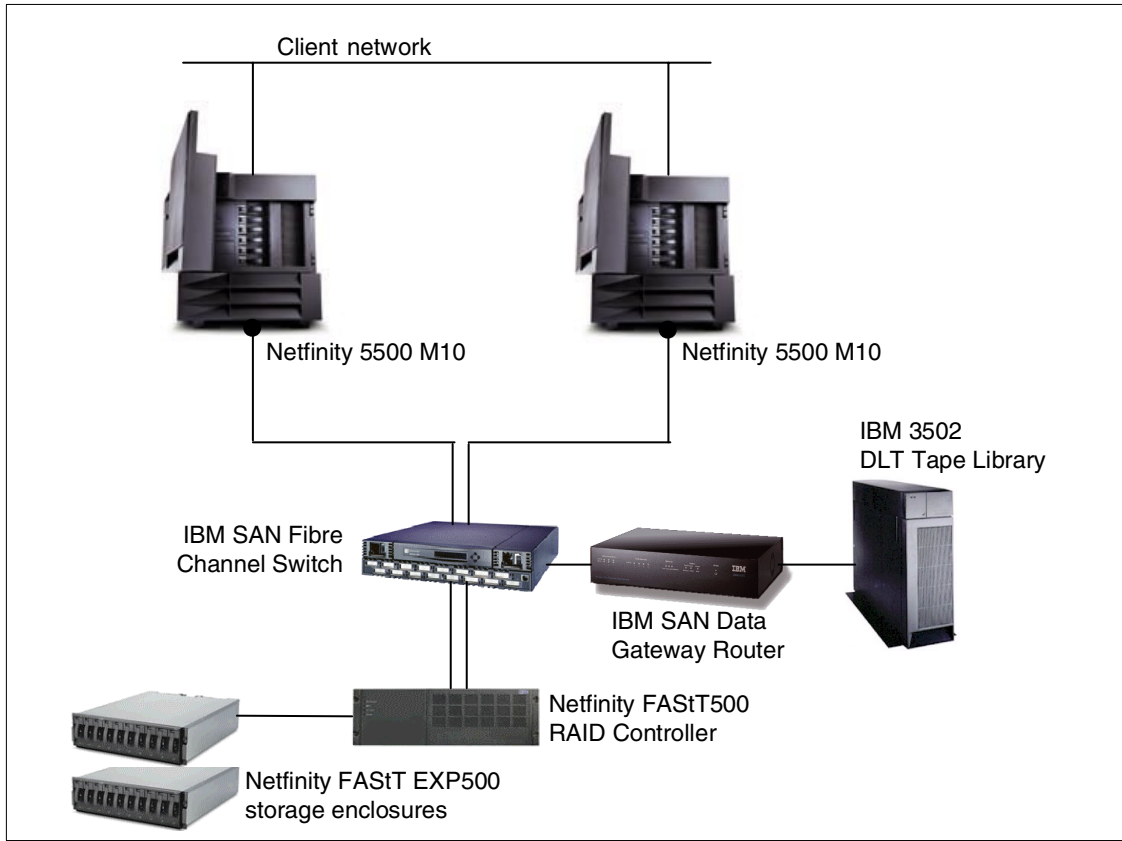


Figure 6. Typical multihost Netfinity FASt500 disk subsystem configuration

The specifics of this configuration are shown in Table 4:

Table 4. Hardware/software requirements for a multihost Netfinity FC disk subsystem

Component	Requirements / recommendations
Software	<ul style="list-style-type: none"> <li>• Microsoft Windows NT 4.0 Server</li> <li>• Microsoft Windows 2000 Server or Advanced Server</li> <li>• Novell NetWare 4.2 or 5.0</li> </ul>
Servers	<ul style="list-style-type: none"> <li>• Two Netfinity 5500 M10 servers, each configured with two internal 36.4 GB hard drives for RAID-1 hardware mirroring of the network operating system</li> </ul>

Component	Requirements / recommendations
Interconnects	<ul style="list-style-type: none"> <li>• One Netfinity 10/100 Ethernet Adapter to connect to the public LAN</li> <li>• Two Netfinity FAST Host Adapters</li> <li>• One IBM SAN Fibre Channel Switch, 2109-S16</li> <li>• One IBM SAN Data Gateway Router, 2108-R3S</li> <li>• Two Netfinity Fibre Channel short-wave GBICs</li> <li>• Netfinity Fibre Channel Cables</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• Netfinity 5500 integrated ServeRAID adapters</li> <li>• One Netfinity FAST500 RAID Controller</li> <li>• Two Netfinity FAST EXP500 storage expansion units</li> <li>• 20 36.4 GB Fibre Channel Hard Disk Drives</li> <li>• IBM 3502 DLT Tape Library</li> </ul>

Key features of the above configuration are:

- The Netfinity 5500 M10 servers are each configured with a Netfinity 10/100 Ethernet Adapter for connection to the public LAN. They also come as standard with an integrated ServeRAID adapter, which we use to support two internal hard disk drives using RAID-1 hardware mirroring for the network operating system.
- Netfinity FAST Host Adapters are installed in the Netfinity 5500 M10 servers. Two adapters may be installed in a single server as a fault-tolerant pair if desired.
- The IBM SAN Fibre Channel Switch connects the host bus adapters to the Netfinity FAST RAID Controller. This 16-port switch allows flexibility in Fibre Channel configurations where it may be cascaded with additional switches to support up to 16 million devices. In simpler configurations, where the switch may not be required, the redundant RAID controllers may be connected directly to the host bus adapters.
- Hub-to-hub or switch-to-switch Fibre Channel connections may be up to 10 km in length, and a maximum of 500 m for other Fibre Channel devices.

### 3.2.3 Two-node cluster configurations

Figure 7 illustrates a basic clustered configuration using Fibre Channel disk subsystems. It uses two Netfinity 5500 M10 servers for failover redundancy, connected to the Netfinity FAST500 RAID Controller with its redundant control units.

If one of the servers, cables or controllers fails, the system will remain operational. In addition, another Fibre Channel host adapter may be added to each of the servers for additional redundancy. This configuration is typical for implementations of Microsoft Cluster Server (MSCS) and Novell Cluster Services (NCS).

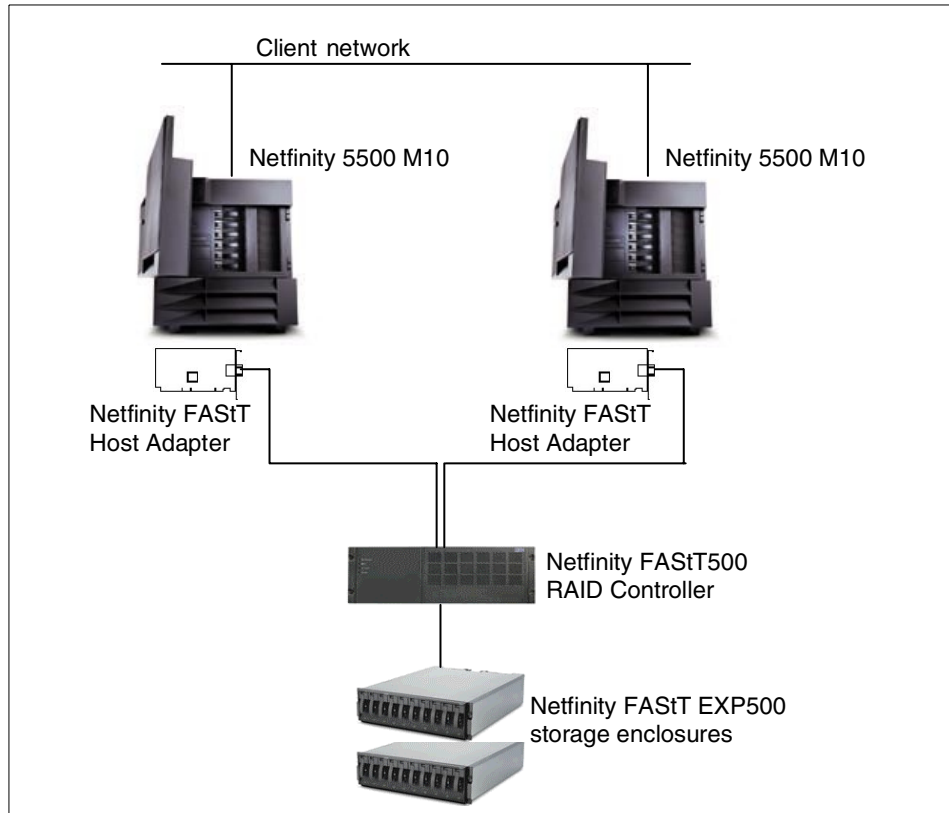


Figure 7. Netfinity FAST disk subsystem for two-node clusters

### 3.2.4 A high-capacity Fibre Channel subsystem

To implement a high-capacity Netfinity Fibre Channel disk subsystem configuration, Netfinity servers may be connected to a Netfinity Fibre Channel Switch, using Netfinity Fibre Channel PCI Host Bus Adapters, supporting multiple RAID controllers to allow access to a large amount of storage. Storage partitioning can be used to isolate and protect each server's private storage areas, even though they may co-exist on the same RAID subsystem.

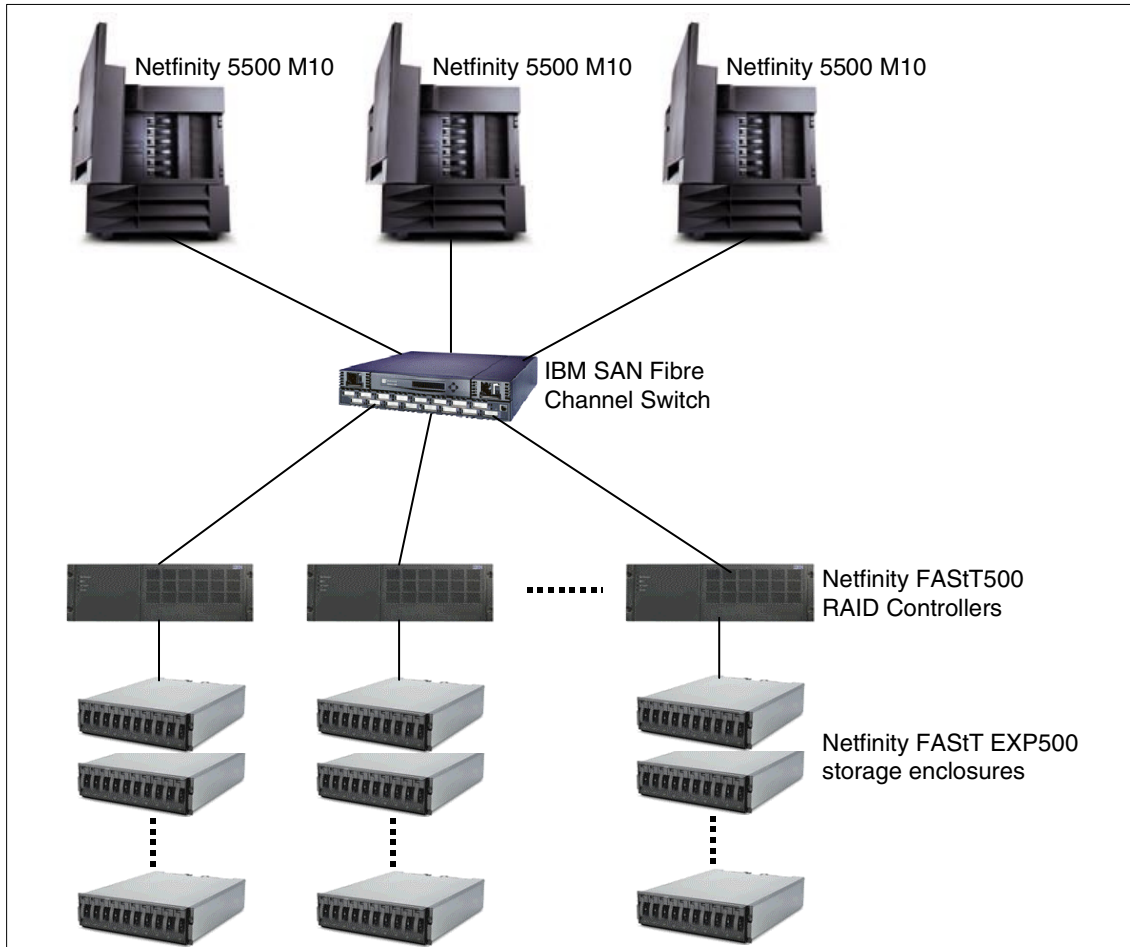


Figure 8. High capacity Netfinity FASTT disk subsystem configuration

The specifics of this configuration are shown in Table 5:

Table 5. Hardware/software requirements for a high-capacity FC disk subsystem

Component	Requirements / recommendations
Software	<ul style="list-style-type: none"> <li>• Microsoft Windows NT 4.0 Server</li> <li>• Microsoft Windows 2000 Server or Advanced Server</li> <li>• Novell NetWare 4.2 or 5.0</li> </ul>
Servers	<ul style="list-style-type: none"> <li>• Netfinity 5500 M10 configured with two internal 36.4 GB hard drives for RAID-1 hardware mirroring of the network operating system</li> </ul>

Component	Requirements / recommendations
Interconnects	<ul style="list-style-type: none"> <li>• One Netfinity 10/100 Ethernet Adapter per server to connect to the public LAN</li> <li>• One or two Netfinity FAStT Host Adapters per server</li> <li>• One IBM SAN Fibre Channel Switch, 2109-S16</li> <li>• Netfinity Fibre Channel cables</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• Netfinity 5500 integrated ServeRAID adapters</li> <li>• Multiple FAStT500 RAID Controller Units</li> <li>• Multiple Netfinity FAStT EXP500 storage expansion units</li> <li>• Multiple 36.4 GB Fibre Channel Hard Disk Drives (based on RAID-5 configuration)</li> <li>• Multiple Netfinity 42U rack enclosures to house the hardware</li> </ul>

- The Netfinity 5500 M10 servers are each configured with a Netfinity 100/10 EtherJet PCI Adapter for connection to the public LAN. They also come as standard with an integrated ServeRAID adapter which we use to support two internal hard disk drives using RAID-1 hardware mirroring for the network operating system.
- Each server also has one Netfinity FAStT Host Adapter to connect to the IBM SAN Fibre Channel Switch.
- Each controller supports two Fibre Channel arbitrated loops with a maximum of 11 storage expansion units per loop. Each of the storage units is able to house 10 Fibre Channel 36.4 GB hard disk drives, giving 110 disks for each loop. The total number of disks supported in a controller is 220 disks in a fully redundant loop. This adds up to 8 TB per controller (based on 220 x 36.4 GB). Operating system constraints on capacity, the number of LUNs supported, or drive letter allocation may limit the usable capacity that can be achieved.
- To increase high availability on the shared disks, RAID-5 may be used. RAID-5 arrays are configured using Netfinity FAStT Storage Manager software. Arrays may be created spanning across both loops within the same controller. If a RAID-5 array is created for each storage expansion unit of 10 hard disk drives, and each expansion unit is allocated a hot-spare disk, this provides a usable capacity equivalent to that of eight disks for each 10 disks attached to the controller, or approximately 45 TB.

A disk subsystem such as that shown in Figure 8 should not be attached to a single server since it will not be able to generate enough disk traffic to make the configuration effective. In other words, the RAID subsystems will be under

utilized. Note also that the switch provides multiple point-to-point connections, so multiple servers can simultaneously communicate through the switch to different devices, all at 100 MBps. With only one server and multiple RAID devices, the connection between the server and the switch will limit the potential throughput of the entire disk subsystem.

If you use multiple adapters then you will have multiple fiber links to the switch (assuming the rest of the hardware remains the same). This also implies that the switch will need to zone the ports so that each adapter does not connect to the same device. Typically we would expect at least two adapters per server, if only for redundancy. One adapter would be configured to access all of the A controllers and the other adapter would be configured to access all of the B controllers. The switch would use zoning to group the connections appropriately.

With one server in such a configuration, connecting two FASt500 RAID Controller Units would provide good performance and capacity flexibility. The type of application, I/O size and expected throughput rates would also have to be considered to accurately gauge where any bottlenecks might arise. In general, however, two RAID units should perform well in most application configurations.

If multiple independent servers are to be used with storage partitioning, then additional RAID subsystems may also be considered. This would be addressed more as a classical SAN configuration with the benefits that SAN brings to the table (such as consolidation and centralized management). However, if the multiple servers are all part of the same cluster then the cluster should really be treated as a single server configuration. In failover condition, you are back to the single server situation and could find that the surviving server cannot adequately provide application resources and data throughput to the storage subsystem.

---

### **3.3 Netfinity disk and tape pooling**

As part of a drive towards consolidating network resources, it can be advantageous for cost and management purposes to provide a single, large set of hard disks and tape drives. IBM FAStT products provide the ability to do this as shown in Figure 9:

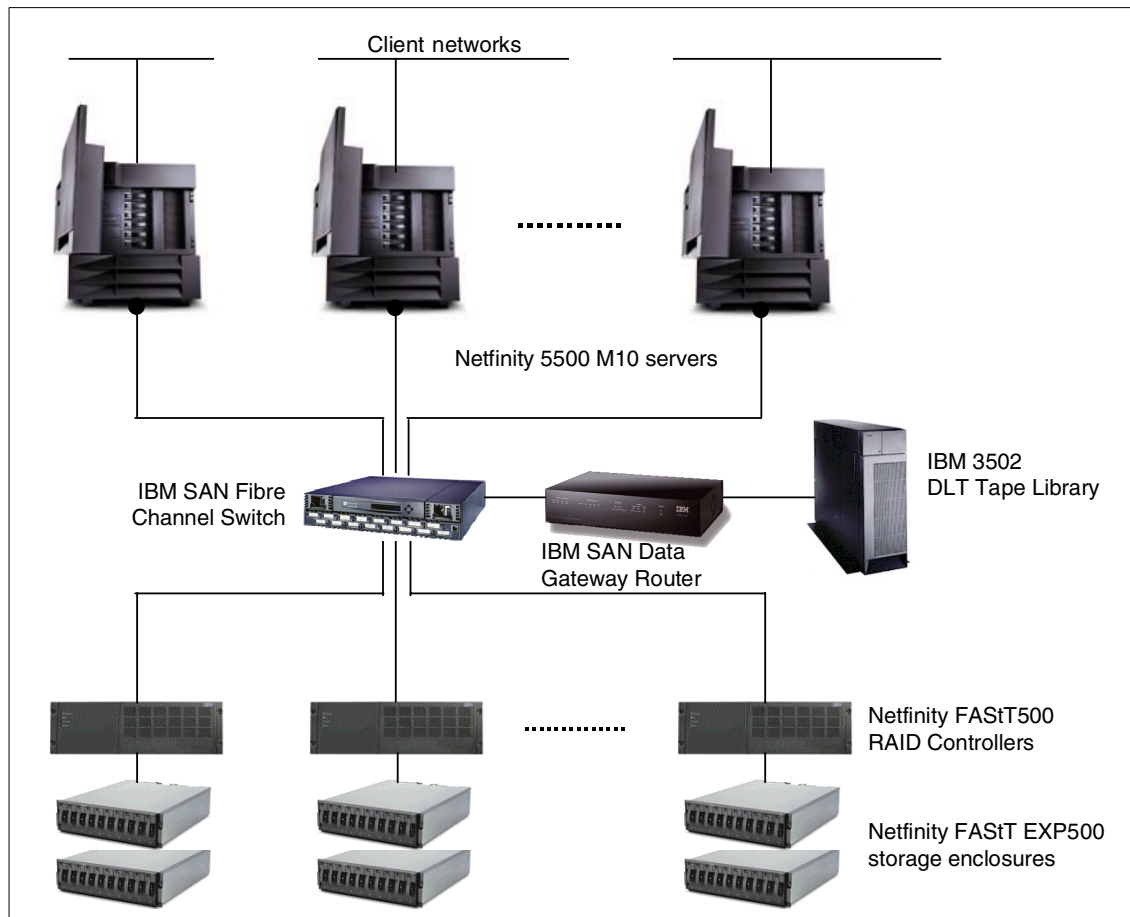


Figure 9. Netfinity disk and tape pooling

This scenario allows multiple servers in diverse locations to utilize a common pool of SAN-attached disk and tape storage devices in a non-clustered environment. A configuration such as this reduces backup time, helps improve disaster protection, and allows both centralized management and a phased approach in implementing a storage area network (SAN).

Disk storage resources are pooled within a disk subsystem or across multiple disk subsystems, and capacity is assigned to independent file systems supported by the operating systems on the servers.

The specifics of this configuration are shown in Table 6:

Table 6. Hardware/software requirements for Netfinity disk and tape pooling

Component	Requirements / recommendations
Software	<ul style="list-style-type: none"> <li>• MS Windows NT 4.0 Server</li> <li>• Netfinity Fibre Channel Storage Manager 7.0</li> <li>• Tivoli Storage Manager 3.7</li> </ul>
Server	<p>One Netfinity server configured with two internal 36.4 GB hard drives for RAID-1 mirroring of the network operating system. The list of supported servers is as follows:</p> <ul style="list-style-type: none"> <li>• Netfinity 5000</li> <li>• Netfinity 5500, 5500 M10, 5500 M20</li> <li>• Netfinity 5600</li> <li>• Netfinity 7000 M10</li> <li>• Netfinity 8500R</li> </ul>
Interconnects	<ul style="list-style-type: none"> <li>• One IBM SAN Fibre Channel Switch 2109-S16</li> <li>• One IBM SAN Data Gateway Router 2108-R3S</li> <li>• One Netfinity FAStT Host Adapter per server (two per server for improved availability).</li> <li>• Netfinity Fibre Channel cables</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• One Netfinity ServeRAID-4L adapter per server for internal disks</li> <li>• Two 36.4 GB Ultra3 SCSI hard disk drives per server</li> <li>• Multiple Netfinity FAStT500 RAID Controllers</li> <li>• Multiple Netfinity FAStT EXP500 Storage Expansion Units</li> <li>• Multiple Fibre Channel Hard Disk Drives</li> </ul> <p>The supported tape libraries are as follows:</p> <ul style="list-style-type: none"> <li>• IBM 3502 DLT Tape Library</li> <li>• IBM Magstar 3570</li> <li>• IBM Magstar 3575</li> </ul>

- Multiple Netfinity servers are connected to the public LAN through the Netfinity 10/100 Ethernet network adapters. Each server is configured with a Netfinity ServeRAID-4L adapter to support two internal hard disk drives using RAID-1 hardware mirroring for the network operating system.
- A Netfinity FAStT Host Adapter is installed in each server, and connected to the IBM SAN Fibre Channel Switch. An IBM SAN Data Gateway Router is connected to the tape library to provide access through the switch.
- One or more disk subsystems such as the Netfinity FAStT500 RAID Controller Units may be connected to the switch to provide access to shared disk storage resources.



- To share the disk storage resources, RAID arrays may be defined using the Netfinity FAST Storage Manager, and allocated to individual servers based on their requirements. Tape backup software, such as Tivoli Storage Manager, may be used to share tape libraries or other tape resources across multiple servers.

### 3.4 Netfinity remote backup, archiving and recovery

Fibre Channel disk subsystems offer the ability to implement new ways to protect your data against system and other failures. An example is given in the following figure:

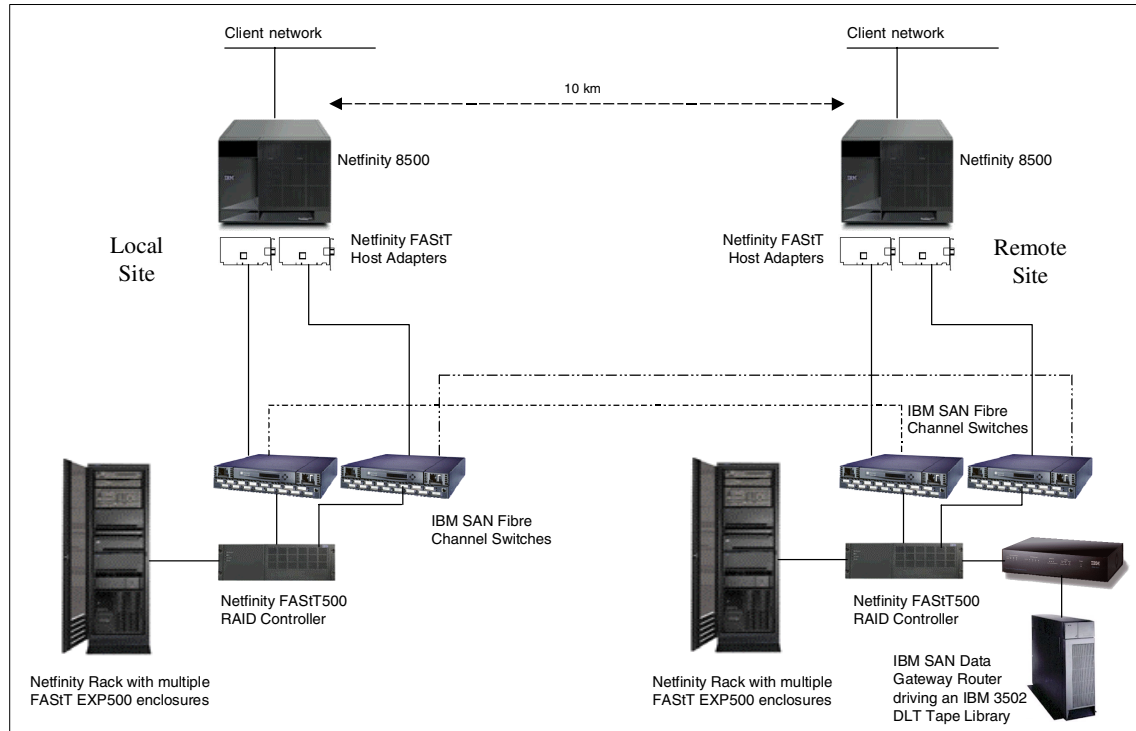


Figure 10. Netfinity remote backup, archiving and recovery

Figure 10 illustrates a Netfinity remote backup, archiving and recovery solution in a non-clustered environment. The long-wave Fibre Channel links between the two sites allow separation distances of up to 10 km, providing excellent protection against disasters. If the local (production) site (to the left in the figure) fails, the remote site can resume the operation by restoring backup data from the tape library.

The specifics of this configuration are shown in Table 7:

Table 7. Hardware/software requirements for Netfinity remote backup, archiving and recovery

Component	Requirements / recommendations
Software	<ul style="list-style-type: none"> <li>• MS Windows NT 4.0 Server</li> <li>• Netfinity Fibre Channel Storage Manager 7.0</li> <li>• Tivoli Storage Manager 3.7</li> </ul>
Server	<p>Two Netfinity servers configured with two internal 36.4 GB hard drives for RAID-1 mirroring on the network operating system. The list of supported servers is as follows:</p> <ul style="list-style-type: none"> <li>• Netfinity 5000</li> <li>• Netfinity 5500, 5500M10, 5500M20</li> <li>• Netfinity 5600</li> <li>• Netfinity 7000M10</li> <li>• Netfinity 8500R</li> </ul>
Interconnects	<ul style="list-style-type: none"> <li>• One Netfinity 10/100 Ethernet network adapter for public LAN per location</li> <li>• Two Netfinity FASt Host Adapters per location</li> <li>• Two IBM SAN Fibre Channel Switches (2109-S16), providing redundancy at each location</li> <li>• One IBM SAN Data Gateway Router 2108-R3S</li> <li>• Netfinity Fibre Channel cables</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• One Netfinity ServeRAID-4L adapter for internal disks</li> <li>• Two 36.4 GB Ultra3 SCSI hard disk drives</li> <li>• Multiple Netfinity Fibre Channel EXP500 storage expansion units</li> <li>• 10 Fibre Channel Hard Disk Drives per storage expansion unit</li> </ul> <p>The supported tape libraries are as follows:</p> <ul style="list-style-type: none"> <li>• IBM 3502-x14 DLT Tape Library</li> <li>• IBM Magstar 3570-cxx</li> <li>• IBM Magstar 3575-Lxx</li> </ul>

- The Netfinity server at the local site is connected to the disk storage through a pair of IBM SAN Fibre Channel Switches in a redundant configuration. A similar configuration exists at the remote site.
- Each switch is connected to its counterpart (up to 10 km away) at the remote location, which provides remote backup, archiving, and recovery capability. Netfinity Fibre Channel long-wave GBICs are used to connect the pairs of switches together through two fiber optic cable links, again providing redundancy.

- The remote site is connected to the tape library through the IBM SAN Data Gateway Router connected to the switch. The tape library is operated using Tivoli Storage Manager running in the remote Netfinity server to perform remote LAN-free backup of the disk resources at the local site.
- The remote site also has disk resources available to it. Should the production site fail, the remote site may recover data from the tape library to its own disk resources, and resume operation from there until the production site can be repaired.

---

### **3.5 Netfinity high-availability solutions using Microsoft Cluster Server**

The next scenario illustrates a clustering environment that allows server-based applications to be made highly available by linking two servers or nodes running Microsoft Windows NT 4, Enterprise Edition or Microsoft Windows 2000 Advanced Server using their clustering technology called Microsoft Cluster Server (Services in Windows 2000) (MSCS). This configuration provides high availability, and reduced planned or unplanned downtime.

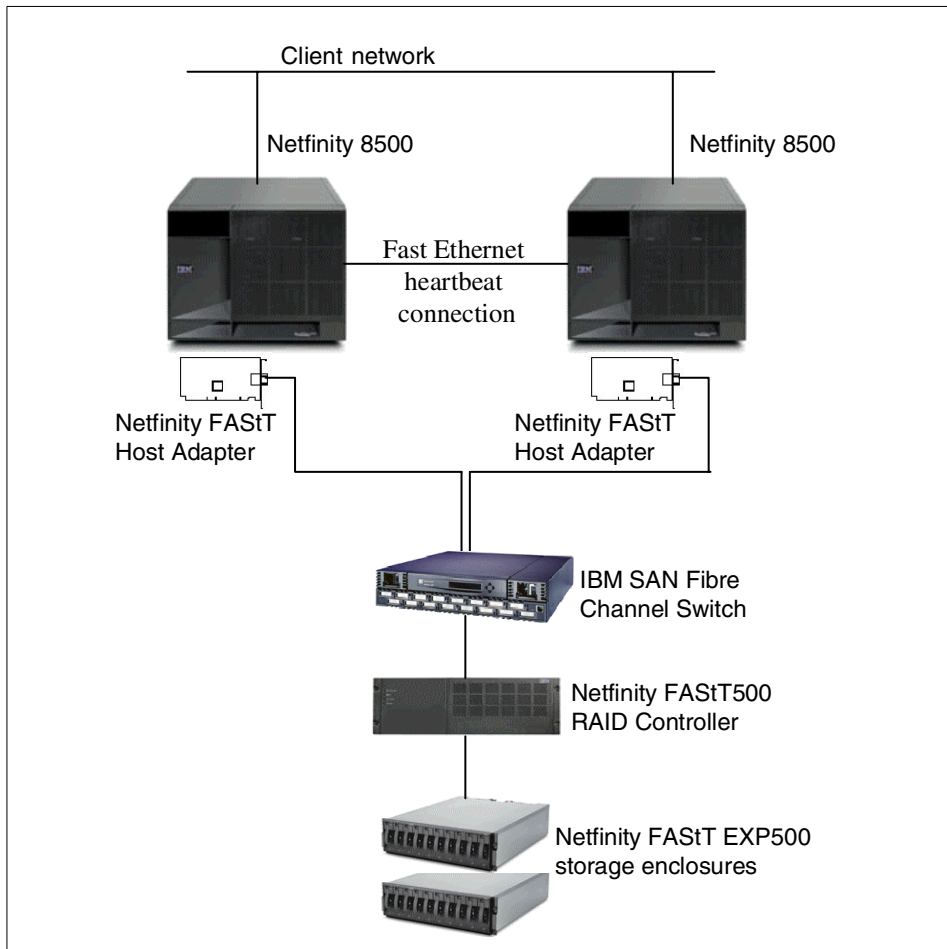


Figure 11. Two-node MSCS cluster using Fibre Channel disk subsystems

The specifics of this configurations are shown in Table 8:

Table 8. Hardware/software requirements using Fibre Channel subsystems

Configuration	Requirements / Recommendations
Software	Microsoft Windows NT 4.0 Enterprise Edition with Service Pack 5 or greater, or Microsoft Windows 2000 Advanced Server
Server	Two identical Netfinity servers certified for clustering. Each server has two identical local hard disks for RAID-1 mirroring of the network operating system.

Configuration	Requirements / Recommendations
Interconnects	<ul style="list-style-type: none"> <li>• Four Netfinity 10/100 Ethernet network adapters to provide connections to the public LAN and for the MSCS heartbeat (other approved interconnects could be used for the heartbeat connection)</li> <li>• Two Netfinity ServeRAID-4L Adapters for internal disks</li> <li>• Two Netfinity FASt Host Adapters</li> <li>• One IBM SAN Fibre Channel Switch 2109-S16</li> <li>• Netfinity Fibre Channel cables</li> </ul>
Storage	<ul style="list-style-type: none"> <li>• One Netfinity FASt500 RAID Controller</li> <li>• Two Netfinity Fibre Channel EXP500 Storage Expansion Units</li> <li>• 10 36.4 GB Fibre Channel Hard Disk Drives per storage expansion unit</li> </ul>

#### Heartbeat connection

Certified components must be used for the MSCS heartbeat connection. We recommend either the Netfinity 10/100 Ethernet Network Adapter or Netfinity 100/10 EtherJet PCI Adapter. If the server comes with an integrated Ethernet adapter, it may be used only to connect to the public LAN, as the integrated adapters are not certified for use as the MSCS heartbeat connection.

Also note that the MSCS heartbeat connection must be point to point, since connection through a hub is not supported.

Although the configuration shown in Figure 11 increases the availability of server resources to your clients, there are several single points of hardware failure that can still cause loss of service that could be avoided. These include:

- The IBM SAN Fibre Channel Switch
- The Netfinity FASt500 RAID Controller
- The Netfinity FASt EXP500 Storage Expansion Units
- Cable connections between the switch and the controller, and controller to the storage expansion units

Any failure in these devices or connections will cause both of the servers to lose access to the storage system. This can be avoided by implementing a fully redundant Fibre Channel disk subsystem. The FASt500 RAID Controller contains a pair of redundant disk controllers as standard and can be used to connect an alternate path to the storage units and to the switch. A second switch is then added to protect against a switch failure. Finally, we

have also added a second host adapter to each server to avoid failover in the event of an adapter failure. This configuration is shown in the following diagram:

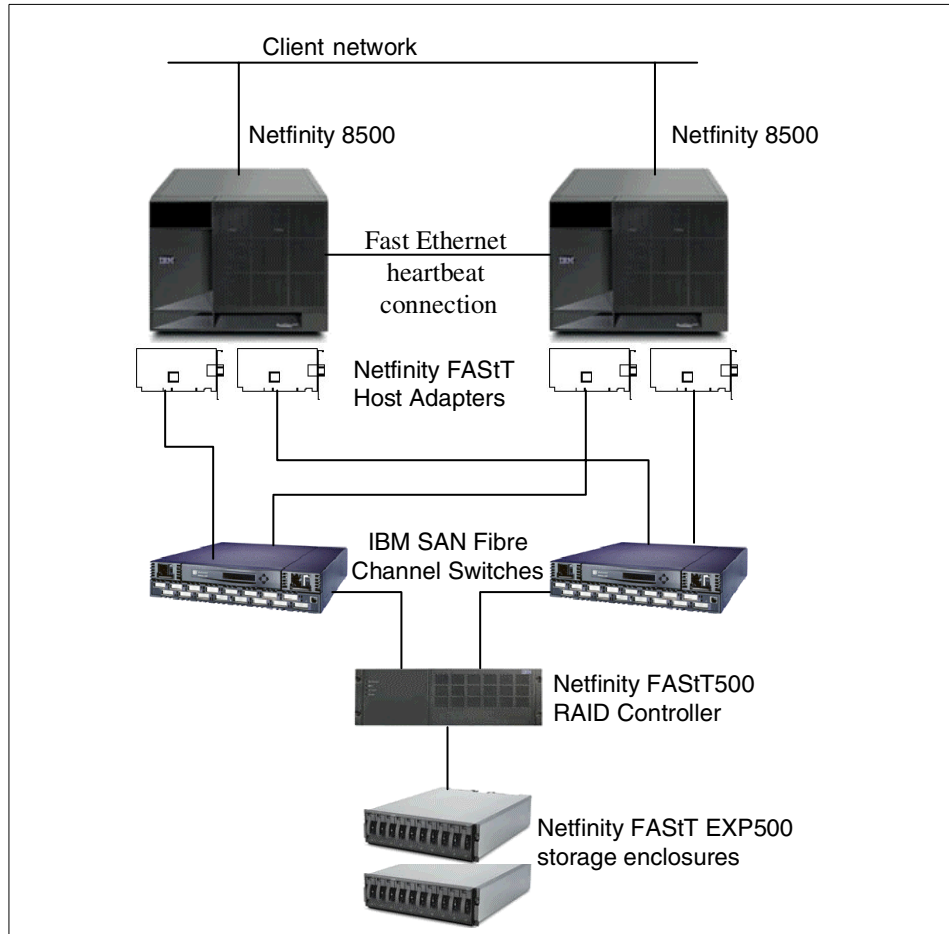


Figure 12. Two node MSCS cluster with a fully redundant Fibre Channel disk subsystem

### Clustering in Microsoft Windows 2000

Clustering is substantially enhanced in the Microsoft Windows 2000 Advanced Server and Datacenter Server operating systems. Apart from improved availability and manageability, these new versions have increased scalability by supporting SMP servers that support a maximum of eight processors for Advanced Server and 32 processors in Datacenter Server. Increased memory capacity is also supported: up to 8 GB of RAM in Advanced Server and 64 GB in Datacenter Server.

The Windows 2000 Advanced Server supports a two-node cluster, as in Windows NT 4 Enterprise Edition. Windows 2000 Datacenter Server will support a four-node cluster. Note that Windows Datacenter Server has not yet been released and details are therefore subject to change.





---

## Part 2. ServeRAID SCSI subsystems



---

## Chapter 4. Introduction to ServeRAID

The chapters in this part of the redbook describe the ServeRAID hardware and how to configure, install, and use ServeRAID adapters in IBM Netfinity servers. New to this fourth edition of the redbook are the following topics:

- Features of the ServeRAID-4H, 4M, 4L and ServeRAID-3HB members of the ServeRAID family of adapters
- The current version of the Windows-based configuration utility, called the *ServeRAID Manager*
- RAID level 5E, supported by the BIOS and firmware Version 3.50 and higher on ServeRAID-3 and ServeRAID-4 adapters
- Spanned array RAID levels 00, 10, 1E0 and 50, supported on ServeRAID-4 adapters
- Integration and management with Netfinity Director

Also, many existing topics from earlier editions of this book have been updated with new information.

---

### 4.1 Netfinity ServeRAID hardware

In this section of the book, we take a closer look at the capabilities of the range of Netfinity ServeRAID SCSI adapters. If you need a brief overview of the major features of these adapters, Table 9 on page 77 summarizes much of the discussion that follows.

The new ServeRAID-4 family consists of three adapters:

- ServeRAID-4H, the high-end offering.

The ServeRAID-4H is a 64-bit PCI adapter that provides four Ultra3 160/m SCSI channels, with two internal and four external SCSI connectors. It supports RAID-0, 1, 1E, and 5E logical drives, and, in addition, RAID-00, 10, 1E0 and 50 logical drives using spanned arrays.

- ServeRAID-4M, the mainstream ServeRAID adapter.

This is a 64-bit PCI adapter with two Ultra3 160/m SCSI channels. Both channels are available for internal or external connections. It supports the same RAID levels as the ServeRAID-4H.

- ServeRAID-4L, the entry level ServeRAID adapter.

ServeRAID-4L is ideally suited to entry-level environments due to the following:

- It supports a single Ultra3 160/m SCSI channel with both an internal and an external connector.
- It has a reduced cache size in comparison to the ServeRAID-4M.
- Its cache is not backed up by a battery.

The ServeRAID-3 family of adapters is also covered in this book. It has three members: ServeRAID-3HB, 3H, and 3L:

- **ServeRAID-3HB**

This is a 64-bit PCI adapter with three Ultra2 SCSI channels. It offers one internal and three external connectors. Supported logical drive RAID levels include RAID-0, 1, 1E, 5, and 5E. It has cache battery-backup implemented as standard. Other enhancements were introduced in BIOS and firmware update Version 3.50, which may also be applied to ServeRAID-3H:

- Elimination of the need for the quorum arbitration cable interconnect in a Microsoft Cluster Server (MSCS) clustering environment.
- Implementation of an adaptive read-ahead cache.
- Support for RAID-5 enhanced (RAID-5E) logical drives.
- Implementation of the FlashCopy function.
- Performance and rebuild recovery enhancements.

- **ServeRAID-3H**

ServeRAID-3H is essentially the same adapter as 3HB, excluding the cache battery-backup option. By adding this cache option, ServeRAID-3H becomes equivalent to the 3HB version.

- **ServeRAID-3L**

This adapter offers entry-level RAID functionality. It is a 32-bit PCI adapter with one Ultra2 SCSI channel. One internal and one external SCSI connector are available, and the controller supports the same RAID levels as the other ServeRAID-3 family members when BIOS and firmware Version 3.50 or higher are used.

---

## 4.2 ServeRAID features and options

Netfinity ServeRAID adapters offer a number of features to give you flexibility in configuring your disk subsystem. We now examine these in some detail.

### 4.2.1 Arrays and logical drives

To configure usable disk space on ServeRAID-attached disks, you must first create *RAID arrays* and *logical drives*. RAID (redundant array of independent

disks) is the technology that groups several disk drives into an array that you can define as one or more logical drives. Each logical drive then appears to the operating system as a single physical drive (for example, Disk 0, Disk 1 and so on, in Windows NT Disk Administrator).

When you group multiple physical disk drives into arrays and logical drives, the ServeRAID controller is able to transfer data from these multiple disk drives in parallel, thereby yielding much higher data transfer rates than that of a single disk. In addition, some RAID configurations will tolerate the failure of one, or even two, disks without loss of data (hence the term redundant). For more information about RAID levels and the options available with the ServeRAID adapters, see 4.2.2, “RAID levels supported by ServeRAID adapters” on page 48.

ServeRAID adapters allow the RAID arrays and logical drives to span multiple SCSI channels within a single adapter. This allows for larger logical drive capacities and, potentially, greater performance levels, as the I/O requests can be distributed evenly across SCSI channels.

ServeRAID-4 adapters introduce a new feature, called *spanned arrays*. These are basically arrays of arrays. Spanned arrays support RAID-00, -10, -1E0 and -50 logical drives. There are two important benefits of spanned arrays:

- Spanned arrays can incorporate a much higher number of physical disk drives than a standard RAID array. Standard arrays are limited to a maximum of 16 disk drives. A spanned array can comprise up to 60 physical disk drives, which can then be defined as a single logical drive. In other words, all disk drives on all SCSI channels of a ServeRAID-4H adapter can be used to form a single logical drive.
- You can achieve higher availability. A standard RAID-5 array can tolerate only a single disk failure, whereas a RAID-50 spanned array can tolerate multiple disk failures. It is important to note, however, that only specific combinations of multiple disks failures can be tolerated (one drive per RAID-5 array). An example of an application of this enhanced redundancy, you can easily configure a RAID-50 logical drive across several external disk enclosures so that the disk subsystem remains operational even when an entire external enclosure fails.

As a summary, the following is a list of the capacities of RAID arrays and logical drives as supported by the ServeRAID family:

- Up to 15 drives per channel (depending on the capabilities of the disk enclosures used).
- RAID arrays spanning multiple channels on the same adapter.

- Up to 8 RAID arrays per adapter.
- Up to 16 hard disks per RAID array for any stripe-unit size.
- Logical drives of RAID-0, 1, 1 enhanced (1E), 5 and 5 enhanced (5E) using ServeRAID-3, and, additionally, RAID-00, 10, 1E0 and 50 using ServeRAID-4.
- Up to 8 logical drives per adapter.
- Up to 16 physical disk drives per array in a non-spanned array on ServeRAID-3 and ServeRAID-4.
- Up to 60 physical disk drives per array in a spanned array on ServeRAID-4.
- Up to 12 ServeRAID-3 or -4 adapters per server (depending on the capabilities of the server) with BIOS and firmware 3.50 and higher. (This is for Windows NT and Windows 2000; other operating systems support up to eight adapters.)

**Note:** Devices using multiple SCSI logical unit numbers (LUNs) are not supported by the ServeRAID adapters.

## 4.2.2 RAID levels supported by ServeRAID adapters

You can choose among several different RAID levels for your logical drives. Each different RAID level provides different performance, fault-tolerant, and disk space efficiency characteristics. For example, customers who wish to provide fault tolerance for their disk subsystem, and are also concerned about disk space efficiency, might want to use RAID-5 logical drives. Alternatively, customers who wish to achieve the highest performance they can get, and possibly do not require fault tolerance, might select RAID-0.

### 4.2.2.1 Hardware or software RAID?

ServeRAID adapters provide RAID support at the hardware level. It is also possible to use software RAID implementations, which can be provided by the operating system or by third-party utilities. For example, most network operating systems support partition mirroring, which is equivalent to RAID-1. However, hardware RAID, as provided by IBM's ServeRAID adapters, has the following advantages over software RAID:

- All RAID calculations are performed by the adapter and therefore place no load on the system processor.
- All read and write accesses to redundant information (mirrored data, RAID-5 parity) are transparent to the operating system when using

ServeRAID. With software RAID, the operating system must handle all additional disk accesses, which decreases performance.

- Hot-pluggable disk drives are not supported by software RAID. This means the server needs to be shut down in order to replace a failed disk drive. With hardware RAID, replacement can be done while the server stays up and running.
- Recovery procedures after a disk drive failure are complex when using software RAID. For example, you have to use Windows NT Disk Administrator to reestablish the fault-tolerant disk environment. With ServeRAID, you simply replace the disk drive and the adapter handles the complexity for you.
- Hardware RAID is more flexible: you can use any available RAID level for your operating system. This is usually not the case with software RAID. For example, the Windows NT RAID-5 equivalent, Stripe Set with Parity, cannot be used for the bootable partition.
- Using logical drive migration, it is very easy to add new disk space to hardware RAID environment or even change RAID levels. Again, this is usually not supported by software RAID.

#### 4.2.2.2 RAID-0

As shown in Figure 13, sequential data blocks are evenly distributed (or *striped*) across all disk drives. This is excellent for performance, since the data accesses on different disk drives can be done in parallel. Performance improves as the number of disk drives used in the array increases.

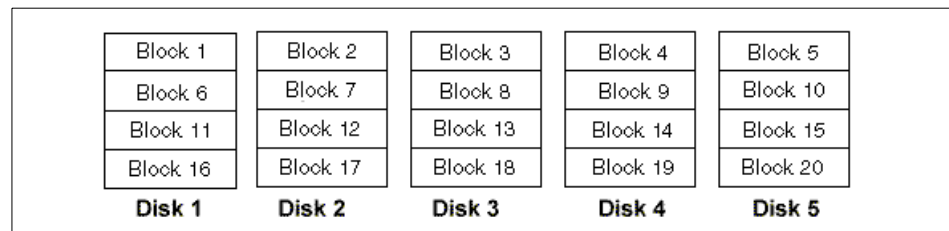


Figure 13. RAID-0 data organization

Among all of the supported RAID levels, RAID-0 offers the best performance from the disk subsystem, but does not provide any fault tolerance. It is obvious from Figure 13 that if, say, Disk 2 fails, blocks 2, 7, 12, and 17 will be lost. Partition information, such as file and directory allocation data, is spread across all disk drives, and the data in the entire logical drive will not be recoverable.

In fact, the probability of data loss due to a disk drive failure in a RAID-0 logical drive is even higher than that for a single disk drive. This is because, assuming a fixed failure rate for any single drive, the failure rate for an n-drive array will be n times the rate for the single drive.

As a result, RAID-0 is not widely used in typical customer environments. In most cases, customers are motivated to use fault tolerance to protect the availability of their data rather than seek the ultimate possible performance. However, RAID-0 can be useful in special cases, when extremely fast disk subsystem performance is required and data availability is not so important.

Many operating systems have a software equivalent of RAID-0. For example, Windows NT Stripe Sets work in the same manner as RAID-0, but at a lower performance level than can be achieved with ServeRAID.

ServeRAID adapters allow a RAID-0 logical drive to span up to 16 physical disk drives.

#### 4.2.2.3 RAID-1

RAID-1 is a simple and yet efficient fault-tolerant RAID implementation. It uses exactly two physical disk drives. The data is mirrored between the drives, so that at any given moment they contain identical data. If a drive fails, the data it held can still be accessed from its mirror copy and the server remains operational.

Data read performance is better than that of a single disk drive. All the data is available on both disk drives, so while one drive is being accessed for reading, the ServeRAID adapter can simultaneously read the next block of data from the other disk drive.

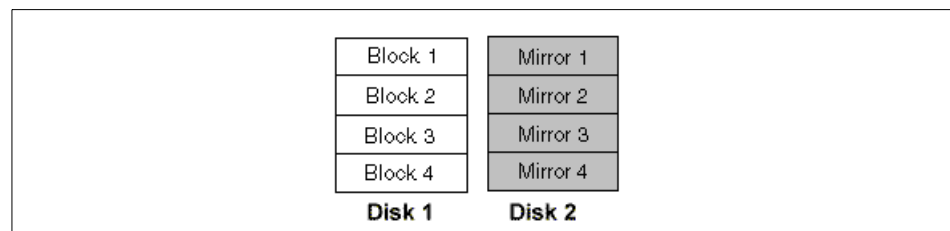


Figure 14. RAID-1 data organization

The biggest drawback of RAID-1 is cost. Since all data is mirrored, only 50% of the total physical storage capacity is available for data.



The software equivalent to RAID-1, mirroring of partitions, is available in many network operating systems.

#### 4.2.2.4 RAID-1 Enhanced

This RAID level, sometimes termed RAID-1E, combines the striping of sequential data blocks (as in RAID-0) with mirroring (as in RAID-1). The idea is to keep the fault tolerance of RAID-1 and the performance advantage of RAID-0. As you can see in Figure 15, sequential data blocks are evenly distributed across all the disk drives and the same is true for mirrored data blocks.

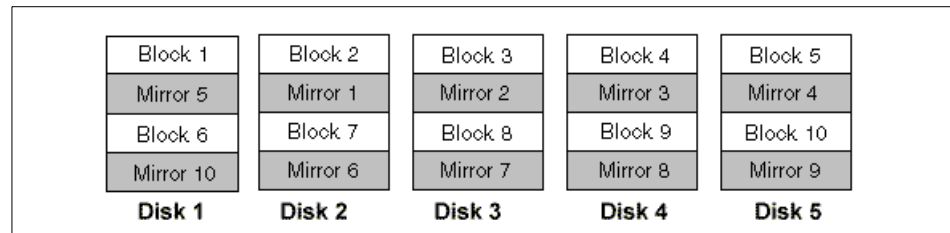


Figure 15. RAID-1 Enhanced data organization

This approach works very well and RAID-1 Enhanced ranks among the fastest of the fault-tolerant RAID levels. Performance increases with the number of disk drives used.

It does, however, have the same cost implications as RAID-1: because all data is mirrored, only 50% of storage space will be available for data.

Using ServeRAID adapters, you need a minimum of three and up to 16 physical disk drives for RAID-1 Enhanced logical drives.

#### 4.2.2.5 RAID-5

This RAID level is probably the most widely used in customer environments. It provides fault tolerance, and performs relatively well. Perhaps its biggest advantage is that it is more cost effective than either RAID-1 or RAID-1 Enhanced.

As you can see in Figure 16 on page 52, a checksum is calculated for each stripe (set of sequential blocks) and also written to the disk drive. Checksum blocks are evenly distributed across the disk drives to improve performance. The checksum data has to be accessed whenever data is written to the disks. If they were kept on a single dedicated disk drive (creating a RAID-4 array), then this drive would become a performance bottleneck. This is why RAID-4 is rarely implemented by RAID adapter manufacturers.

RAID-5 is very cost effective. The equivalent of one disk drive capacity is sacrificed to provide fault tolerance. For example, if you use five disk drives, 80% of the total physical capacity can be used for data as illustrated in Figure 16:

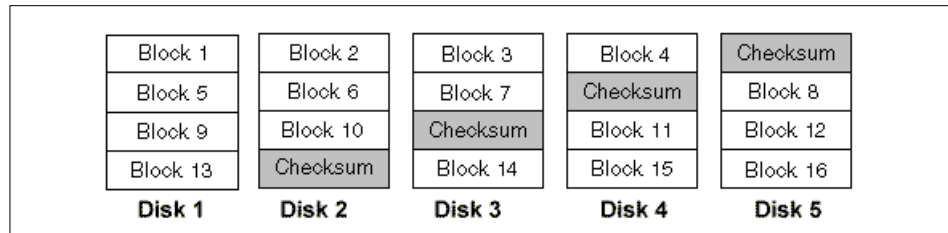


Figure 16. RAID-5 data organization

Data read performance is very good and it approximates RAID-0. When reading, only data blocks are accessed and they are distributed in a manner very similar to RAID-0. Writing performance is a bit slower. To write a block of data, the following must happen:

1. The old data and checksum blocks have to be read.
2. The new checksum is calculated using the data in the old data block, the new data block, and the old checksum.
3. The new data block and the new checksum are written to the disk drives.

Note that an operation that would be a single write in a RAID-0 array requires two reads and two writes in a RAID-5 array.

Efficient caching can greatly reduce the impact of this process, but writing performance remains relatively low compared to the other RAID levels discussed. However, in a typical customer environment, there will be a lot more reading of data than writing. RAID-5 is optimized for reading so it will usually provide reasonable performance. Performance increases with the number of disk drives used.

If a disk drive fails, the data can still be accessed. However, the performance is impacted. Instead of simply reading the data block on a failed disk drive, the following process must take place:

1. All of the other data blocks in the stripe must be read.
2. The checksum block for the stripe must be read.
3. The missing data is calculated using the checksum and the other data blocks.

So, instead of only one disk access, the ServeRAID adapter has to read the blocks from all surviving disk drives. When a RAID-5 logical drive is operating with a physical drive failure, it is said to be in a *critical* state, since the failure of another drive will result in loss of data. From the discussion above, it is apparent that performance when in a critical state deteriorates as the number of drives in the logical drive increases.

ServeRAID adapters also support the use of a *hot-spare*. When a hot spare is configured, the data on a failed disk drive will automatically be rebuilt to the hot spare disk drive, while the server remains operational, thus returning the array to a non-critical state. Without a hot spare, the failed disk drive has to be replaced as quickly as possible, because RAID-5 logical drives cannot tolerate another disk drive failure.

An example of a software RAID-5 implementation would be Stripe Sets with Parity, available in Windows NT.

You must use at least three physical disk drives in order to configure a RAID-5 logical drive. ServeRAID adapters support a maximum of 16 physical disk drives in a single RAID-5 array.

#### 4.2.2.6 RAID-5 Enhanced

This RAID level, also known as RAID-5E, was introduced with Version 3.50 of ServeRAID BIOS, firmware and utilities. It is supported on ServeRAID-3 and ServeRAID-4 adapters and is functionally equivalent to a RAID-5 array with a hot spare. In a RAID-5E array, however, the hot spare drive is distributed across the drives of the array, providing better performance. To illustrate this, consider Figure 17 below, which shows a standard RAID-5 array with a hot spare disk drive. In normal operation, the hot spare does not participate in the array, remaining idle unless a drive in the array fails.

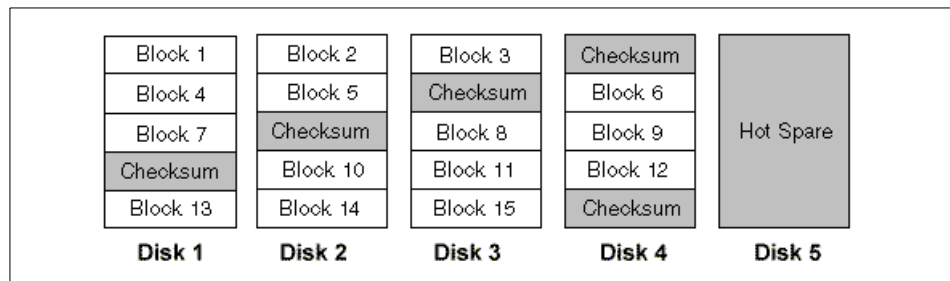


Figure 17. RAID-5 data organization with a hot spare

By incorporating Disk 5 within the array and distributing the space for the hot spare function across the drives of the array, you can get increased performance, because you now have one extra drive providing parallel disk access. This way of organizing the data is shown in Figure 18. Observe that we no longer have a disk drive that acts as a dedicated hot spare; instead the hot spare disk space is spread evenly across all disk drives. The hot spare space is kept unoccupied, so that the data can be automatically rebuilt if a disk drive fails.

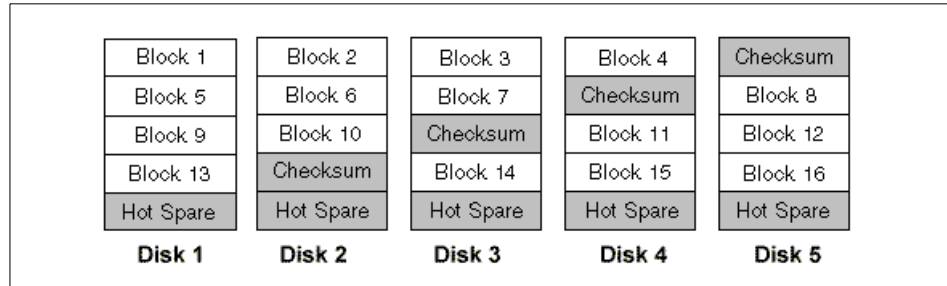


Figure 18. RAID-5 Enhanced data organization

RAID-5 Enhanced offers better performance than RAID-5. On top of that, it can actually tolerate two disk drive failures. When the first disk drive fails, the data will be rebuilt and the logical drive will convert to ordinary RAID-5. The hot spare space ensures this can be done. The RAID-5 logical drive can now tolerate another disk drive failure without losing access to the data. There is one proviso: the second drive must not fail before the data rebuild process fully completes.

RAID-5 Enhanced requires a minimum of four disks and supports a maximum of 16 disks. The overall capacity of the array is reduced by the capacity of two disk drives.

There are two limitations you should be aware of when using RAID-5E arrays. First, only one logical drive is supported in a RAID-5E array. This is not usually a problem since, for performance reasons, we recommend that arrays normally have only one logical drive defined within them. The second limitation is that the hot spare function in a RAID-5E array is dedicated to that array. A normal hot spare (as in Figure 17) can be used by any array controlled by the adapter on which it is hosted.

#### 4.2.2.7 RAID-00

RAID-00 is supported by ServeRAID-4 adapters and provides striping of data blocks across several RAID-0 *sub-logical drives*. The main reason RAID-00

and the other RAID-x0 levels have been implemented on ServeRAID-4 is to increase the number of physical disk drives that can be incorporated into a single logical drive. These new levels allow you to use all disk drives on all of a single ServeRAID-4 adapter's SCSI channels in a single logical drive. This means that logical drives can now use up to 60 physical disks. Figure 19 shows the data organization for a RAID-00 logical drive:

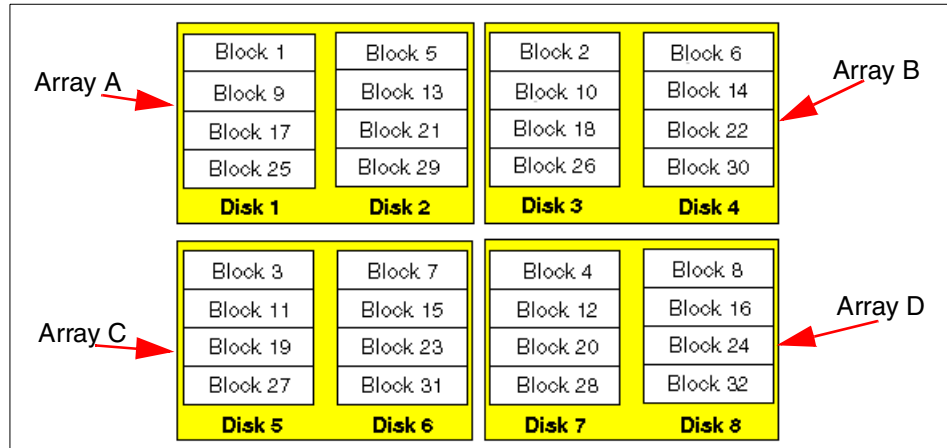


Figure 19. RAID-00 - striping across RAID-0 sub-logical drives

In this particular example, we have created four arrays (using two physical disks for each array), and then created a spanned array across them and finally, we created a RAID-00 logical drive in the spanned array.

#### 4.2.2.8 RAID-10

RAID-10 is the second new RAID level offered by ServeRAID-4 adapters. It provides striping of data across several RAID-1 sub-logical drives. This way, you achieve the performance of RAID-0 and also fault tolerance of RAID-1. Another benefit is that the number of physical disk drives can be larger than with RAID-1 logical drives, which are limited to two physical disks. Because sub-logical drives use mirroring, you can only use 50% of the total disk capacity for data. Figure 20 on page 56 shows the way data is organized in a RAID-10 logical drive:

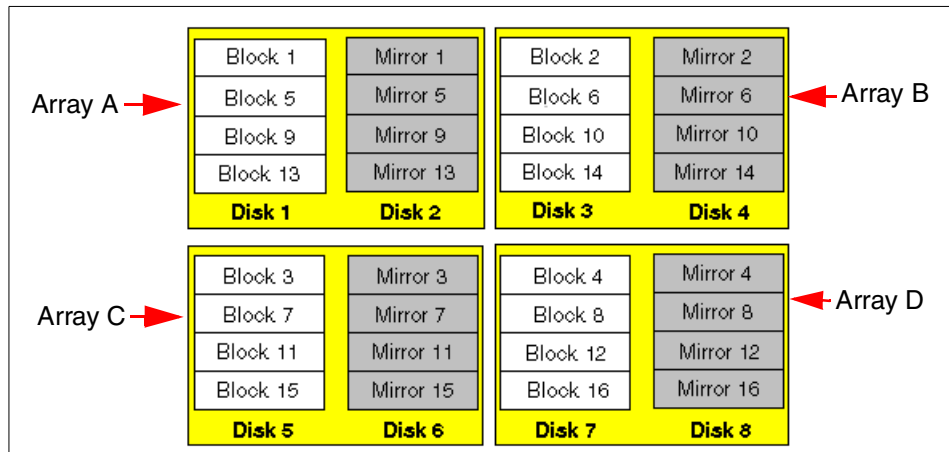


Figure 20. RAID-10 - striping across arrays with RAID-1 sub-logical drives

You can configure RAID-10 through the Spanned Arrays option of the ServeRAID configuration tool. Follow these steps:

1. Define several two-drive arrays (these will contain RAID-1 sub-logical drives).
2. Define a spanned array across your two-drive arrays.
3. Create one or more RAID-10 logical drives in the spanned array. This will implicitly create RAID-1 sub-logical drives on the two-drive arrays and data will be striped across those sub-logical drives.

#### 4.2.2.9 RAID-1E0

The third new RAID level offered by ServeRAID-4 adapters is called RAID-1E0. It stripes data across several RAID-1E sub-logical drives. The principle is the same as for the other RAID-x0 levels already discussed. RAID-1E0 mirrors each data block, so 50% of total capacity is available for data storage. It offers better fault tolerance than plain RAID-1E. A disk drive can fail in each of the sub-arrays and the disk subsystem still remains operational. However, if several disk drives fail in any single sub-array, the RAID-1E0 logical drive is no longer accessible.

Figure 21 shows a RAID-1E0 implementation.

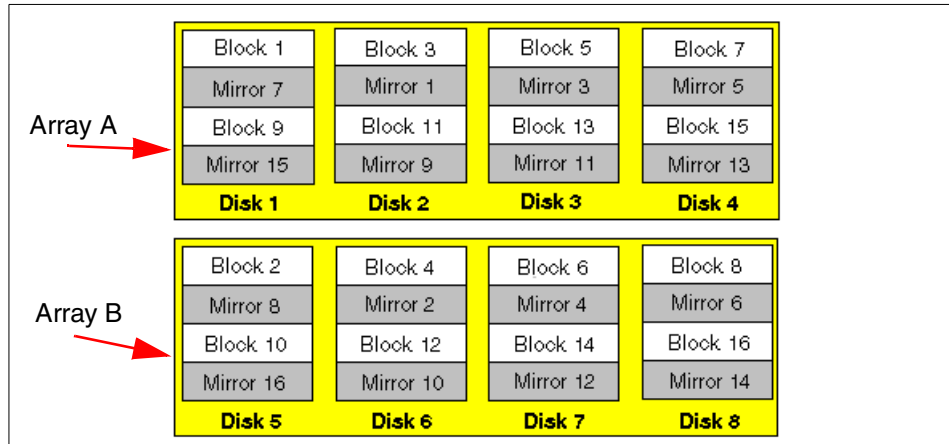


Figure 21. RAID-1E0 - striping across arrays with RAID-1E sub-logical drives

In this example, we have created two arrays, A and B, with four physical disk drives in each. Then, we have created a spanned array across A and B and a RAID-1E0 logical drive in that array.

#### 4.2.2.10 RAID-50

RAID-50 is built in a similar fashion to the other RAID-x0 levels and is also only supported on ServeRAID-4 adapters. The sub-logical drives are created as RAID-5 arrays. A minimum of three disks and a maximum of 16 can be used for each sub-array. The usable data capacity varies according to the number of disk drives in sub-arrays and to the number of sub-arrays themselves. Since each sub-array uses a RAID-5 sub-logical drive, the capacity of one disk drive per each sub-array is sacrificed for fault tolerance. The overall fault tolerance of a RAID-50 spanned array is better than that for a standard RAID-5 array, because the subsystems can tolerate a single disk drive failure in each sub-array without losing access to your data. Note, however, that a second disk drive failure in any single sub-array will make the logical drive inaccessible.

Figure 22 on page 58 shows an example of a RAID-50 logical drive, created as a spanned array with two sub-arrays A and B. Five physical drives are used in each sub-array:

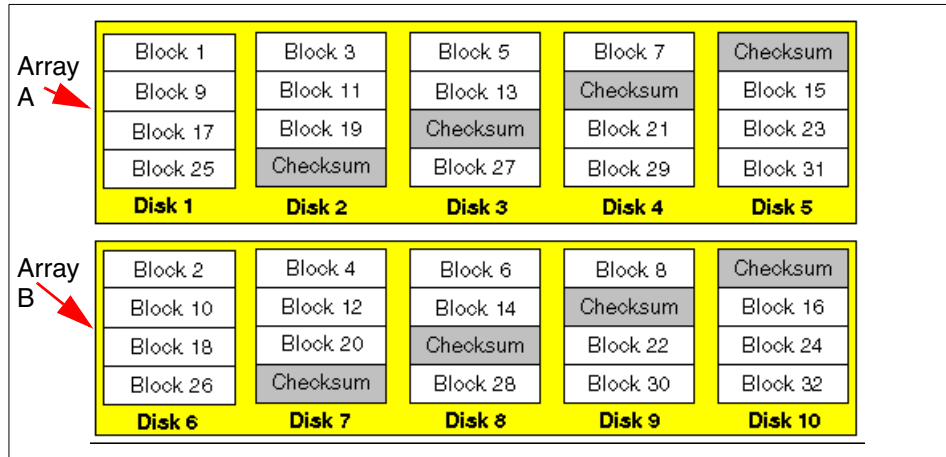


Figure 22. RAID-50 - striping across arrays with RAID-5 sub-logical drives

### 4.2.3 ServeRAID-4H adapter and Ultra3 160/m SCSI

This section discusses the significant new features implemented in ServeRAID-4 adapters, many of which are a result of support for the Ultra3 160/m SCSI interface.

#### Ultra3 160/m SCSI

Ultra3 160/m SCSI provides a subset of the full Ultra3 SCSI specification. It supports features that include Double Transition clocking, CRC and Domain Validation. It does not include some Ultra3 SCSI features, such as Packetization and Quick Arbitration.

For information about SCSI 3 specifications, see:

<http://www.t10.org/scsi-3.htm>

Both the data and address buses in the PCI interface of ServeRAID-4H adapter are 64-bits wide. The adapter provides four Ultra3 160/m SCSI channels. Up to 15 devices are supported on each channel, to give a total of up to 60 devices per adapter. All four channels have connectors available on the adapter backplate to allow attachment to external storage enclosures. Two of the channels also have internal connectors on the adapter for connection to drives inside the host system. A single channel must not be connected to both internal and external drives.



The maximum theoretical throughput for each channel is 160 MBps. Ultra3 160/m uses the same clock frequency as Ultra2 SCSI, but data transfers occur on both rising and falling edges of the clock signal, effectively doubling the throughput. This feature is called *Double Transition (DT)* clocking.

**Note:** Double Transition clocking requires low-voltage differential (LVD) signalling. On a single-ended SCSI bus, clocking will revert to *Single Transition* mode, with a maximum throughput of 80 MBps.

If you use a mixture of Ultra3 and Ultra2 devices on an LVD-enabled SCSI bus, the Ultra2 devices will operate at Ultra2 speed (80 MBps), but Ultra3 devices can still communicate at Ultra3 speed (160 MBps).

In addition, Ultra3 160/m SCSI can use a cyclic redundancy check (*CRC*) to ensure data integrity and is therefore far more reliable than older SCSI implementations that only support parity control.

*Domain Validation* is another feature of Ultra3 160/m SCSI. It is performed during the SCSI bus initialization and the intent is to ensure that devices on the SCSI bus (=domain) can reliably transfer data at the negotiated speed. Only Ultra3 capable devices can use Domain Validation.

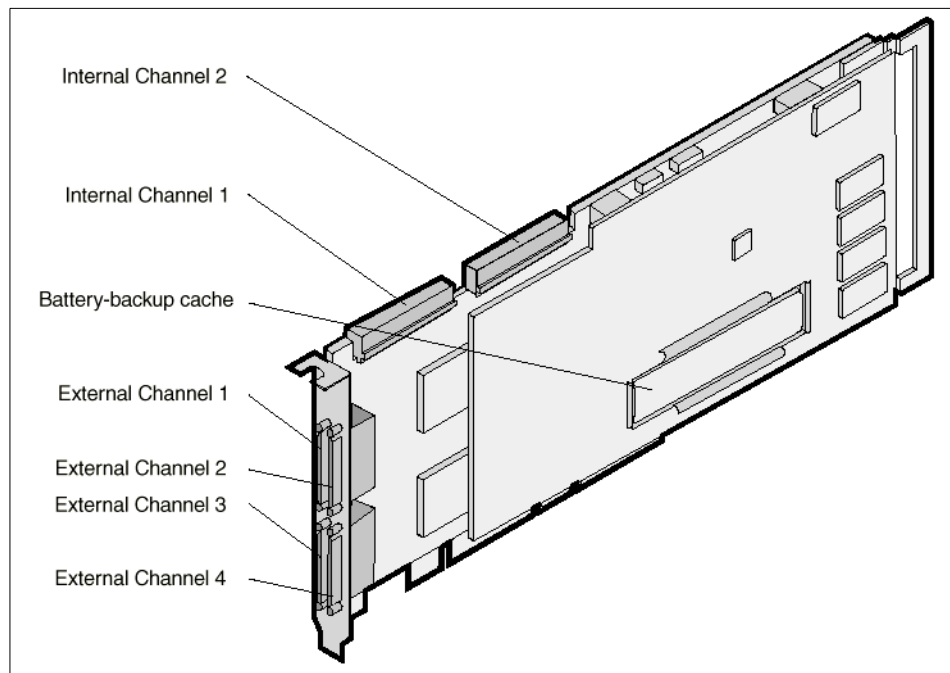


Figure 23. ServeRAID-4H

The ServeRAID-4H adapter, shown in Figure 23 above, uses Power PC 750 processor running at 266 MHz. It has a cache size of 128 MB and the cache backup battery is standard. The cache is not mirrored.

A new feature supported by ServeRAID-4H is *spanned arrays*, which implement logical drive RAID levels 00, 10, 1E0 and 50. These are discussed in 4.2.2, “RAID levels supported by ServeRAID adapters” on page 48.

Certain limitations apply when spanned arrays are used:

- Logical drive migration of spanned arrays is not supported.
- Spanned arrays are not supported in clustering environment.
- Spanned arrays are not supported when you install two ServeRAID-4H adapters as a fault-tolerant pair.

Finally, the ServeRAID-4H adapter also handles SCSI disconnects and reconnects more efficiently than the predecessor ServeRAID-3 adapters.

#### **4.2.4 ServeRAID-4M adapter**

The ServeRAID-4M is a 64-bit PCI adapter that offers two Ultra3 160/m SCSI channels. Both channels can be attached either internally or externally (but not simultaneously), and each can support up to 15 devices. The adapter uses an Intel 960RN processor and has 64 MB of battery-backup cache.

#### **4.2.5 ServeRAID-4L adapter**

The 64-bit PCI ServeRAID-4L adapter is the entry-level member of the ServeRAID-4 family. It provides one Ultra3 160/m SCSI channel that can be attached either internally or externally (but not simultaneously). Up to 15 devices are supported. This adapter also uses an Intel 960RN processor but differs from the ServeRAID-4M by having smaller cache (16 MB), which does not have the battery-backup feature.

#### **4.2.6 ServeRAID-3HB adapter**

The ServeRAID-3HB adapter provides three Ultra2 SCSI channels and has one internal and two external connectors, as shown in Figure 24. The internal connector is a standard 68-pin SCSI-2 F/W connector and the external connectors are standard 0.8 mm VHDCI (very high density connector interface) connectors. These connectors support both single-ended and Low Voltage Differential Signaling (LVDS) SCSI interfaces.

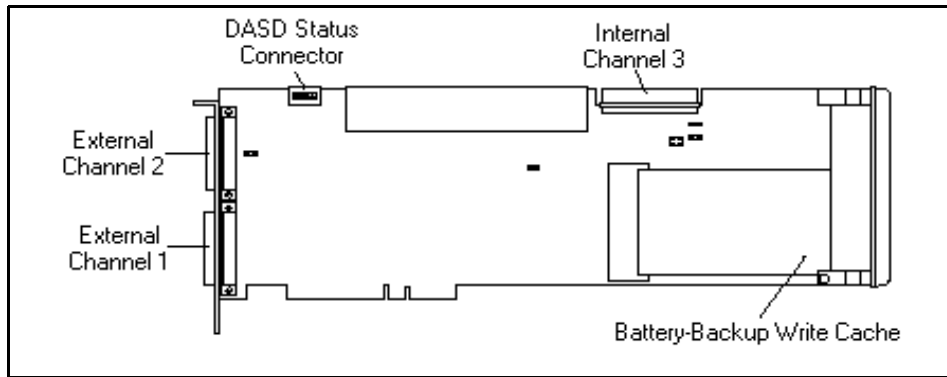


Figure 24. The ServeRAID-3HB adapter (component-side view)

Each channel of the ServeRAID-3HB adapter can support up to 15 devices, for a total of 45 devices. Channels 1 and 2 can only be connected through the external connectors. Channel 3 is connected either via the internal connector or externally, by fitting the IBM ServeRAID Channel 3 Cable Option Kit (supplied with the adapter). In this way, all three channels may be used to connect to external devices. The adapter with the cable option kit installed is depicted in Figure 25.

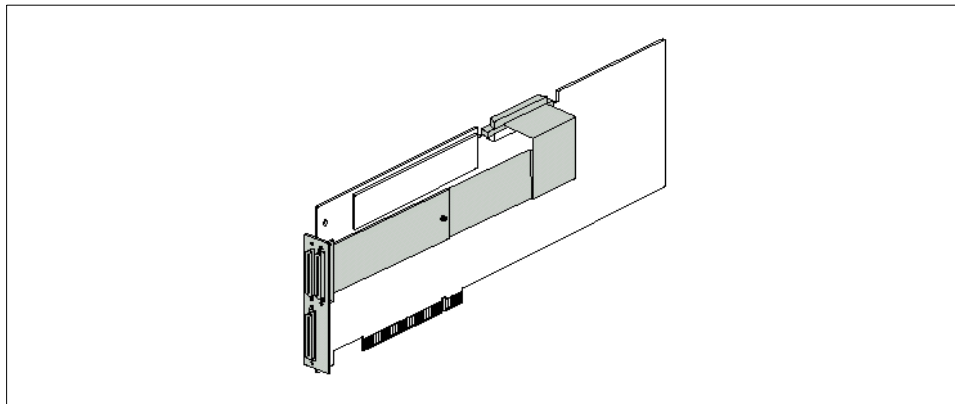


Figure 25. IBM ServeRAID Channel 3 Cable Option Kit

#### 4.2.7 ServeRAID-3L adapter

The ServeRAID-3L adapter provides a single Ultra2 SCSI channel that supports up to 15 devices. This channel has both internal and external connectors, as shown in Figure 26 on page 62. However, you cannot use both the internal and external connector at the same time.

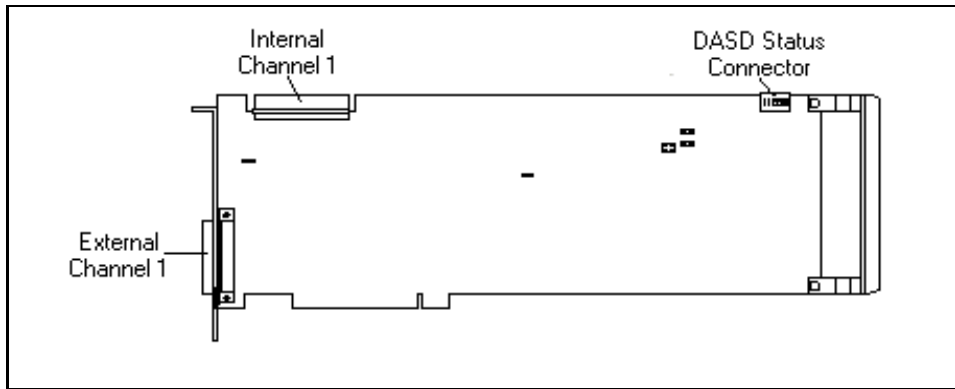


Figure 26. The ServeRAID-3L adapter (component-side view)

Just as for the ServeRAID-3HB, the ServeRAID-3L's external connector is an industry-standard 0.8 mm VHDCI (Very High Density Connector Interface) connector. This connector supports both Low Voltage Differential Signaling (LVDS) SCSI and single-ended SCSI cables.

## 4.2.8 Older ServeRAID adapters

### 4.2.8.1 ServeRAID-3H adapter

The ServeRAID-3H Adapter was the forerunner to the ServeRAID-3HB. The battery-backup cache was not included as standard, but was available as an option.

The adapter had three Ultra2 SCSI channels and each could support up to 15 devices. Channels 1 and 2 were available externally and channel 3 internally. With the IBM ServeRAID Channel 3 Cable Option Kit, all three channels were available for external connections.

### 4.2.8.2 ServeRAID II adapter

The ServeRAID II adapter offered three Ultra SCSI channels through three internal and two external connectors.

Each channel of the ServeRAID II adapter could support up to 15 devices, for a total of 45 devices. The SCSI channels 1 and 2 could be connected either through the external connectors or through the internal connectors but not both. Channel 3 was connected via the channel 3 internal connector or, by using a third channel cable/bracket kit, it could be connected externally, allowing all three channels to be used to connect to external devices.

**Note:** The third channel cable/bracket kit was not included with the ServeRAID II and had to be ordered separately.

#### **4.2.8.3 ServeRAID adapter**

The IBM ServeRAID SCSI Adapter had three Fast/Wide SCSI-2 channels. Each SCSI channel could support up to 15 devices, giving a total of 45 devices per adapter.

Internal connectors were available for all three SCSI channels, but only SCSI channel 1 was available externally.

SCSI channels 2 and 3 could also be made available for external connections by using the 16-bit SCSI Internal Bulkhead Cable option. This option routed cables from the internal channel connectors to knockout ports in the server chassis.

#### **4.2.9 LVDS SCSI connectivity**

ServeRAID-4 and ServeRAID-3 adapters support the connection of Low Voltage Differential Signaling (LVDS) SCSI devices as well as standard single-ended (SE) SCSI devices. For performance reasons, you should avoid mixing SE and LVDS SCSI devices on the same channel.

With the advent of Ultra2 SCSI speeds, permissible cable lengths that will support single-ended devices have become too short for any practical use (approximately 75 cm). To compensate for this, a new connection standard has been developed which uses the signaling aspects of the older SCSI differential standard without its associated costs.

Differential SCSI technology transmits each signal along two cables rather than one. The signals transmitted along a pair of cables are identical but in opposite polarity to each other. Using this technique, cable lengths of up to 12 m can be achieved, thanks to the improved immunity to noise and interference signals.

To implement an LVDS SCSI solution, the adapter, cabling, disks and disk enclosures must all support LVDS.

**Note:** The ServeRAID II and ServeRAID adapters do not support LVDS SCSI devices.

#### 4.2.10 “Optimal” SCSI speed

When you configure the ServeRAID adapters using the Windows-based configurator, you have the option to specify the SCSI speed of each of the SCSI channels on the adapters.

The speed option Optimal was introduced with ServeRAID-3H and ServeRAID-3L. It is also available for the ServeRAID-3HB and ServeRAID-4 adapters. When selected, it lets the adapter determine the highest transfer speed, based on the types of SCSI drives and storage enclosures in use. The Optimal setting is not available for ServeRAID II nor for the original ServeRAID adapter.

#### 4.2.11 64-bit PCI data path

ServeRAID-4, ServeRAID-3HB and ServeRAID-3H adapters have a 64/32-bit PCI data path, allowing them to be installed in either a 32-bit or a 64-bit PCI slot in the server.

The maximum theoretical throughput of a 32-bit 33 MHz PCI bus is 132 MBps. A 64-bit 33 MHz PCI bus can handle up to 264 MBps. When all three channels on the adapter are operating at Ultra2 SCSI speeds (80 MBps theoretical maximum), which is the case for ServeRAID-3HB and ServeRAID-3H, a 32-bit PCI slot may not be sufficient to carry burst transfers at these speeds, so a bottleneck may occur on the PCI bus. The situation is made worse with the ServeRAID-4H adapter with its four channels, each of which is capable of 160 MBps. For this reason, you should install your ServeRAID adapters in 64-bit slots whenever possible.

64-bit PCI slots are implemented in many of the mid-range and high-end Netfinity servers, and are likely to be offered in additional systems in the future.

#### 4.2.12 ServeRAID adapter cache

Each of the ServeRAID adapters has cache memory installed. The ServeRAID-4 adapters use ECC SDRAM memory; all other adapters use 60 ns EDO memory. Table 9 on page 77 lists the cache sizes of each of the ServeRAID adapters. The caches operate only when the adapter write policy for a logical drive is set to write-back (WB) mode. They offer no additional benefit for logical drives that are set to write-through (WT) mode.

Write-back cache operates by indicating to the operating system that a write to disk is complete before the actual write to the drive has occurred. The data is temporarily stored in the cache and written out to disk some time later. This

can offer greater performance and data throughput, but there is an exposure to data loss in the event of a power failure, since data not yet written to disk will be lost unless the cache memory is protected by battery. Having battery-backup for the cache means that this risk is reduced.

#### Clustering configurations

If you are using your server in a clustered configuration, you must not operate any shared logical drives in write-back (WB) mode, as data could be lost in a failover situation.

When one server fails and the surviving server in the clustered pair takes over control of the shared drives, there is a possibility that the failing server's most recently processed data will be lost. If write-through cache is used instead, this problem is eliminated.

There are two battery-backup cache options currently available for the ServeRAID-3H and ServeRAID II adapters (the battery-backup cache is standard on ServeRAID-4H, 4M and ServeRAID-3HB). Both of these options protect data held in the adapter's write-back cache from being lost in the event of a server or RAID adapter failure. Once the fault is rectified, the cache option allows the data to be restored to the server. In addition, these options contains high-speed cache memory to optimize RAID performance.

#### 4.2.12.1 32 MB battery-backup cache option

This option is compatible with the ServeRAID-3H and ServeRAID II adapters. It has 32 MB of battery-backed EDO memory. This memory mirrors the adapter's on-board cache.

**Note:** The battery-backup feature only works with logical drives that are configured to be in write-back mode.

When this option is installed on the ServeRAID II adapter, it will only mirror the 4 MB standard cache installed on the adapter and will not provide any additional cache memory.

In the event of a power failure, the battery will maintain the data in the cache for approximately 10 days. The option can also be used when an adapter fails. You simply remove the cache option from the failed adapter and install it in a functioning one. When you power the server on, the data in the cache will be flushed onto the disk drives.

During normal powered operation, the battery will be maintained in a continuously charged state. The battery has a life expectancy of about two years.

#### 4.2.12.2 8 MB battery-backup cache option

The 8 MB battery-backup cache option provides a battery-backup cache for the ServeRAID II adapter. It may be of interest to customers requiring a further measure of data protection in their system.

**Note:** Installing the 8 MB option does not increase the cache size from 4 MB to 8 MB. Only the first 4 MB is used, and is mirrored to the on-board 4 MB to provide an extra level of redundancy.

##### Enabled by default?

If you install the 8 MB battery-backup cache option on a ServeRAID II adapter and you are using Version 2.30 of the BIOS and firmware, the battery-backup feature will *not* be enabled by default. The feature is, however, enabled by default if you are using Version 2.40 of the BIOS and firmware.

#### 4.2.13 I<sub>2</sub>O enabled

Over the past several years, significant advancements have been made in the area of CPU performance. However, the I/O bus has not kept pace with the speed increases of the CPU. If an I/O device requires data transfer, the CPU must pause its processing to satisfy the request.

I<sub>2</sub>O (for Intelligent Input/Output) is a feature that off-loads I/O processing from the CPU to an additional supporting processor. By relieving the CPU of this I/O burden, overall system performance is boosted.

With the I<sub>2</sub>O feature on the ServeRAID-4 and ServeRAID-3 adapters, once the CPU requests the transfer, the adapter manages the transfer without involving the CPU on the server. The server must be I<sub>2</sub>O ready in order to support this function.

**Note:** All models of Netfinity are I<sub>2</sub>O ready. The term *I<sub>2</sub>O ready* means that the system has the intention of being certified for I<sub>2</sub>O compliance. Compliance testing, when defined, will be performed by the operating system vendor. *I<sub>2</sub>O ready* is a statement that the I<sub>2</sub>O functionality that will be supported is in place and the system is ready for I<sub>2</sub>O compliance testing.



#### 4.2.14 Active PCI support

The Active PCI feature lets you remove, replace, and add the ServeRAID-4, ServeRAID-3 or ServeRAID II adapters without first having to power down the server. See 5.4, “Active PCI support” on page 106 for more details.

Active PCI support is especially effective when combined with the fault-tolerant pair feature so that if one adapter fails, the second adapter will take over all disk I/O functions and you can replace the failed adapter while keeping the server running. See 5.5, “Configuring a fault-tolerant pair” on page 114 for more details.

To use the hot-swap capabilities, you need:

- A server with active PCI slots
- Operating system support
- Device driver support

#### 4.2.15 Fault-tolerant adapter pair

This feature allows you to install a pair of ServeRAID adapters to provide a redundant connection from the server to an external storage enclosure. With this configuration, if one of your ServeRAID adapters fails you still have access to the enclosure. The disk drives attached to the internal backplane in the server cannot be connected to the fault-tolerant adapter pair. This feature is supported on ServeRAID-4, ServeRAID-3 and ServeRAID II adapters.

You can refer to 5.5, “Configuring a fault-tolerant pair” on page 114 for more information on how to configure the adapters for this feature.

#### 4.2.16 Hot-swap rebuild

This function, when enabled, will allow a defunct disk to be automatically rebuilt as soon as it is replaced with a new disk. If the function is disabled and a defunct disk is replaced, the rebuild operation must be manually started through ServeRAID Manager or another ServeRAID utility.

You can change the value of this parameter using the `IPSSSEND HSREBUILD` command. It is available for all members of the ServeRAID family including the IBM ServeRAID SCSI Adapter using firmware V2.23 or later.

**Note:** If a logical drive migration (LDM) function is currently executing, the rebuild operation will commence after the LDM is finished.

#### 4.2.17 Data scrubbing

RAID-1 and RAID-5 logical drives will prevent data loss in case of a disk failure, but only when the duplicate data or RAID parity is correct. You would normally expect this to be the case, but physical disk media defects could cause inconsistencies between the actual data and duplicate or parity data. Inconsistencies can also occur when the server is not shut down properly. For example, the server is simply powered off while the ServeRAID controller is writing data to a RAID-5 logical drive. The new and changed data block might have been written to the disk successfully, but not the updated parity block. If the battery-backup cache is not used, this will cause inconsistency and possible data loss upon physical disk failure.

For this reason, you are strongly advised to synchronize all RAID-1 and RAID-5 logical drives after any improper power-off of the server. You also have the option of running a scheduled synchronization using Netfinity Manager.

The need to synchronize RAID arrays is not always fully appreciated, so, to provide a greater degree of protection, the IBM ServeRAID II Ultra SCSI adapter (firmware V2.30 or later) introduced data scrubbing. This feature is also provided on ServeRAID-4 and ServeRAID-3 adapters. Data scrubbing involves continuously checking for errors in all sectors of RAID-1 and RAID-5 logical drives while your system is running. This function executes as a low-priority background task within the adapter and eliminates the need for periodic scheduled synchronization.

In ServeRAID II, synchronization or data scrubbing involves the recalculation and rewriting of either the RAID parity (RAID-5) or the duplicate data (RAID-1).

With ServeRAID-3 adapters, the algorithm is more efficient than that used in ServeRAID II, but the result is the same — you still avoid the need to perform periodic synchronizations using Netfinity Manager.

The data scrubbing feature in ServeRAID-3 adapters uses a process to look for media defects. If a media defect is found, the adapter reconstructs and repairs the data (RAID-1 and RAID-5). For RAID-5, this process does not update the RAID parity unless a media defect is found in the RAID parity stripe unit. For RAID-0, ServeRAID cannot reconstruct data if there is a media failure. Sometimes the drives themselves can recover data from areas that have media damage. This is the only data recovery available in RAID-0 arrays.

While the data scrubbing process is occurring, system-initiated read and write operations are processed normally. The data scrub feature is transparent to the user and completely handled by the adapter as already mentioned. Since normal read and write operations are prioritized over data scrubbing, it can take a considerable amount of time to check the entire disk space. Therefore you might still consider manual synchronization after each improper power-off in order to reduce the exposure to potential data loss.

**Note:** If you have a ServeRAID or ServeRAID II adapter with V2.23 firmware, you should still synchronize your RAID-5 arrays weekly, as data scrubbing is not available on these adapters. For ServeRAID II adapters, we recommend you upgrade to the latest firmware to get this feature.

#### **4.2.18 Autosync**

On ServeRAID adapters, it is essential that each new RAID-5 logical drive is synchronized (not only initialized) immediately after creation. If this step is not performed, parity data may not be written correctly and data loss could occur in case of a physical disk failure.

Autosync eliminates the need to manually synchronize RAID-5 logical drives before storing data. The adapter will start the “initial” synchronization when you define the logical array. This process is done in the background so that the user can proceed with installation or running the system. Autosync is an installation tool.

Autosync is supported on all ServeRAID-4 and ServeRAID-3 adapters. It requires V2.4 or later of the BIOS and firmware code on ServeRAID II adapter and is not supported on the older IBM ServeRAID SCSI Adapter.

Just as with data scrubbing, while the autosync process is occurring, system initiated read and write operations are processed normally. The autosync feature is transparent to the user and completely handled by the adapter.

#### **4.2.19 Configuration data stored in multiple locations**

Configuration data contains all information needed to access the data on logical drives. This information includes:

- SCSI IDs and channels of physical disk drives in each array
- Hot-spare drive definitions
- Logical drive sizes and their RAID levels
- Stripe unit size and other parameters

The configuration data is critical to the operation of your ServeRAID adapter. If it is lost, it will not be possible to access the data stored on the disks anymore.

IBM RAID adapters prior to ServeRAID stored their configuration data in Flash EEPROM on the adapter itself. On those adapters it was extremely important to save the configuration data to a file on a diskette. If the RAID adapter failed and had to be replaced, this would allow you to regain access to the data on your disks by restoring the configuration from the diskette.

ServeRAID adapters improve the resilience of the disk subsystem by storing the vital configuration data in multiple locations. The adapter automatically stores configuration information in three locations:

1. Flash EEPROM on the adapter
2. Non-volatile RAM (NVRAM) on the adapter
3. On a reserved area on each disk drive in online (ONL) or rebuild (RBL) state

If a ServeRAID adapter fails and has to be replaced, the configuration data can simply be recovered from the disk drives. This eliminates the need to have a backup of configuration on a diskette. There are also other uses of this feature:

- The disk drives can be reordered and repositioned in different bays and SCSI channels without confusing the adapter.
- It is possible to transfer the disk drives from one controller or server to another.

If you still want to save the ServeRAID configuration to a file, you can do so by using the `IPSEND BACKUP` command. With the `IPSEND RESTORE` command you can then recover the configuration, if needed.

#### **4.2.20 ServeRAID utilities**

The following ServeRAID utilities are available:

- ServeRAID Manager.
- The RAID Manager tool within Netfinity Manager.
- The ServeRAID Manager tool within Netfinity Director.
- The GUI-based ServeRAID Configuration Utility on a bootable CD.
- The Mini-Configuration Utility, built into the adapter. This is flash EEPROM-based and available at boot time.

It is important to be familiar with these utilities and know when to use them.

#### **4.2.20.1 ServeRAID Manager Version 3.60**

ServeRAID Manager Version 3.60 runs on Microsoft Windows NT, Windows 95/98, OS/2 and Novell NetWare Version 5 systems, either locally on the server or remotely using TCP/IP. The tool is used during normal operation of the server. Administration functions can be performed online with either minimal or no down time to the server. These functions include:

- Adding or removing disk drives using logical drive migration
- Increasing logical drive space using logical drive migration
- Creating and deleting arrays
- Creating new logical drives
- Rebuilding critical logical drives after a physical disk drive failure
- Configuring for clustering environment

This utility is discussed in detail in 5.3, “ServeRAID Manager” on page 92.

#### **4.2.20.2 ServeRAID Manager Version 4.0**

At the time of writing, ServeRAID Manager Version 4.0 is in beta phase. It will contain all the functionality from the earlier version and, additionally, these enhancements:

- Windows 2000 support
- Red Hat Linux support
- SCO OpenServer 5.0.5 support
- Spanned array logical drive levels RAID-00, 10, 1E0 and 50
- The agent will run as a service
- SNMP trap support
- Active PCI hot replace support on Windows NT

#### **4.2.20.3 ServeRAID Configuration Program**

The CD-ROM-based ServeRAID Configuration Program is usually used for the initial setup of your adapter, but is also used for selected recovery operations. The bootable CD-ROM is based on the Windows 95 shell and has an Explorer-style interface. All actions in the GUI are initiated from the pull-down or pop-up menus. See 5.2.1, “ServeRAID Configuration Program” on page 82 for more details.

The diskette-based configuration utility is no longer available. The final version was 3.50; however, it is not recommended for use with adapter BIOS

Version 3.60 and higher. You should always use the same version of utilities and BIOS.

#### **4.2.20.4 Mini-Configuration Utility**

The Mini-Configuration Utility, programmed into the adapter's firmware, offers a limited set of configuration utilities and can be accessed by pressing Ctrl+I at boot time. It resides in flash EEPROM on the ServeRAID adapter. The utility is discussed in depth in 5.2.2, "ServeRAID Mini-Configuration Utility" on page 89.

### **4.2.21 Logical drive migration**

One of the strongest management features of the ServeRAID adapters is logical drive migration (LDM), which offers unrivaled disk subsystem flexibility. The following functions are offered:

- Changing the RAID levels of logical drives in an array.
- Adding hard disks to an array and increasing logical drive capacity.
- Adding hard disks to an array and increasing the available free space.

These features enable you to reconfigure logical drive structures online, with little impact on users. LDM is discussed in depth in 5.3.2, "Logical drive migration (LDM)" on page 94.

### **4.2.22 Command-line utilities**

ServeRAID adapters offer the following command-line utilities:

- IPSEND
- IPSMON

These commands allow you to perform various tasks on ServeRAID adapters by entering the appropriate command. These commands can help you with configuration, troubleshooting, multiple server roll-out, and so forth.

#### **4.2.22.1 IPSEND**

IPSEND is a utility providing a command line interface for performing various tasks on ServeRAID adapters. The utility also allows you to build commands into batch files, which can be very useful for creating multiple configurations for a system rollout exercise.

For example, you could create arrays, create logical drives, and initialize and synchronize logical drives from a batch file on a bootable diskette. You could then go on to initiate an automated operating system download. This process

could even be done over a LAN or WAN, perhaps to a remote branch office. Automating this type of operation, and running the batch file overnight, for example, to configure systems with no user intervention required can simplify the task and minimize disruption.

Refer to the README.TXT file on the ServeRAID Command Line Programs diskette or to the *ServeRAID-3HB, ServeRAID-3H and ServeRAID-3L Ultra2 SCSI Controllers Installation and User's Guide* for more information about the format, options and syntax of the IPSEND command.

### **IPSMON**

IPSMON is a utility that monitors ServeRAID adapters for failed drives, Predictive Failure Analysis (PFA) warnings, rebuilds, synchronizations and logical drive migrations. If any of these occurs, a message is logged to the display and/or a log file.

Being able to read and understand the log entries generated by IPSMON is a very important part of recovering an array when one or more drives are marked *defunct* (DDD). Using the log, you can determine in what order drives went defunct, and, if multiple drives fail, which one is the out-of-sync drive.

More information about IPSMON can be found on the README.TXT file on the ServeRAID Command Line Programs diskette or in the *ServeRAID-3HB, ServeRAID-3H and ServeRAID-3L Ultra2 SCSI Controllers Installation and User's Guide*.

## **4.2.23 FlashCopy**

FlashCopy was introduced with ServeRAID BIOS, firmware and utilities Version 3.50. It is an IPSEND subcommand that copies the contents of a logical drive to another logical drive. It actually creates a snapshot impression of the source logical drive on the target logical drive. When the command completes, the target logical drive will contain the exact image of the source logical drive at the time of issuing the FlashCopy command. There is no need to wait until the data image transfer completes, you can start using the target logical drive immediately after issuing the command.

### **Windows Only**

Note that the FlashCopy function is only available for Windows NT and Windows 2000 systems.

A typical use of FlashCopy would be for a tape backup or drive-cloning for a multi-server rollout.

Consider the following guidelines and restrictions:

- The source logical drive can be used normally while the FlashCopy operation is in progress.
- You only can have one operating system partition per logical drive.
- The source and target logical drives must be on the same ServeRAID adapter. They do not need to be (and usually will not be) in the same array.
- The logical drive size must be smaller than, or equal to, 124 GB.
- For optimal performance, the source and target logical drive sizes should be the same.
- With BIOS and firmware Version 3.60, you can run up to four concurrent FlashCopy commands.

The IPSSSEND FlashCopy options are:

#### **MAP**

This command will identify how ServeRAID logical drives map to operating system drive letters. It also indicates the status of FlashCopy for each logical drive.

#### **BACKUP**

Creates a snapshot of the source to the target logical drive and copies all the data as a background task. You can use this to clone the source logical drive. Data can be read from both drives and written to the source drive. The target drive can be written to once the copy process has completed.

#### **NOBACKUP**

Creates a snapshot of the source logical drive on the target logical drive, but does not copy any data. This can be used as a temporary source for a tape backup. Data is copied to the target disk only when new data is to be written to the source disk. Data can be read from both drives and written to the source drive. You should not write data to the target drive until the FlashCopy link is broken using the STOP option.

#### **STOP**

Ends the FlashCopy BACKUP process or breaks the FlashCopy NOBACKUP link.

#### **DELETE**

This option deletes the array. It is very useful when you want to move your target logical drive to another system, perhaps when doing a multi-server rollout. You would create an exact image of the source



logical drive on the target using FlashCopy BACKUP, then you would delete the array with the target logical drive from the ServeRAID configuration. Finally, you would transfer the disk drives to another server and use FlashCopy IMPORT.

You should be careful when Write-Back cache policy is used on the target logical drive. If you delete the target array too soon, the data might not yet be entirely flushed from the cache. Therefore it is a good idea to use Write-Through mode for the target logical drive.

#### **IMPORT**

As described in the above paragraph, this option adds an array and logical drive to an existing ServeRAID configuration.

**Note:** you cannot use this option in a cluster environment.

For exact syntax and more information on FlashCopy, refer to the relevant section of the ServeRAID-3 and ServeRAID-4 User's Guides.

#### **4.2.24 BIOS and firmware**

Occasionally, IBM will release new levels of the ServeRAID adapter BIOS and firmware. These may be for some or all members of the ServeRAID family, and can be either for enhancements to existing functions or to add new functions.

It is important to understand the difference between firmware and BIOS. The BIOS is the software that runs each time the server is booted. Many adapter cards have an adapter BIOS that sits in flash ROM on the adapter card. At system boot time, the system BIOS polls the system for active adapters and, if it finds an adapter BIOS, it loads and runs it. The ServeRAID BIOS banner at boot time shows this is happening. The BIOS code runs on the server CPU and is the same level for every member of the ServeRAID family. At the time of writing, the latest level of BIOS is Version 3.60 and Version 4.0 is in beta phase.

Firmware also resides in flash memory on the adapter card and runs on the ServeRAID adapter processor. This controls all of the functions on the card and controls the flow of data between the server's CPU (BIOS, device driver, and application functions) and the disk drives that are attached to the adapter.

Each of the ServeRAID adapters runs different levels of firmware. For example, the latest firmware levels at the time of writing are:

- ServeRAID adapter: V2.25.01
- ServeRAID II adapter and integrated controller: V2.88.13
- ServeRAID-3 adapters: V3.60.21
- ServeRAID-4 adapters: V4.00 beta

**Note:** For a complete change history of BIOS and firmware updates, see the README.TXT file on the BIOS and firmware update diskette.

#### **4.2.25 Feature comparison**

For ease of reference, we have summarized the major features of each member of the ServeRAID family of adapters in the following table:

Table 9. ServeRAID adapters: feature comparison

Feature	ServerRAID-4H	ServerRAID-4M	ServerRAID-4L	ServerRAID-3HB	ServerRAID-3H	ServerRAID-3L	ServerRAID II
Internal connectors	2	2	1	1	1	1	3
External connectors (standard and maximum)	4/4	2/2	1/1	2/3	2/3	1/1	2/3
External connector type	VHDCI	VHDCI	VHDCI	VHDCI	VHDCI	VHDCI	VHDCI
Cache size (MB)	128	64	16	32	32	4	4
Cache battery-backup	Yes	Yes	No	Yes	Option	No	Option
PCI bus	64/32-bit 33 MHz	64/32-bit 33 MHz	64/32-bit 33 MHz	64/32-bit 33 MHz	64/32-bit 33 MHz	32-bit 33 MHz	32-bit 33 MHz
“Optimal” SCSI setting	Yes	Yes	Yes	Yes	Yes	Yes	No
LVDS support	Yes	Yes	Yes	Yes	Yes	Yes	No
I <sub>2</sub> O ready	Yes	Yes	Yes	Yes	Yes	Yes	No
Active PCI support	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fault-tolerant pairing	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hot-swap rebuild	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data scrubbing	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Autosync	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Logical drive migration	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Command line utilities	Yes	Yes	Yes	Yes	Yes	Yes	Yes

---

## 4.3 External storage enclosures

Current high-end Netfinity servers offer a relatively low number of internal disk drive bays. These bays should primarily be used for the disk drives containing the operating system. The trend towards running bigger and more sophisticated enterprise applications on Intel-based servers is creating demand for larger amounts of disk space than can be realistically hosted within a system unit. Such applications, or at least large application databases now commonly reside on external disk drives, housed in externally connected storage enclosures. Shared storage clustering implementations (Microsoft Cluster Server, for example) and fiber-attached disk subsystems make external storage enclosures even more important.

### 4.3.1 Netfinity EXP200 Storage Expansion Unit

The Netfinity EXP200 Storage Expansion Unit supports up to 10 half-high Ultra2 SCSI hot-swap disk drives, at a maximum data transfer rate of 80 MBps. It requires Ultra2 SCSI cables, which may be up to 20 meters long.

EXP200 hot-swap disk drives use a tray, compatible with the Netfinity 8500 and 5600 servers, and made of aluminum. The tray is designed to virtually eliminate vibrations that have been observed to cause soft data errors when several less sturdy drives are contained in a single enclosure. This greatly reduces the number of retry operations and thereby provides better performance.

Two SCSI backplanes are contained in this expansion unit. The backplanes can be separated and connected to different SCSI buses, or they can be joined to form a single SCSI bus using a switch at the back of the unit. When the backplanes are separated, the order of the SCSI IDs for the disk drive bays alternates, as shown in Figure 27:

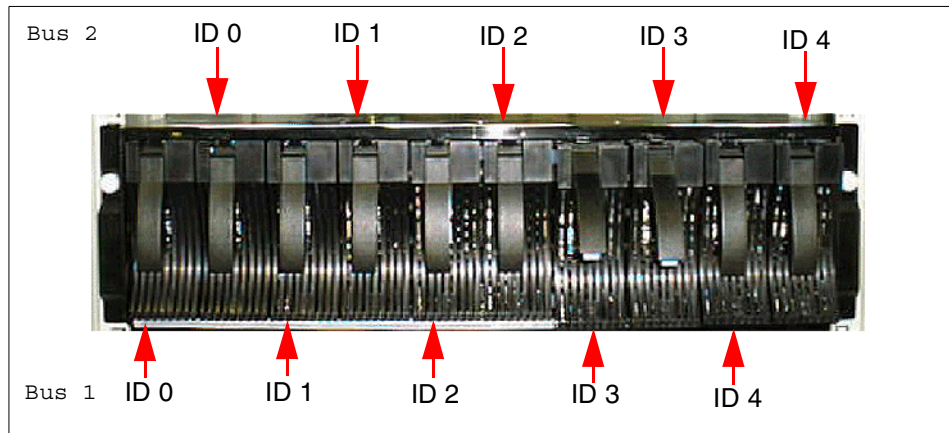


Figure 27. EXP200 SCSI IDs - two SCSI buses

When the backplanes are joined into a single bus, the following happens:

- Only one backplane uses its SCSI ID to communicate and SCSI ID of the other backplane is released.
- The drive bays' SCSI ID order changes from alternating to sequential.

The EXP200 supports 3U rack or tower usage. A switch at the back is used to select the appropriate mode.

The enclosure contains one hot-swap power supply. You can install optional redundant hot-swap power supply for higher availability. The two fans are hot-swappable and redundant.

#### 4.3.2 Netfinity EXP300 Storage Expansion Unit

The Netfinity EXP300 Storage Expansion Unit is a follow-on to the EXP200 enclosure. Significant enhancements in comparison to its predecessor include:

- Ultra3 SCSI 160/m capability

The unit supports a maximum data transfer rate of 160 MBps and SCSI cable lengths up to 25 meters are possible.

- Up to 14 slim hot-swap disk drives in converged trays

EXP200 and older storage enclosures (EXP15 and EXP10) supported up to 10 disk drives and this meant that some SCSI bus IDs remained unused. Now, all 16 SCSI IDs on the bus may be used: 14 for the disk drives, one for SCSI controller and one for the hot-swap backplane.

Therefore the maximum storage capacity that ServeRAID adapters can support in practice is significantly increased.



*Figure 28. EXP300*

Other features of the EXP300 enclosure are similar to those of the EXP200 unit:

- Rack or tower installation
- Hot-swap and redundant fans
- One hot-swappable power supply is supplied as standard
- An optional redundant hot-swappable power supply
- Single or dual SCSI bus

---

## Chapter 5. Implementing ServeRAID subsystems

In this chapter we discuss configuration and administration of ServeRAID adapters using the utilities that are provided with the adapters. We also examine the adapter management tools and integration with Netfinity Director and Netfinity Manager. The more advanced features of the ServeRAID family, such as Active PCI support and fault-tolerant adapter pairs, are also covered.

---

### 5.1 Utilities

Two different sets of utilities are available for ServeRAID adapters:

- Configuration utilities, available on the bootable ServeRAID CD-ROM or in Flash EEPROM, which are used to configure the ServeRAID adapter, and also to perform troubleshooting in cases when the operating system cannot run on the server.
- ServeRAID Manager, which can be used during normal operation of the server. This utility provides online monitoring of the status of the ServeRAID adapter and its attached disk drives and enclosures, and also provides alerting and recovery capabilities. This tool helps to minimize server downtime by allowing you to perform many of the tasks available in the configuration utilities, such as adding new disk drives, creating new arrays and logical drives, and logical drive migration, while the server is running.

Before ServeRAID Manager was made available, the ServeRAID Administration and Monitoring Utility was used to perform the tasks above.

To be able to work effectively with ServeRAID adapters, it is of vital importance that you are familiar with all these utilities.

**Important!**

Make sure that the versions of ServeRAID BIOS, firmware, the driver and all configuration and administration utilities you use are at the same level. All components at a particular level are grouped into a package and you should always download and use the entire package, not just a single element.

---

## 5.2 Configuration utilities

As already mentioned, the following configuration utilities are available:

- Bootable CD-ROM-based ServeRAID Configuration Program (graphical user interface)
- Mini-Configuration Program, accessible at boot time by pressing <Ctrl-I>

The bootable diskette-based DOS ServeRAID Configuration Utility is not available anymore. The final version of this utility was 3.50 and it never supported the RAID-5E logical drives.

We now take a closer look at these configuration aids.

### 5.2.1 ServeRAID Configuration Program

The information presented here applies primarily to Version 3.60 of the ServeRAID Configuration Program because Version 4.0 was in beta phase at the time of writing. Important enhancements in Version 4.0 will be the ability to configure spanned arrays and RAID-00, 10, 1E0 and 50 logical drives, but we have given information about these new levels when it was available.

The CD-ROM-based ServeRAID Configuration Program resides on the ServeRAID Support CD, supplied with the ServeRAID adapter. It is also available on the *Netfinity Setup and Installation CD*, which is included in the ServerGuide package.

**Note:** for ServerGuide versions prior to 5.0, the ServeRAID Configuration Program resided on the Hardware Guide CD.

It is also possible to download the ISO image of the ServeRAID Support CD. The image file can be found on IBM PC technical support Web page:

<http://www.pc.ibm.com/support>

For Version 3.60, the URL is:

[ftp://ftp.pc.ibm.com/pub/pccbbs/pc\\_servers/00n9126.iso](ftp://ftp.pc.ibm.com/pub/pccbbs/pc_servers/00n9126.iso)

To start the utility, insert the appropriate CD-ROM into the CD-ROM drive, and restart the server. If the CD from the ServerGuide package is used, you will have to navigate through the ServerGuide menus:

- Select **Run Netfinity setup programs and configure hardware**
- Select **Custom configuration** and the appropriate operating system
- Select **Run ServeRAID Configuration**



There are three panels in the Configuration Program window as shown in Figure 29. You select the object you wish to work on in the tree view and the information about that object is displayed in the main panel:

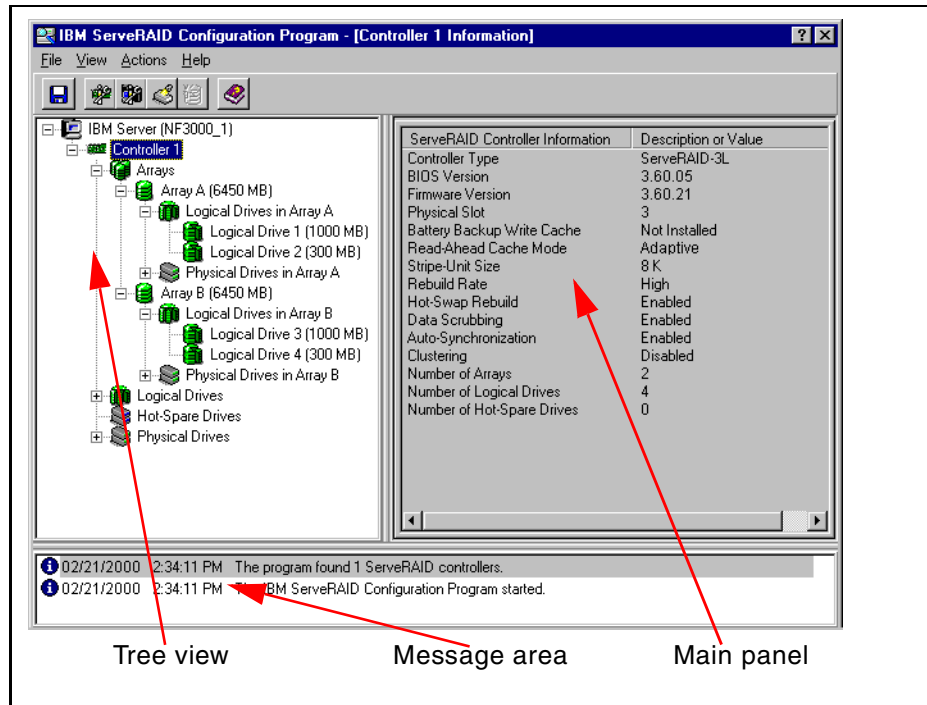


Figure 29. ServeRAID Configuration Program in information mode

The program runs in two modes. If the program detects unconfigured ServeRAID adapters, it will start in the *configuration mode*. If all adapters are configured, the program will start in the *information mode*.

- *Configuration mode* — Multiple panels of instructions will prompt the user to configure arrays, logical drives and hot spares. Figure 30 on page 84 shows an example of configuration mode.
- *Information mode* — Information relevant to the currently selected object will be displayed. When this mode is active, you can use the functions available from the menu and tool bars to customize settings for the controllers. Figure 29 shows the utility in information mode.

For more information about the configurator, see Chapter 3, “Using the configuration programs” of the *ServeRAID-3H, ServeRAID-3HB, and ServeRAID-3L Ultra2 SCSI Adapters Installation and User’s Guide*.

### 5.2.1.1 Configuration mode

The configuration mode has two configuration paths as shown in Figure 30.

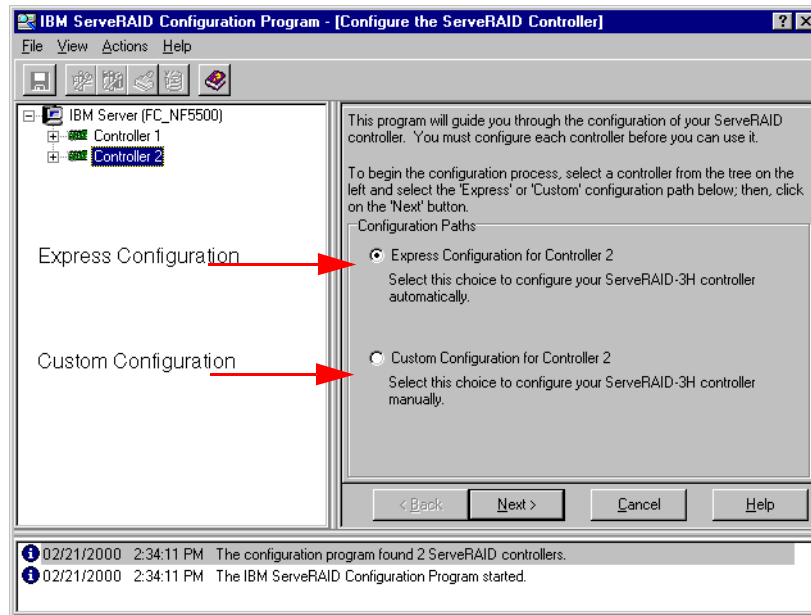


Figure 30. Configuration mode

- The *express* configuration path provides a quick and easy way to automatically configure arrays and logical drives, and creates the most efficient configuration based on the capacity and number of the available drives.

Express configuration groups up to 16 *ready* drives of the same capacity into one disk array and defines one logical drive for each array. It defines the size of the logical drive based on the amount of free space available and it assigns the highest RAID level possible, based on the number of physical drives available (that is, RAID-5 for 3-16 drives, RAID-1 for 2 drives, RAID-0 for 1 drive). If there are more than 16 drives of the same capacity, it will configure the first 16 for the first array, the next 16 for another array, and so on.

When there are four or more ready drives of the same capacity, the express path will define one of them as a hot spare. If more than one group of drives contains more than four drives, only one hot-spare is created of the largest drive size.

**Note:** Express configuration will not create RAID-5E logical drives.

- The *custom* configuration path allows you to configure your ServeRAID subsystem manually.

Using this path, you can configure your arrays, logical drives and hot spares as you wish. The configurator will warn you if your configuration has some undesirable features (for example, defining a 9.1 GB hot spare for an array of 18.2 GB disks), or when there are unused ready drives or unallocated free space.

### 5.2.1.2 Creating Arrays

If you choose the custom configuration path, you can manually select disks to form arrays and logical drives. Figure 31 below shows the window that allows you to create arrays. To add one or more ready drives to an array, right-click the ready drive icon and click **Add to New Array**. Alternatively, drag the disk icon in the tree view and drop it on the array icon in the main panel. After a drive is added to an array, its status changes from *Ready* to *Online*.

When you set the first drive to Add to New Array, Array A is automatically created. When you add the next drive, the configurator gives you the option to add it to Array A or to create another new array.

You can also select several drives at once and, with drag and drop, add them into the array.

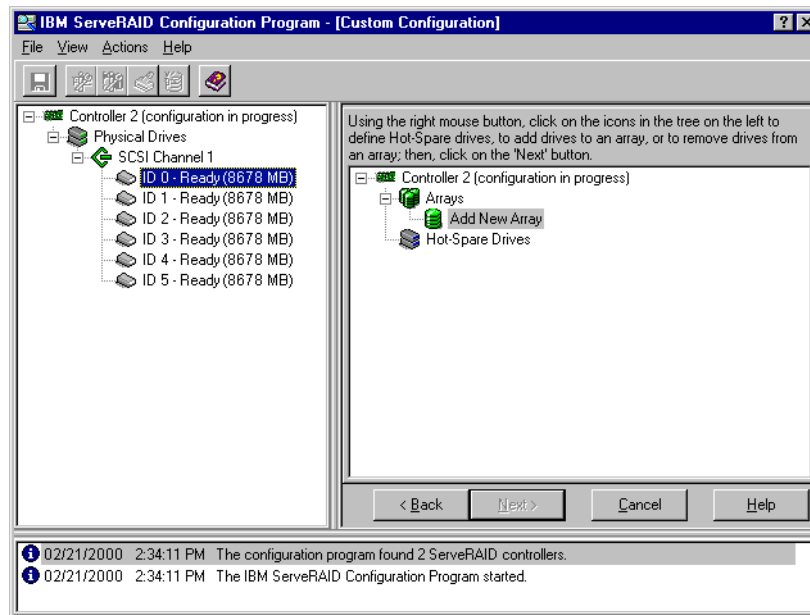


Figure 31. Creating arrays

### 5.2.1.3 Hot-spare

To mark a drive as a hot spare, right-click the disk icon and click **Set Drive State to Hot-Spare**. You can also mark a drive as a hot spare by dragging it on to the **Hot-Spare Drives** icon.

### 5.2.1.4 Creating Logical Drives

The next step is to create one or more logical drives in the array. For performance reasons, it is preferable to configure only one logical drive per array. Click the **Next** button to see Figure 32, which allows you to create logical drives.

You can create a single logical drive in an array or you can subdivide it into several logical drives by clicking on **Add Logical Drive**. Each logical drive appears to the operating system as a physical hard disk drive. You can have up to eight logical drives per ServeRAID adapter.

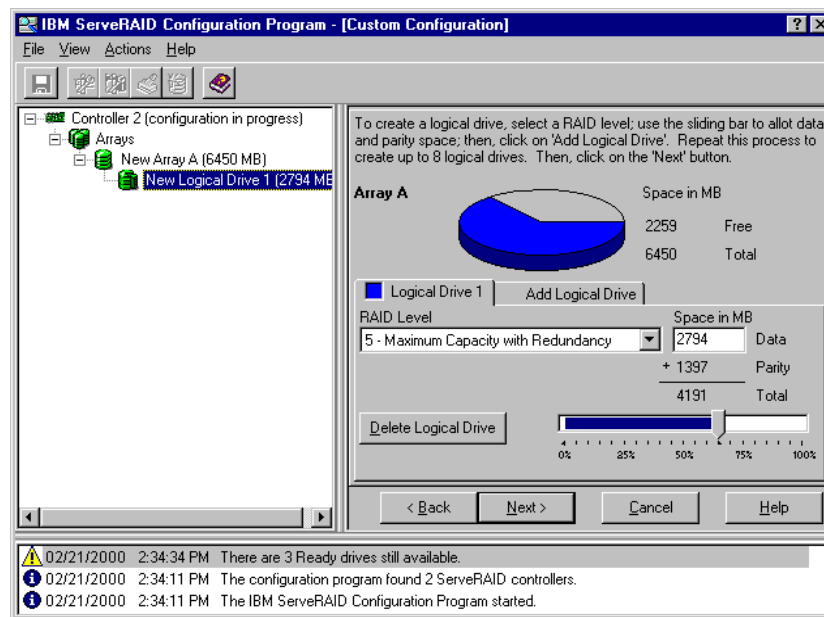


Figure 32. Creating logical drives

Now you select the RAID level for the logical drive and by moving the sliding bar from right to left you can change its capacity.

Once you have created all the logical drives you need, click the **Next** button. Figure 33 then appears, giving you a summary of the configuration you have selected.

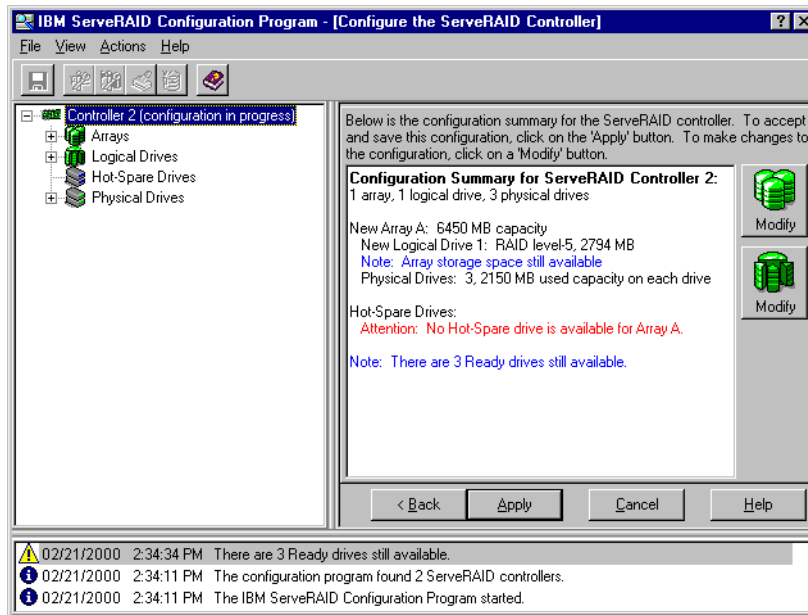


Figure 33. Configuration summary

Examine any informational (blue) and warning (red) messages in the configuration summary window, as these may suggest that you need to change your configuration, perhaps to improve fault-tolerance by creating a hot-spare as in Figure 33.

You can accept this configuration by clicking **Apply**. If you have more arrays or logical drives to create, you can click one of the two **Modify** buttons to return to the configuration panel:

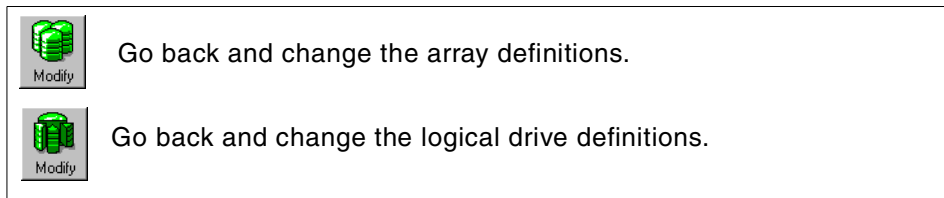


Figure 34. Modify buttons

### 5.2.1.5 Configuration Complete

Once you have configured your RAID arrays and logical drives, you will be returned to information mode, where you can view the devices just defined. This next figure gives an example of a completed configuration:

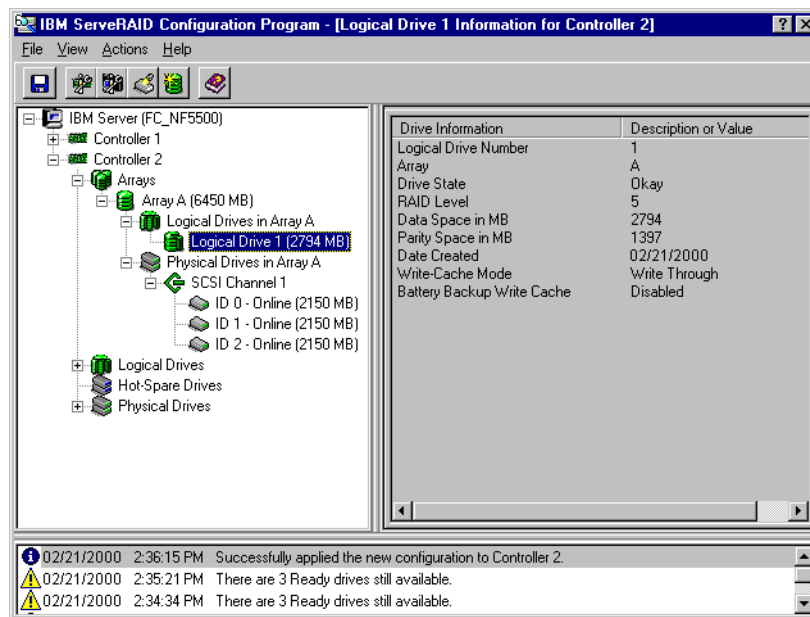


Figure 35. Configuration complete

For more information about how to configure your arrays and logical drives, see Chapter 3 of the *ServeRAID-3H, ServeRAID-3HB, and ServeRAID-3L Ultra2 SCSI Adapters Installation and User's Guide*.

### 5.2.1.6 Information Mode

You can click any of the objects in the tree view on the left-hand side of the main window in Figure 35 to display information about that particular object in the main panel on the right.

You can also perform actions on the selected object by either clicking the **Actions** pull-down menu or by right-clicking the object. The Actions menu is context-sensitive and the list of valid actions will change in accordance with the object selected.

For example, right-clicking the Controller object produces a pop-up window similar to that in Figure 36:

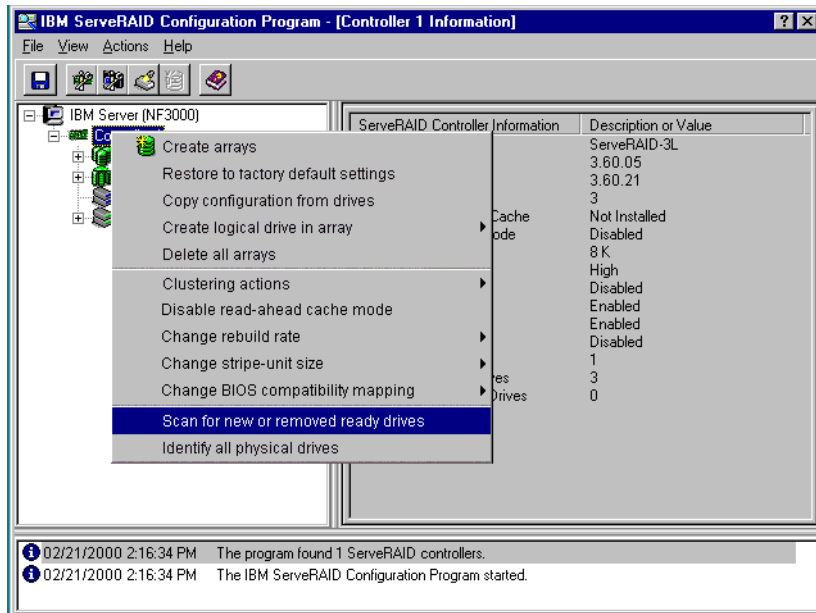


Figure 36. Actions available for the Controller object

For more information, see Chapter 3 of the *ServeRAID-3H, ServeRAID-3HB, and ServeRAID-3L Ultra2 SCSI Adapters Installation and User's Guide*.

## 5.2.2 ServeRAID Mini-Configuration Utility

The Mini-Configuration Utility is a BIOS-based program that offers the ability to display your adapter settings and to perform a limited set of configuration functions without using the Windows CD-ROM-based configuration utility.

When a BIOS/firmware update is applied to the ServeRAID adapter, this will also update the Mini-Configuration Utility.

By design, the utility is always used offline, requiring that the server be taken down. To access the Mini-Configuration utility, start or restart your server. During the server's boot sequence, you will see the following message:

Press <Ctrl+I> for MiniConfig Utility

Press Ctrl+I to start the utility. If you have more than one adapter installed, you can select which adapter to work with and then the Main Menu appears (Figure 37 on page 90).

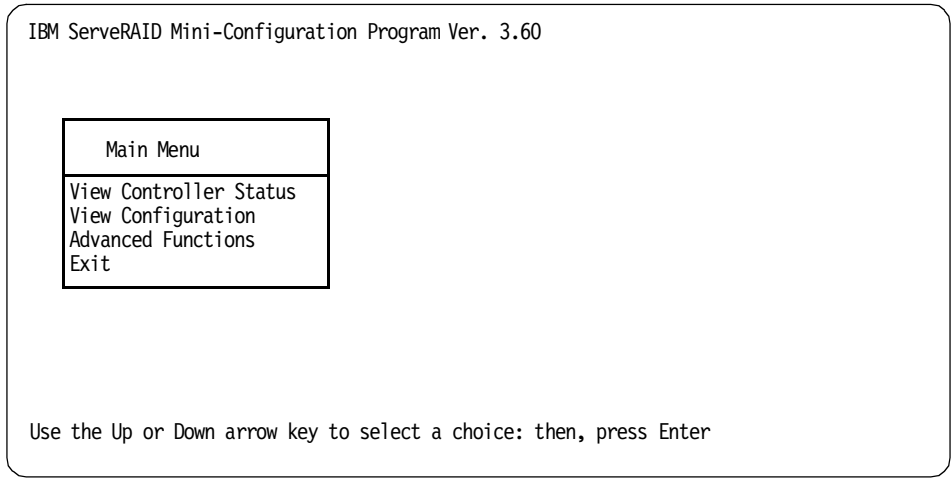


Figure 37. ServeRAID Mini-Configuration utility

As well as viewing both the adapter status and the logical drive configuration, you can perform certain administrative functions. Selecting **Advanced Functions** from the menu gives you the following:

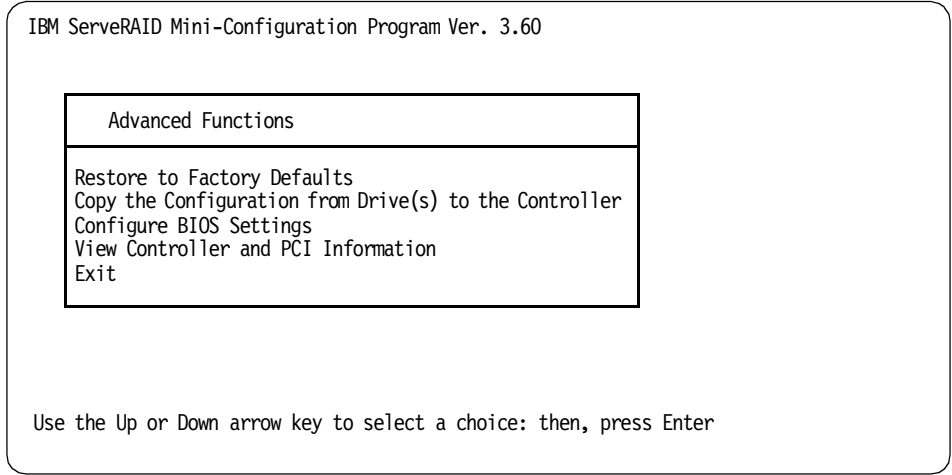


Figure 38. Advanced functions

The options you can select here are:

- Restore to Factory Defaults  
This option resets the adapter configuration to factory defaults, and places all accessible drives in the ready (RDY) state. The RAID configuration in



the adapter Flash ROM will be erased, but no user data on the disk drives will be lost. To regain access to the data you must either restore the configuration from the drive(s) or from the backup diskette.

**Note:** This choice will not change any of the controller parameters (stripe unit size, rebuild rate, data-scrubbing and so on)!

- Copy the Configuration from Drive(s) to the Controller

This option reads the configuration from the drives and copies it to NVRAM and EEPROM on the adapter. It overrides the current configuration stored in the adapter. You would typically use this function when installing a replacement adapter in a system that previously had a working ServeRAID configuration.

- Configure BIOS Settings

This option lets you configure the adapter's BIOS settings (see Figure 39).

- View Controller and PCI Information

This choice displays data about the adapter hardware and the PCI registers.

### **ServeRAID BIOS configuration**

Among other features, these BIOS settings allow you to enable bootable CD-ROM support, although this is usually not used on the current models of Netfinity servers as they use IDE-attached CD-ROM drives.

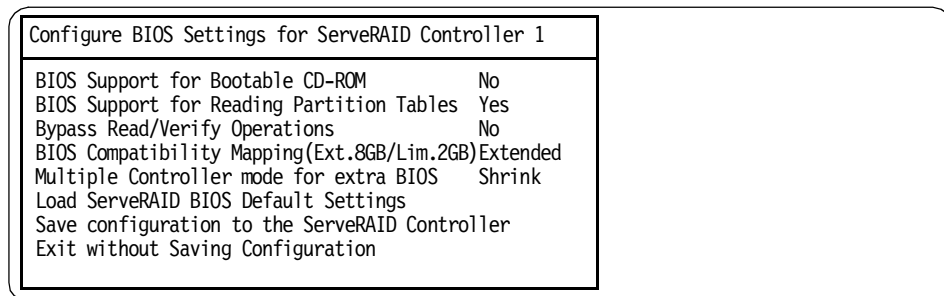


Figure 39. ServeRAID BIOS Configuration

The options you can select here are:

- BIOS Support for Bootable CD-ROM

This option allows you to use bootable CD-ROMs in your system. Set to Yes if you want to do this.

- BIOS Support for Reading Partition Tables

This option allows you to decide how drive partitioning will be handled. This can be done by the operating system or by the ServeRAID adapter. The default is *Yes*, which means that the ServeRAID adapter will control drive partitioning.

**Note:** We recommend that you set this option to *Yes*. Setting to *No* would normally be used by IBM service personnel for debug purposes.

- Bypass Read/Verify Operations
- BIOS Compatibility Mapping

BIOS Compatibility Mapping can be set to *Limited* or *Extended*. Selecting *Limited* informs the ServeRAID BIOS that the system supports 2 GB or smaller hard disk drives. The *Extended* setting informs the ServeRAID BIOS that the system supports 8 GB or smaller hard disk drives.

Unless you are migrating from older PCI or Micro Channel adapters, you would normally set this to *Extended*.

- Multiple Controller mode for extra BIOS

The Multiple Controller mode has two settings, *Erase* and *Shrink*. When the parameter is set to *Erase*, redundant copies of the ServeRAID BIOS are erased. When the parameter is set to *Shrink*, the extra copies of the ServeRAID BIOS are removed from memory, but stored for future use.

To ensure that you will have a copy of the ServeRAID BIOS available if your active copy becomes defective or unavailable, leave the Multiple Controller parameter set to *Shrink*.

- Load ServeRAID BIOS Default Settings

---

### 5.3 ServeRAID Manager

ServeRAID Manager replaces the ServeRAID Administration and Monitoring Utility. The most apparent difference between the utilities is in the graphical user interface used by the former software, which now matches the ServeRAID Configuration Program.

The previous utility could be installed from the ServerGuide Application Guide CD-ROM, or from a diskette available on the IBM PC support Web page. ServeRAID Manager can be installed from the ServeRAID Support CD or it can be downloaded from the IBM PC Support Web page:

<http://www.pc.ibm.com/support>

The current version of the utility at the time of writing is Version 3.60 and is supported for use on the following operating systems:

- Windows 95/98/NT
- Novell NetWare
- OS/2
- SCO UnixWare

ServeRAID Manager Version 4.00, which is in beta phase at the time of writing, will bring the following enhancements:

- Support for Windows 2000, Red Hat Linux and SCO OpenServer.
- Support for spanned arrays and RAID-00, 10, 1E0 and 50 logical drives.
- The agent will run as a service.
- SNMP trap support.
- Active PCI hot-replace support in Windows NT.

### 5.3.1 Creating arrays and logical drives

You can create new arrays and logical drives in exactly the same manner as when using the ServeRAID Configuration Program, described in 5.2.1, “ServeRAID Configuration Program” on page 82. But there is one important difference: ServeRAID Manager allows you to perform these tasks while the server is up and running. With the Active PCI capabilities of current midrange and high-end IBM Netfinity servers, it is even possible to hot-add additional ServeRAID adapters, connect them to the disk drives in the expansion enclosures and create new arrays and logical drives, all without the need to bring the server offline. Use the **Scan for new or removed ready drives** action for the Controller object to discover newly connected disk drives, as shown in Figure 40:

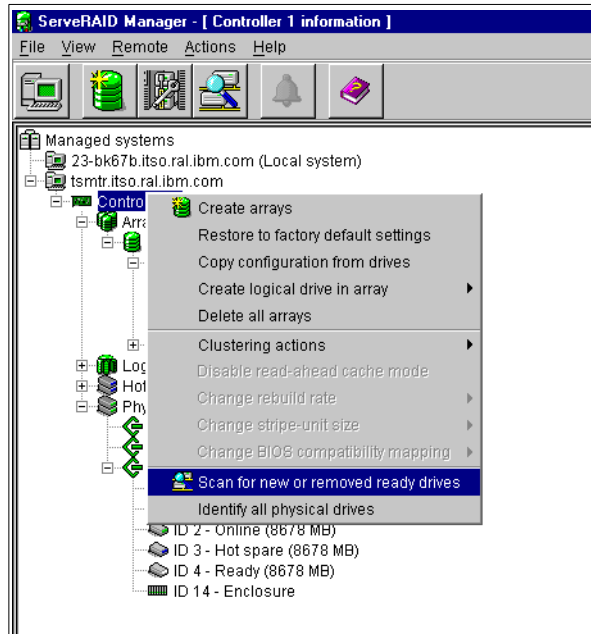


Figure 40. Scan for new or removed ready drives

Once the newly found disk drives appear in the tree display, you can start adding new arrays and logical drives. Alternatively, the new disks could be used to increase the capacity of existing arrays by using logical drive migration, which is explained in the next section.

### 5.3.2 Logical drive migration (LDM)

This is one of the most powerful and flexible features of IBM's ServeRAID adapters. It allows you to increase the capacity of existing arrays and logical drives and also to change their RAID level. ServeRAID Manager allows you to perform these actions while the server is up and running, with only minor performance degradation for users during the process.

This ability was introduced with the original ServeRAID adapter. Previously, altering disk storage required lengthy downtime, as migration of any type was a rather complicated process. For example, to modify the RAID level of a logical drive used to involve a number of time-consuming steps:

1. Back up your existing data from the logical drive.
2. Delete the logical drive, and perhaps the owning array if new drives have to be introduced, as would be necessary to migrate from RAID-1 to RAID-5.

3. Create new arrays and logical drives.
4. Format the storage space for the operating system being used.
5. Restore your data.

LDM simplifies this process significantly. Now disk storage changes and reconfigurations can be carried out online, in much shorter times, with the possibility of data loss or corruption minimized.

You start the logical drive migration process by selecting the array in the tree with the right mouse button. This will display the Actions menu for that particular array, with LDM being one of the items. This is shown in Figure 41 on page 96.

**Note:** LDM can be performed only if the following conditions are met:

- At least one logical drive is available, that is, a maximum of seven logical drives currently exist. During migration, one logical drive is internally created for temporary usage and its state is set to System (SYS). When migration is complete, this logical drive will be removed.
- The source logical drive (the logical drive you wish to perform an LDM migration on) is in the OK (OKY) state.

If a disk drive fails during LDM and you are migrating between fault-tolerant RAID levels, the process will continue and eventually complete. You must then replace and rebuild the failed disk drive.

LDM will also recover from a power failure. If during an LDM migration the power is lost to the server, LDM will simply restart the migration as soon as power is restored, with no data corruption.

LDM can be performed in three different ways, as shown in Figure 41 on page 96:

- You can change the RAID level of existing logical drives.
- You can increase the size of free space, which can then be used to create additional logical drives.
- You can increase the sizes of existing logical drives.

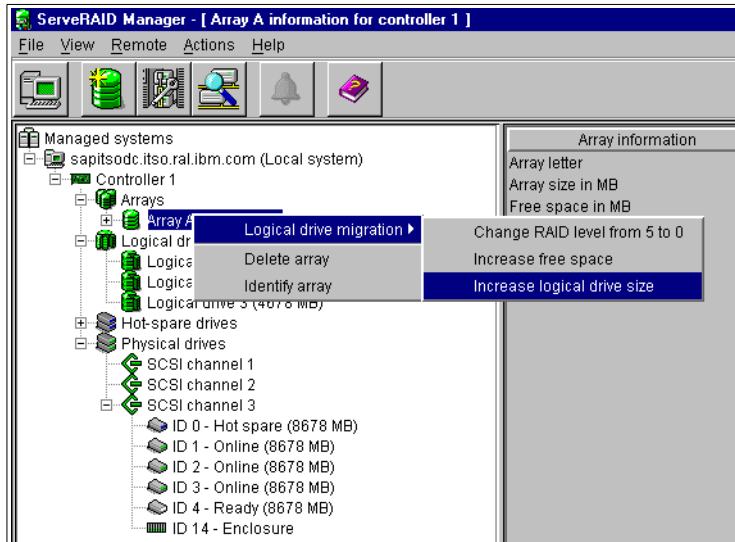


Figure 41. Logical drive migration options

### Changing the RAID level

Changing the RAID level of logical drives could prove invaluable from a system flexibility viewpoint. For example, suppose you have a system containing a RAID-1 logical drive over two physical disk drives. You want to add another disk drive and convert to RAID-5. Without LDM, this would require a system reconfiguration from scratch and a lengthy downtime. Using LDM, you can easily change logical drive RAID level in your array from RAID-1 to RAID-5, without interrupting the network and with no down time.

LDM supports the following RAID level changes:

- Change from RAID-0 to RAID-5 by adding one hard disk drive.
- Change a two-drive RAID-1 to RAID-5 by adding one hard disk drive.
- Change from RAID-5 to RAID-0 by removing one hard disk drive.

#### Important

The RAID level will be changed on all logical drives in the selected array. Therefore, all logical drives in the array must be at the same RAID level and they will all be converted in the same way.

If your array contains logical drives of different RAID levels, it will not be possible to change any RAID levels using LDM.

To convert from RAID-0 to RAID-5 requires an additional disk to be added to the RAID-0 array. After you select the ready disk drive you wish to add, the summary of new configuration will appear, as in this example:

Below is the configuration change initiated by the logical-drive migration function. To save this configuration, click 'Apply.' To cancel the changes, click 'Cancel.'

Configuration summary for ServeRAID controller 1, array A

Logical drive	Size	New size	Level	New level
1	2000	2000	0	5
2	2000	2000	0	5
3	4678	4678	0	5
Free space	8678	17356	None	None

Figure 42. Summary of LDM configuration changes

Changing the RAID level of logical drives is transparent to the operating system, which observes no apparent change to its disk configuration because the sizes of all logical drives remain the same. Therefore, no shutdown and reboot is required.

### **Increasing the free space in an array**

By selecting this option, the existing logical drive sizes remain unchanged after adding new physical disk drives. As shown in Figure 43, the free space increases. You should create new logical drives in order to utilize the newly added disk space. It is important to realize that the existing logical drive structure changes, even though their sizes do not. All the data blocks belonging to existing logical drives are restriped to include the new disk drives. This is illustrated in Figure 43:

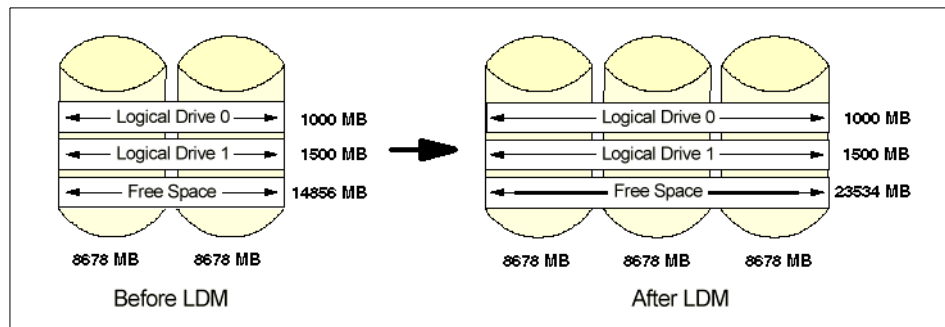


Figure 43. Increasing free space in the array

After creating new logical drives, you should create and format partitions with appropriate operating system utility. A reboot may be required at this point,

depending on the operating system that you use. For example, Windows NT Disk Administrator will not be able to detect new logical drives without shutting down and rebooting.

### **Increasing the logical drive sizes**

A second approach to integrating one or more new hard disks into an existing array is to increase the size of all logical drives in the array. Each of the logical drives increases in size in the same proportion, as you can see in Figure 44:

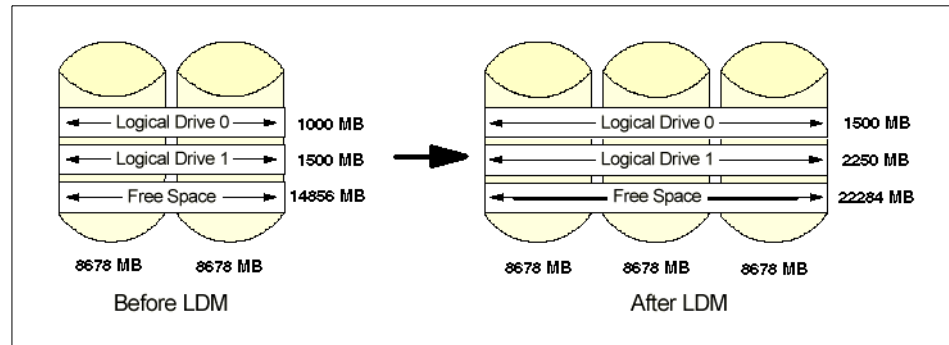


Figure 44. Increasing logical drive sizes

After the LDM process completes, the enlarged logical drives will appear to the operating system as larger physical drives. The extra disk space can be used for creating additional partitions or, alternatively, you can use Partition Magic or a similar tool to expand existing partitions.

**Note:** some operating systems will require a shutdown and reboot before detecting the additional disk capacity. For example, Windows NT Disk Administrator will only report the increased capacities after shutting down and rebooting. Note that the increased disk size is shown as free space:



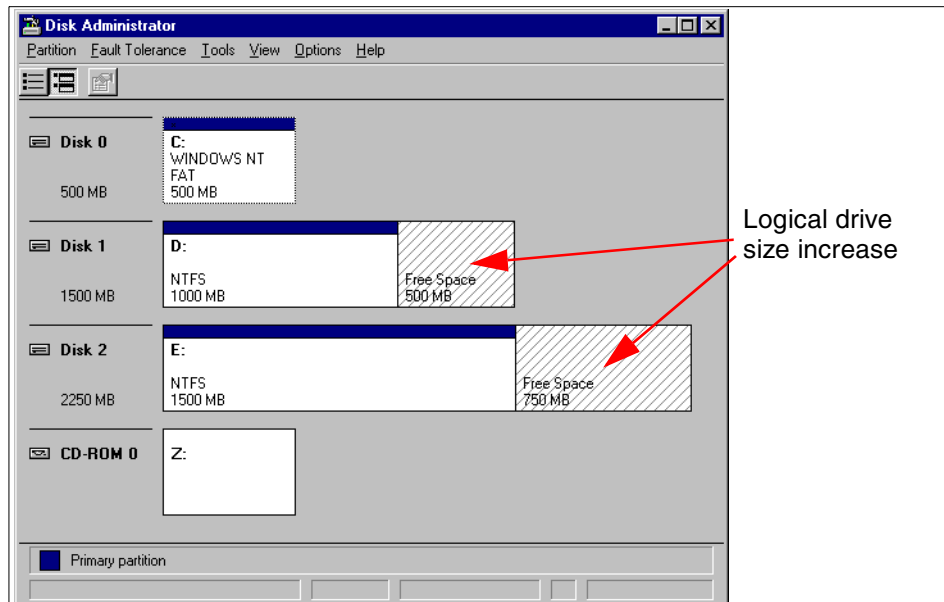


Figure 45. Windows NT Disk Administrator after increasing logical drive sizes

### 5.3.3 Recovering from physical disk drive failure

Another important task of the ServeRAID Manager utility is rebuilding failed disk drives. Only failed disks within fault-tolerant RAID logical drives can be rebuilt, that is RAID levels 1, 1E, 5, and 5E.

Two possible scenarios exist:

- A hot-spare disk drive is not defined.

The rebuild process can only start after the failed disk drive is replaced. The Hot-swap Rebuild parameter determines how the rebuild process will begin. If this parameter is *enabled*, the rebuild will start automatically as soon as the replacement drive is inserted into the drive bay. If the parameter is *disabled*, then you have to initiate the rebuild process manually using ServeRAID Manager. Once the process starts, you can check the rebuild status using its progress indicator.

- A hot spare disk drive is defined.

In this case, the hot spare disk drive will automatically be used as a replacement drive, and any failed disk attached to the controller (assuming the failed disk belongs to a fault-tolerant logical drive) will be rebuilt to this drive. When the failed disk drive is physically replaced, the new drive becomes the new hot spare drive. Note that this means that whenever a

disk drive failure occurs and hot spare is used, the stripe order for the array will change.

If the Hot-swap Rebuild parameter is *enabled*, the replacement drive will automatically become the new hot spare drive. If the parameter is *disabled*, this will have to be done manually after replacing.

**Hot spare restriction**

Note that a hot spare drive cannot be used if its capacity is smaller than that of the failed disk drive.

Figure 46 shows the ServeRAID Manager window during the rebuild process:

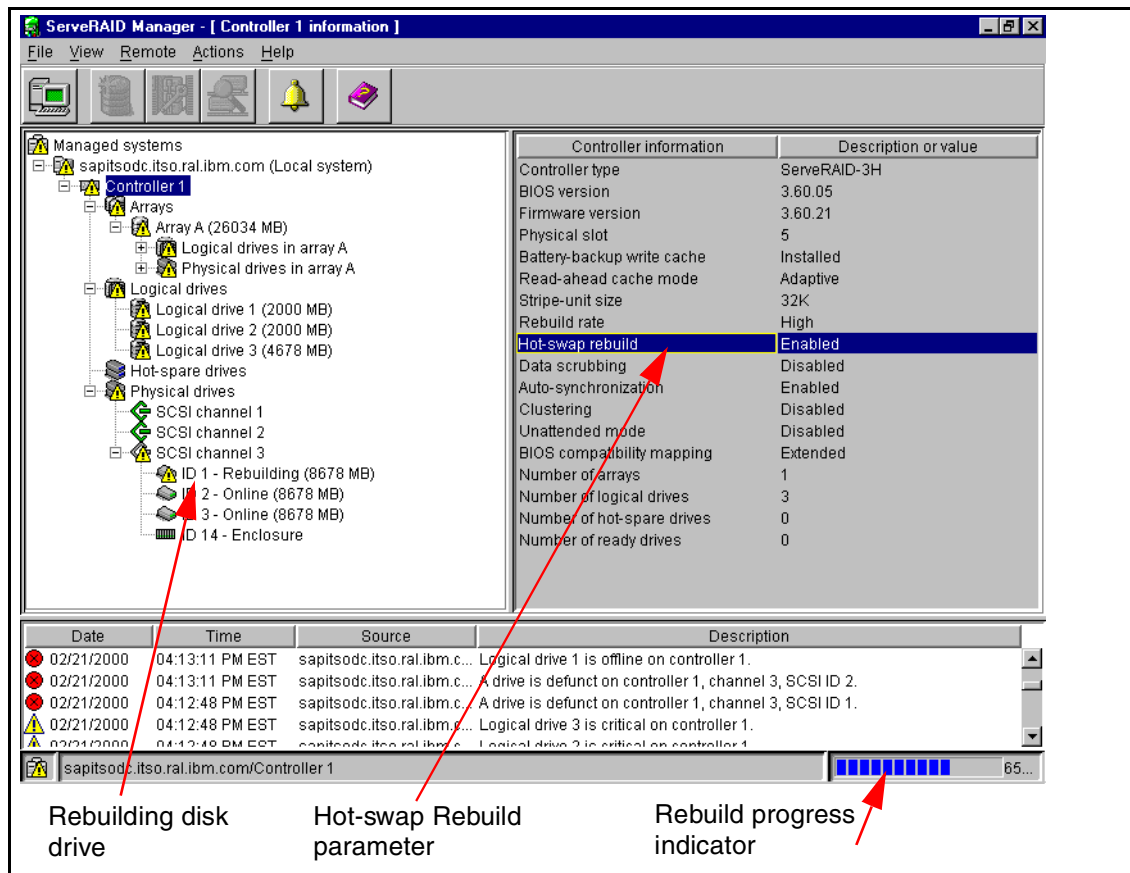


Figure 46. Rebuilding a disk drive

### ***Recovering from multiple failed disk drives***

All RAID levels apart from RAID-0 (and hence RAID-00) provide protection against a single disk drive failure. If more than one disk drive fails, however, the data on the disk subsystem will become inaccessible and the applications or even the entire server will go down. RAID-5E can be an exception, because it can tolerate two failed disk drives as long as there is enough time between the two disk failures for RAID-5E logical drive compression to RAID-5 to complete successfully.

If multiple disk drives fail in an array where the operating system resides, this will disable the server. You will have to use the bootable CD with the ServeRAID Configuration Program to repair the disk subsystem. However, if the failures occur in an array that contains applications and data, but not the operating system, it will be possible to use the ServeRAID Manager to manage the array. Applications and data in other arrays will remain accessible and the server will still be operational to some degree.

Physical failures of more than one disk drive in a short period of time are extremely rare. When multiple disk drive failures occur, this is in many cases the result of a problem elsewhere in the system. For example, the SCSI bus connections or termination could be bad, or a SCSI device might be malfunctioning and causing interference on the bus. Due to problems of this nature, the ServeRAID adapter could be unable to communicate with disk drives correctly and therefore falsely change their states to *defunct*. If this happens to more than one disk drive, the array will not function anymore.

In most cases, only one disk drive is really malfunctioning, and it is usually the one that goes to the defunct state first. It may be possible for logical drives in the array to be brought back online by keeping the first defunct drive in a failed state and reverting all other defunct drive states to *online*. ServeRAID Manager or the ServeRAID Configuration Program can be used to achieve this. Doing so will bring RAID-1 and RAID-5 logical drives from the *Offline* to *Critical* state, which means you can access the data and the server should again be operational at this point.

When the first drive goes to the defunct state (whether it really has failed or is a result of a masked problem), it can no longer participate in the array, while the other drives continue to be accessed for reading and writing. This means that the first defunct drive will not be in-sync with other drives anymore. If it only appears to have failed, you could simply revert it back to the online state, but if you do so you may corrupt your data. This is because its data is likely to be out-of-sync with the other drives in the array. To avoid this problem, the first defunct drive must always be rebuilt.

Data on other defunct drives remains in sync; as soon as one of them went into defunct state, all disk access was stopped immediately. So they can simply be brought back to the online state.

This means you will only be able to successfully recover from multiple defunct drives if you can identify the drive that failed first. ServeRAID Manager and IPSPMON utilities can provide this information, but only if they are up and running on the server. This is a very strong argument to have the ServeRAID utilities installed and operational. This information can also be found in the ServeRAID event log file. It is a good idea to save the logs to a text file in such a case. Do not clear the log after saving it to a file so it is available in original form for further troubleshooting. Customers should contact IBM technical support if a multiple disk drive failure occurs.

Figure 47 shows an example of multiple failed disk drives. In the figure, drives with SCSI IDs 1, 2 and 3 appear to be defunct. The message area in the lower part of the window can tell you the order of the drive failures. In our example, the disk drives have failed in this order: SCSI ID 2, then 1, then 3.

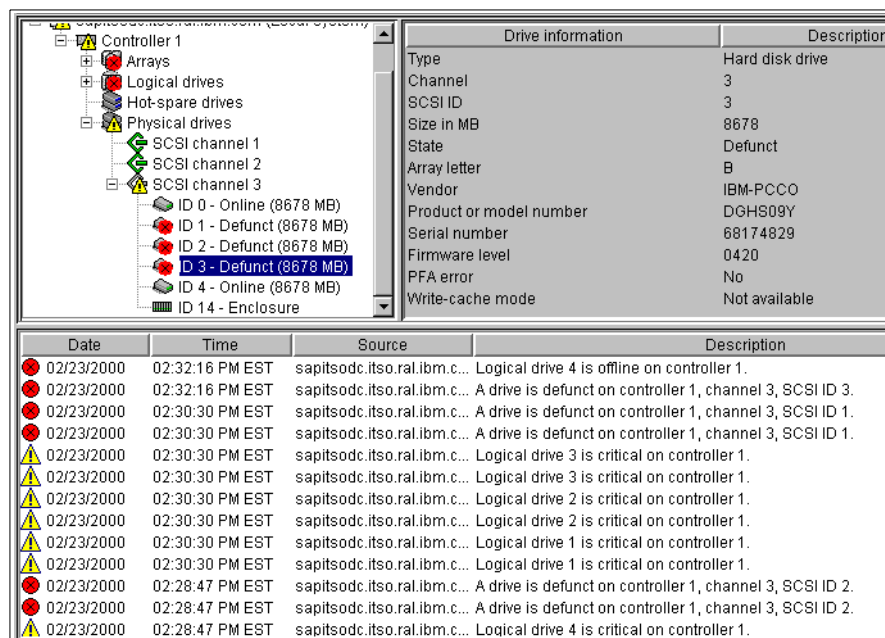


Figure 47. Multiple failed disk drives

A proper way to recover would be to set the drives with SCSI IDs 3 and 1 to online, as you can see in Figure 48:

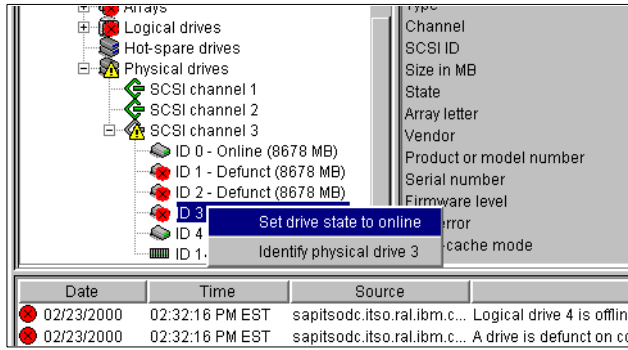


Figure 48. Bringing the disk drive online

The next step is to replace and rebuild the remaining defunct disk drive. Since this is now the only defunct disk drive in the array, ServeRAID Manager will not give you an option to set the drive state to online. Only the Replace drive and rebuild option will be available (see Figure 49).

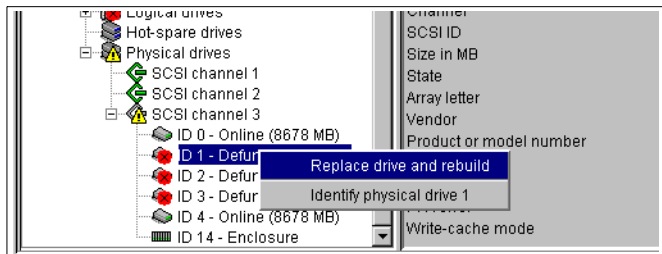


Figure 49. Replacing and rebuilding the disk drive

**Note**

The above discussion applies only when multiple drives appear to be defunct, but only one disk drive has actually failed. If two or more disk drives fail, you will have to replace them and restore your data from a backup.

If, however, you can recover from multiple defunct disk drives using the above procedure, meaning that the drives are functional, you should perform all necessary troubleshooting actions to determine the cause. If not, you will probably experience similar incidents in the future.

### 5.3.4 Remote system management

This feature allows you to use a computer to manage the disk subsystems on remote servers having ServeRAID adapters. Network administrators will often want to perform such management from their own workstation. ServeRAID Manager must be up and running both on the server and the workstation. To reduce the memory footprint, you can run the ServeRAID Manager on the server as an *agent*, that is, without the graphical user interface. To run ServeRAID Manager as an agent, execute RAIDAGNT.BAT in the ServeRAID Manager installation directory (\PROGRAM FILES\RAIDMAN by default).

**Note:** You should use the ServeRAID Manager GUI to configure and set up the ServeRAID Manager Agent notification list and security list first. Do not attempt to run the ServeRAID Manager GUI and ServeRAID Manager Agent at the same time because they share the same TCP/IP port number.

ServeRAID Manager uses TCP/IP for communication between the managing workstation and the servers with ServeRAID adapters. The default port used by the software is 34571, but you can change this in the User Preferences window, which is accessed by clicking **File > User Preferences**:

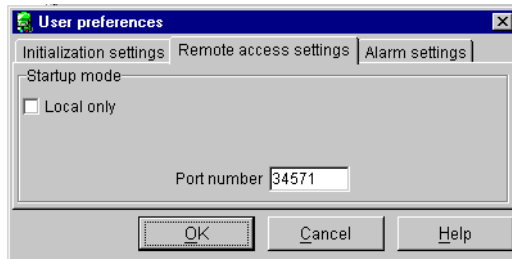


Figure 50. TCP/IP port number for remote management

By default, security is enabled, requiring a valid user name and password in order to establish a connection to the server. Therefore, you must create at least one user name and password combination before you can connect from your workstation. If you are not concerned about security at this level, it may be disabled.

The Security Manager tool inside ServeRAID Manager (see Figure 51) allows you to create, modify and delete user names and passwords.

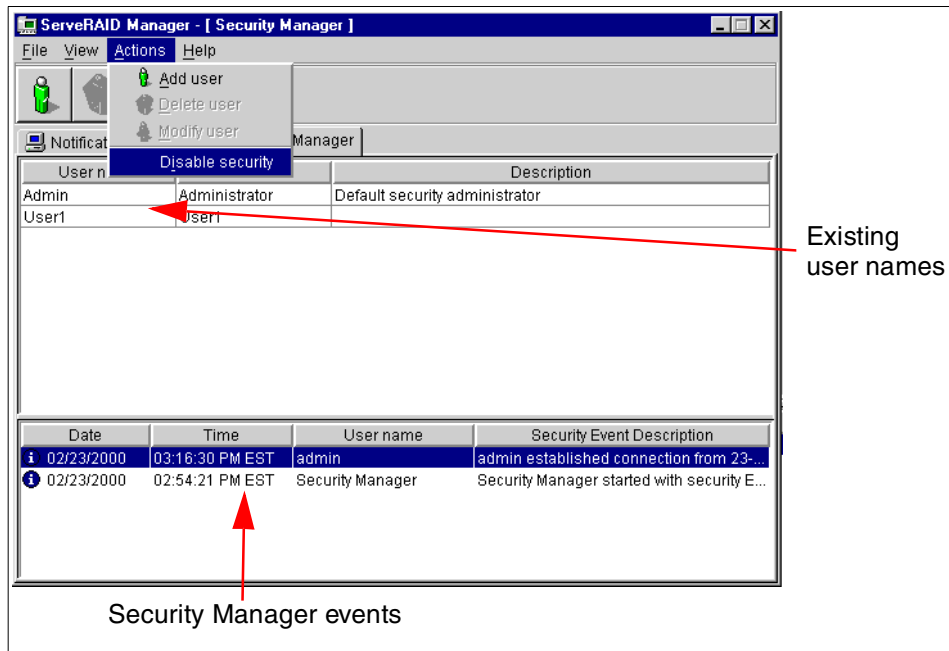


Figure 51. Security Manager

On the workstation, select **Add remote system** in the Remote menu and complete the following dialog box:

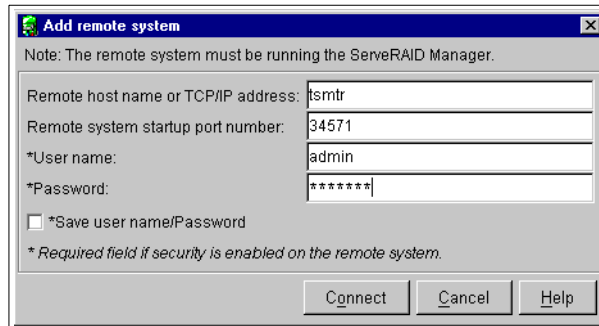


Figure 52. Add remote system

If the hostname or IP address, port, user name, and password are all valid, connection will be established and you will be able to manage the ServeRAID adapters in the same way as you would for locally attached disks. You can be connected to several servers at the same time.

Notification Manager allows you to specify which systems will be notified about events that occur on the local system. You simply add the hostnames or IP addresses of those systems. ServeRAID Manager must be running on all of the systems you specify.

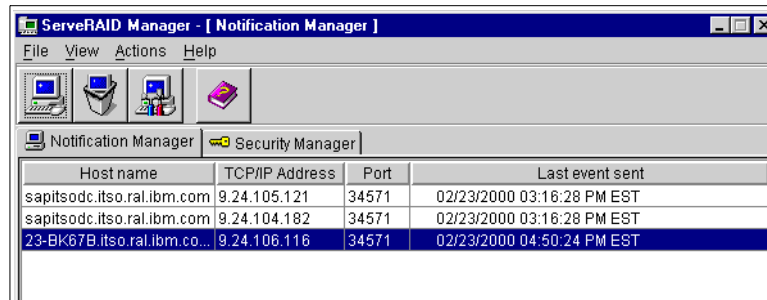


Figure 53. Notification Manager

ServeRAID Manager generates and maintains four log files, which can prove very useful during recovery procedures:

- RaidEvt.LOG - this file contains all events from ServeRAID Manager's event viewer.
- RaidNot.LOG - this is Notification Manager's event viewer.
- RaidSec.LOG - this is Security Manager's event viewer.
- RaidErr.Log - this file contains ServeRAID Manager-generated Java messages.

These files reside in the ServeRAID Manager directory. New events will be appended to the files until their size reaches 200,000 bytes. At this point, the file's extension is renamed to .OLD and a new file is created. If a .OLD file already exists, it will be deleted.

---

## 5.4 Active PCI support

The current midrange and high-end Netfinity servers have a number of PCI slots that conform to the Active PCI specification. These slots allow the insertion or removal of specified PCI adapters while the server is up and running in order to minimize downtime. Active PCI functionality is supported on all ServeRAID-4, ServeRAID-3 and ServeRAID II adapters.

The IBM solution is a combination of hardware, drivers, and management applets that let you manage power and control circuitry to safely add or



remove PCI adapters. With ServeRAID adapters (other than the original ServeRAID adapter), you can:

- Add a new ServeRAID adapter while the server is running and configure it without rebooting (called a *hot-add*).
- Remove an existing ServeRAID adapter while the server is running and replace it with another identical adapter (called a *hot-swap*).

In conjunction with this hot-plug function, the ServeRAID adapters can be installed in a fault-tolerant redundant pair configuration. You can achieve the highest level of availability when you use both redundancy and hot-pluggability. This way, if an adapter fails, its partner will take over, keeping the server operational. With hot-pluggability, you can now replace the failed adapter while the server remains operational.

See 5.5, “Configuring a fault-tolerant pair” on page 114 for more information on fault-tolerant pair support.

#### 5.4.1 Active PCI software and hardware components

The IBM PCI hot-plug solution is made up of the following software components:

- Fault-tolerant device drivers
- A Desktop Management Interface (DMI) agent
- The IBM PCI hot-plug solution

The diskette image files containing these components can be downloaded from the IBM support Web page at:

<http://www.pc.ibm.com/support>

The installation procedure is simple. You have to replace the existing device drivers with the fault-tolerant ones and you then run **Setup** from the DMI agent and IBM PCI hot-plug solution diskettes. A reboot will be required after installing the device drivers and the DMI Agent. The components should be installed in the above order. Make sure you review the readme file on each diskette; there you can find the latest details about the installation.

ServeRAID driver Version 3.60 and above is fault-tolerant, so you only need to install the DMI agent and IBM PCI hot-plug solution.

**Note:** If you want to use a fault-tolerant pair of ServeRAID adapters, you have to install an additional component: the IBM ServeRAID DMI Component Service.

This component is required for the Fault Tolerant Management Interface to be able to work with a fault-tolerant ServeRAID adapter pair. If you only need the Active PCI functionality and you do not plan to use ServeRAID adapter pairs, you do not need to install this component.

In order to use the IBM PCI hot-plug solution, you need the following hardware:

- A Netfinity server with Active PCI slots.
- PCI adapters that support the Active PCI specification (ServeRAID adapters, Netfinity 10/100 Fault Tolerant Ethernet Adapter, Token Ring 16/4 PCI Adapter 2 and so on). A list of supported adapters can be found at:


<http://www.pc.ibm.com/us/compat/hotplug>

#### 5.4.2 The hot-plug tools

Hot-plug services install the following tools that help you monitor and configure the Active PCI slots and the devices in those slots:

- Netfinity hot-plug system tray pop-up menu
- PCI hot-plug controls
- PCI hot-plug Wizard
- PCI Hot-Swap Wizard

##### ***Netfinity hot-plug system tray pop-up menu***

The Netfinity hot-plug system tray pop-up menu gives you quick access to information about the Active PCI slots in the server. To access the menu, click the Netfinity hot-plug icon on the task bar  2:01 PM with the left mouse button. The system tray pop-up menu will appear as shown in Figure 54:

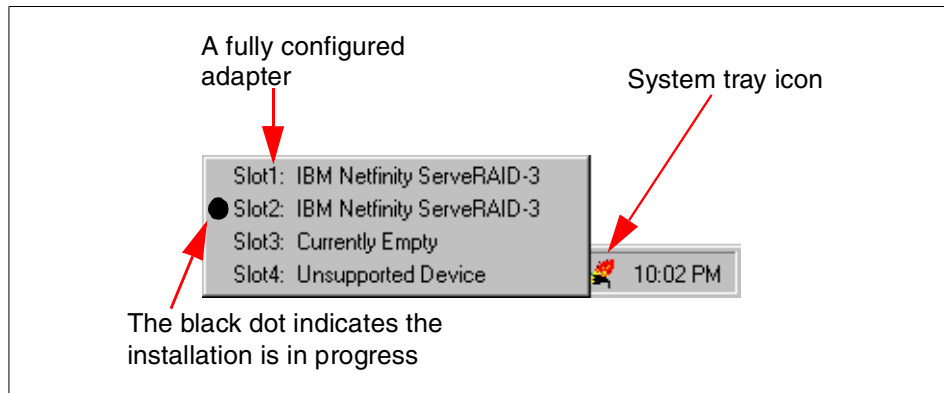


Figure 54. System tray pop-up menu

As you can see from Figure 54, slot 2 is marked with a black dot, indicating it contains an adapter that is pending installation. A pending adapter is not fully configured and requires user interaction to finish the installation. To complete configuration for this adapter, click the adapter name to start the IBM hot-plug Wizard as described in “PCI Hot Plug Hardware Wizard” on page 110.

This pop-up menu provides two functions:

- It allows you to resume a previously suspended hot-add operation
- It shows you the adapters that are installed in the hot-swap slots

### **PCI hot-plug controls**

This applet lets you monitor the adapter and slot power status for each Active PCI slot.

**Note:** Version 1 of this applet also lets you manually turn the power to a slot on or off. This function was removed in Version 2.

You can open the applet by right-clicking on the hot-plug icon  2:01 PM on the Taskbar or by selecting the **Hot-plug Controls** icon from the Control Panel as shown in Figure 55:

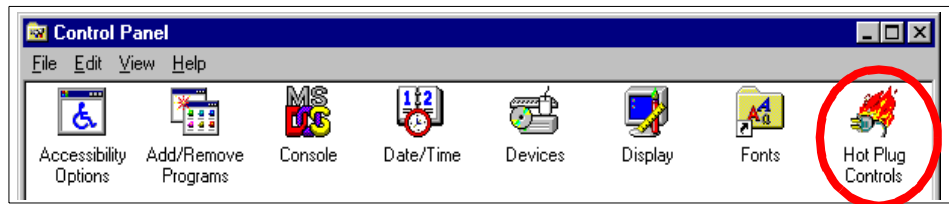


Figure 55. Hot Plug Controls icon in the Control Panel

Using either method, the Hot Plug Slot Properties window appears as shown in Figure 56 below.

**Note:** You cannot open both the Hot Plug Slot Properties and the Hot-Swap Wizard windows at the same time.

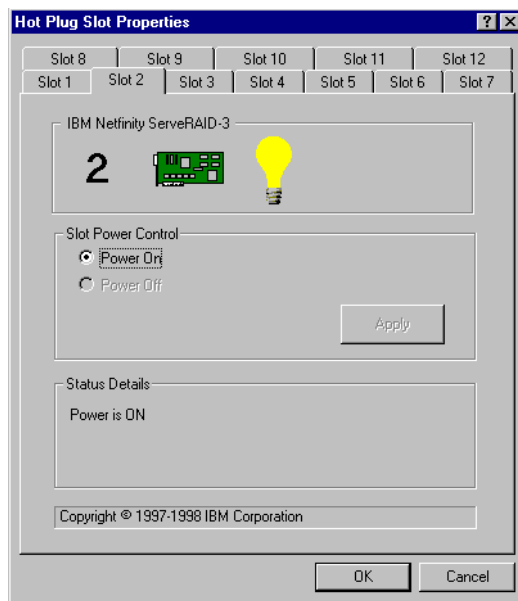


Figure 56. Hot Plug Slot Properties

The utility allows you to control the power for vacant PCI slots, but you cannot power-off an occupied slot.

### **PCI Hot Plug Hardware Wizard**

The PCI Hot Plug Hardware Wizard automatically starts when you insert a supported adapter into one of the hot-plug PCI slots. The wizard will guide you through the configuration and startup of the adapter.

If you choose to cancel the adapter configuration, you can restart this Wizard with the system tray pop-up menu as described in , “Netfinity hot-plug system tray pop-up menu” on page 108.

Here are some general rules and guidelines:

- After a hot-add, there is no need to reboot the server, even if Windows NT asks you to do so.
- When you attempt to hot-add a ServeRAID adapter in a hot-plug system that has already had the adapter installed and then removed, hot-add will fail.

The administrator is strongly advised to mark all SCSI drivers which have no devices connected as Disabled using **Control Panel > Devices** and then reboot the system so that these changes can take effect. Doing so will prevent the driver from being started and allow the hot-add to proceed normally.

- Hot-remove is not supported with Windows NT 4.0. That is, you cannot remove a configured adapter without also replacing it with another adapter.

This is because Windows NT 4.0 does not allow stopping the device drivers for network or SCSI adapters once they have been started. Since the drivers can't be stopped, the hardware can't be removed without causing problems.

- We strongly recommend that you perform any hot-plug function while system I/O activity is low.

During the process of hot-adding or hot-swapping an adapter, there is a window of time where the machine can become unresponsive while the new hardware is initialized. This delay varies, depending upon the hardware and its current configuration. During this time, services on the machine will not be available and this could cause errors in any applications that may time out.

To hot-add a new ServeRAID adapter perform the following steps:

1. Make sure that you are logged on as a user with administrative rights.
2. Insert the ServeRAID adapter firmly into a vacant Active PCI slot. Use the plastic guides to ensure the adapter is properly seated.
3. Connect any applicable adapter cables.
4. Lower the black plastic tab and close the adapter retention latch.
5. The power LED next to the slot will turn on and the Hot Plug Wizard will start automatically as shown in Figure 57:



Figure 57. Hot-adding of a ServeRAID adapter

6. Follow the instructions in the utility until the adapter is started successfully. click **Finish** to end the process.

**Note:** Do not reboot. Even if prompted, you do not need to reboot.

Once the process completes, you can use ServeRAID Manager to create new arrays and logical drives on newly attached disk drives.

If a problem occurs during installation, refer to the Event Viewer for information about the installation process.

#### ***PCI Hot-Swap Wizard - replacing a ServeRAID adapter***

You have to use this utility to hot-replace a ServeRAID or another supported PCI adapter.

**Note:** You must use an adapter of the same hardware level as a replacement. For example, you cannot remove a ServeRAID II adapter and replace it with a ServeRAID-3HB adapter using the hot-swap process.

Although it is not necessary, we recommend that both adapters have the same level of firmware and the same level of BIOS on them. You should, however, be sure to connect the disk drives to the same SCSI channels to which they were connected on the old adapter. The replacement adapter will obtain the RAID configuration from the disk drives and the operation will fail if the configuration does not match the information from the drives.

During a hot-swap of a stand-alone ServeRAID adapter (that is, an adapter that is not part of a fault-tolerant pair) or the *active* adapter of the ServeRAID fault-tolerant pair, any I/O performed on the adapter during this operation will fail. The applications attempting to use the adapter will get an error code informing them that the adapter is busy.

**Note:** The hot-swap of a single ServeRAID adapter that contains the operating system boot partition is not supported.

Before you start, make sure that you are logged on as a user with administrative rights.

1. Open the Hot-Swap Wizard by clicking on **Start > Programs > IBM PCI Hot Plug Applications > IBM PCI Hot-Swap**.
2. The Hot-Swap Wizard will notify you of any non-hot-pluggable PCI adapters that it finds in Active PCI slots. Next, it will display a list of supported PCI adapters that can be hot-swapped (Figure 58).



Figure 58. Hot-Swap Wizard - adapter selection

An adapter will only appear on the list if the fault-tolerant driver is installed.

**Note:** If you have several identical adapters in your server, make sure you select the correct one.

3. The wizard will now guide you through the process of swapping the adapter. It is extremely important to read and follow the wizard on-screen instructions carefully. The wizard will tell you exactly what to do: when to

insert the new adapter, open or close the slot latch and connect the SCSI cables. An example is shown in Figure 59. If you do not follow the steps precisely, the operation will fail.



Figure 59. Hot-Swap Wizard - replacing a ServeRAID adapter

Once the process is finished, you can start using the replaced adapter immediately.

---

## 5.5 Configuring a fault-tolerant pair

With this feature, you can configure a ServeRAID adapter pair and connect both adapters to the same storage enclosure in order to provide access to the disk drives even after one of the adapters has failed.

You can use this feature on ServeRAID-4, ServeRAID-3 and ServeRAID II adapters. The original ServeRAID adapter and the ServeRAID controllers integrated onto the planar of some Netfinity systems are not supported.

**Note:** You can only use a fault-tolerant pair of ServeRAID adapters with disk drives installed in external disk drive enclosures. Disks connected to internal server backplanes are not supported.

The hardware setup is similar to the cabling of a Microsoft Cluster Server-based cluster setup except that both adapters are installed in the same server. The disk subsystem should be connected to the same SCSI channel on both adapters.



### 5.5.1 Failover

The fault-tolerant pair is configured as an active/passive arrangement. The *active* adapter has all RAID arrays, logical drives, and hot spares defined to it and the *passive* adapter does not have any devices defined.

When a failover occurs, all configured devices are transferred to the passive adapter, which then becomes the active adapter.

You can install a fault-tolerant pair of ServeRAID adapters into non-active or active PCI slots. Using active PCI slots will increase the availability of your system because you can avoid server downtime while replacing the failed adapter.

A failover will occur whenever the ServeRAID device driver is unable to send a command to the active adapter, and the active adapter does not respond to a reset command. Specifically, a failover will occur in the following situations:

- A failover request was initiated manually through the Fault Tolerant Management Interface applet or ServeRAID Manager.
- The active adapter fails.
- The power to the slot where the active adapter is installed is turned off.

**Note:** An automatic failover will only occur after I/O is attempted to the failing active adapter.

A failover will *not* occur in the following situations:

- The system's SCSI cables are loose.
- Failure of an attached disk (you use disk arrays to protect against this).
- Failure of an external disk enclosure (you can use disk arrays across multiple enclosures to protect against this).

### 5.5.2 Guidelines and restrictions

Here are some general rules and guidelines:

- A fault-tolerant pair cannot be part of an MSCS shared cluster.
- A ServeRAID controller integrated into the system's motherboard can be disabled if required.
- Only the active adapter in a fault-tolerant pair can be managed by the ServeRAID Manager. The passive adapter is greyed out.
- Configuring a fault-tolerant pair requires a reboot.
- All logical drives defined must have cache configured as write-through.

- When using multiple fault-tolerant pairs in a system, each set of adapters must have a unique set of names. The maximum number of fault-tolerant adapter pairs is determined by the maximum supported adapters in your system.
- All shared SCSI channels among the paired adapters must have non-conflicting SCSI IDs.
- Both cards in the pair must have their host names and partner names configured correctly.
- All logical drives must have unique shared logical disk IDs, which are set up only on the first adapter.
- If both adapters have logical drives configured, they cannot be configured as a fault-tolerant pair.

### 5.5.3 Installation

There are three scenarios where you can use a ServeRAID fault-tolerant pair:

1. Configure two new ServeRAID adapters as a fault-tolerant pair and install Microsoft Windows NT on a logical drive in the enclosure connected to the adapter pair.

**Note:** In order to gain fault tolerance for the operating system, NT must be installed on a shared logical drive residing in the enclosure.

2. Add two adapters to your system where Windows NT is already installed on internal hard disks (that is, not on a logical drive in the enclosure). In this scenario, the operating system will not be protected from a failed ServeRAID adapter.
3. Add a second ServeRAID adapter to your system where Microsoft Windows NT is already installed on a logical drive residing in the enclosure. The existing and new adapter can then be configured to form a fault-tolerant pair.

**Note:** All of the above scenarios will require the server to be powered off to configure the adapter pair. The offline configurator is used to configure the SCSI IDs and merge IDs.

The overall steps in the configuration are:

1. Install the Fault Tolerant Management software
2. Power the server and enclosure off
3. Install the adapter(s)
4. Connect the enclosure to the first adapter

5. Boot the ServeRAID Configuration Program CD
6. Configure the first adapter
7. Configure the second adapter
8. Assign merge IDs
9. Connect the enclosure to the second adapter
10. Restart the server

***The Fault Tolerant Management Interface applet***

The Fault Tolerant Management Interface is a part of the PCI Hot Plug Solution and it installs from the same diskette and at the same time as Hot Plug and Hot-Swap Wizards. It is a Control Panel applet, which lets you do the following:

- View the current status of the fault-tolerant pair.
- Manually force the active adapter to failover, thereby transferring control of the RAID arrays and logical drives to the previously passive adapter.

The applet requires the ServeRAID fault-tolerant device driver and the IBM fault-tolerant DMI agent to be installed. An additional component, the *IBM ServeRAID DMI Component Service*, is required for the Fault Tolerant Management Interface to be able to work with a fault-tolerant ServeRAID adapter pair.

The installation of these components is discussed in 5.4.1, “Active PCI software and hardware components” on page 107.

All the required diskette image files can be obtained from the IBM support Web page at:

<http://www.ibm.com/pc/support>

When the Fault Tolerant Management Interface applet is installed, a new icon is placed in the Windows NT Control Panel as shown in Figure 60:



Figure 60. Fault-Tolerant Management Interface applet icon

### 5.5.3.1 Configuring the adapter pair

To configure the adapters follow these steps:

1. Insert the ServeRAID Configuration Program CD.
2. Shut down the server and the disk enclosure and power them off.
3. Install both adapters. If the first adapter is already installed then just install the second adapter.
4. If not already connected, connect the disk enclosure to the first adapter. Leave the second adapter unconnected for now. The enclosure will be connected to the second adapter in step 19.
5. Power on the enclosure and server and when the ServeRAID BIOS messages appear on the screen, press Ctrl+I to start the Mini-Configuration Utility.
6. For each adapter installed in the server, set Multiple BIOS mode to *Shrink*. This parameter is set by clicking **Advanced Functions**, then **Configure BIOS Settings**.
7. Exit the Mini-Configuration Utility, restart the server and boot from the ServeRAID Configuration CD.
8. If there are no logical drives configured (that is, you are using Scenario 1 or Scenario 2 as described in 5.5.3, "Installation" on page 116), you will need to initialize the first adapter. If you are using Scenario 3, skip to step 12.
9. If necessary, cancel out of the array configuration function to return to the adapter view function.

**Note:** In this example, controllers 2 and 3 as shown in Figure 61 will be the fault-tolerant pair:

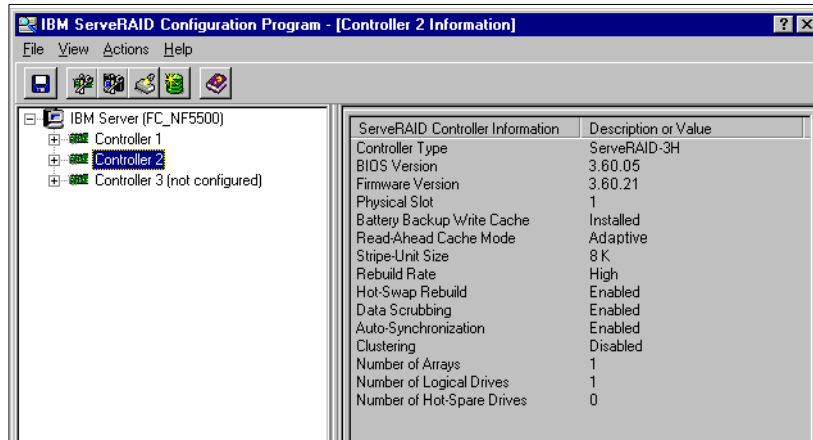


Figure 61. Configuring fault-tolerant pair

10. Right-click the first adapter ("Controller 2" in our server) and select **Restore to Factory Defaults**. Confirm the operation.
11. Configure arrays and logical drives, as described in 5.3.1, "Creating arrays and logical drives" on page 93.
12. Configure the SCSI IDs and host IDs of the first adapter in the pair and the merge IDs for each logical drive defined to the first adapter. Merge ID assignment is very important here; it determines which logical drives will be participating in the fault-tolerant environment. Follow these steps:
  - a. Right-click the first adapter and click **Configure for Clustering**.
  - b. Enter values for the host IDs:
    - Controller Host ID
    - Cluster Partner Host ID

For our configuration in Figure 62 on page 120, we typed in PAIR-A1 and PAIR-A2, respectively. Note that these values are case sensitive.
  - c. Set the SCSI ID for each SCSI channel on this adapter to 6 (the default is 7). Both adapters will appear on the same SCSI bus and you must make sure that their SCSI IDs do not clash.
  - d. For each logical drive, check **Shared** and specify a unique merge ID for each one in the range 1-8 (you can have up to eight logical drives defined to an adapter). These settings are shown in Figure 62:

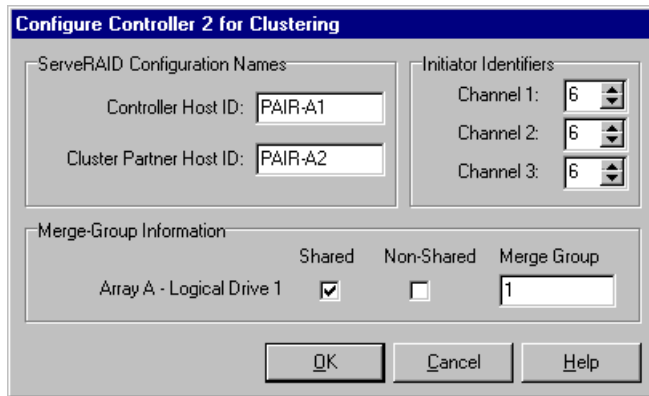


Figure 62. Configure for clustering: entering parameters

13. Initialize the second adapter by right-clicking it ("Controller 3" in our server) and select **Reset to Factory Defaults**.
14. Right-click the second adapter and select **Configure for Clustering**.
15. Change the Controller Host ID and Cluster Partner Host ID values to be the reverse of those you specified in step 12. In our case, these are set to PAIR-A2 and PAIR-A1, respectively. Ensure the Initiator Identifiers are set to 7 for each channel. Your configuration window should be similar to Figure 63:

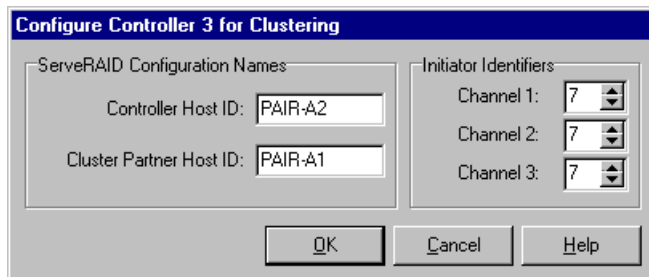


Figure 63. Configuring the second adapter

**Note**

Once you return to the main configuration window, the second adapter will still be marked as "not configured". This is normal.

16. Make sure *Unattended Mode* is **Enabled** on both adapters.
17. Exit the configuration utility and remove the CD-ROM from the drive.

18. Power off the server and then the external enclosure.
19. Connect the second SCSI cable from the other external enclosure connector to the second adapter. Ensure you use the same SCSI connector on the second adapter as you did on the first adapter. That is, the shared bus must use the same channel on each adapter.
20. Power on the enclosure, then the server.

When the operating system restarts, the two adapters are now configured as a fault-tolerant pair.

#### 5.5.4 Working with the fault-tolerant pair

You can use two utilities to monitor and manage the operation of a fault-tolerant ServeRAID pair:

- The Fault-Tolerant Management Interface applet in Control Panel
- ServeRAID Manager

##### ***Using the Fault Tolerant Management Interface applet***

When you start the applet, the following window appears:

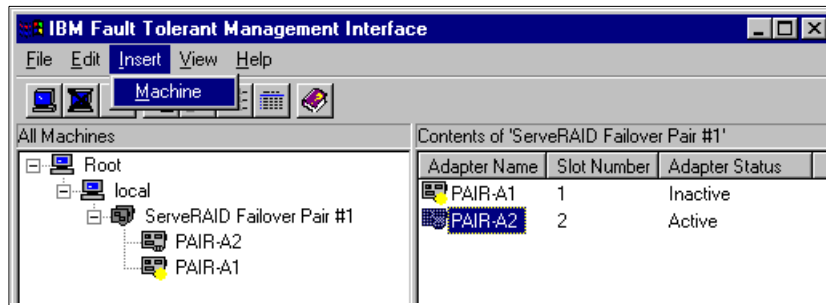


Figure 64. IBM Fault Tolerant Management Interface

You can see all configured ServeRAID pairs. Note that you can also connect to other servers and display the ServeRAID pairs configured on them. To do this, select menu item **Insert -> Machine**, as shown in Figure 64. You will be able to enter the other server's machine name or browse for it using the following window:

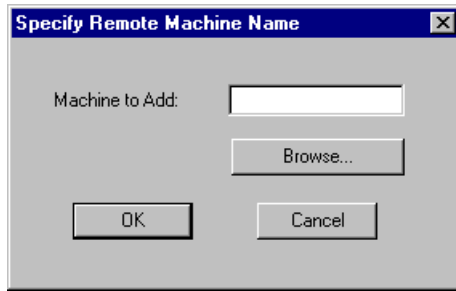


Figure 65. Adding remote server

The applet shows you the status of both adapters in a pair. You can perform a manual failover of the adapters by right-clicking the active adapter and selecting **Force Failover**, as shown in Figure 66:

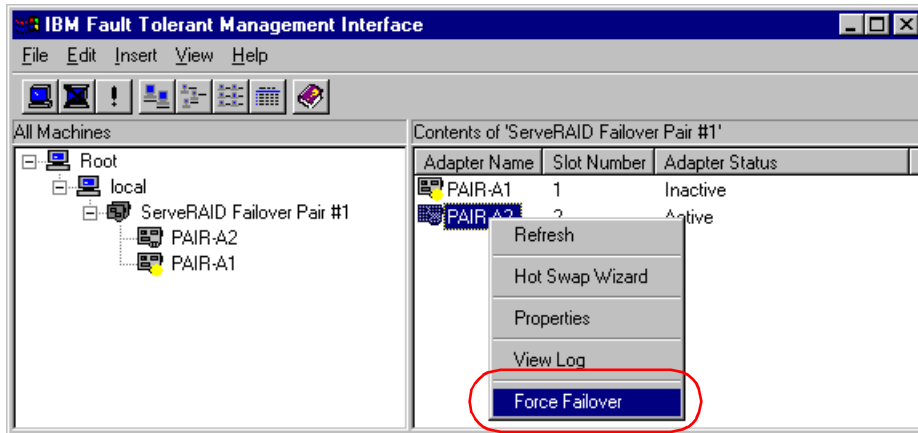


Figure 66. Manual failover

Note that you can also launch the Hot-Swap Wizard from this menu. Usually you do not want to do this for your active adapter. Using this option starts the Hot-Swap Wizard to replace an adapter that has failed.

The **View Log** option will start the Windows NT Event Viewer. All ServeRAID pair related events are written to this log.

### ***Fault-tolerant pairs in ServeRAID Manager***

ServeRAID Manager is another tool you can use to monitor and manage the pairs of ServeRAID adapters. Once a fault-tolerant pair has been configured, it appears in ServeRAID Manager as shown in Figure 67:



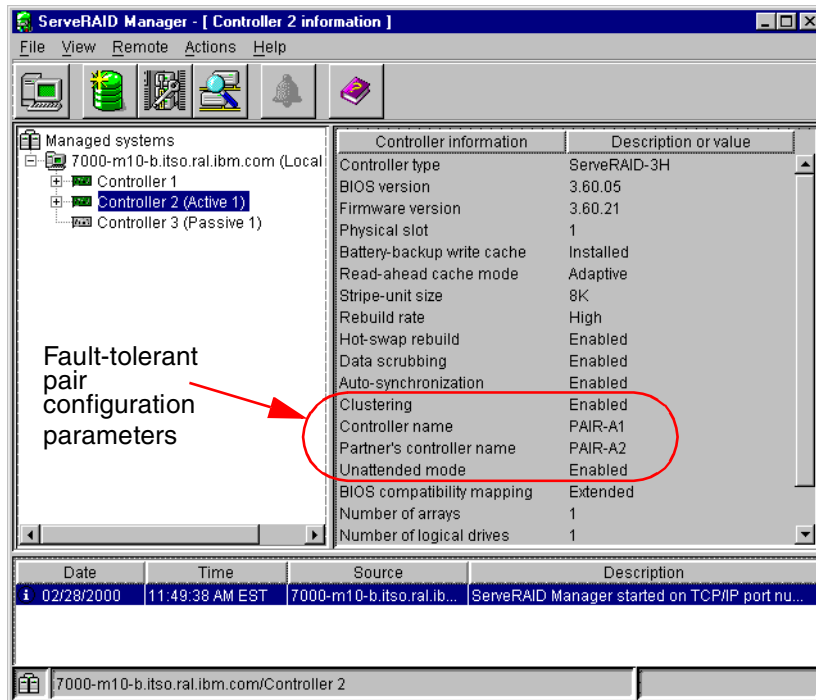


Figure 67. Fault-tolerant pair in ServeRAID Manager

You can fully manage the physical disk drives, arrays, and logical drives on the active adapter. The passive adapter, however, is greyed out and you cannot do anything other than display controller information.

ServeRAID Manager allows you to perform a manual failover. To do this, right-click the active adapter and then select **Clustering actions -> Fail from Active to Passive**. This action is available only when ServeRAID adapter pairs are configured, and is shown in Figure 68:

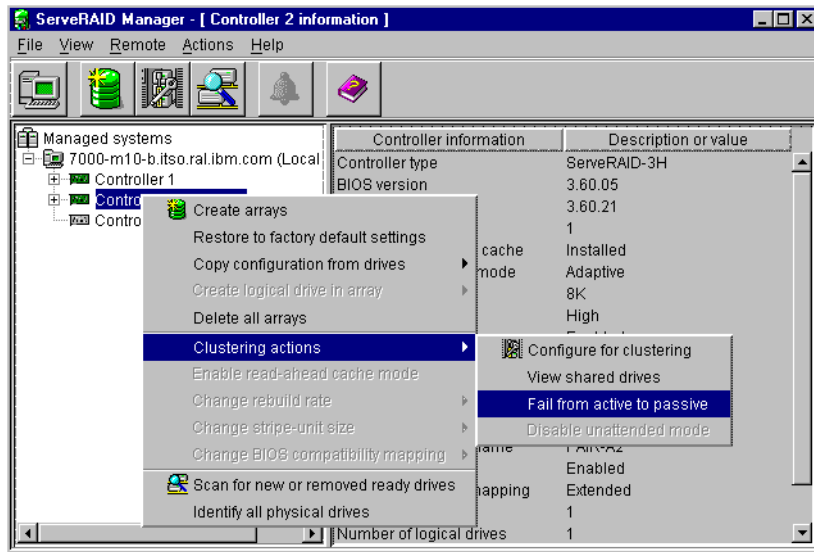


Figure 68. Manual failover using ServeRAID Manager

## 5.6 Clustering with ServeRAID

ServeRAID adapters, and SCSI subsystems in general, support a two-node shared disk cluster environment. Limitations on bus-length and device addressing make SCSI impractical for clusters with more nodes. If you do wish to use more than two nodes in a cluster, other technologies such as Netfinity's SSA or Fibre Channel disk subsystems need to be used.

In a ServeRAID-based cluster, the shared disk drives have to reside in external storage enclosures, and are connected to a ServeRAID adapter in each node, as shown in Figure 69. The configuration is similar to that for a fault-tolerant pair. In both cases you connect the two ServeRAID adapters to a common set of disks. There are differences, however:

- When using a fault-tolerant pair, one adapter is active and the other is passive. That is, only one adapter has access to the drives.
- When using clustering, both adapters can actively access the disks. The clustering software allocates ownership of logical drives to one node or the other and only the owning node can access a specific drive.

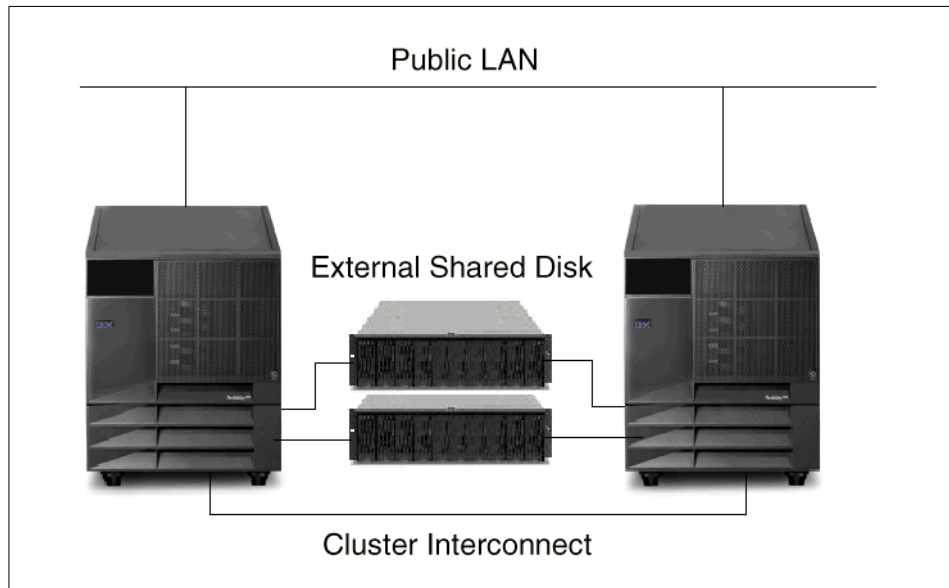


Figure 69. Two-node cluster with ServeRAID adapters

### **Operating system**

Each node has the operating system installed on its own local disk drives. These disk drives are not a part of clustered storage and may not even be attached to ServeRAID adapter. However, if they are attached to the ServeRAID adapter that is also connected to the shared storage, you must assign non-shared merge IDs to the operating system logical drives.

### **SCSI Initiator IDs on the shared SCSI bus**

The ServeRAID adapters will be connected to either an EXP200 or EXP300 external storage enclosure, using external SCSI cables. You should make sure that the two backplanes in the storage enclosure are configured as a single SCSI bus. It is also mandatory to connect the enclosure to the same SCSI channel on both adapters.

Taking a closer look at a single, shared SCSI channel, the devices on the SCSI bus are:

- The two ServeRAID adapters
- The backplane in the external storage enclosure
- The disk drives in the external storage enclosure

For correct operation, all devices on a SCSI bus must have unique SCSI IDs, so the ServeRAID adapters have to use different SCSI initiator IDs. Typically, we set the SCSI ID of the first adapter to 6 and leave the second adapter's ID at the default setting of 7, as previously illustrated in our discussion about configuring a fault-tolerant pair (5.5.3.1, "Configuring the adapter pair" on page 118).

#### ***Unattended mode***

You must enable unattended mode on both adapters. The active node that currently owns the disk drives will detect them as *online* and the standby node will not be able to access them at all; it will mark them as *defunct*. If unattended mode is disabled (the default), the standby node will not boot into the operating system, but rather stop booting at ServeRAID POST message and wait for user input.

#### ***Arrays and logical drives***

When creating arrays and logical drives, keep in mind that only one logical drive per array is allowed in a clustering configuration. Also, RAID-5 logical drives will not fail over if they are in a critical state. Therefore, it is highly recommended to implement a hot spare disk drive in clustering environment. A hot spare in a clustered environment is allocated to a specific adapter. This means that you must therefore assign a hot spare drive for each node.

#### ***Quorum disk drive***

When using Microsoft Cluster Server, the quorum log should reside on a RAID-1 logical drive. Locating the quorum on a RAID-5 logical drive can cause problems in the case of a disk failure, since the quorum resource will not be able to fail over until the failed disk drive is rebuilt to a hot spare or the drive is replaced and rebuilt. This could potentially disable the cluster if the active node failed.

#### ***Write-through cache policy***

Do not use a write-back cache policy for the shared logical drives. Doing so will cause data loss when failover occurs. There is no way to transfer dirty data in the ServeRAID adapter cache from the failing node to the surviving one. Therefore, all shared logical drives should use a write-through cache policy.

#### ***Clustering parameters***

You must specify the clustering parameters and again this is similar to fault-tolerant pair configuration:

- Specify the host adapter name and the partner adapter name.
- Assign shared merge IDs for all shared logical drives.

- Assign non-shared merge IDs for all non-shared logical drives on a ServeRAID adapter pair. For example, this could be the logical drives local to the server, which contain the operating system.

**Note:** you only create arrays and logical drives on one node. Also, the shared merge IDs are only assigned on the first node. The second node will pick the configuration and merge IDs up from the first node. But you must still create non-shared merge IDs for local logical drives on the second node.

#### ***IBM ServeRAID Windows NT Cluster Solution Diskette***

In the Microsoft Cluster Server environment, after the Microsoft Cluster Server is installed on both nodes, you must apply the IBM ServeRAID Windows NT Cluster Solution diskette. Do not forget to do it on both nodes. This will add the new *ServeRAID disk resource* into the MSCS environment. It will also create a group and a resource of this type for each shared logical drive.

When installing Microsoft Cluster Server in a ServeRAID environment, it is important to use the */localquorum* switch. The quorum resource can be moved to the shared logical drives only after installation completes because the correct resource type (ServeRAID disk resource) is available only at the end of the installation procedure.

For a detailed description of clustering solutions and installation procedures in a ServeRAID environment, see *IBM Netfinity High Availability Cluster Solutions Using the IBM ServeRAID-3H and IBM ServeRAID-3HB Ultra2 SCSI Controllers - Installation and User's Guide*, available for download at:

<http://www.pc.ibm.com/support>

To assist you with planning a clustered environment based on Netfinity servers, we recommend *Netfinity Clustering Planning Guide*, SG24-5845. This IBM Redbook discusses Microsoft Cluster Server and other clustering solutions for Windows NT and other operating systems.

---

## **5.7 Boot-time messages**

During its power-on self test (POST), the ServeRAID adapter compares stored configuration information with the configuration that it finds on the SCSI bus. If a discrepancy exists, one or more status messages appear after POST completes, but before the operating system loads and you will be given a list of options to select from. Possible situations that ServeRAID will detect include:

- New drives have been installed.
- Previously configured drives are missing.
- Previously configured drives are not in their configured location.
- A new adapter or imported drives have been detected.

***New drives have been installed***

In this situation, you will see the following message:

NN new Ready drive(s) found

This is an informational message only. Your next step would be to use the configuration or administration utilities to create arrays and logical drives or perform logical drive migration.

***Previously configured drives are missing***

If drives have been removed, or have failed and cannot be detected, you will see this message:

NN OnLine drive(s) not responding

If this is not what you expected, you should check that all your disk drive enclosures are powered on and all hot-swap disk drives are properly seated. Possible choices here are:

**F2** Detailed information

This selection will tell you which disk drives do not respond.

**F4** Retry the command

You would select this, for example, if you forgot to turn on an external enclosure, or a hot-swap drive was not seated properly.

**F5** Change the configuration and set the drive(s) defunct

You might press F5 when, in a RAID-5 or RAID-1 configuration, a drive has started in defunct (DDD) mode. In this instance, you can continue to boot the server in critical mode to enable you to replace the faulty drive while the server is running.

**F10** Continue booting without changing the configuration

***Previously configured drives are not in their configured location***

When the adapter detects that a previously configured drive is present, but the drive is in a new location, the following message appears:

NN OnLine Drive(s) has been rearranged

Here you can select from among the following:

**F2** Detailed information

**F4** Retry the command

You might press F4 if the drives were relocated temporarily, after you have moved the drives back to their original locations.

**F5** Change the configuration and set the drive(s) defunct

Usually you would not want to select this option for this particular message.

**F6** Change the configuration and accept the rearrange

This will make the new drive positions valid and is the normal response for an intentional drive relocation.

**F10** Continue booting without changing the configuration

***A new adapter or imported drives have been detected***

Unlike the New drives installed message, here the adapter has detected that the drives installed in the server are new to this machine, but were part of a different ServeRAID adapter configuration. You will see this message:

NN OnLine Drive(s) found with mismatch Configuration

From here, you have the following choices:

**F2** Detailed information

**F4** Retry the command

This selection would usually not be appropriate in this situation.

**F5** Change the configuration and set the drive(s) defunct

This selection is also not normally appropriate here.

**F7** Import configuration information from drive

Press F7 when you replace a faulty adapter or when you move an

entire disk array from one system to another without also moving the adapter.

**F10** Continue booting without changing the configuration

---

## 5.8 ServeRAID command-line utilities

Two utilities, IPSSSEND and IPSMON, are provided with IBM's ServeRAID adapters, and can be used to control the adapters from a command prompt on the following operating systems:

- Microsoft Windows NT
- Windows 2000
- Novell NetWare 3.12, 4.1X and 5.0
- IBM OS/2 Warp Server and OS/2 LAN Server
- SCO OpenServer 5.0.X
- SCO UnixWare
- Linux

IPSSSEND provides a rich set of subcommands to configure and control ServeRAID adapters and can be of great assistance in the following activities:

- Server roll-out  
Use the BACKUP, RESTORE, INIT, INITSYNC, SYNCH and COPYLD subcommands.
- Error recovery  
Use the GETSTATUS, REBUILD, SETSTATE and UNBLOCK subcommands.
- Problem isolation and debug  
Use the ERASEEVENT, GETEVENT and SELFTTEST subcommands.
- ServeRAID configuration  
Use the DRIVEVER, GETCONFIG, HSREBUILD, READAHEAD and UNATTENDED subcommands.
- Logical drives copy  
Use the FLASHCOPY subcommand.

The IPSMON command is much simpler and enables monitoring and logging of all relevant ServeRAID events to a log file and/or monitor.



### 5.8.1 IPSSSEND subcommands

Each operating system has a different set of commands for IPSSSEND. For example, the IPSSSEND COPYLD command is available only in DOS. In order to get the complete list of IPSSSEND commands available for each respective operating system, run the IPSSSEND program without any parameter in that particular operating system.

For example, from a Windows NT command prompt, type in IPSSSEND and you will see the list of subcommands shown Figure 70:

```

C:\IpsAdm>ipssend

Licensed Material - Property of IBM Corporation
IBM ServeRAID Command Line Interface v3.60.08
Copyright (C) IBM Corporation 1996 - 1999
All Rights Reserved
US Government Restricted Rights - Use, Duplication, or Disclosure
Restricted by GSA ADP Schedule Contract with IBM Corporation

Usage: IPSSSEND <Command> <Param 1> ... <Param N>
Help : IPSSSEND <Command> for specific help on any command.

  Command  | Param 1 | Param 2 | Param 3 | Param 4 | Param 5
  -----  | -
AUTOSYNC   | :Controller:Logical Drive|:NOPROMPT|
BACKUP     | :Controller:Filename    |:NOPROMPT|
DEUINFO    | :Controller:Channel     |:SCSI ID |
DRIVEUER   | :Controller:Channel     |:SCSI ID |
ERASEEVENT | :Controller:Options     |
GETCONFIG  | :Controller:Options     |
GETEVENT   | :Controller:Options     |
GETSTATUS  | :Controller:            |
HSREBUILD  | :Controller:Options     |
INIT       | :Controller:Logical Drive|:NOPROMPT|
INITSYNC   | :Controller:Logical Drive|:NOPROMPT|
FLASHCOPY  | :Controller:Options     |
MERGE      | :Controller:Merge ID    |
READAHEAD  | :Controller:Options     |
REBUILD    | :Controller:Channel     |:SCSI ID |:New Channel|:New SCSI ID
RESTORE    | :Controller:Filename    |:NOPROMPT|
SETSTATE   | :Controller:Channel     |:SCSI ID |:New State
SYNCH      | :Controller:Scope       |:Scope ID|
UNATTENDED| :Controller:Options     |
UNBLOCK    | :Controller:Logical Drive|
UNMERGE    | :Controller:Merge ID    |
  
```

Figure 70. Help for IPSSSEND

To get a more detailed help screen on how to use a particular command, you can type IPSSSEND and the command. For example, IPSSSEND HSREBUILD will produce Figure 71:

```

C:\IpsAdm>ipssend hsrebuild

Usage: IPSSEND HSREBUILD <Controller> <Options>
Controller --> Number of controller (1 to 12)
Options    --> ON  Enable Hot Swap Rebuild
           ?   Display status of Hot Swap Rebuild Feature

The HSREBUILD command is used to set the ServeRAID controller hot swap
rebuild feature on. Use the ? to display the current status of the hot
swap rebuild feature.

C:\IpsAdm>_

```

Figure 71. Help for IPSSEND HSREBUILD

For more information, refer to a detailed description of IPSSEND and IPSMON in Chapter 7 of *ServeRAID-3H, ServeRAID-3HB, and ServeRAID-3L Ultra2 SCSI Controllers Installation and User's Guide*.

---

## 5.9 Managing the subsystem: Netfinity Director and Netfinity Manager

ServeRAID Manager is a useful tool for configuring, monitoring and troubleshooting the ServeRAID environment. However, Netfinity Director and Netfinity Manager provide the following additional functions:

- Alert handling and automatic alert actions
- Integration into higher-level management platforms
- A common set of tools for all system management tasks

We will discuss integration with both Netfinity Director and Netfinity Manager in the following sections. Netfinity Director is a new tool that is likely to replace Netfinity Manager in the future. However, at the time of writing, Netfinity Manager remains a vital system management tool for Netfinity servers.

### 5.9.1 Netfinity Director integration

Netfinity Director implementation requires the following three components:

- Netfinity Director Server
- Netfinity Director Management Console
- Netfinity Director Client (or Agent)

This differs from Netfinity Manager, which operates in a peer-to-peer management environment, without the need for a server.

The Netfinity Director Server performs all the management tasks and the Management Console provides the user interface. The Client must be installed on every computer that you wish to manage. Managed Netfinity servers would typically have only the client component installed.

When you install the Netfinity Director Server, this will also install the console and the client onto the same machine, but you can install the management console separately on another machine. This would typically be the network administrator's computer. The console will connect to the Netfinity Director Server using TCP/IP. Figure 72 shows a window with the Netfinity Director installation options:

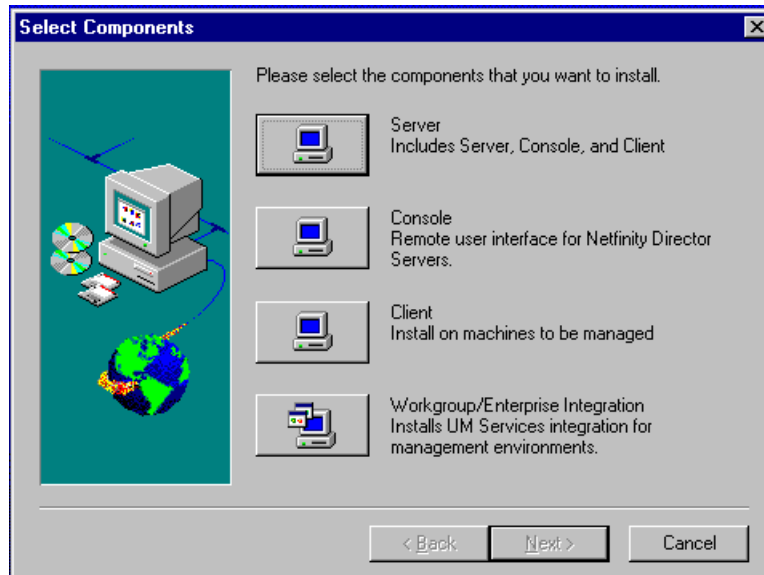


Figure 72. Installation options for Netfinity Director

The standard Netfinity Director installation does not include ServeRAID support. To be able to manage ServeRAID adapters through Netfinity Director, you must also install the IBM UM Server Extensions (UMSE), a component of the Life Cycle Tools (LCT) for servers. Make sure you install the UMSE on all client servers that you wish to manage, and any system that you will use as the management console. The UMSE code can be obtained free of charge at this URL:

[http://www.pc.ibm.com/ww/netfinity/systems\\_management/nfdir.html](http://www.pc.ibm.com/ww/netfinity/systems_management/nfdir.html)

Installing this code provides support for the following:

- Cluster Systems Management
- Advanced System Management adapters and processors
- ServeRAID adapters
- Capacity management

The installation and usage of Netfinity Director Server, console, clients, and Life Cycle Tools are covered in detail in *Netfinity Director - Integration and Tools*, SG24-5389.

Once all the required components are installed, you can access ServeRAID Manager from the Netfinity Director Console, as shown in Figure 73:

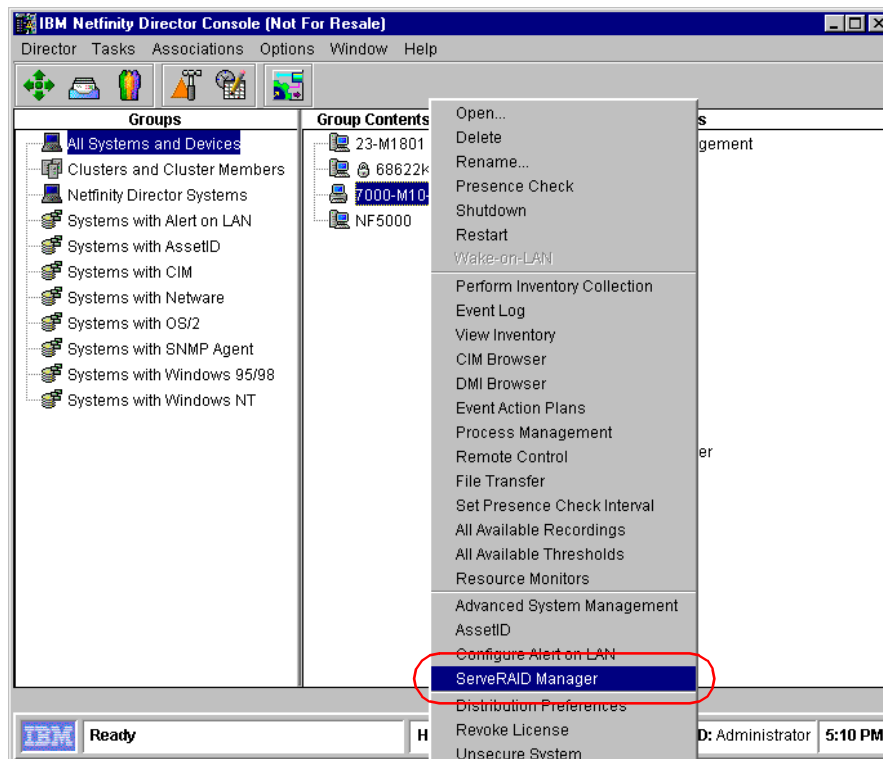


Figure 73. Netfinity Director Console - accessing the ServeRAID Manager

**Note:** When launching ServeRAID Manager from within the Netfinity Director, you will only be able to manage the ServeRAID adapters in the currently selected system. As you can see in Figure 74, the **Add remote system** icon

is greyed out. Additionally, no network connectivity messages appear when you start the Netfinity Director ServeRAID Manager. This does not mean functionality is limited, since you can easily connect to another system through the Netfinity Director Console.

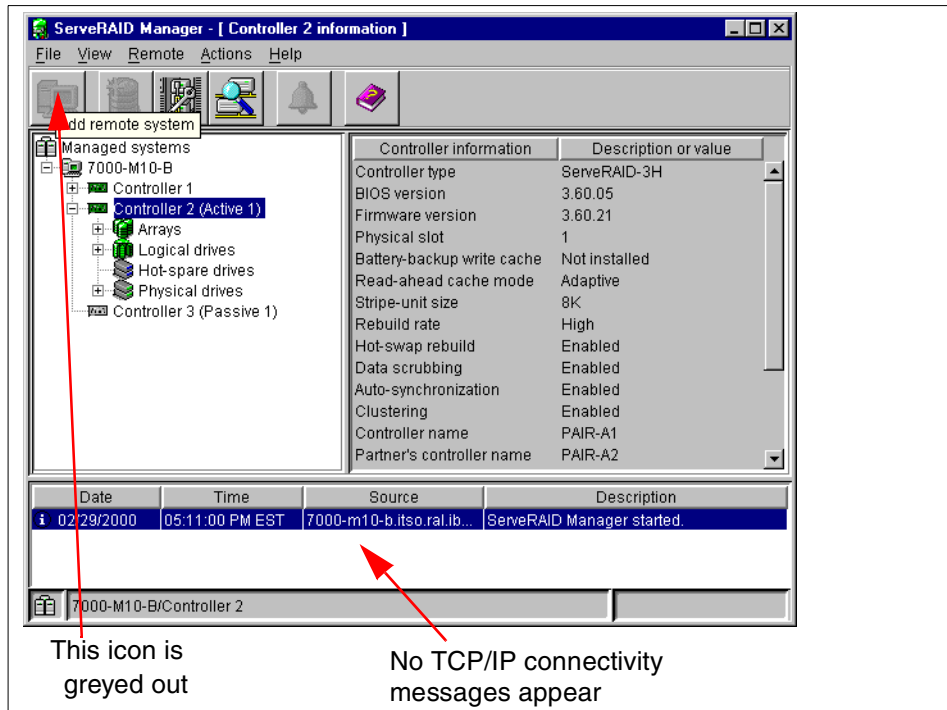


Figure 74. Netfinity Director ServeRAID Manager

In all other respects, operation is the same as for the native ServeRAID Manager. You can add new disk drives, create arrays and logical drives, and perform logical drive migration and recovery tasks, such as replacing and rebuilding failed disk drives.

Netfinity Director will receive notifications about all ServeRAID adapter events, such as disk drive failure, LDM process status, rebuild process status, and so forth. It will put those events into the system Event Log. Figure 75 on page 136 shows an example of ServeRAID events in the log:

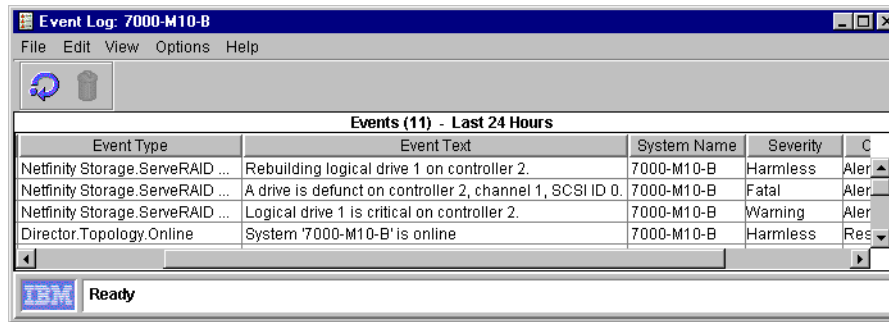


Figure 75. Netfinity Director: Event Log entries

You can configure automatic actions, or responses, to these events. For example, if the Netfinity Director receives notification of a disk drive failure, you can configure it to page the network administrator automatically. Use the **Event Action Plans** (see Figure 76) option on the Netfinity Director tasks list to configure those actions.

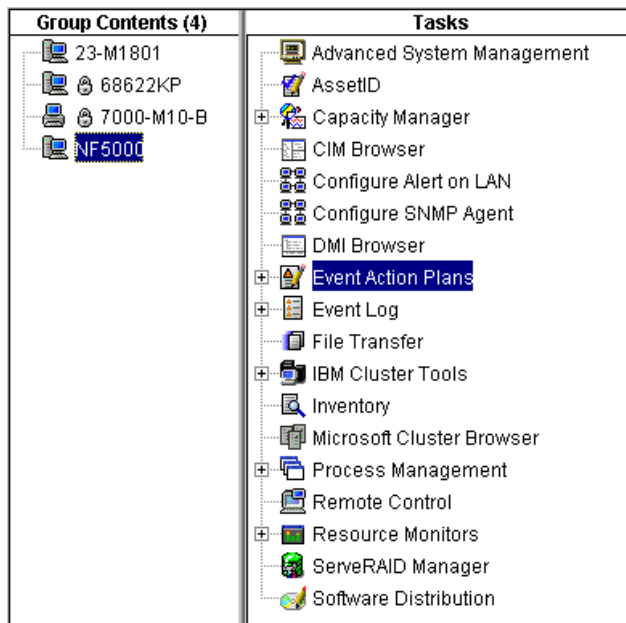


Figure 76. Netfinity Director - Event Action Plans

Detailed information about configuring actions and aspects of Netfinity Director is available in *Netfinity Director - Integration and Tools*, SG24-5389.

## 5.9.2 Netfinity Manager integration

For many years, IBM Netfinity Manager has been the primary system management tool for IBM Netfinity and PC Servers, and also IBM desktops and ThinkPads. Netfinity Manager operates in a peer-to-peer environment between managers and clients. The manager code is usually installed on a network administrator's workstation and the client code (Client Services for Netfinity) is typically installed on all the systems you want to manage. Various operating systems are supported for either manager or client operation, or both:

- Windows NT (manager or client)
- Windows 95/98 (manager or client)
- NetWare (client only)
- OS/2 (manager or client)

A manager can act as a client to another manager. This feature is especially useful when connecting through modems. It allows you to manage all computers in a network, not just the one you are connected to with the modem. Communication among managers and clients can use a number of different communication protocols as indicated in this list:

- TCP/IP
- NetBIOS
- IPX
- SNA
- Serial communication

Netfinity Manager can be installed from the ServerGuide Applications CD, which is included in the ServerGuide package, which is shipped with each Netfinity server.

Netfinity Manager provides several services to assist you in managing ServeRAID adapters. These are:

- RAID Manager
- System Information
- System Monitors
- Alert Actions

### **RAID Manager**

You can use this service to monitor the ServeRAID adapter's operation, replace and rebuild failed disk drives, synchronize logical drives and assign or remove a hot spare drive. However, you cannot create new arrays and logical drives or perform logical drive migration with this tool. You must use ServeRAID Manager for these tasks.

To access this service, double-click the **RAID Manager** icon in the Netfinity Manager window.



Figure 77. RAID Manager icon

The Netfinity RAID Manager window will open, showing you the following three items:

- The server containing the disk drives
- All ServeRAID adapters installed
- Existing logical (or virtual) drives

Initially, the picture in the window will usually not be entirely correct. The following discrepancies might exist:

- The picture does not match your server.
- No physical disk drives are visible in the picture. The reason is that, by default, the RAID Manager service will display the disk drives connected to SCSI channel 1 of the first ServeRAID adapter. If no disk drives are connected to that channel, none will be displayed.
- No external disk drive enclosures are displayed.

Figure 78 shows such an example. The server is Netfinity 7000-M10, presented correctly in the figure; however, no disk drives are visible. In addition, the EXP200 external disk drive enclosure, which is connected to the ServeRAID adapters 2 and 3, does not appear in the window.



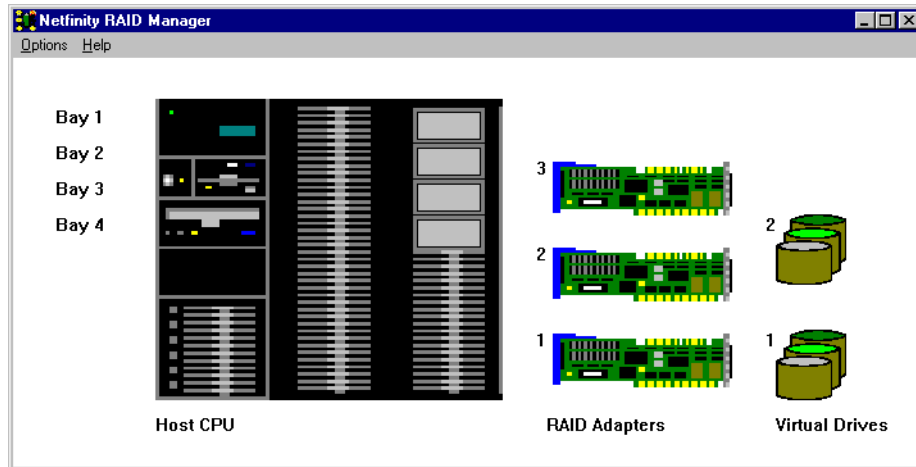


Figure 78. Netfinity RAID Manager - default view

You have to configure the disk drive enclosures in order to get the correct view. Select **Options -> Configure Enclosures**. This allows you to do two things:

- Identify the SCSI channel that is connected to the internal backplane in the server in order to display the internal disk drives.
- Select **Options -> Add Enclosure** and specify all external disk enclosures connected to the server.

Following these steps, the window is refreshed and is now correct, as you can see in Figure 79 on page 140. The internal disk drives and the EXP200 disk enclosure are now displayed.

It is very important to configure the enclosures properly. Until you can see the disk drives and enclosures, you will not be able to manage the ServeRAID environment.

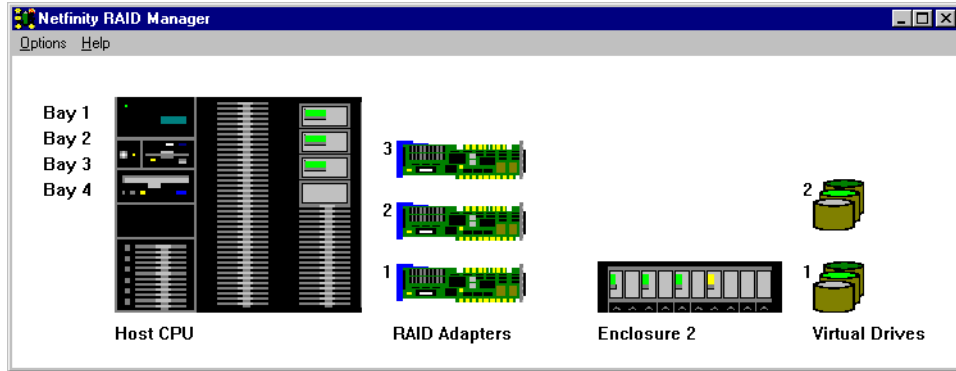


Figure 79. Netfinity RAID Manager - configured enclosures

You can perform various actions on disk drives, adapters and logical drives using this interface. Right-clicking an object will display the actions menu available for that particular object. Figure 80 shows the actions available for a ServeRAID adapter. We have highlighted **Backup Configuration** (which saves the configuration data to a file on a diskette) to demonstrate this. The Netfinity RAID Manager is the only utility that lets you do this, but this facility is not of any practical value since there is no way to restore the configuration from the file. This option is present only for historic reasons (it used to be available in the DOS diskette-based ServeRAID configuration utility, now replaced by the CD-based ServeRAID Configuration Program).

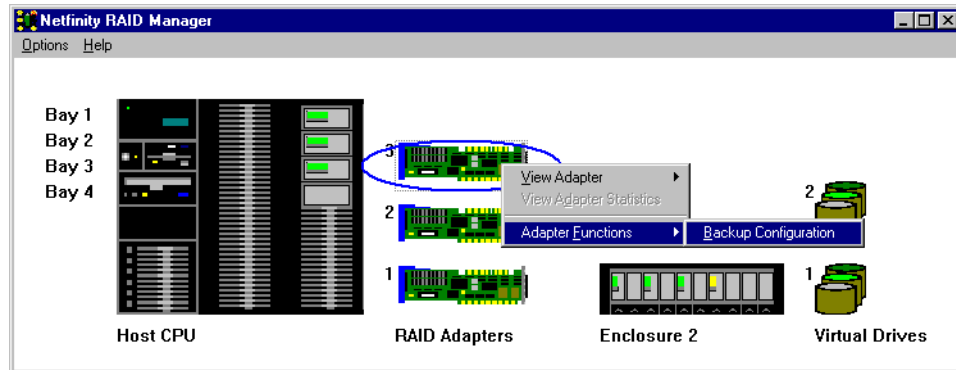


Figure 80. ServeRAID adapter actions

If a disk drive fails, it will be indicated in the RAID Manager window. The critical logical drive (termed a Virtual Drive in this tool) will be marked by a red exclamation mark and Netfinity Manager will display an alert. You can now replace the failed disk drive, as shown in Figure 81. If the *Hot-swap*

*Rebuild* parameter is enabled, the rebuild starts automatically as soon as the drive is replaced. Otherwise, you have to start the rebuild process manually, using the action menu for the disk drive.

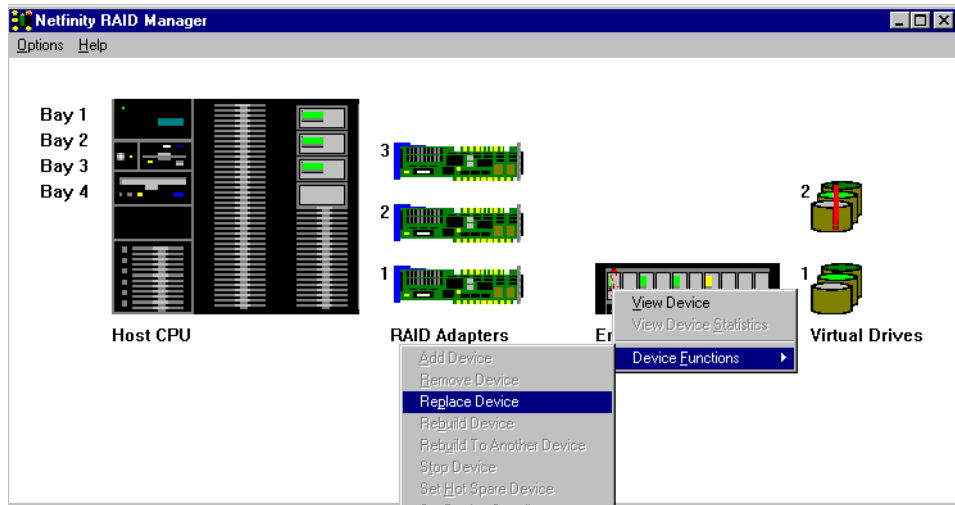


Figure 81. Replacing a failed disk drive

During the rebuild process, a progress indicator will be shown. You will not be able to access any other RAID Manager functions until the rebuild finishes.

### **System Information**

You can use this service to find out detailed information about the hardware components and operating system in the server. By double-clicking the **RAID System** icon, shown in Figure 82, you will be able to access information about your ServeRAID adapters, logical drives and physical drives.



Figure 82. RAID System icon from the System Information tool

Clicking **RAID Adapter Information** displays information about the total number of physical and logical drives, and the number of both defunct physical drives and critical logical drives.

Double-clicking **Virtual Devices** will show details of all the configured logical drives. However, System Information only gathers and displays data. In order to perform any actions on the logical drives, such as synchronizing, you must use the RAID Manager service.

Clicking **Physical Devices Information** shows all the disk drives attached to each of the SCSI channels of a ServeRAID adapter. An example is given in Figure 83. You can double-click any of the disk drives and detailed information about the drive will be displayed.

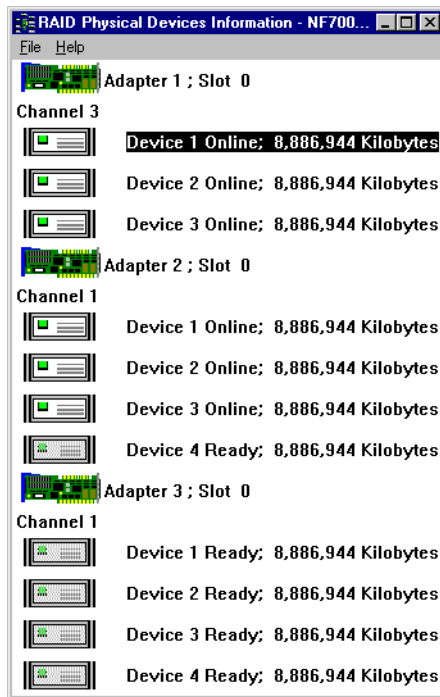


Figure 83. RAID Physical Devices Information window

### **System Monitors**

Monitoring system components is one of the most useful and powerful features of Netfinity Manager. You can monitor both statistics and the status of ServeRAID logical and physical drives. For example, Figure 84 shows the RAID statistics available for logical drives. A threshold can be defined for each monitored object, so you will be alerted whenever something is operating out of predefined parameters. You may also configure automated responses to these alerts, and doing so will be discussed in next section.

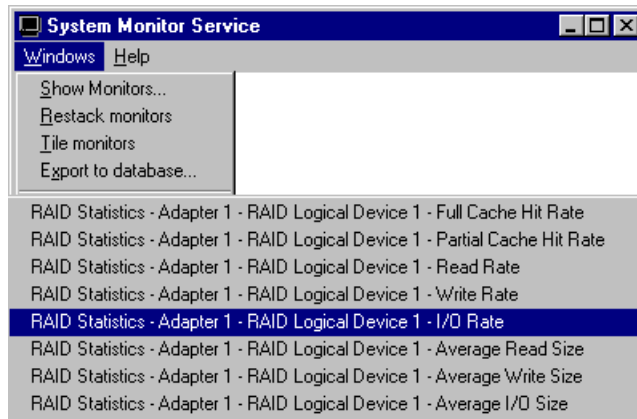


Figure 84. ServeRAID logical drive statistics monitors

As mentioned, you can also monitor the status of both physical and logical drives. By default, alerts on changes of states are enabled. For example, if a disk drive fails and one or more logical drives go to critical or offline state, an alert will appear. Figure 85 shows the current status of all logical and physical drives in our test system.

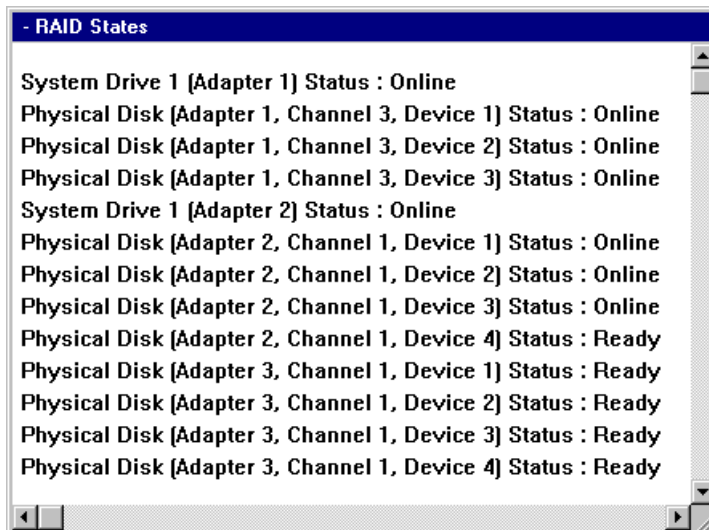


Figure 85. States of physical and logical drives

### **Alerts and actions**

A number of actions are available as automated responses to Netfinity Manager alerts. The following list identifies those that are most often used:

- Notify user with pop-up message
- Add alert to log file
- Add event to event log
- Forward alert to another system
- Play a .wav file
- Set or clear error condition for sending system
- Send alert to a pager
- Execute command
- Send alert as an e-mail

To configure alert responses, double-click the **Alert Manager** icon:



**Alert Manager**

*Figure 86. Alert Manager icon*

When the Alert Manager service starts, it displays the log of all alerts received. The log might be fairly large, so you have an option to customize the display of alert log. By selecting the **Alert Log Views** pushbutton, you can filter by *date and time* or by *alert profiles*. For example, instead of seeing all alerts, you might prefer to see only the ServeRAID related alerts. An example Alert Log with several ServeRAID-related alerts is shown in Figure 87.

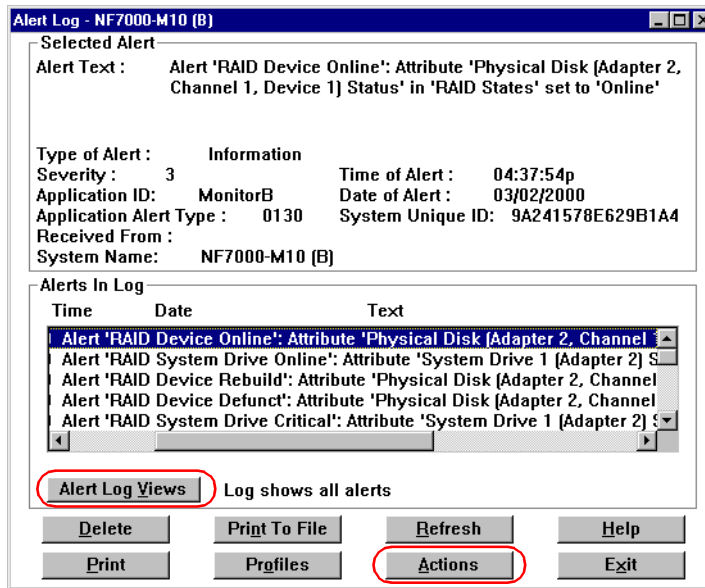


Figure 87. Alert Manager displaying the Alert Log

To configure responses to alerts, select the **Actions** pushbutton.

You can find detailed information on configuring alert actions in the *Netfinity Manager User's Guide*, shipped with the product.

## 5.10 Performance considerations

Ultimately, all data must be retrieved from and stored to disk. Disk accesses are usually measured in milliseconds, whereas memory and PCI bus operations are measured in nanoseconds or microseconds. In other words, disk operations are typically thousands of times slower than PCI transfers, memory accesses, and LAN transfers. This explains why the disk subsystem can easily become the major bottleneck for any server configuration. A detailed understanding of disk subsystem operation is critical for effectively solving many server performance problems.

Other aspects of your server's configuration can also negatively affect overall system performance. For a detailed discussion of performance tuning, see *Tuning Netfinity Servers for Performance*, SG24-5287. Although focused on Windows 2000 and Windows NT, much of the material in the book can be applied to all server operating systems.

### 5.10.1 Factors affecting ServeRAID performance

Many factors affect RAID subsystem performance. The most important considerations when configuring the IBM ServeRAID adapter are:

- Your RAID implementation
- The number of physical drives you will have available
- Disk drive performance
- Logical drive structure
- Firmware levels
- Stripe size
- SCSI bus configuration
- Write-back cache policy

### 5.10.2 RAID subsystem planning

You should spend some time planning your server's RAID configuration. Careful selection of appropriate numbers of disks and the RAID levels used can significantly affect disk subsystem performance. To illustrate the potential impact of poor planning, Figure 88 on page 147 illustrates the performance differences among RAID-0, RAID-1E and RAID-5 for a two-way Pentium II server configured with four 7200 RPM Fast/Wide SCSI-2 drives, implemented using the IBM ServeRAID adapter. The workload consisted of 50% reads and 50% writes. The chart shows the RAID-0 configuration delivering about 127% greater throughput than RAID-5 and 58% greater throughput than RAID-1E.

Remember, though, that RAID-0 has no fault tolerance and is therefore best utilized for read-only data when downtime for possible backup recovery is acceptable. RAID-1E or RAID-5 should be selected for applications requiring fault tolerance. RAID-1E is usually selected when the number of drives is low (fewer than six) and the price for purchasing additional drives is acceptable. RAID-1E offers about 42% more throughput than RAID-5. These performance considerations should be understood before selecting a fault-tolerant RAID strategy.



### RAID levels: performance versus cost

**Note:** These measurements were taken using a 50/50 read/write workload. In a more read-intensive environment, the performance difference between the various RAID levels would be less apparent because the read performance of RAID-1E and RAID-5 is much better than their write performance.

To review the relative costs of the different RAID levels, see 4.2.2, “RAID levels supported by ServeRAID adapters” on page 48, and Table 10 on page 148.

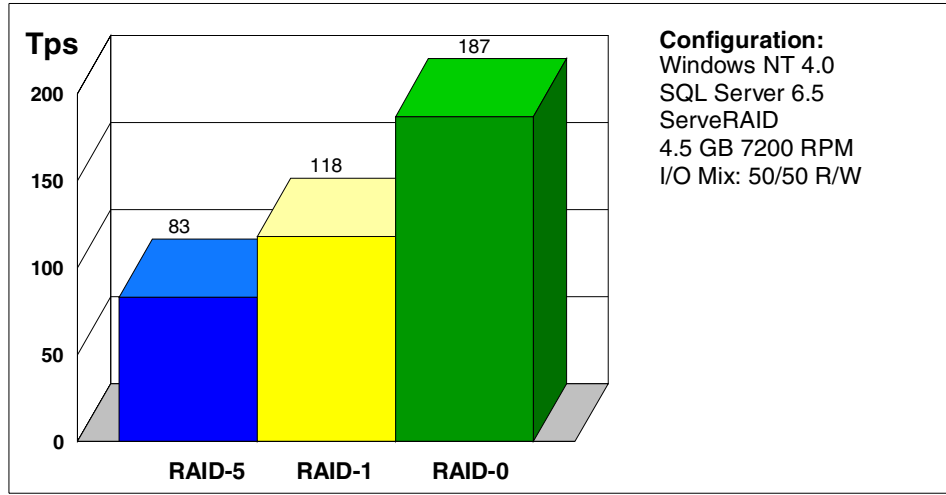


Figure 88. Comparing RAID levels

In many cases, RAID-5 is the best choice because it provides the best price versus performance combination for configurations requiring capacity greater than five or more disk drives. RAID-5 performance approaches RAID-0 for workloads where the read/write ratio is high. Servers executing applications that require fast read access to data and high availability in the event of a drive failure should employ RAID-5.

Table 10 shows a summary of the performance characteristics of the RAID levels commonly used in array controllers:

Table 10. Summary of RAID levels performance

RAID level	Data capacity <sup>1</sup>	Sequential reads <sup>2</sup>	Sequential writes <sup>2</sup>	Random reads <sup>2</sup>	Random writes <sup>2</sup>
Single Disk	n	6	6	4	4
RAID-0	n	10	10	10	10
RAID-1	n/2	7	5	6	3
RAID-1E	n/2	5	4	7	6
RAID-5	n-1	7	7 <sup>3</sup>	7	4
RAID-5E	n-2	8	8 <sup>3</sup>	8	5
RAID-10	n/2	10	9	7	6

**Notes:**

1. In the data capacity column, *n* refers to the number of equally sized disks in the array.
2. 10 = best, 1=worst. You should only compare values within each column. Comparison between columns is not valid for this table.
3. With write-back cache enabled

### 5.10.3 Number of drives

The number of disk drives in a RAID array significantly affects performance because each drive contributes to total subsystem throughput. Capacity requirements are often the only consideration used to determine the number of disk drives configured in a server. Throughput requirements are often not well understood and are completely ignored. Capacity is used because it is easily estimated and is often the only information available.

The result is a server configured with sufficient disk space, but insufficient data throughput to keep users working efficiently. There is a temptation to purchase high-capacity drives because they have a lower price per byte compared to smaller drives. Selecting a few large drives instead of a larger number of smaller drives reduces the total system price. This decision can, however, result in disappointing performance.

It is difficult to accurately specify server application throughput requirements when attempting to determine the disk subsystem configuration. In addition, subsystem throughput measurements are complex. To express a user requirement in terms of “bytes per second” would be meaningless because

the disk subsystem's byte throughput changes as, for example, a database grows and becomes fragmented, or as new applications are added to the server.

The best way to understand disk I/O and users' throughput requirements is to monitor an existing server. Tools such as the Windows NT Performance Monitor can be used to examine the logical drive queue depth and disk transfer rate. Logical drives that have an average queue depth much greater than the number of drives in the array will make the system wait for data. This indicates that performance would benefit by adding drives to the array.

#### **Adding Drives**

In general, adding drives is one of the most effective changes that can be made to improve server performance.

Measurements show that server throughput for most server application workloads increases as the number of drives configured in the server is increased. Performance is usually improved for all RAID levels. Figure 89 shows the effect of adding drives to a RAID-0 array in our test environment. Similar gains can be expected for all I/O-intensive server applications such as file serving, Lotus Notes, Oracle, DB2, and Microsoft SQL Server.

Performance will continue to improve as drives are added until another server component becomes a bottleneck. In general, most servers are configured with an insufficient number of disk drives and this should be the first area considered when poor performance is a problem.

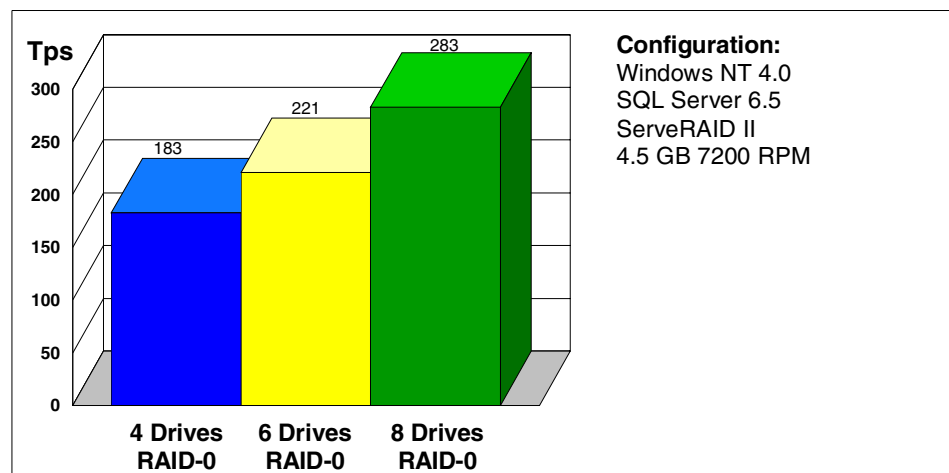


Figure 89. Improving performance by adding drives to arrays

#### Rule of Thumb

For most server workloads, when the number of drives in the array is doubled, server throughput will improve by about 50 percent until other bottlenecks occur.

With the IBM ServeRAID family of adapters, resolving performance problems of this type is simplified since you can use the logical drive migration feature to add drives to existing arrays without disrupting users or losing data. Refer to 5.3.2, “Logical drive migration (LDM)” on page 94 for more details of this feature.

#### 5.10.4 Drive performance

A disk is made up of multiple platters coated with magnetic material that stores your data. The entire platter assembly is mounted on a spindle that revolves these disks around the central axis. A head assembly mounted on an arm moves across the platter surface (linear motion) to read the data stored on the magnetic coating of the platter.

Drive performance has a significant impact on overall server throughput. There are four major components to the time it takes a disk drive to execute and complete a user request:

- Command overhead

This is the time it takes for the drive’s electronics to process the I/O request, and depends on whether it is a read or write request and whether or not the command can be satisfied from the drive’s buffer. This value is of the order of 0.1 ms for a buffer hit to 0.5 ms for a buffer miss.

- Seek time

This is the time it takes to move the drive head from its current cylinder location to the target cylinder. As the radius of the platters used in modern drives has been decreasing, and drive components have become smaller and lighter, average seek times have been decreasing. The average seek time for most current disk drives used in servers today is usually 5-7 ms.

- Rotational latency

Once the head reaches the target cylinder, the time it takes for the target sector to rotate under the head is called the rotational latency. The average latency is half the time it takes the drive to complete one rotation, so it is inversely proportional to the RPM value of the drive:

- 5400 RPM drives have a 5.6 ms rotational latency
  - 7200 RPM drives have a 4.2 ms rotational latency
  - 10,000 RPM drives have a 3.0 ms rotational latency
- Data transfer time

This value depends on the *media data rate*, which is how fast data can be transferred from the magnetic recording media, and the *interface data rate*, which is how fast data can be transferred between the disk drive and disk controller (that is, the SCSI transfer rate).

The media data rate improves as a result of greater recording density and faster rotational speeds. A typical value is 0.8 ms. The interface data rate for Ultra3 160/m SCSI is 160 MBps. Assuming 4 KB I/O transfers (which are typical for Windows NT Server), the interface data transfer time is 0.025 ms. As you can see, with the latest SCSI technology the interface data transfer time is becoming very low.

Obviously, the significant values that affect performance are the seek time and the rotational latency. This is especially true for random I/O (which is usually predominant for a multi-user server). Seek times are related to the physical characteristics of the drive's head mechanics, but reductions are expected to continue as the physical drive attributes become smaller.

Reducing latency can be achieved by techniques such as *rotational positioning optimization* (RPO), whereby I/O requests are reordered to match the sequence that the data is laid out on the drive's surface. Tests have shown that 7200 RPM drives with RPO have performance results similar to those running at 10,000 RPM.

For sequential I/O (such as with servers with small numbers of users requesting large amounts of data) or for I/O requests of large block sizes (for example, 64 KB), the data transfer time does become important in comparison with seek time and latency, so the use of Ultra2 SCSI and Ultra3 SCSI can have a positive effect on overall subsystem performance in such cases.

It should be borne in mind that caching and read-ahead is employed on the drives themselves, too, so the time taken to perform the seek and rotation is masked to some extent, depending on the precise way in which data is being accessed. In this case, the data transfer time can become more significant.

The easiest way to improve disk performance is to increase the number of data requests that can be made simultaneously. This is achieved by using

many drives in a RAID array and spreading the data requests across all drives as described in 5.10.3, “Number of drives” on page 148.

Table 11 summarizes the raw performance data for three of IBM’s high-end drives.

*Table 11. Comparing 10000 RPM and 7200 RPM drives*

<b>Disk drive</b>	<b>Capacity</b>	<b>RPM/ latency</b>	<b>Seek time</b>	<b>Buffer size</b>	<b>Media data transfer rate</b>
Ultrastar 36LP	18.3 GB	7200/ 4.2 ms	4.17 ms	4 MB	248-400 Mbps
Ultrastar 36LZ	18.3 GB	10K/ 3.0 ms	2.99 ms	4 MB	280-452 Mbps
Ultrastar 72ZX	73.4 GB	10K/ 3.0 ms	2.99 ms	16 MB	280-473 Mbps

### **5.10.5 Logical drive configuration**

Using multiple logical drives on a single physical array is convenient for managing the location of different files types. However, depending on the configuration, it can significantly reduce server performance.

When you use multiple logical drives, you are physically spreading the data across different sections of the array disks. If I/O is directed to each of the logical drives, the disk heads have to seek further across the disk surface than they do when the data is stored on one logical drive (see Figure 90). Using multiple logical drives greatly increases seek time and can decrease performance by as much as 25%.

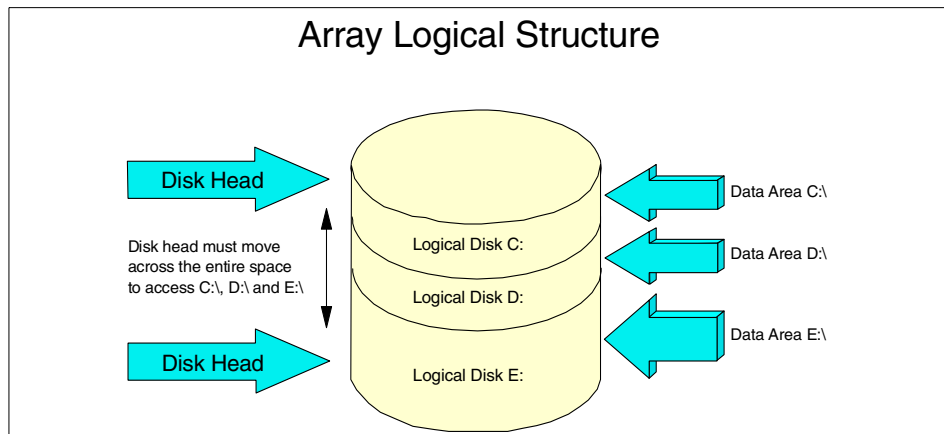


Figure 90. Logical drive structure

The fastest configuration is a single logical drive for each physical RAID array. Instead of using logical drives to manage files, create directories and store each type of files in a different directory. This may significantly improve disk performance by reducing seek times because the data will be as physically close together as possible.

If you really want or need to partition your data on multiple drives, you should configure multiple RAID arrays if possible, rather than configuring multiple logical drives in one RAID array.

### 5.10.6 Stripe unit size

Striping is the process of storing data across all the disk drives that are grouped in an array.

The granularity at which data from a file is stored on one drive of the array before subsequent data is stored on the next drive of the array is called the *stripe unit* (also referred to as *interleave depth*). For the ServeRAID adapter family, the stripe unit size can be set to 8 KB, 16 KB, 32 KB, or 64 KB.

The collection of these stripe units from the first drive of the array to the last is called a *stripe*.

The stripe and stripe unit are shown in Figure 91:

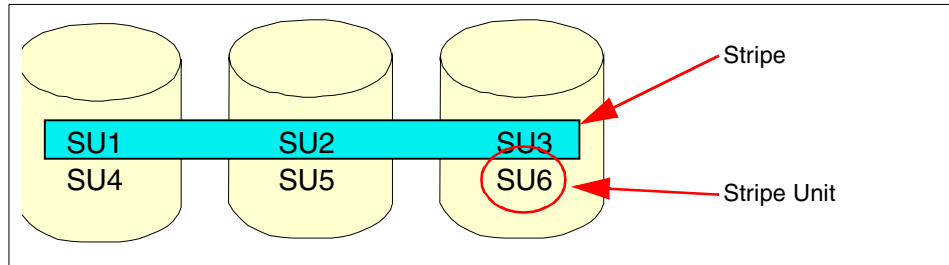


Figure 91. Stripe and stripe units

Using stripes of data better balances the I/O requests within the logical drive. On average, each disk will perform an equal number of I/O operations, thereby contributing to overall server throughput. Stripe size has no effect on the total capacity of the logical disk drive.

The selection of stripe size affects performance. In general, the stripe size should be approximately as large as the median disk I/O request size generated by your server applications.

If the stripe size is set too small, a performance degradation occurs when a server application requests data that is larger than the stripe size, because two or more drives must be accessed for the I/O request. Ideally, only a single disk I/O occurs for each I/O request.

Alternatively, selecting too large a stripe size can reduce performance because the system is not making best use of the multiple drives in the array. It can also be a problem with RAID-5 in particular, where additional stripe units must be read to calculate a checksum during writing. Use too large a stripe, and extra data must be read each time the checksum is updated.

Selecting the correct stripe size is a matter of understanding the typical request size performed by your particular application. Few applications use a single request size for each and every I/O request. Therefore, it is not possible to always have the ideal stripe size. However, there is always a best-compromise stripe size that will result in optimal I/O performance.

Windows 2000 or Windows NT 4.0 Performance Monitor can monitor disk I/O request sizes and help you determine the proper stripe size. Using Performance Monitor, select:

- Object: Physical Disk
- Counter: Avg. Disk Bytes/Transfer
- Instance: the drive that is receiving the majority of the disk I/O



As an example, the trend value for this counter is shown as the thick line in Figure 92. The running average is shown as indicated.

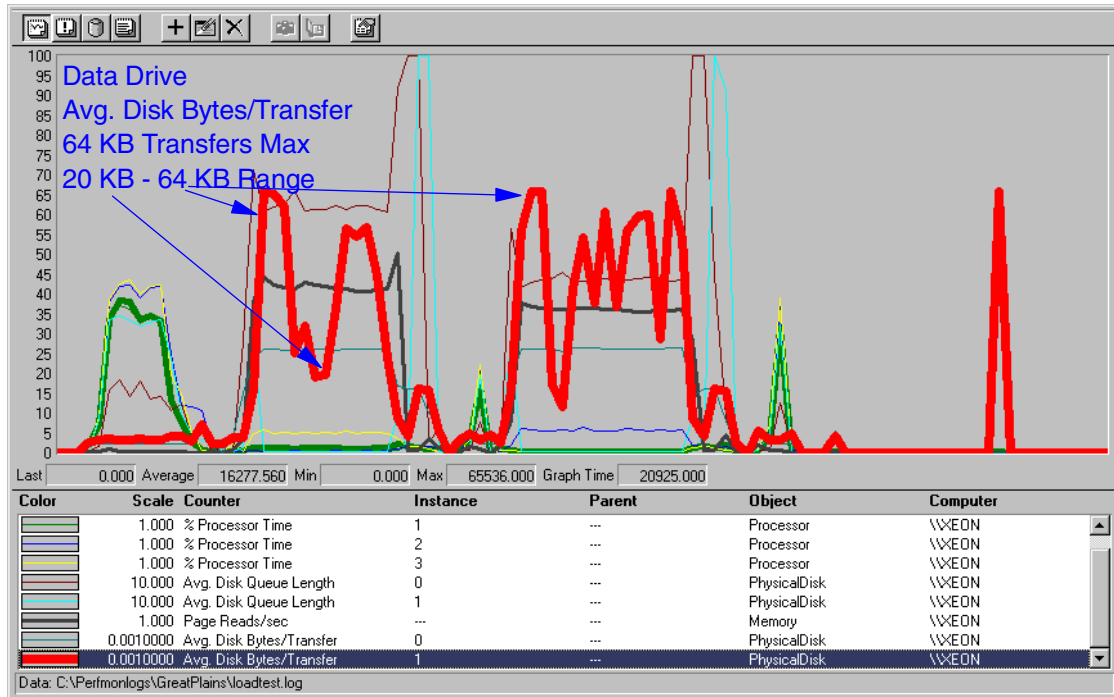


Figure 92. Average I/O size

Figure 92 represents an actual server application. As you can see, the application request size (represented by Avg. Disk Bytes/Transfer) varies from a peak of 64 KB maximum to about 20 KB for the two run periods.

This particular server was configured with an 8 KB stripe size, which produced poor performance. Increasing the stripe size to 16 KB would improve performance and increasing the stripe size to 32 KB would increase performance even more. The simplest technique would be to place the time window around the run period and select a stripe size that is approximately as large as the average size shown in the running average counter.

### Monitoring physical disk activity

If you wish to monitor physical disk activity for Windows NT Server 4.0, the command DISKPERF -Y must be executed and the server must then be restarted. (Keeping this setting on all the time uses some 2-3% of the server's CPU but if your CPU is not a bottleneck, this is irrelevant and can be ignored.)

In Windows 2000, the physical disk counters are always enabled.

Alternatively, you might want to refer to Table 12 below to determine the adequate stripe size. This table gives figures based on performance testing by IBM's server development group. While not definitive, it provides a useful starting point for typical applications.

Table 12. Stripe size for various applications

Applications	Stripe size
Groupware (Lotus Notes, Exchange and so forth)	16 KB
Database Server (Oracle, SQL Server, DB2 and so forth)	16 KB
File Server (NetWare)	16 KB
File Server (NT)	16 KB
Web Server	8 KB
Video File Server	64 KB
Other	8 KB
<b>Notes</b> <ul style="list-style-type: none"><li>• SQL Server 7.0 uses 8 KB I/O blocks but experiments have shown that performance can usually be improved by using double the I/O block size (that is, 16 KB).</li><li>• Oracle uses multiple block sizes: 2 KB, 4 KB or 8 KB. While using 16 KB is not the optimum for all cases, it is also not significantly slow either. Further I/O analysis on specific customer data may determine that 8 KB or 16 KB block sizes may produce better performance.</li></ul>	

### 5.10.7 SCSI bus organization

The SCSI bus organization of drives on a multi-bus controller (such as ServeRAID) does not significantly affect performance for most server workloads.

For example, in a four-drive configuration, it doesn't matter whether you attach all drives to a single SCSI bus or if you attach two drives each to two different SCSI buses. Both configurations will usually have identical disk subsystem performance. This applies to applications such as database transaction processing, which generate random disk operations of 2 KB or 4 KB. The SCSI bus does not contribute significantly to the total time required for each I/O operation, as each I/O operation usually requires drive seek and latency times. Therefore, the sustainable number of operations per second is reduced, causing SCSI bus utilization to be low.

For a configuration that has six or more drives, or one that runs applications that access image data or other large sequential files, performance improvement can be achieved by using a balanced distribution of drives across the multiple SCSI buses of your ServeRAID adapter.

### 5.10.8 Write-back cache operation

The ServeRAID adapter default is to operate the cache in *write-through* mode. In this mode, each disk write operation is passed to a disk drive before the device driver informs the operating system that the I/O has completed. In *write-back* mode, the disk controller performs the write operation into the adapter cache buffer, informs the operating system that the write operation has completed, and the ServeRAID adapter performs the physical disk write some time later.

A common misunderstanding is that write-back improves overall performance for server workloads simply because it allows a disk write operation to be completed sooner. Write-back does indeed usually improve performance, but mainly because the controller can gather multiple write operations together and perform disk operations more efficiently.

All server operating systems employ main memory as a very large disk cache. The operating system or server application usually performs a disk write that is issued to a disk cache in main memory, without actually doing a disk operation. In this case, the user's application will be informed that the disk write was complete long before the disk controller will perform the write operation.

Some applications such as transaction processing use a log disk to store database updates. A log drive is where a log of all database activities are stored. In the event of a system failure, the logs are used to recover data. No transaction can complete without being written to the log disk. In this case, using write-back cache mode in a disk controller can slightly improve

performance. However, the log information must be protected with a battery backup write-back cache.

As previously mentioned, most performance gains obtained from using write-back mode are derived from the ability of the disk controller to merge multiple write commands into a single disk write operation. This is particularly true for RAID-5 logical drives, because the controller must update the checksum information for each data update. Write-back mode allows the disk controller to keep the checksum data in adapter cache and perform multiple data updates before actually updating the checksum information on the disk.

#### **Use battery-backup**

If you do plan to use write-back mode on the ServeRAID, we recommend you use the battery-backup option, which ensures no cached data is lost in the event of system or adapter failure.

### **5.10.9 Write-back versus write-through cache**

Most people think that write-back mode is always faster because it allows data to be written to the disk controller cache without waiting for disk I/O to complete. Figure 93 on page 159, however, illustrates how disk subsystem performance varies with the applied load for each mode of cache operation and reveals that write-back operation does not always equate to optimal performance.

Although this is usually the case when the server is lightly loaded, it may not be true for a heavily loaded server. As the server becomes busy, the cache fills completely, causing data writes to wait for space in the cache before being written to the disk. When this happens, data write operations slow to the speed at which the disk controller can free up space in the cache. If the server remains busy, the cache is flooded by write requests, resulting in a bottleneck. In this case, cache acts as an added overhead. This problem can be largely avoided by increasing the cache capacity.

In write-through mode, write operations do not wait in cache memory that must be managed by the processor on the RAID adapter. When the server is lightly loaded (the left-hand zone in Figure 93), write operations take longer because they cannot be quickly stored in the cache. Instead, they must wait for the actual disk operation to complete. Thus, when the server is lightly loaded, throughput in write-through mode is generally lower than in write-back mode.

However, when the server becomes very busy (the right-hand zone in Figure 93), I/O operations do not have to wait for available cache memory. They go straight to disk, and throughput is usually greater for write-through than in write-back mode.

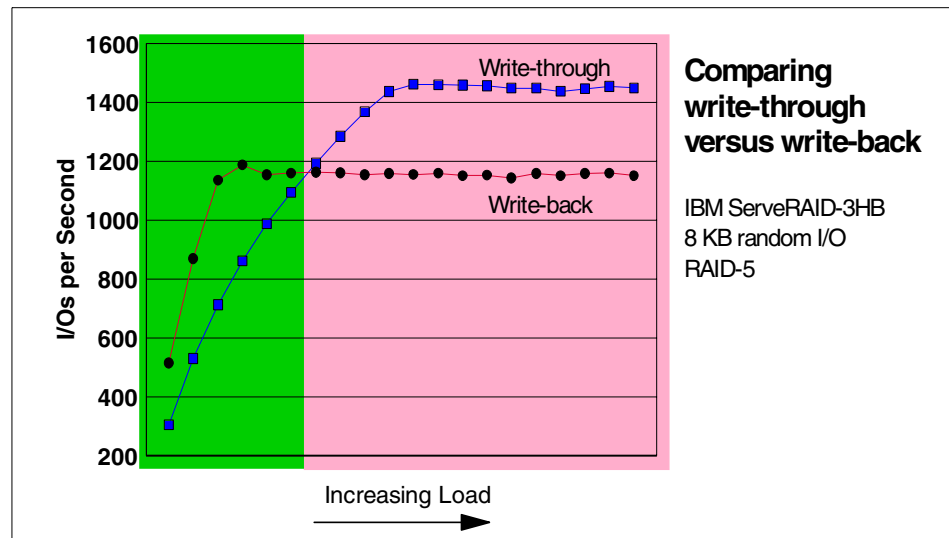


Figure 93. Comparing write-through and write-back modes under increasing load

**Configuration tip**

If the disk subsystem is very busy, then use write-through mode for increased performance.

If the disks are configured correctly, and the server is not overloaded, write-back mode is usually the better choice.

**5.10.10 RAID adapter cache size**

IBM performance tests show that the ServeRAID-3HB adapter with 32 MB of cache typically outperforms other vendors' RAID adapters with 64 MB of cache for most real-world application workloads. Once the cache size is above the minimum required for the job, the extra cache usually offers little additional performance benefit.

The cache increases performance by providing data that would otherwise be accessed from disk. However, in real-world applications, total data space is so much larger than disk cache size that, for random operations, there is very

little statistical chance of finding the requested data in the cache. For example, a 50 GB database would not be considered very large by today's standards. A typical database of this size might be placed on an array consisting of seven or more 9-GB drives. For random accesses to such a database, the probability of finding a record in the cache would be the ratio of 32 MB/50 GB, or approximately 1 in 1,600 operations. Double the cache size, and this value is decreased by half, still a very discouraging hit-rate. You can easily see that it would take a very large cache to increase the cache hit-rate to the point where caching becomes advantageous for random accesses.

As we said in 5.10.8, "Write-back cache operation" on page 157, in RAID-5 mode, significant performance gains from write-back mode are derived from the ability of the disk controller to merge multiple write commands into a single disk write operation. This is because the controller must update the checksum information for each data update. Write-back mode allows the disk controller to keep the checksum data in adapter cache and perform multiple updates before completing the update to the checksum information contained on the disk. In addition, this does not require a large amount of RAM.

In most cases, disk array caches can usually provide high hit rates only when I/O requests are sequential. In this case, the controller can pre-fetch data into the cache so that on the next sequential I/O request, a cache hit occurs. Pre-fetching for sequential I/O requires only enough buffer space or cache memory to stay a few steps ahead of the sequential I/O requests.

This can be done with a small circular buffer. Technically, the cache size needs to grow as a percentage of the number of concurrent I/O streams supported by the array controller. The original ServeRAID adapters supported up to 32 concurrent I/O streams, so 32 MB of cache was deemed enough to provide a high-performance hit rate for sequential I/O. For newer RAID adapters, the number of outstanding I/O requests can be as high as 128; thus, these adapters will have proportionally larger caches. This is why ServeRAID-4H has 128 MB of cache, ServeRAID-4M has 64 MB, and ServeRAID-4L has 16 MB.

Most people do not invest the time to think about how cache works and often conclude that "bigger is always better." The drawback is that larger caches take longer to search and manage. This can slow I/O performance, especially for random operations, since there is a very low probability of finding data in the cache. During periods of light loading, very little cache memory is required.

In identical hardware configurations it will take more CPU overhead to manage 64 MB of cache compared to 32 MB, and even more for 128 MB. The

point is that bigger caches do not always translate to better performance unless they are managed by more powerful CPUs. ServeRAID-4H has a 266 MHz Power PC 750 with its own 1 MB L2 cache. This CPU is approximately 5-7 times faster than the 40 MHz CPU used on ServeRAID-3HB. As a consequence, ServeRAID-4H can manage the larger cache without running slower than ServeRAID-3HB.

Furthermore, the amount of cache is, ideally, related to the number of drives attached. Typically cache hits are generated from sequential read-ahead. You do not need to read ahead very much to have 100% hits. With more drives attached to the adapter, there are more I/O streams to prefetch. ServeRAID-4H has four SCSI buses that support up to 56 drives compared to 42 for ServeRAID-3H.

### **5.10.11 Device drivers**

Device drivers play a major role in disk subsystem performance. A device driver is a low-level piece of system software is written to control a specific device. Most device drivers are vendor specific and can often be downloaded from the Web. Installing the correct device driver for specific hardware is very important. Selecting an incorrect device driver (if it works at all) can cause poor performance or data loss.

Device drivers are also specific to the operating system. Current operating system installation CD-ROMs usually contain the drivers for widely used adapters. However, drivers for specialized adapters, such as ServeRAID, typically have to be installed from diskettes.

For example, the Windows 2000 CD-ROM contains the Version 3.50 of the ServeRAID driver and will allow you to install the operating system onto ServeRAID-attached disks. We recommend you install Windows 2000 using that driver, then, once the installation is complete, upgrade to the latest driver.

The version of the device driver can also have a performance impact. Some of the critical performance problems can be resolved simply by upgrading to the latest version of the appropriate device driver.

### ServeRAID Version 3.60

Disk counters within Windows NT Performance Monitor will always display a value of zero, if ServeRAID driver Version 3.60 is installed. This problem will be fixed in the new version of the device driver. The following workaround is available until then. As usual, be careful when editing the Windows registry.

1. Start REGEDT32.EXE.
2. In the following key - change the value of "Group:REG\_SZ:" from "filter" to "SCSI CDROM class":

Hkey\_Local\_Machine\System\CurrentControlSet\Services\ipsperf

3. Restart system.

### 5.10.12 Firmware

ServeRAID firmware Version 3.50 provided a significant improvement in performance in comparison with earlier versions. It included many optimizations that resulted in a system level gain in performance as much as 20-25% for typical server workloads. This firmware version was also used to introduce RAID-5E to the ServeRAID-3 family of adapters.

Version 3.50 of the firmware and device driver also introduced Adaptive Read Ahead algorithms that turn the read-ahead function on and off based upon the demands of the active workload. In other words, whenever the adapter firmware detects transfers that would benefit from read-ahead, the option is dynamically turned on. If the I/O workload changes so that read ahead reduces the overall performance, it is turned off. This feature reduces the complexity of configuring an array for maximum performance by automating the setting of the read-ahead parameter.

Version 3.50 also improved performance by optimizing I/O for RAID-1E. This feature improves performance by better balancing physical I/O operations between the mirror drive pairs. The net gain in performance for RAID-1E is as much as 66%.

Version 3.60 introduces additional performance enhancements, including:

- Refined instruction path length

The feature significantly improves the performance of the ServeRAID-3 family of adapters when executing cache hit operations. Since many customers make purchase decisions by running small data size



benchmarks, the design lab could not ignore performance obtained while accessing the majority of data from adapter cache. Version 3.60 offers greater performance by restructuring the executed code for a better fit and to stay resident in the L1 processor cache. The on-board CPU now runs significantly faster by reducing the CPU wait times for slower memory accesses.

- Greater concurrent I/O

This feature enables the ServeRAID-3 family of adapters to have up to 128 concurrent outstanding I/O operations. This change increases performance for configurations that utilize a large number of disk drives. Allowing a larger number of concurrent outstanding I/O operations enables the disk drives to optimize I/O by reordering seek operations.

- Removed the eight-drive limitation for 32 KB and 64 KB stripe sizes

Removing the limitation of eight physical drives for 32 KB and 64 KB stripe sizes lets you have configurations of up to 16 physical disks for applications that require large block transfers. Applications such as video and image serving can now use larger arrays. These larger arrays provide both greater capacity and increased throughput.

In general, customers can expect to see as much as a 20-25% improvement in throughput for average business applications from these modifications.

Figure 94 shows the gains obtained for typical random I/O server applications. The specific configuration is 8 KB block, 67% read, and 33% write, random transactions.

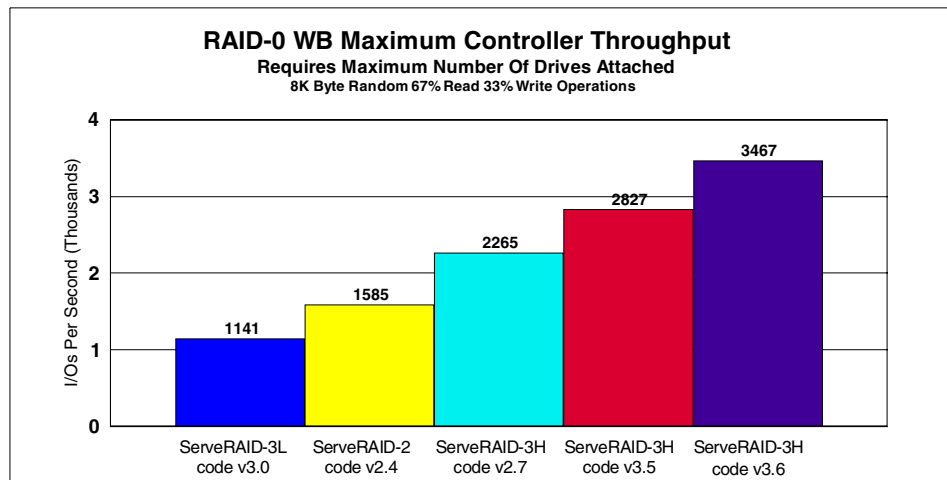


Figure 94. Maximum ServeRAID family RAID-0 throughput performance

### 5.10.13 SCSI bus transfer rate

The SCSI bus transfer rate can range from 20 MBps (F/W SCSI-2) to 160 MBps (Ultra3 160/m). For large sequential disk I/O workloads (such as a video or image server, or power desktop users such as CAD/CAM) the speed of the bus can play an important role in the disk subsystem's performance.

However, for more typical server applications, the I/O characteristic is mostly random, and the performance gain achieved with faster SCSI bus speeds is minimal. The reason for this is that the time it takes to transfer the data across the SCSI bus is insignificant compared to the time it takes to access the disk (the seek time plus the drive latency plus controller overheads). For today's hard disks, typical drive values are as follows:

- Disk controller overhead: < 1 ms
- Disk drive seek operation: 7 ms
- Disk drive latency: 4 ms

As you can see in Table 13, the performance gain becomes more significant with larger block sizes.

Table 13. Percentage gains for random I/O with different SCSI speeds

Block Size	Bus Speed	Bus Transfer Time	Access Time	Gain
4 KB	20 MBps	0.2 ms	12.2 ms	Baseline
	80 MBps	0.05 ms	12.05 ms	1.2%
	160 MBps	0.025 ms	12.025 ms	1.4%
32 KB	20 MBps	1.6 ms	13.6 ms	Baseline
	80 MBps	0.4 ms	12.4 ms	8.8%
	160 MBps	0.2 ms	12.2 ms	10.3%

---

## Part 3. Fibre Channel subsystems



---

## Chapter 6. Introduction to Fibre Channel

Fibre Channel (FC) is the premier storage solution for businesses that need reliable, cost-effective and scalable information storage and delivery at high speeds. The technology also allows physical separation of the storage subsystem from the host server over greater distances than more traditional storage attachment methods such as SCSI.

These attributes make Fibre Channel a prime candidate for implementing storage area networks (SANs). SANs are generating a great deal of interest from customers who are creating enterprise-class systems based on Netfinity servers. You can read about SANs in a Netfinity environment in Part 5, "Storage area networks (SANs)" on page 277. While SANs themselves are still developing in this marketplace, Fibre Channel is rapidly becoming established as a storage interconnect technology.

Briefly, these are the major features of Fibre Channel disk subsystems:

- Operation at 100 MBps allows the implementation of high-performance storage subsystems.
- Support for large databases and data warehouses due to the ability to attach many more drives to a single FC controller.
- Attachment over large geographical distances allows implementation of effective storage backup and recovery systems.
- Support for multiple host attachment simplifies the implementation of server clusters.
- It is simple to increase capacity as application and business needs for storage grow.

After an introduction to Fibre Channel technology, this chapter describes the hardware and infrastructure components that are available for implementing Fibre Channel solutions on IBM Netfinity servers.

Chapter 7, "Implementing Fibre Channel disk subsystems" on page 199 describes the software tools available to manage these storage solutions. It introduces the Netfinity Fibre Channel Storage Manager and describes how to perform important tasks such as creating a logical drive, adjusting cache operation, and partitioning the storage space.

---

## 6.1 IBM's implementation of Fibre Channel

Fibre Channel is an open standard communications and transport protocol as defined by the American National Standards Institute (ANSI Committee X3T11). The Fibre Channel protocols can operate over copper and fiber optic cabling, the latter at distances of up to 10 kilometers. IBM's implementation of Fibre Channel utilizes fiber optic cabling, which we will refer to as Fibre Channel cabling or FC cabling in this book.

### Fibre or Fiber?

Fibre Channel was originally designed to support fiber optic cabling only. When support for copper cabling was added, the committee decided to keep the name but change the spelling from fiber to fibre. When referring to fiber optic cabling, the correct American English spelling, fiber, should be used.

Fibre Channel is essentially a network infrastructure, and can carry a number of different communications protocols, including Internet Protocol (the IP in TCP/IP) as more usually found in Ethernet or token-ring networks. However, Fibre Channel is not limited to networking communication protocols and the standards describe its use for SCSI command protocols among others.

### 6.1.1 Fibre Channel topology

There are various ways in which you can connect your server to a Fibre Channel disk subsystem. In general these fall into one of three categories: arbitrated loops, switched connections, or point-to-point connections.

An arbitrated loop is realized with hubs linking up to 126 ports or nodes on a single fibre loop, shared among all nodes. The maximum of 126 ports stems from using only a subset of the 24-bit addressing scheme. Communication can only take place between a single pair of nodes at any point in time. Nodes that wish to use the fiber are subject to an arbitration algorithm, which determines which node can have access to the media.

A switched topology provides a dedicated path between any two devices in the entire Fibre Channel fabric allowing for over 16 million devices (24-bit addressing). Multiple simultaneous connections between pairs of devices are possible with the maximum number of connections being dependent upon the capability of the Fibre Channel switch. Each node or device has the full media bandwidth of 100 MBps available to it.

The third topology is a point-to-point connection, where two devices are hard-wired to each other, sharing a dedicated communications channel. This is rarely used as the relative expense of Fibre Channel is unlikely to be justifiable. This is because the full capacity of the link is not typically used in a point-to-point configuration.

All three topologies can exist in a single setup. However, you should keep in mind that an arbitrated loop can only be connected to a single switched topology, and at only one point.

In order to understand the advantages of the Fibre Channel protocol, and why it has been chosen by IBM for SAN environments in preference to other technologies, we now briefly outline its structure.

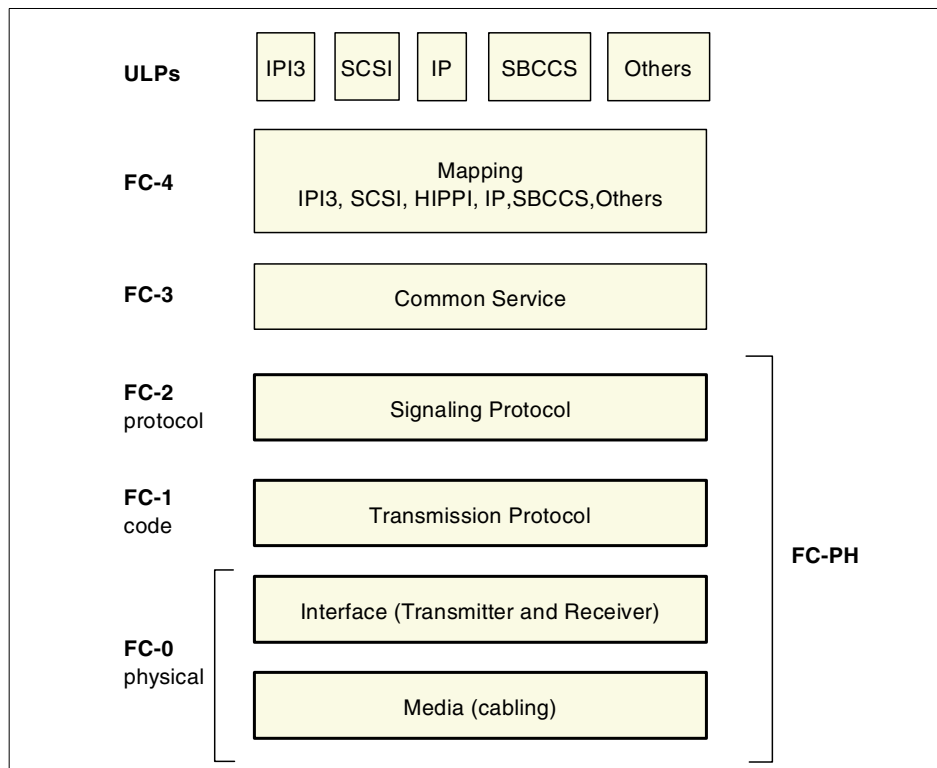


Figure 95. Fibre Channel protocol layers

As you can see in Figure 95, Fibre Channel, in common with other communication protocols, can be considered as having several distinct layers that perform different functions in creating the overall communication protocol.

The layers in the diagram are broadly divided into lower, physical layers and upper layers. Fibre Channel functions are implemented at different levels of this stack as follows:

**Physical Layers:**

- **FC-0** defines physical media and transmission rates. These include cables and connectors, drivers, transmitters and receivers.
- **FC-1** defines encoding schemes. These are used to synchronize data for transmission.
- **FC-2** defines the framing and flow protocols. These are self-configuring and support arbitrated loop and other topologies.

**Upper Layers:**

- **FC-3** defines common services for nodes. An example of a defined service is *multicast*, which delivers a single transmission to multiple destinations.
- **FC-4** defines the upper layer protocol mapping. Protocols such as FCP (SCSI), FICON, and IP can be mapped to the Fibre Channel transport service.
- **ULPs** (Upper Layer Protocols) are associated with IPI and SCSI command sets, HIPPI data framing, IP, and other upper level protocols.

The most significant difference between the Fibre Channel protocol and other established network protocols, such as TCP/IP, is that the error detection and recovery, flow control, and transport control is handled in the physical layers. This allows these functions to be implemented in either hardware, firmware, or a combination of both. This means that higher performance levels are achievable in comparison with protocols that implement such functions at higher levels in the stack, and which are therefore usually processed by the host and therefore much slower. Furthermore, it also leads to a faster transfer of data from host memory to a host-bus adapter.

Figure 96 shows the structure of a frame, the basic packet of data used in a Fibre Channel connection. A single frame can have a size of up to 2112 bytes. Frames are sequenced to transfer larger amounts of data. Each sequence can be up to 65536 frames in length. Depending on the type of application requesting data transfer, multiple sequences can be grouped to transfer as much as 128 MB of data, allowing for streaming of data between two nodes or devices.



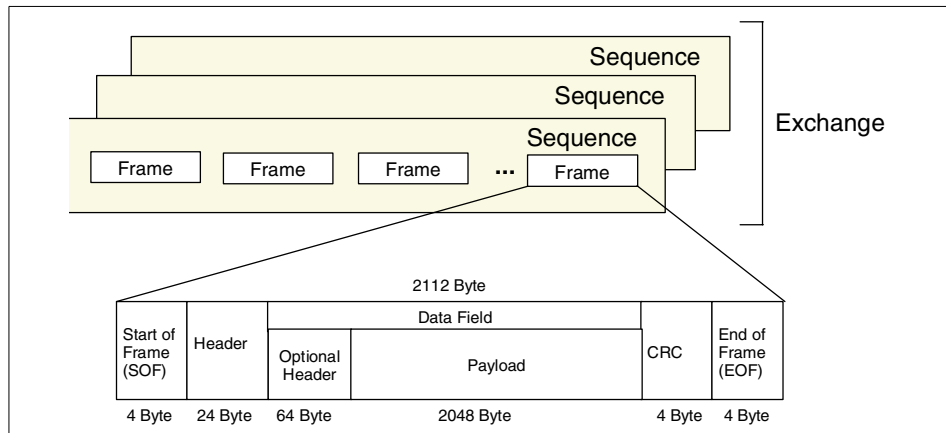


Figure 96. Fibre Channel frame and sequence structure

### What is a frame?

A Fibre Channel application that utilizes the upper layer protocols of the stack will see a sequence as the smallest unit of data and sometimes call it a “frame”. So, do not get confused if you find that the “frame” setting in a Fibre Channel application is bigger than 2148 bytes and varies in size.

Combining the above attributes of the Fibre Channel protocol leads to an efficient, high performance, reliable, and scalable protocol that is optimally designed for SAN environments.

For further information on the Fibre Channel standard, we encourage you to have a look at the ANSI committee’s official Web site, which can be found at:

<http://www.t11.org>.

## 6.2 Netfinity Fibre Channel hardware

The physical infrastructure of a Netfinity Fibre Channel storage installation is realized using fiber optical cables, gigabit interface converters (GBICs), media interface adapters (MIAs), hubs, and switches and is referred to as the interconnect *fabric*. The logical structure of the fabric can be described as topology, although these terms are often used interchangeably.

The fabric interconnects a variety of devices, including, hosts (servers), Fibre Channel RAID controllers, storage enclosures, and gateways. In the following

sections, we describe these hardware elements that are the components of a complete Fibre Channel storage solution.

### 6.2.1 Fibre Channel cabling and components

The usable length of a fiber cable is determined by the wavelength, power, and *mode* of the laser-emitting device. The mode of a laser refers to a specific spatial distribution or pattern of light within the fiber medium.

Short-wave laser emitters (780 to 850 nm) use multi-mode cables that have an inner diameter of 50 to 62.5 microns (1 micron = 1 millionth of a meter). The light can enter the cable in multiple modes. The many light beams tend to lose shape as they move down the cable. This loss of shape is called *dispersion* and limits the distance for multi-mode cable to a maximum of 275 to 500 meters.

Long-wave laser emitters (1310 nm) employ single-mode cables that have an inner diameter of 7 to 9 microns. This type of cabling supports up to 10 kilometers. Its distance is limited by the power of the laser at the transmitter and by the sensitivity of the receiver.

Table 14 lists the fiber optical cabling types supported by IBM.

Table 14. Cable types supported by IBM

Diameter (microns)	Mode	Laser type	Distance
9	Single-mode	Long-wave	<=10 km
50	Multi-mode	Short-wave	<=500 m
62.5	Multi-mode	Short-wave	<=175 m

IBM cabling uses the duplex-SC connector as shown in Figure 97:

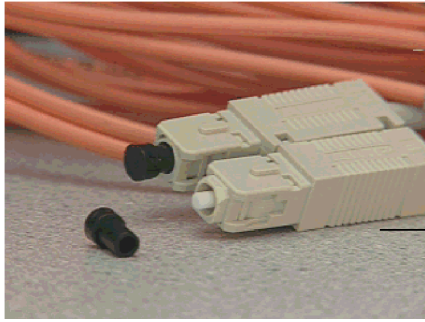


Figure 97. IBM Fiber Cables with duplex-SC connectors

Gigabit interface converters (GBICs), as shown in Figure 98, are hot-pluggable devices that convert a serial optical signal into a serial electrical signal or vice versa. They form the interface between fiber optical cables and devices on a fiber loop. Both short-wave and long-wave versions are available.

Short-wave GBICs can be installed on any device in your Fibre Channel loop. Long-wave GBICs would typically be used in the Fibre Channel hub or switch to interconnect remote sites, even though they may also be installed in the managed hub and the FAStT500 controller unit mini-hubs. You select the appropriate cabling and matching GBICs depending on the distance you require between nodes on the Fibre Channel loop.



Figure 98. Short-wave GBIC (left) and a media interface adapter (MIA).

Media interface adapters (MIAs), also shown in Figure 98, are transceivers that convert between optical cable and a DB-9 copper interface, fulfilling a similar function to GBICs. MIAs are only used on the 3526 Fibre Channel

RAID controller unit. The new FAStT500 controller unit uses only GBICs for its interfaces.

The IBM Fibre Channel hub, shown in Figure 99, is a 7-port FC-AL hub, designed to provide connectivity and simplify loop cabling. It comes with four short-wave GBICs as standard, so you will need to purchase additional GBICs to connect more than four devices. To allow for expansion beyond seven connections, hubs can be cascaded once, that is, you can have a maximum of two hubs between any two devices in the fibre loop. Since each hub has seven ports, you can configure up to a maximum of 37 ports for the attachment of hosts or controller units. All hub-attached devices are on the same loop, however, so it is recommended that you attach no more than two controller units to ensure good performance.



*Figure 99. IBM Netfinity Fibre Channel hub*

The hub can also be used to extend the distance between hosts, controller units and storage expansion units by using long-wave GBICs. These support long-wave cable lengths of up to 10 km. Long-wave GBICs could only be used in the Fibre Channel hub in 3526-based Fibre Channel subsystems, so two hubs were required to achieve distances above 500 m. The newer FAStT500 Controller can utilize long-wave GBICs, but switches and hubs would normally still be used for long-distance connections.

For further information on Fibre Channel nomenclature and labelling we recommend the ANSI committee's official Web site:

<http://www.t11.org>

## 6.2.2 Fibre Channel PCI adapters

The IBM Netfinity Fibre Channel PCI adapter (Figure 100) is an intelligent, high-performance, DMA busmaster host adapter designed for high-end systems.

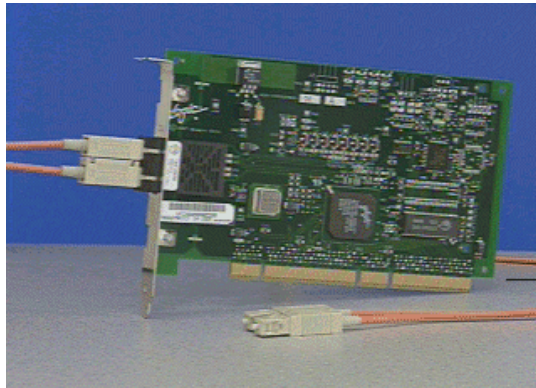


Figure 100. The IBM Fibre Channel PCI adapter

This Fibre Channel adapter is a 64-bit (66 MHz and 33 MHz) PCI card but it can be used in a 32-bit PCI slot. The fiber optic media connector supports only short-wave cabling (50 micron).

The following network operating systems are currently supported:

- Windows NT 4.0 and 2000
- NetWare 4.1x and 5.x
- SCO UnixWare 7.0 or 7.1

The card is hot-pluggable using Active PCI slots under Windows NT 4.0 and Windows 2000. When two Fibre Channel PCI cards are used in a single server for redundancy, Active PCI guarantees server uptime after a single component failure, enabling you to change the defective card without downing the server.

The newer Netfinity FAST Host Adapter, supports additional protocols, such as IP, and attaches the FAST500 RAID Controller Unit to Netfinity servers thorough a suitable fabric.

The adapter BIOS includes a configuration utility called Fast!UTIL, which is accessed by pressing Alt+Q during the adapter BIOS initialization. This utility allows you to set many parameters that control the adapter's operation, scan for devices on the Fibre Channel loop, and format attached drives along with

other operations. The *IBM Netfinity FAStT Host Adapter Installation and User's Handbook*, shipped with the adapter, gives details of the changes you can make using Fast!Util.

### 6.2.3 The Fibre Channel Controller Unit 3526

The Fibre Channel Controller Unit 3526 provides Fibre Channel optical attachment to a host server, and six SCSI channels are available for connection to external disk enclosures. The RAID controller converts the incoming Fibre Channel optical data (through MIAs, see 6.2.1, "Fibre Channel cabling and components" on page 172) to an electrical signal, performs RAID calculations, and then directs the appropriate SCSI commands to the low-voltage differential SCSI (LVDS) channels.

The controller unit is a rack-mounted device that connects to and controls disks installed in Netfinity EXP10, EXP15 and EXP200 storage enclosures. As illustrated in Figure 101 on page 177, it features redundant hot-swappable power supplies and fans, and, with the optional Netfinity Fibre Channel Failsafe RAID Controller, hot-swappable redundant RAID controller cards. Also included within the controller is 128 MB of cache, protected by battery backup.

In the simplest configuration, the RAID controller unit is connected directly to a Fibre Channel PCI adapter installed in a server. For more complex configurations, connection will be to a Fibre Channel hub or switch. The RAID controller unit has two Fibre Channel connections for each of the RAID controllers installed in the unit (one controller is standard, the other is optional). As mentioned previously, only short-wave connections are supported by the controller unit, which means that the cable to the host adapter or hub can be up to 500 meters in length.

#### Controller terminology

The RAID controllers are sometimes referred to as RAID controller modules or blades.

Six independent Ultra2 LVD SCSI channels allow connection to up to six external storage enclosures. The RAID controllers provide the functions necessary to let you configure RAID arrays using the disks in the external enclosures.

This product has been available since the end of 1998, and we do not cover it in detail here. The management software has been merged with the version available for the new FAStT500 controller unit as described in 7.3, "Migrating

from SYMlicity Storage Manager” on page 203. All other product features and hardware maintenance guidelines can be taken from the installation and user manuals.

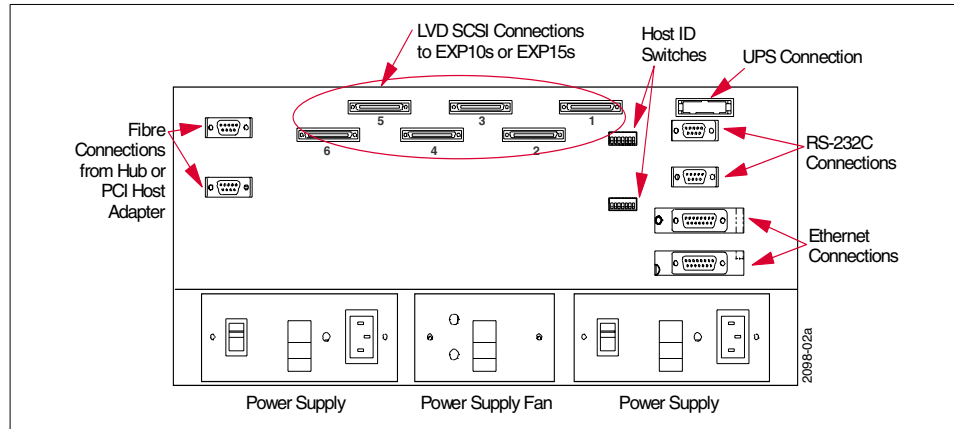


Figure 101. Fibre Channel RAID Controller Unit (Rear Panel)

## 6.2.4 The Netfinity FAST500 RAID Controller

The recently announced IBM Netfinity Fibre Array Storage Technology (FAST) products now offer customers the option to implement an adapter-to-drive Fibre Channel storage subsystem. In the predecessor product, Fibre Channel connections were maintained from the host adapter to the RAID controller, but the disk themselves were Ultra2 SCSI drives. The new FAST 9.1 GB, 18.2 GB, and 36.4 GB drives, spinning at 10,000 RPM, are native Fibre Channel drives, offering the best available performance.

The Netfinity FAST500 RAID Controller is a high performance, rack-mountable RAID controller that has fiber interfaces to both the server host and the attached disk drives. It is the central building block for Netfinity storage in a SAN environment, enabling the separation of hosts from storage, and provides a single point of management for all attached storage.

The controller unit supports up to two controllers, each controller being an independent RAID engine that can be configured in different modes serving the needs of a specific storage configuration. Each controller has 256 MB of data cache that is protected by battery backup. In addition, the controller unit offers complete redundancy features both on the host and drive interfaces.

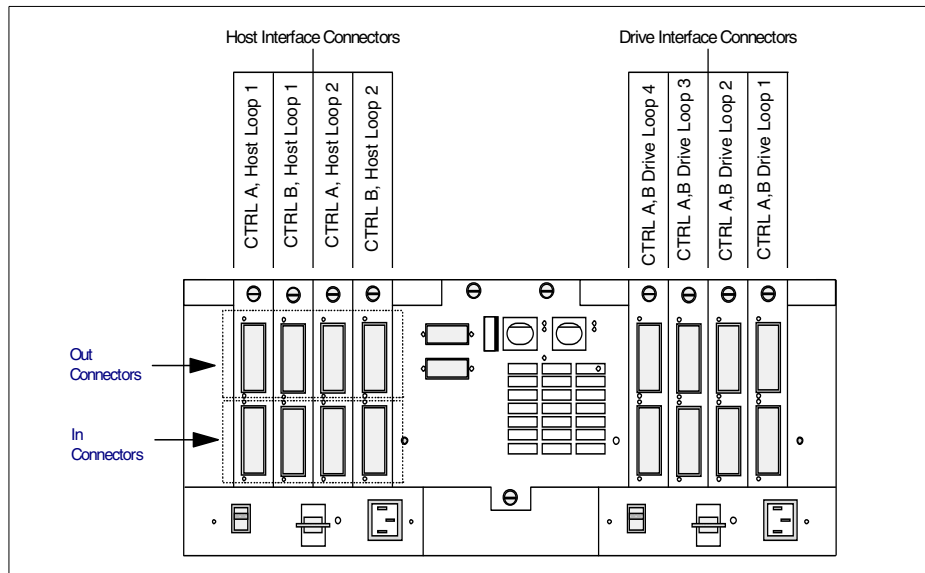


Figure 102. Rear view of the Netfinity FAST500 RAID Controller

This is accomplished through the integration of mini-hubs into the external interfaces of the controller unit. Figure 102 shows the host interface and drive connectors at the rear of the controller unit. Each interface card functions as a mini-hub and accommodates two ports in which GBICs are installed, allowing connection of the fiber cables. The Netfinity FAST500 RAID Controller comes with two mini-hubs on both the host and drive interface side as standard.

The host interface connectors connect to either Controller A or B. Redundancy is achieved through the second controller, assuming that the host can provide an alternative path to it through a second host adapter card.

The drive interface connectors provide redundant paths to the drive enclosures. The drive expansion unit itself, which is described in 6.2.9, “The IBM FASTT EXP500 Storage Expansion Unit” on page 185, provides path redundancy in the same way as the controller unit, through the integration of mini-hubs, providing two fiber loops to the expansion units.

In order to better understand the cabling requirements of the controller unit to the host and the drive expansion units, a logical view of the connectors at the rear of the controller unit to the internal controllers is shown in Figure 103:



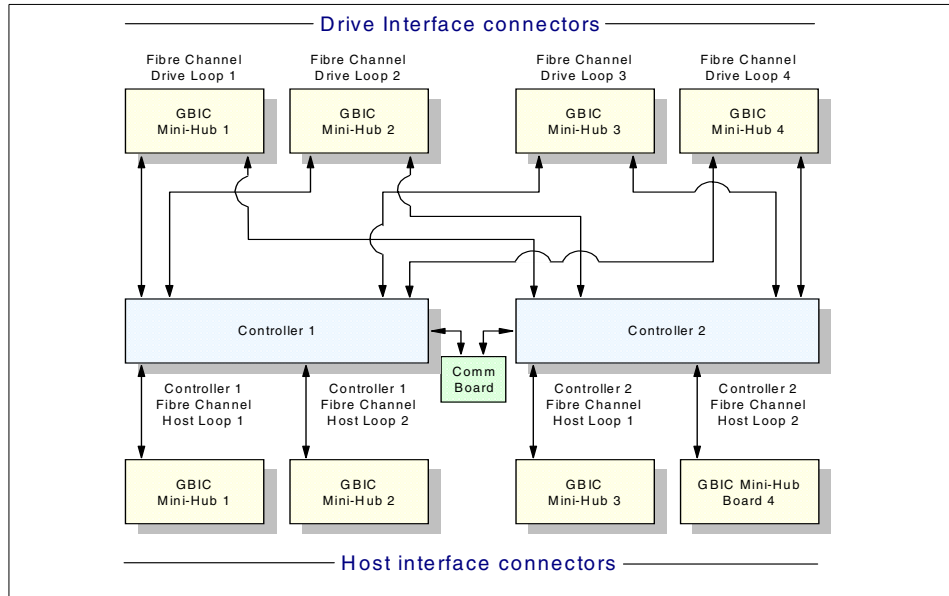


Figure 103. Logical view of the Netfinity FASt500 RAID Controller

Each mini-hub in the drawing represents two GBIC connectors and therefore one row of connectors as seen from the rear of the unit in Figure 102. You can easily match the physical ports at the rear of the controller with the drawing in Figure 103 by comparing the host and drive loop numbering. Note that each line represents a duplex fiber connection.

To ensure full redundancy, you must connect each host to each controller board. Without using external hubs or switches you can connect up to four redundant hosts each with two PCI Fibre Channel Cards directly to the controller unit. The drive interfaces allow the attachment of up to 11 drive enclosures per pair of interface cards in a fully redundant configuration. This allows a total of 22 EXP500s to be connected to the controller unit. The limitation in number of total drive enclosures stems from the Fibre Channel Arbitrated Loop protocol, which allows a maximum of 126 devices per loop.

### 6.2.5 Fibre Channel controller RAID levels

The implementation of RAID as supported by the two Netfinity Fibre Channel controllers (the 3526 and the Netfinity FASt500 RAID Controller) differs from the implementation used by the ServeRAID SCSI adapters and SSA adapters.

Both Fibre Channel controller units offer RAID levels 0, 1, 3 and 5. RAID-1 arrays can consist of more than two disk drives as long as the number of drives is even.

Table 15 summarizes the maximum number of disk drives that can be configured into a single array, depending on the Fibre Channel controller unit used. These numbers are the total numbers of disks in a single RAID array, they do not reflect the total usable space. Using the Netfinity FAStT500 RAID Controller controller unit as an example, a RAID-5 array can have a maximum of 30 disks, providing you with 29 disks' capacity as usable space, whereas a RAID-1 array will give you a usable space of only 15 disks' capacity.

Table 15. Maximum number of disk drives per array

RAID level	Controller Unit 3526	Netfinity FAStT500 RAID Controller
0	20	30
1	30	30
3	20	30
5	20	30

Logical drive migration is also supported by the Fibre Channel controllers. Migrations are possible to and from each RAID level with few restrictions (as an example, you cannot migrate a five-disk RAID-5 array to RAID-1 because RAID-1 requires an even number of drives).

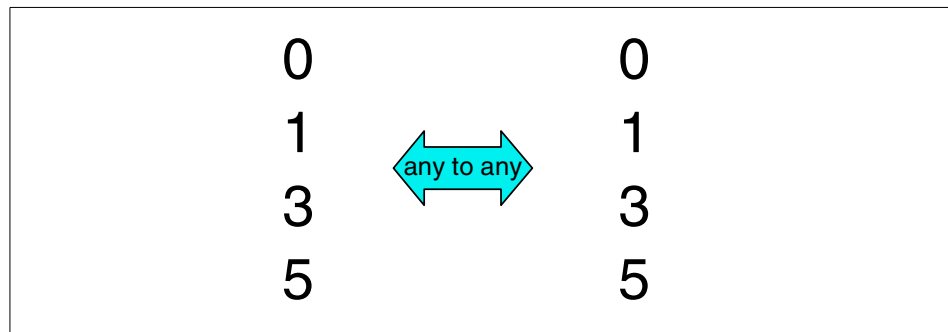


Figure 104. RAID level migrations

Your data is still accessible during the migration process, although some effect on subsystem performance may be observed, depending upon the

nature of concurrent activity. Once the migration process is started it cannot be stopped. The array being migrated must be in an optimal state.

The following series of diagrams (Figure 105 to Figure 109) show the way in which data is distributed across the disks in array. For the sake of simplicity, we have assumed that a single logical drive has been configured within each array.

RAID-0 stripes the data across all drives in the array. This provides excellent performance, but no redundancy. A failing disk results in loss of all data in the array so RAID-0 is usually reserved for fairly static data that is easy to recover from a backup or by reinstalling. Here is how the data is distributed across the disks in the array:

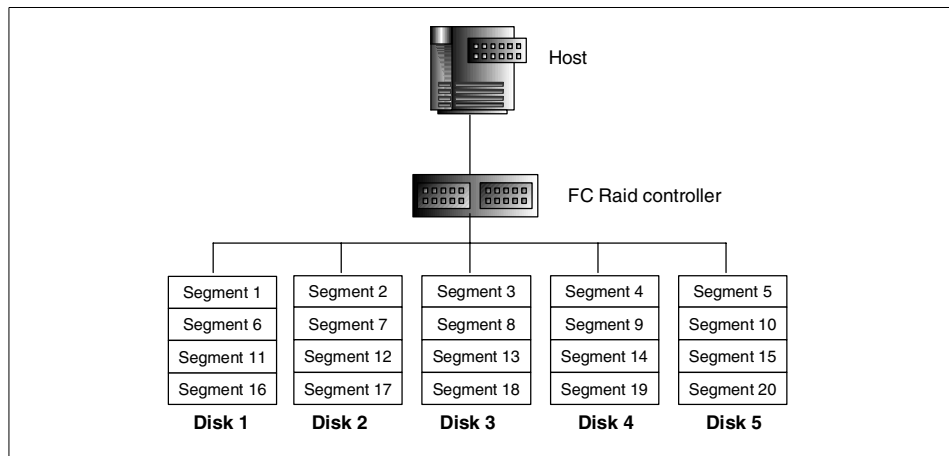


Figure 105. RAID-0 data organization

RAID-1 is essentially disk-mirroring. In its purest form, only a single drive is mirrored, limiting the maximum capacity available to that of a single drive:

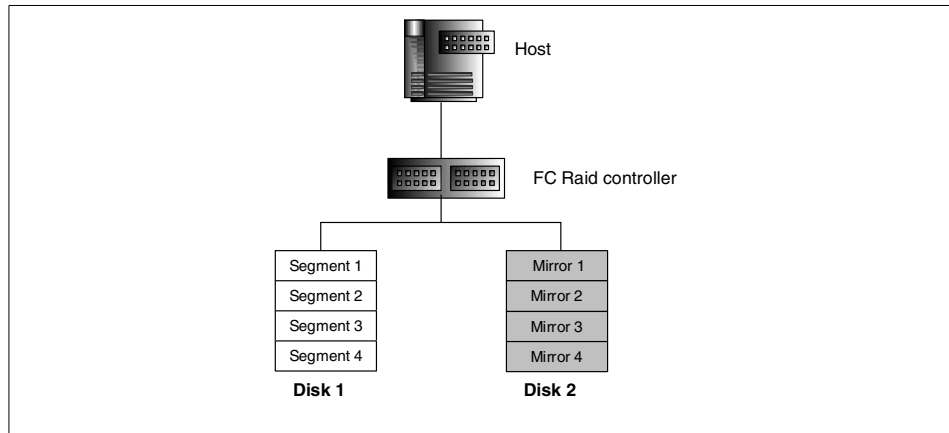


Figure 106. RAID-1 with only two drives

However, modern RAID controllers usually provide a trivial extension to this, allowing the creation of much larger logical drives:

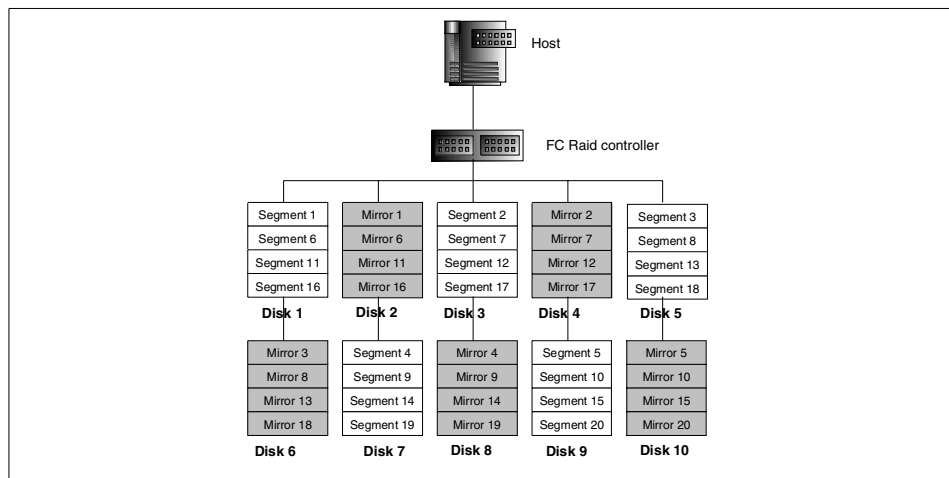


Figure 107. RAID-1 with more than two drives (sometimes called RAID-10 or RAID-0/1)

The Netfinity FASt500 Controller Unit also supports RAID-3. Originally, the Berkeley definitions of RAID levels specified that RAID-3 arrays stripe bytes across the disks. More recently, the RAID Advisory Board definitions allow for block striping across the drives, which is the way RAID-3 is implemented in

the FAStT500 subsystem, and hence there is essentially no difference between RAID-3 and RAID-5.

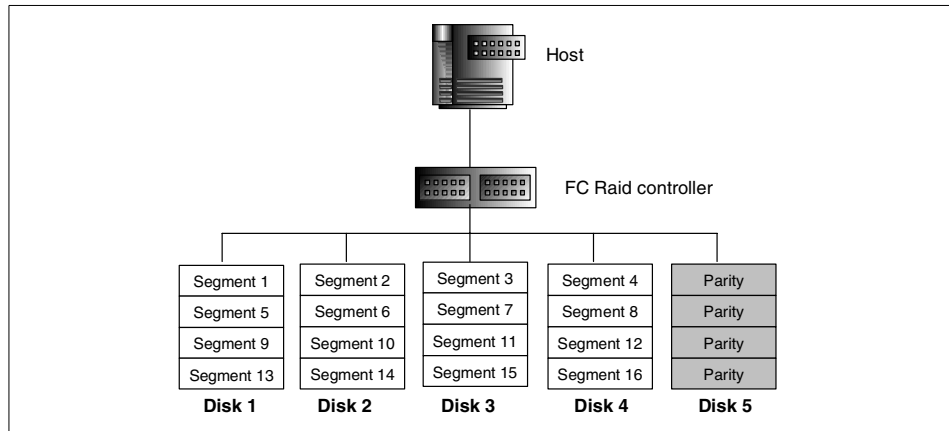


Figure 108. RAID-3 data organization

Finally, we show how RAID-5 arrays have their data distributed across the disks in the array. RAID-5 is probably the most commonly used RAID level as it offers a good compromise between cost, performance and redundancy. A single drive can be lost without interrupting service from the server.

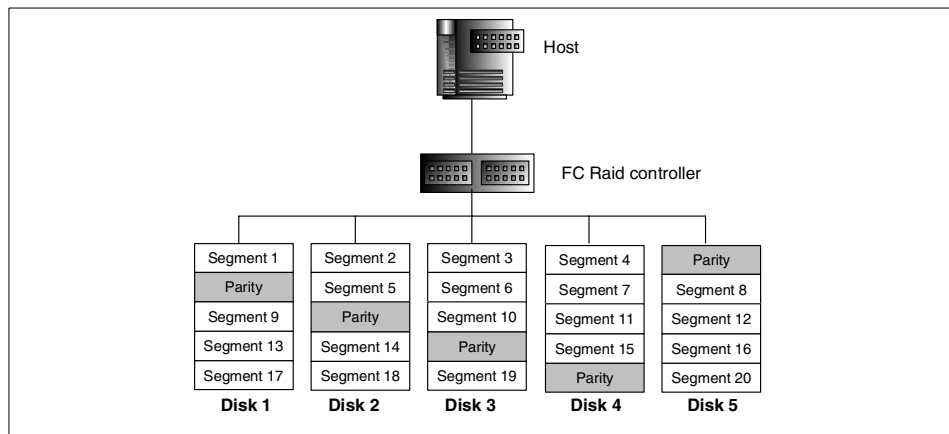


Figure 109. RAID-5 data organization

## 6.2.6 The IBM Netfinity FAStT200 RAID/Storage Unit

The IBM Netfinity FAStT200 RAID/Storage Unit is the latest in IBM's range of Fibre Channel storage subsystem components. It is a highly integrated unit

that includes up to two RAID controllers and 10 drive bays in a single, 3U, rack-mountable enclosure, and is shown in Figure 110.

The Netfinity *FAST200R* RAID/Storage Unit offers a high degree of redundancy with hot-plug redundant power supplies, fans and RAID controllers. When less redundancy is acceptable, the *FAST200* RAID/Storage Unit has a single controller and power supply, both of which can be augmented with a second redundant component should your needs change.



Figure 110. Netfinity *FAST200* RAID/Storage Unit

The *FAST200* units offer similar RAID functions to the *FAST500* Controller and utilize disks compatible with the *FAST EXP500* disk enclosure. The following list highlights the key features of these units:

- Flexible data storage configurations with multiple host and drive buses.
- Support for mirroring, redundant loops, and pooling of disks and tape drives.
- High availability features as standard or as options.
- High performance - bandwidth up to 166 MBps.
- Highly scalable - supports up to 30 FCAL disks with redundant loops by attaching external *FAST EXP500* enclosures.
- Standard Netfinity *FAST* Storage Manager manages up to four storage partitions.

- Support for RAID levels 0, 1, 3, 5, and 10.
- Data storage up to 10 km (6.2 miles) away, providing additional protection from catastrophic occurrences.
- Support for long- and short-wave Fibre Channel environments with GBICs and optical cabling.

### 6.2.7 The IBM Netfinity EXP200 Storage Expansion Unit

This new storage enclosure offers SCSI-based connections and can be connected to the 3526 controller unit.

For further information we refer you to 4.3.1, “Netfinity EXP200 Storage Expansion Unit” on page 78.

### 6.2.8 The IBM Netfinity EXP300 Storage Expansion Unit

At the time of writing, this expansion unit is not supported in Fibre Channel configurations. It may be attached to ServeRAID-4 family, and ServeRAID-3 family adapters only.

### 6.2.9 The IBM FAST EXP500 Storage Expansion Unit

The IBM FAST EXP500 is a disk drive expansion unit with Fibre Channel interfaces both externally to the controller unit and for the disk drives themselves. Figure 111 shows a physical view of the unit:

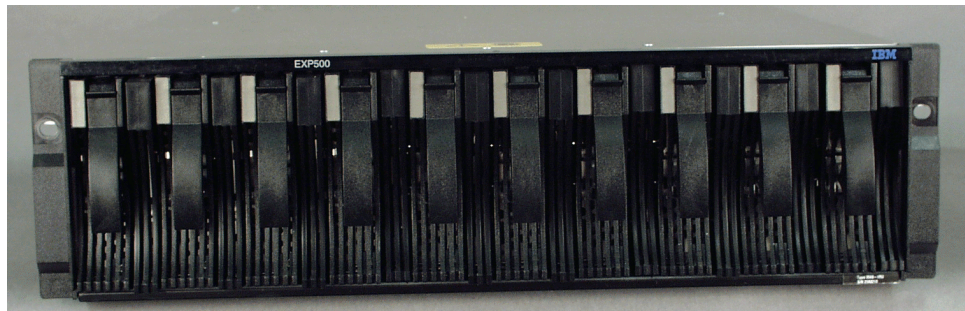


Figure 111. The IBM Netfinity FAST EXP500

The EXP500 can house up to 10 half-high or slim-line 40-pin Fibre Channel drives and features redundant fibre loop support. A redundant EXP500 fibre loop consists of one or more expansion units connected to a host or controller using two sets of fiber cable. Figure 112 on page 186 shows a logical view of redundant loop design.

Two Enclosure Services Monitor (ESM) boards are contained within an EXP500 unit. Each ESM board has two GBIC ports, one for incoming and one for outgoing Fibre Channel cables. Both connections are full duplex and they fulfill the same function but are utilized differently in the way we recommend you cable them to the controller unit and other drive enclosures.

Each hard drive has two interfaces that connect, one each, to the two internal loops of the EXP500. If one of the loops experiences a link problem, the hard drives are still accessible through the other loop. This is also true for single drive interface (that is, ESM) failures. If a single drive interface fails, it can be bypassed and the drives may still be accessed through the other loop.

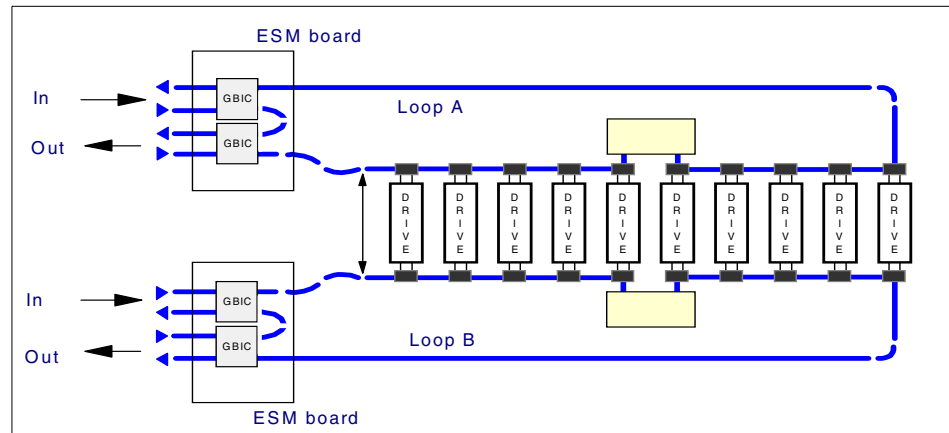


Figure 112. Redundant loop diagram of FAST EXP500

In Figure 113, we show a view of the back panel of the EXP500 enclosure, pointing out, in particular, the locations of the GBICs and the ESM assemblies:



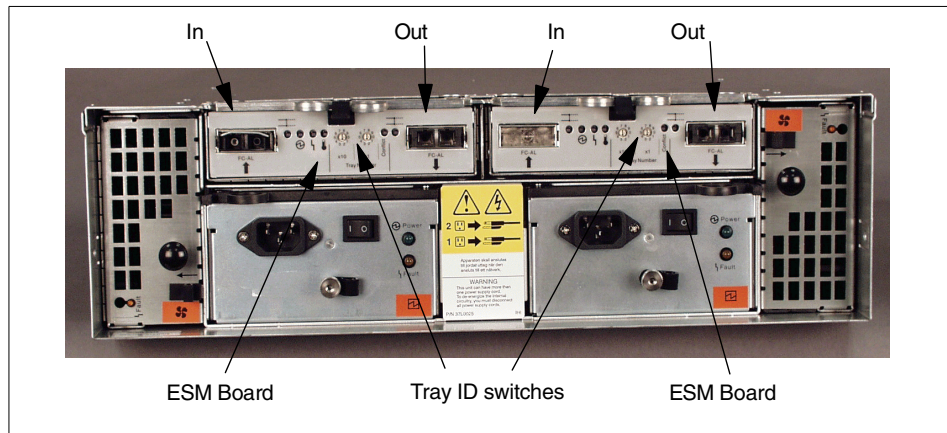


Figure 113. Rear view of the FAStT EXP500 unit

#### Setting the tray ID

When attaching a new EXP500 to a controller unit, remember to set the Tray ID using the rotary switches at the back of the unit on each ESM board. You can set any ID between 00 and 99.

The IDs on both ESM boards in a single EXP500 must be set to the same value. Setting the IDs correctly is especially important when connecting to an existing controller unit, changing an ESM board, or when troubleshooting an existing installation. Keep in mind that the Tray ID must be unique for each expansion unit attached to a Fibre Channel Controller Unit.

When changing the tray ID of an expansion unit, the controller unit must be rebooted (by cycling the controller power).

Detailed usage and maintenance instructions are provided in the product documentation: *IBM Netfinity EXP500 - Installation and User's Handbook*, shipped with the unit. This manual is especially recommended when you install disk drives into the EXP500 for the first time and want to learn more about the various control LEDs and switches used to provide information and to configure the unit.

#### 6.2.10 Cabling requirements for the EXP500

Both the Netfinity FAStT500 RAID Controller and the EXP500 storage enclosure offer full Fibre Channel interface integration. This introduces new

concepts in cabling these components, especially as the total solution is designed to be fully redundant by avoiding any single point of failure.

The drive interface connectors of the controller unit, as seen in Figure 102 on page 178, have outgoing (top row) and incoming (bottom row) connectors. You will also see incoming and outgoing interfaces on the ESM boards of the EXP500 as in Figure 113 on page 187. When installing a Netfinity Fibre Channel disk subsystem, we recommend that you always cable outgoing ports to incoming ports and vice versa. This is not required for correct operation of the hardware as there is no physical difference between the incoming and outgoing ports; each interface is a full Fibre Channel loop with two optical fiber cables. It is merely a designation for allowing easier installation of the cabling. Following a convention, as we have suggested, makes cabling consistent and, therefore, easier to follow and debug when necessary.

As you can see in Figure 114, to attach a single EXP500 unit to the controller, we have followed the simple rule above by connecting the outgoing connectors on the Netfinity FAStT500 RAID Controller to the incoming connectors on the EXP500. Furthermore, in order to provide full redundancy, both Fibre Channel loops within the EXP500 are cabled to the controller unit. Each mini-hub on the RAID controller represents a single Fibre Channel loop which is connected to both controllers as shown previously in Figure 103 on page 179.

In order to provide redundancy, the second drive loop in the EXP500 is connected to a second mini-hub. Referring to Figure 102 on page 178, you can see that drive loops 3 and 4 are utilized in this example.

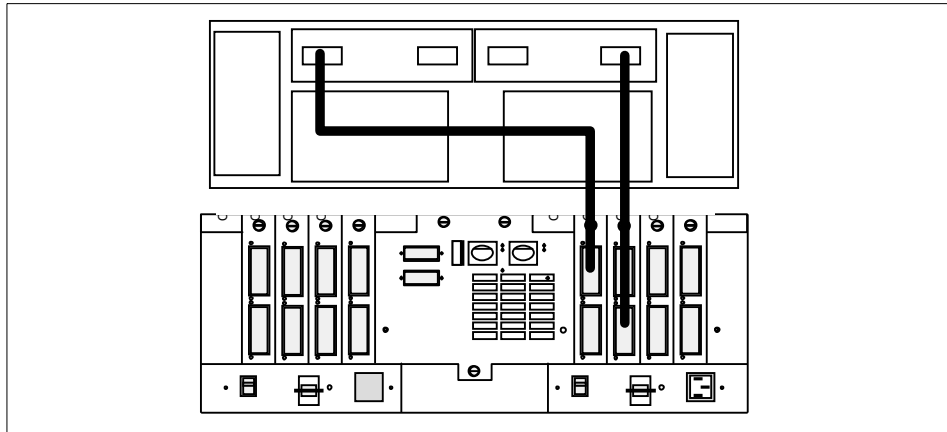


Figure 114. Single EXP500 connected to a Netfinity FAST500 RAID Controller

When additional drives are required, multiple EXP500 storage enclosures are attached to the Fibre Channel loop, and the same simple rule should be followed. However, there is a slight variation which we will now describe.

Configure your first loop from drive loop 4 (out) at the controller unit to the top EXP500 (in), then daisy-chain the remaining EXP500s (out to in) as shown in Figure 115 on page 190. It is important to remember that this is a complete, operational Fibre Channel loop, not a string of connections. The redundant loop is then connected, starting at drive loop interface 3 (in) coming from the bottom EXP500's second ESM board (out), again daisy-chaining the EXP500's (out to in). connector to the drive loop interface 3 (ingoing) of the controller unit. The multiple EXP500 enclosures are cascaded with short fiber cables.

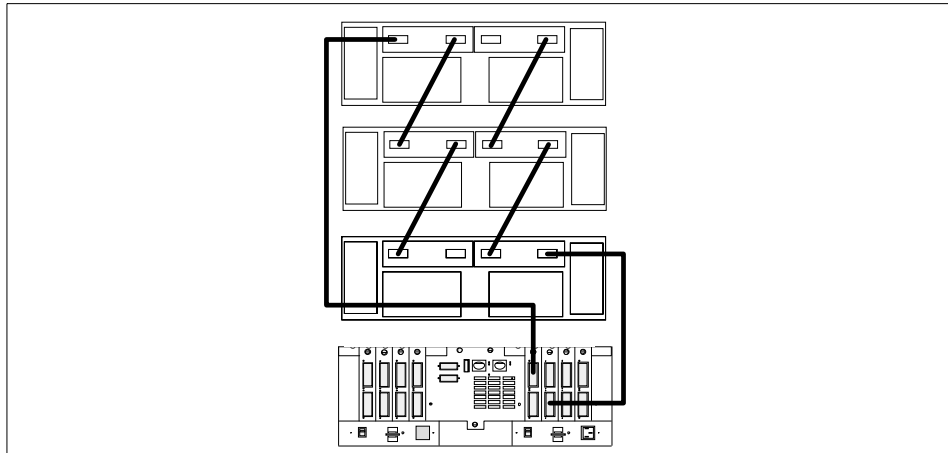


Figure 115. Multiple EXP500 cabling scenario

It is worth reiterating that this is a redundant configuration. Two loops, supporting up to 11 EXP500 storage enclosures, have been configured. There is a significant benefit to cabling the units as we have described. Should one of the enclosures fail in such a way that the controller cannot communicate with drives in enclosures beyond the fault, the controller will still have access to all enclosures except for the one that has failed. If both loops had been cabled in the same direction (either “up” or “down” the stack of EXP500s), then communication with the more remote enclosures would be lost.

In order to extend the previous scenario of going up to a total of 22 storage enclosures, you should follow Figure 116. By repeating the cabling pattern we have just discussed, but using drive loop interfaces 2 and 1, the controller unit can take up to another additional 11 storage enclosures in a redundant configuration.

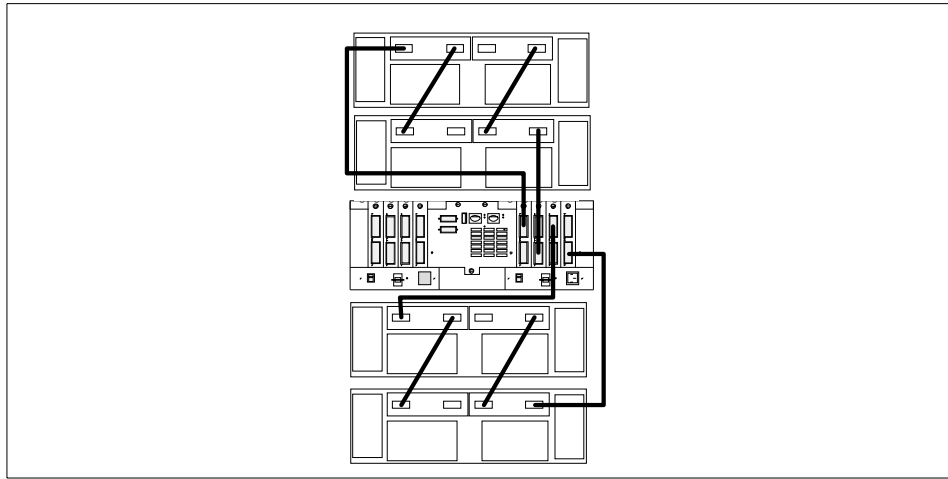


Figure 116. Using four drive mini-hubs allows up to 22 EXP500s to be attached

Attaching hosts to the Fibre Channel control unit is equally simple. Starting with a single host with two PCI Fibre Channel cards as shown in Figure 117, you connect each host adapter to a separate controller using the outgoing host interface connectors of the controller unit. As you can see from Figure 102 on page 178 and Figure 103 on page 179, the above cabling connects both controllers A and B within a single Fibre Channel loop.

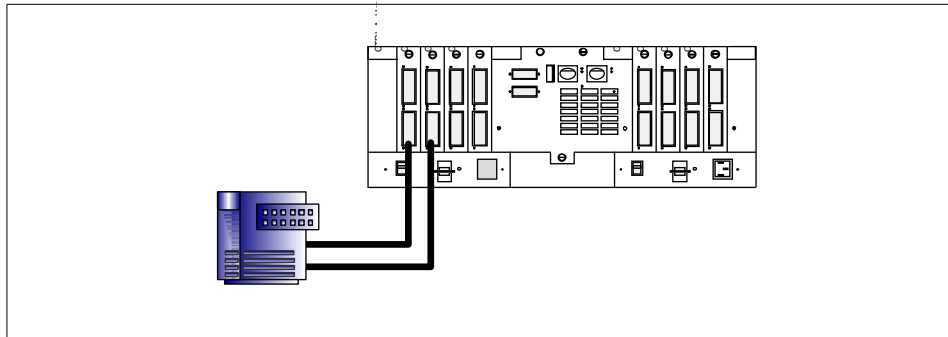


Figure 117. Single host attachment to a Netfinity FAST500 RAID Controller

When planning for storage partitioning (see 7.6.2, “Storage partitioning” on page 219), you will see that logical drives are not assigned to specific PCI Fibre Channel adapters but to the server in which the adapters reside. Bearing this in mind, connecting the adapters to both controllers A and B in

the Netfinity FAStT500 RAID Controller gives you the most redundancy, even though they are in the same Fibre Channel loop.

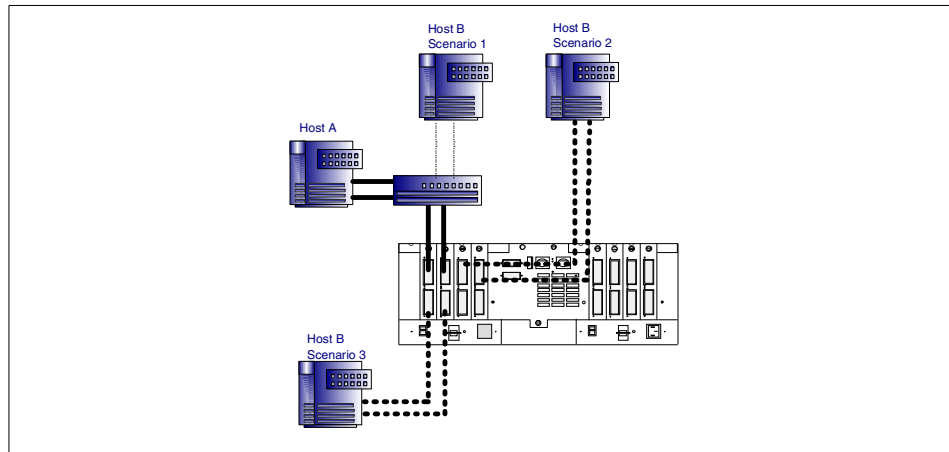


Figure 118. Multiple host attachment to a Netfinity FAStT500 RAID Controller

When going from a single to multiple hosts as in Figure 118, you have three possible scenarios to choose from, depending on the solution you wish to implement.

Scenario 1 cables a second host to the same host interface connectors via a Fibre Channel switch or hub. This is a possible standard configuration when implementing Microsoft Cluster Server, other two-node high availability solutions, or file and print servers. Note that for this configuration to be valid, the switch must be using zoning to map out its ports. For two adapters from each node to connect into a hub or into a switch without zoning, and then from that switch or hub into both controllers may cause unexpected behavior. If zoning is not used, then two switches or two hubs must be used.

In scenario 2 you can utilize the second Fibre Channel loop allowing more hosts to be connected to the controller unit in total. This would be the premier choice for performance conscious installations, since it distributes load onto two separate Fibre Channel loops.

Scenario 3 is similar to the first and utilizes the mini-hubs built into the control unit.

The controller unit allows the total attachment of 64 hosts (128 Fibre Channel PCI cards). Each controller has its own Fibre Channel loop and can therefore

take up to a maximum of 32 hosts, each equipped with two Fibre Channel PCI cards.

The scenarios described are suggestions to help explain the principles behind the new Fibre Channel hardware. Careful consideration should be put in every single configuration you wish to create, depending on the actual requirements of your host applications and the desired performance and redundancy features.

### 6.2.11 The IBM SAN Fibre Channel Switches

The IBM Fibre Channel switches provide high-speed, fault-tolerant interconnectivity in a SAN fabric. The interconnection of IBM and IBM-compatible switches and hubs creates a scalable, fault-tolerant switch fabric, potentially containing hundreds of Fibre Channel ports. The IBM Fibre Channel Switch is offered in eight-port (Model S08) and 16-port (model S16) versions as shown in Figure 119. Both models come with four short-wave GBICs as standard.



Figure 119. The IBM Fibre Channel Switches 2109, Models S08 (top) and S16

Each port connects at 100 MBps and supports both short-wave and long-wave GBICs, enabling you to build a switch topology over distances of up to 10 km. As multiple switches are interconnected or cascaded to build up a fabric, the performance of the fabric can be vastly increased. The latency of the switch is under two microseconds at peak Fibre Channel throughput of 100 MBps. However, this latency could increase when the destination or device is a loop. This would occur if the traffic on the destination loop blocks the output port of the switched connection.

The switch has a non-blocking architecture, meaning that the switch can handle multiple data paths simultaneously. If it makes a connection between two ports, it does not report a busy signal to another port trying to make a connection. Routing is implemented as hardware cut-through, which means that transmission can begin as soon as a frame arrives at the port.

The aggregate bandwidth is 8 x 100 MBps (0.8 GBps) or 16 x 100 MBps (1.6 GBps), for the S08 and S16 models, respectively.

Each port of the switch can operate as an inter-switch port (E\_Port) between switches, as a fabric port (N\_Port) to realize point-to-point links, or as a fabric loop port (FL\_Port) for attachment to a Fibre Channel loop.

Figure 120 shows four of these switches in a ring topology. All traffic going between switches 1 and 2 will be distributed across the two direct links. When traffic is going from switch 1 to 3, the switch software automatically makes all four possible paths available to this specific route, that is two paths through each of switches 2 and 4 respectively.

Additional switches and storage devices can be added to an existing fabric without any disruption. A new switch is integrated into the fabric automatically and requires no manual intervention or configuration.

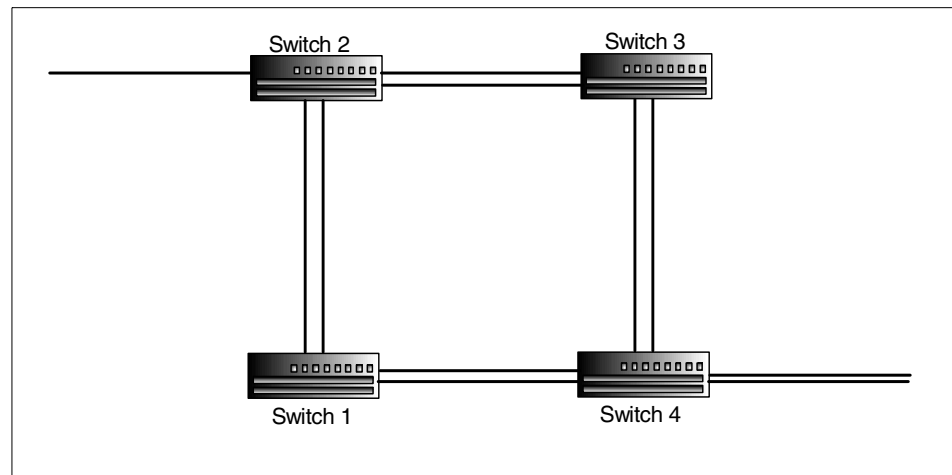


Figure 120. Increasing bandwidth by cascading multiple switches



Whenever a fabric is being designed, the following things should be taken into account when full redundancy and availability are a prime goal of the entire storage solution:

- Fault-tolerant topologies are only feasible with two Fibre Channel PCI adapters in each host connecting to separate switches.
- The fabric design must offer alternative paths in the case of a single link failure.
- If speed is critical, devices that require most of the bandwidth should be put on the same switch. If this is not possible, multiple links between appropriate switches will increase bandwidth.
- The switch fabric allows a maximum of seven hops in routing, so you can have eight switches in between any two points in your fabric. A hop is defined as a link from one switch to the next.
- A maximum of 32 switches can be cascaded.
- A maximum of 8 paths from any node to a target can exist. This is the maximum the switch software can administer.
- Every hop or switch in the fabric adds a maximum of 2 microseconds in signal delay.
- Every kilometer of fiber cable adds 5 microseconds to signal delay
- Cascading switches can create complex fabrics. The IBM SAN Fibre Channel Switch offers *zoning*, which is the ability to specify levels of access to connected ports or devices. This not only simplifies the management of cascaded switch environments but it is equally important in multi-host environments for security reasons.
- A Fibre Channel arbitrated loop can only be connected to a switch at one point in the fabric. Since any port on the switch can act as a FL\_port you can have multiple arbitrated loops connected to a single switch, assuming these arbitrated loops are all independent of each other.

The switch can easily be managed either directly through a Telnet session or by using the IBM StorWatch Switch Specialist via a 10/100 Mbps Ethernet port. The StorWatch software is based on an HTTP service and can therefore be accessed from any Web browser.

The initial setup of the switch requires an IP address to be set for the switch. The 16-port model provides a front-panel LCD with navigation buttons for this purpose. Due to space restrictions, the 8-port model offers a serial DB-9 interface to provide access for setting the initial configuration.

### **6.2.12 IBM SAN Fibre Channel Managed Hub**

The IBM SAN Fibre Channel Managed Hub is a 1U high, rack-mounted, entry-level device for connecting together components in a Fibre Channel fabric. It is designed to support a homogeneous cluster of host servers and storage systems. An option is available to allow configuration as a stand-alone device.

The hub has eight FC-AL ports. Seven of them support fixed short-wave optical media for connecting devices on multimode fiber over distances of up to 500 meters. The eighth port is a gigabit interface converter (GBIC) slot that can be configured for either short-wave or long-wave optical media. Long-wave singlemode fiber can be up to 10 km in length.

An arbitrated loop is logically formed by connecting all eight ports on the hub into a single loop, or the ports can be zoned into several independent arbitrated loops. Each port supports 100 MBps full duplex data transfer. Two Managed Hubs can be cascaded, providing a loop of up to fourteen ports, and the hub can also be attached to the IBM SAN Fibre Channel Switch, providing loop attachment of storage devices.

The StorWatch FC Managed Hub Specialist, included with the Managed Hub, enables you to configure, manage, and service the hub via a Web browser from a workstation over an Ethernet network connection to the Managed Hub. The StorWatch Specialist also provides SNMP messages, traps, and management information bases (MIBs) that can be integrated into existing enterprise structures.

### **6.2.13 The IBM SAN Data Gateway for SCSI**

The IBM SAN Data Gateway is a hardware solution that enables the attachment of SCSI storage systems to Fibre Channel storage networks. It provides one to three short-wave Fibre Channel ports and four Ultra SCSI differential ports for attachment of disk or tape storage.



Figure 121. The IBM SAN Gateway 2108-G07 for SCSI

The short-wave Fibre Channel ports allow attachment using fiber optic cables of up to 500 m in length. This distance can effectively be extended to up to 10 km by incorporating Fibre Channel switches or hubs employing long-wave GBICs. The SAN Data Gateway occupies one Fibre Channel ID for each Fibre Channel interface, and the SCSI target devices are LUNs on the same Fibre Channel ID.

Each Ultra SCSI channel has internal termination, which can be disabled by the management software and automatic speed negotiation. The SCSI Channel IDs are set to 7 and support up to 15 SCSI target IDs and up to 32 LUNs per ID. The SAN Data Gateway is managed through the StorWatch SAN Data Gateway Specialist management software.

Devices attached to the SAN Data Gateway should all be either tape units or disk units; they should not be mixed. The following devices are currently supported:

Disk attachment:

- Enterprise Storage Server
- Versatile Storage Server

Tape attachment:

- Tape Subsystems Magstar 3570 and 3590
- Tape Library Magstar 3575

For detailed product documentation and current support information of non-IBM and non-Intel based servers, we refer you to:

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



---

## Chapter 7. Implementing Fibre Channel disk subsystems

This chapter provides a detailed introduction to the software utilities that enable you to exploit the capabilities of IBM's Netfinity Fibre Channel products.






After describing some of the terminology used in discussing Fibre Channel products, we examine the two basic ways to manage your Fibre Channel disk subsystems. These are direct management and host-agent management. Also included in this chapter are examples showing how Netfinity Storage Manager can control and configure your disk subsystem.



---

### 7.1 Terminology and definitions

The following table summarizes important terms that are used when discussing the Netfinity Fibre Channel products. Where appropriate, we have included the related icon as used in the Fibre Channel management software.

Table 16. Fibre Channel terms

Unconfigured capacity 	The capacity of drives in the storage subsystem that have not been assigned to an array. This space can be used to create new arrays and logical drives.
Array 	A set of drives that are logically grouped together by the controller in a storage subsystem. The RAID level is defined for an array, that is, all logical drives within a single array have the same RAID level.
Logical drive 	A logical drive is a contiguous subsection of an array. Each logical drive is presented to the operating system as a single physical drive.
Free capacity 	A contiguous region of unassigned capacity on a defined array. This space can be used to create one or more logical drives.
Host 	A computer that is able to access the storage subsystem through an I/O data connection.

<p>Host group</p> 	<p>An entity in the storage partition topology that defines a logical collection of hosts that need shared access to one or more logical drives.</p>
<p>Storage subsystem</p> 	<p>A storage entity, managed by the storage management software, that consists of both physical components (such as drives, controllers, and switches) and logical components (such as arrays and logical drives).</p>
<p>Management station</p>	<p>A computer used to manage storage subsystems on the network.</p>
<p>Storage partition</p>	<p>A collection of storage subsystem logical drives that is visible to a host or is shared among hosts that are part of a host group.</p>
<p>LUN</p>	<p>Logical Unit Number. The number a host uses to access a logical drive. Each host has its own LUN address space depending on the operating system executing on the host.</p>

---

## 7.2 Introduction to Netfinity Fibre Channel Storage Manager 7

Netfinity Fibre Channel Storage Manager 7 (SM7) is the management software for the 3526 controller unit and the Netfinity FASt500 RAID Controller and all supported storage enclosures attached to them.

Storage Manager 7 comprises several software components, allowing maximum flexibility for deploying it in different scenarios:

- SM7client

SM7client is installed on a management station, and lets you perform storage management tasks for accessible Fibre Channel storage subsystems. It can be installed on either a remote workstation or on the host to which the subsystem is attached. The SM7client software has been termed a thin client because it provides an interface for storage management based on information supplied by the storage subsystem controllers. When you use SM7client from a management station to manage a storage subsystem, you send commands to the storage subsystem controllers. The controller firmware contains the necessary logic to carry out the storage management commands. The controller is responsible for validating and executing the commands and providing the status and configuration information that is sent back to SM7client. The

SM7client software is what is usually being referred to by the term Storage Manager.

- SM7agent

SM7agent, the host-agent management program, is a software component that you can install on one or more hosts connected to storage subsystems. The host-agent, along with the network connection on the host, provides a network management connection to the storage subsystem. Individual Ethernet connections also exist on each controller for direct management.

A management station can communicate with a storage subsystem through the host that has SM7agent installed. SM7agent receives requests from the management station through the network connection to the host and sends them to the controllers in the storage subsystem through the Fibre Channel I/O connection. Included in the SM7agent package is the SM7devices utility that correlates the logical drives you create using the storage management software with their operating system device names. You should install the SM7agent software on all host computers, even if you plan to manage the storage subsystem directly over the network. You can stop the SM7agent from running using an operating system-specific method. For more information, refer to the *Installation and Support Guide* shipped with the product.

- RDAC

RDAC is an acronym for Redundant Disk Array Controller, which is a software component that comprises a multipath driver and a hot-add utility. This software is installed on the host system and provides redundancy in the case of component failure.

For redundancy, a pair of active controllers is installed in the storage subsystem controller unit. Each logical drive in the storage subsystem is owned by one of these controllers. The owning controller manages the I/O between the logical drive and the application host along the I/O path. When a component in the I/O path fails, such as a cable or the controller itself, the RDAC multi-path driver transfers ownership of the logical drives assigned to that controller to the other controller in the pair.

The hot-add function allows you to register newly created logical drives with the operating system.

Netfinity Storage Manager has two main views: the Enterprise Management window and the Subsystem Management window, both shown in Figure 122 on page 202.

The Enterprise Management window allows you to add and discover the storage subsystems you want to manage. You also define partitions, topology, and logical drive-to-LUN mappings here. A powerful script editor is also available through this interface.

The Subsystem Management window lets you manage a specific storage subsystem and perform any necessary administrative tasks. This is the interface you use to create arrays, logical drives and hot-spares. It also gives you access to alert functions and general information on subsystem components.

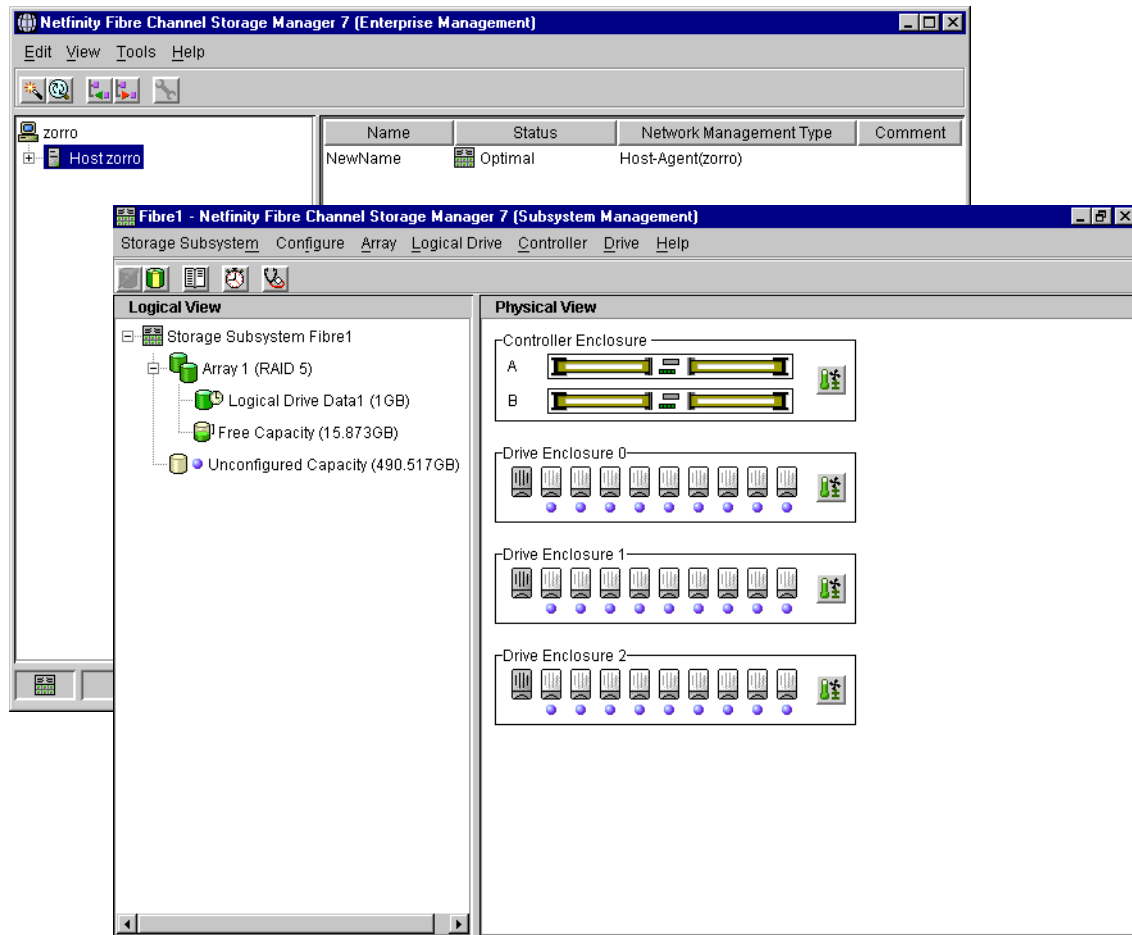


Figure 122. Enterprise Management (rear) and Subsystem Management window (front)



---

### 7.3 Migrating from SYMplicity Storage Manager

If you are already managing a 3526 series controller, it is likely that you still use SYMplicity Storage Manager 6.22, and the Fibre Channel controller itself will have firmware Release 3.x. New versions of both software elements are available with the new product announcements.

An upgrade to Netfinity Fibre Channel Storage Manager 7 controller and firmware Release 4.x will add new features and capabilities to an existing installation:

- Storage Partitioning (four partitions by default on a 3526).
- Support for up to 32 logical units under Windows NT 4 with Servicepack 5.
- Background media scan (data scrubbing).
- Remote management capabilities.
- Improved graphical user interface, which eases storage management tasks considerably.

If desired, you can still use SYMplicity Storage Manager 6.22 and run it in parallel with the newer Netfinity Fibre Channel Storage Manager 7. This is termed a *coexistence installation*. There is one important limitation when operating in this mode, however. You cannot use both versions of the management software to manage a single controller. You must use SYMplicity Storage Manager 6.22 for RAID controllers running firmware Release 3.x, not firmware Release 4.x controllers. Similarly, Netfinity Fibre Channel Storage Manager 7 must only be used with firmware 4.x controllers. Attempts to violate this restriction will fail as version 6.22 cannot communicate with a 4.x firmware controller and SM7 cannot communicate with a controller with firmware 3.x.

A coexistence installation should be considered as a transitional step when adding new Netfinity FAStT500 RAID Controller to an existing configuration using one or more 3526 controllers.

#### **Before migrating**

The host-management method, described in 7.4.2, “Host-agent management” on page 206, requires an access volume for communication with the controllers. The access volume uses one of the allowable LUNs. If the host system has already configured its maximum number of LUNs, you must give up a LUN for use as the access volume.

When performing a firmware upgrade from Release 3.x to 4.x on a 3526 controller unit, an intermediate firmware level (3.1.3) is used. This is primarily to allow the controllers to modify the arrangement of the configuration information stored on the hard disks, ready to support the new features offered by the 4.x firmware.

A consequence of this new data arrangement is that it is not possible to move hard drives used on firmware 3.x controllers to firmware 4.x controllers and preserve their configuration. The hard drives can be used, but all logical drive configuration information and data will be lost and you will have to create a completely new configuration.

**Important: Upgrading firmware from 3.x to 4.x**

A utility called SM7migrate performs the RAID controller firmware upgrade. Upgrading directly from 3.x to 4.x without going through the intermediate level 3.1.3 can cause loss of communication with the storage subsystem controllers. To avoid this problem, follow the process documented in the *IBM Netfinity Fibre Channel Storage Manager for Windows NT Installation and Support Guide*, shipped with the product.

---

## 7.4 Controller management

To allow flexibility in managing your Fibre Channel disk subsystems, there are two independent ways of administering the hardware, called *direct management* and *host-agent management*, respectively. The following sections discuss these two approaches and the implications each has for managing the storage subsystem.

### 7.4.1 Direct management

The 3526 and the Netfinity FASt500 RAID Controller each have two Ethernet ports (one connected to each RAID controller within the unit). Using direct management, you control the Fibre Channel storage subsystem directly over a network connected to the controller unit's two Ethernet ports, each of which must be connected to the network. For security reasons you may want to implement a separate LAN segment to access the disk controllers, although the units do offer encrypted password protection for access.

In order to communicate with the controllers, you must have a DHCP server on this network to assign IP addresses and associate them the MAC addresses of the Ethernet ports. You can find the MAC address either printed on a label on the controller itself or in the profile of the controller unit. To

locate the MAC address within the profile log, look for the logical drive section. Under each logical drive you will find a worldwide name that includes the MAC address as illustrated in Figure 123. The logical drive is associated with a specific controller through the array definition that you find in the profile file as well, giving you the corresponding MAC address of each controller.

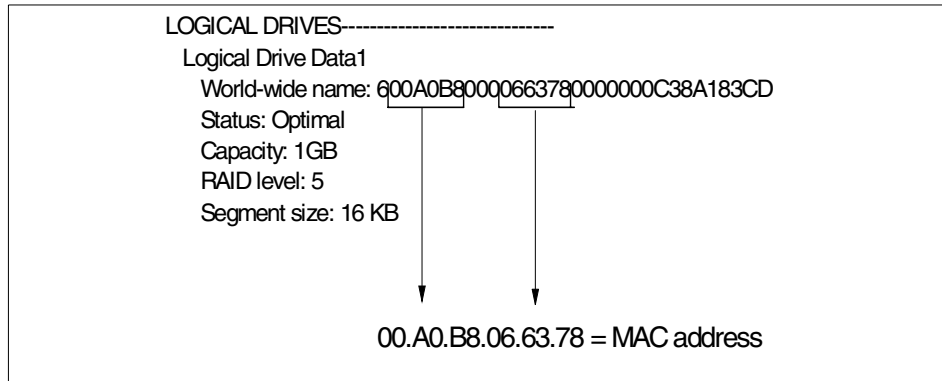


Figure 123. Retrieving the MAC address from the World-wide name in the profile file

Figure 124 shows the basic topology of a directly managed Fibre Channel storage subsystem:

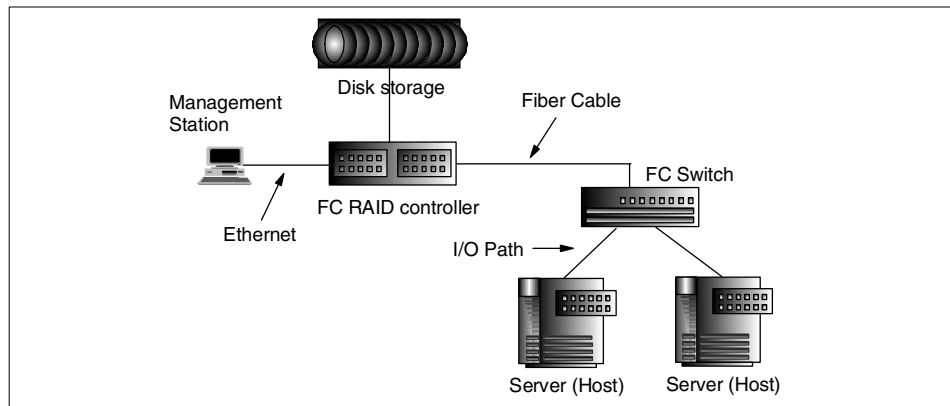


Figure 124. Direct management of the Fibre Channel Storage subsystem

The advantages of managing storage subsystems directly include:

- The Ethernet connections to the controllers enable a management station running Windows NT to manage storage subsystems connected to a host

with an operating system other than those supported by Netfinity Fibre Channel Storage Manager 7.x.

- You do not need to use an Access Volume to communicate with the controllers as you would if you were running the host-agent software. This means that you can configure the maximum number of LUNs supported by the operating system and the host adapter you are using.

The disadvantages of managing storage subsystems directly include:

- It requires two Ethernet cables to connect to both storage subsystem controllers.
- When adding devices, you must specify an IP address or host name for each controller.
- This method requires a number of network preparation tasks (as explained in the *IBM Netfinity Fibre Channel Storage Manager for Windows NT Installation and Support Guide*).

#### **7.4.2 Host-agent management**

Host-agent management utilizes the Fibre Channel cabling between the disk subsystem and an attached host to manage the Fibre Channel controllers. It does not require any additional cabling or setup.

The host system that has a Fibre Channel I/O path to the controller requires SM7agent (the host-agent software) to be installed. Netfinity Fibre Channel Storage Manager may be executed either locally on this host or remotely on a management station connecting to the host through the LAN to which normal clients are attached.

When starting up the Enterprise Management window of the Storage Manager, it allows you to add the host to the management domain. This mode of management is particularly useful when you have multiple controller units in your network, connected to a number of different host systems. The storage manager can discover host agents on the network automatically, or you can manually specify the host connected to the storage system you wish to administer.

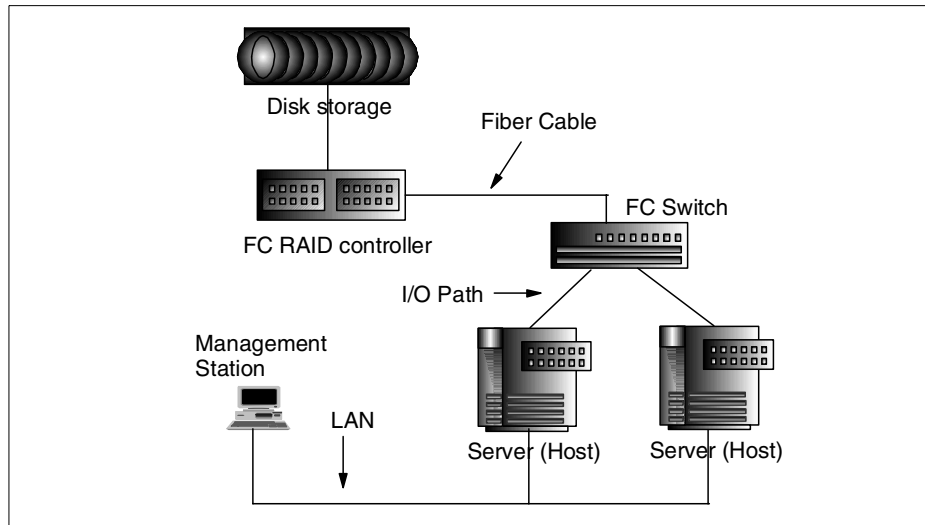


Figure 125. Host-based administration

The SM7client code, which is the name of the storage manager, communicates with the host-agent service, which in turn forwards requests to the controllers. The target controller executes the command and returns a status message to SM7client. The storage manager is essentially a front-end “thin” client to the controller’s “server”.

#### Role of SM7agent

Although the host-agent (SM7agent) does not appear to be needed when administrating the controllers directly, you still have to install this software component on every host connecting to a Fibre Channel Storage system. The SM7agent component also contains the SM7device utility, which is responsible for mapping of operating system device names with the logical drives created on the Fibre Channel controllers.

The implementation of both management methods is described in detail in the *IBM Netfinity Fibre Channel Storage Manager for Windows NT Installation and Support Guide*.

The advantages of managing storage subsystems through the host-agent include:

- You do not have to run Ethernet cables to the controllers.
- You do not need a BOOTP server to connect to the network.
- You do not need to perform many of the preparation tasks required for direct management.
- When adding devices, you only have to specify a host name or IP address for the host rather than the individual controllers in a storage subsystem. Storage subsystems attached to the host are automatically discovered.
- You can add a host to the management domain, and the attached storage subsystems are automatically discovered.

The disadvantages of using the host-agent method include:

- You are limited to configuring one less LUN than the maximum number allowed by the operating system and the host adapter you are using. The host-agent requires a special logical drive called an access volume to communicate with the controllers in the storage subsystem. This access volume uses one of the allowable logical unit numbers (LUNs).
- If you are upgrading controllers from firmware Version 3.x to Version 4.x and your host system has already configured its maximum number of LUNs, you must give up a LUN to be used as an access volume.

---

## **7.5 Managing the controllers with Netfinity Storage Manager**

In addition to providing the basic tools for configuring and allocating disk resources, the IBM Netfinity Fibre Channel Storage Manager allows you to directly manipulate internal controller resources. If you need to update a controller's firmware or make changes to its NVSRAM settings, the storage manager makes this a simple task.

### **7.5.1 Downloading firmware and NVSRAM contents**

In the previous version of the Fibre Channel management software, SYMplicity Storage Manager, to update the firmware and NVSRAM contents required that you download four different files. These were the NVSRAM file and the Bootware, Appware and Fibre Channel firmware files. You also needed a \*.DEF file to ensure that you downloaded the correct set of files. The three firmware files were in the \*.DL format.

This process has been greatly simplified with the new storage manager software. Downloading the NVSRAM and firmware to a controller is now an

option within the Storage Subsystem Management window. All three firmware files have now been packed in a \*.DLP file, which is simply referred to as the controller firmware file.

Keep in mind that the NVSRAM file needs to be downloaded before the firmware file.

The new \*.DLP firmware files cannot be downloaded through the serial interface of the controllers. The new DLP format is a packaged file format which is not understood by the controller interface.

### 7.5.2 Changing NVSRAM settings

A number of settings within the Fibre Channel controllers are manipulated by changing the contents of the NVSRAM. Depending on the precise way you intend to use the controllers, you may need to set the NVSRAM accordingly. Each controller has its own NVSRAM memory space, which has to be configured correctly.

A scripting language is provided to give you a way to modify the NVSRAM settings. The settings are changed by executing custom scripts, using the Enterprise Management window. In the following list, we give a short description of several of the pre-written scripts that are supplied with the product:

networkon.scr	This script enables direct networking to IBM Netfinity FC RAID controllers. When the controller unit is powered up after running this script, the controllers look for a DHCP server on the network to make a request for an IP address to be assigned to them. You would need to run this script if you wish to implement a direct-managed environment.
networkoff.scr	If you are implementing a host-agent managed environment, and wish to disable direct networking of the FC controllers, you would run this script. It prevents the controller from making a DHCP request and, as a result, the boot time of the controllers is reduced.
reseton.scr	Executing this script enables propagated reset for clustered environments. The effect of a propagated reset is very much like a hard reset of the SCSI bus being executed when a host joins or leaves the bus. This setting is required in MSCS or other two-node cluster implementations.

resetoff.scr	To disable the propagated reset function mentioned in the description of the reseton.scr, you would run this script.
softreseton.scr	This script enables soft reset to the Fibre Channel controllers in a multi-node cluster environment. The command fulfills a similar function to the reseton.scr, but the command completes faster. This script would normally be executed in clustering environments with more than two hosts, such as Netfinity Availability Extensions (Cornhusker) and Novell Cluster Services.
softresetoff.scr	To disable the soft reset function mentioned in the description of the softreseton.scr, you would run this script.

For more information about the scripting engine, we refer you to the online help of the IBM Netfinity Fibre Channel Storage Manager for Windows NT.

---

## 7.6 Managing storage

The Subsystem Management window in the IBM Netfinity Fibre Channel Storage Manager application lets you perform all of the common administration tasks required to manage your storage. The menu bar in the window is context-sensitive and whenever an item in the window is selected, all valid actions can also be invoked by right-clicking the object's icon.

To avoid necessary duplication, we do not discuss the more straightforward management tasks, such as creating arrays and logical drives. These procedures are well documented in the product manuals and the IBM Netfinity Fibre Channel Storage Manager online help. In this section we describe the more advanced storage administration tasks.

### **Duplicate controllers?**

Hosts with the host-agent software installed are automatically discovered by the storage management software and appear in the device tree in the Enterprise Management window, along with their attached storage subsystems. A storage subsystem might be duplicated in the device tree if you are managing it through its Ethernet connections and it is attached to a host with the host-agent software installed. In this case, the duplicate storage subsystem icon can be removed from the device tree using the Remove Device option in the Enterprise Management window.



## 7.6.1 Advanced storage administration tasks

A number of useful functions beyond the more usual tasks of creating and allocating arrays and logical drives are available using IBM Netfinity Fibre Channel Storage Manager. Performance tuning and other tasks you may wish to perform once your system is basically installed and running can be undertaken using the facilities available in IBM Netfinity Fibre Channel Storage Manager

### 7.6.1.1 Segment size

When creating logical drives, you can specify the segment size which is the amount of data written to or read from a single hard disk in a logical drive before moving to the next disk in the array. A segment is made up of multiple sectors (blocks of 512 bytes) on the hard disk.

The size of segment you select can have an effect on overall system performance. Similar considerations to those discussed in 5.10.6, "Stripe unit size" on page 153 in regard to SCSI subsystems apply equally well for Fibre Channel subsystems. (For a detailed discussion of performance tuning, see *Tuning Netfinity Servers for Performance*, SG24-5287. Although focused on Windows 2000 and Windows NT, much of the material in the book can be applied to all server operating systems.)

To help you select an appropriate segment size, you are given a choice of File System, Database, or Multimedia in the Expected logical drive usage list box. This is illustrated in Figure 126, where we have selected **File System**:

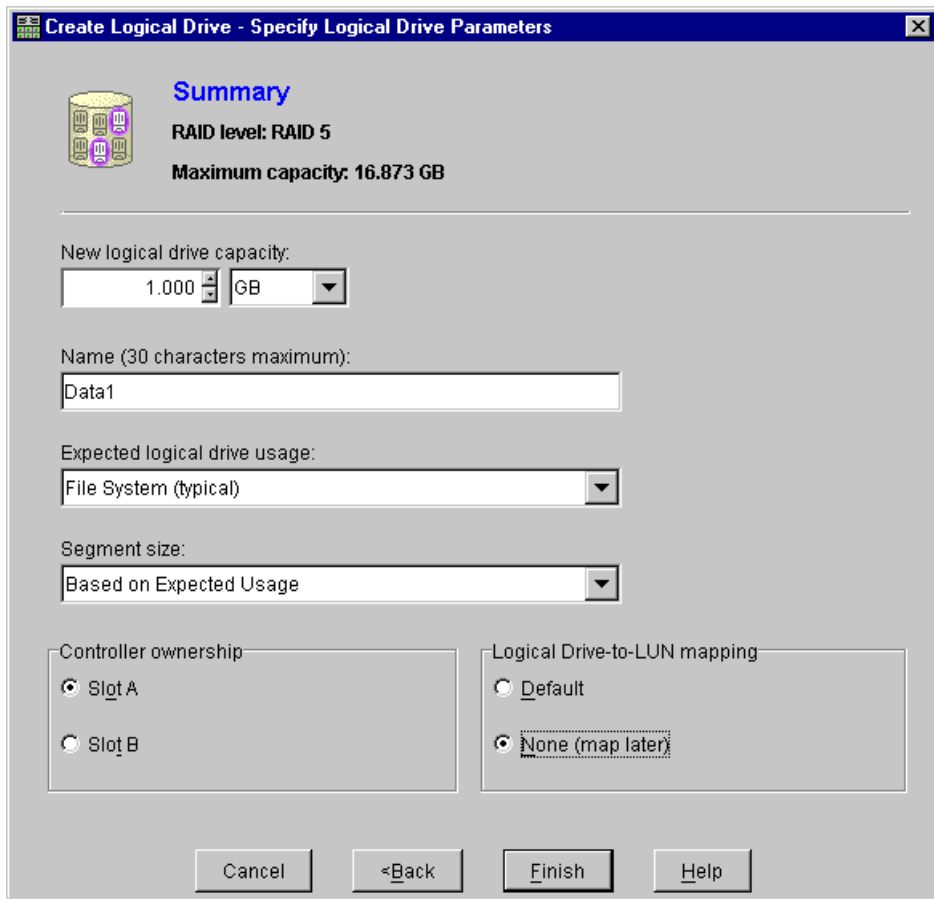


Figure 126. Creating a logical drive and specifying the segment size

For your selection to take effect, you must set the Segment size list box to Based on Expected Usage. Table 17 indicates the actual segment size used for each of the options:

Table 17. Meaning of segment size denominators

Expected logical drive usage	Segment size
File System	16 KB
Database	16 KB
Multimedia	64 KB

If you have, or can gather, detailed information about the typical size of I/O requests made by your server applications, you may prefer to set the segment size directly. You can specify segment sizes of 8, 16, 32, 64, 128 or 256 KB. The segment size is set independently for each logical drive and can be altered while active without compromising data integrity.

#### **Segment size**

The segment size was expressed as a multiple of 512 byte sectors in SYMlicity Storage Manager 6.22. With the introduction of Netfinity Storage Manager and the new 4.x firmware for the 3526 controller, the segment size is now explicitly expressed in KB.

You should also be aware of a possible confusion in terminology. When discussing ServeRAID adapters, segment size is sometime referred to as stripe unit size or simply stripe size.

#### **7.6.1.2 Adding drives without rebooting**

After creating a logical drive using the storage manager, Windows NT 4.0 Server normally needs to be rebooted to be able to access and configure the drive in Disk Administrator. Rebooting can now be avoided by using the new hot\_add utility, which adds new logical drives to the operating system dynamically. The hot\_add utility comes with the storage manager and can be run from a command prompt. To remove logical drives from the operating system once they have been deleted by the storage manager, however, you still need to reboot Windows NT 4.0.

Almost as soon as you have created a logical drive, the Immediate Availability Feature provides access to it. The first and last 10 MB of the drive are low-level formatted and, shortly after this, the drive can be utilized by the operating system and data written to it. A background process continues to perform a low-level format of the remainder of the drive with, perhaps, some minor impact on performance until the format is complete. Low-level formatting ensures the integrity of the drive, since media defects can be found and isolated during this process.

#### **7.6.1.3 Media scan**

With the new firmware (Release 4.0) and IBM Netfinity Fibre Channel Storage Manager, a feature called media scan has been introduced. This feature provides a mechanism for detecting drive media defects before they are found during normal operation, helping to prevent possible data loss.

As an example: if a disk in a RAID-5 array has an undetected media defect and one of the other physical drives in the array fails, any attempt to read data from the segment that contains the media defect, or the corresponding segment on the failed drive, will fail as the data cannot be rebuilt.

This feature is enabled at the storage subsystem level as shown in Figure 127. It must, however, be enabled for specific logical drives as indicated by the informational message in the dialog window:

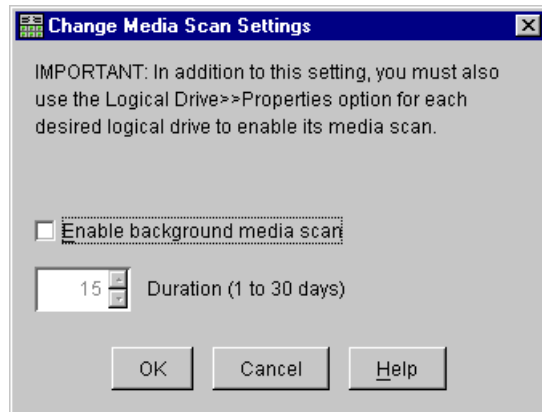


Figure 127. Enabling the media scan feature

#### **Data scrubbing and media scan**

Data scrubbing is another term that is sometimes used for the media scan feature in other RAID disk subsystems.

The duration parameter in Figure 127 lets you specify the total number of days you want to allow for one complete cycle of media scanning. The longer you specify, the lower the priority of the task will be, which, in turn, means higher system I/O performance.

Examining the properties of a single logical drive (Figure 128), you can see how to enable the media scan for a specific logical drive. If you select the redundancy check/repair option, the process will recalculate parity or mirror information as required by the RAID level being used by the logical drive. Any recovered or unrecovered errors encountered during this process, will be logged in the Major Event Log with the message A Parity/Data Mismatch was detected.

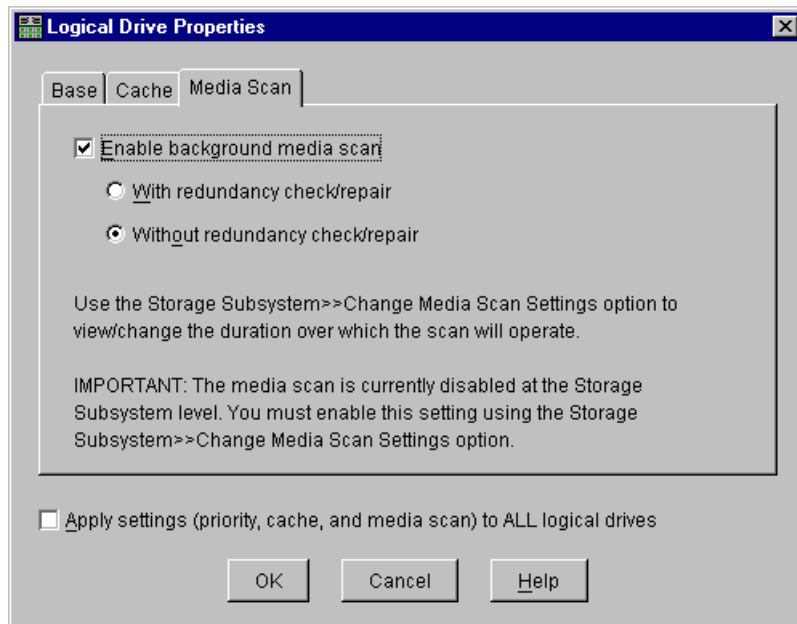


Figure 128. Setting the media scan properties of a logical drive

The rate at which the media scan or a segment size change is performed can be modified for a specific logical drive by setting the relative priority given to drive modification processes. This is set through the Properties dialog window of the specific logical drive, as you can see in Figure 129:

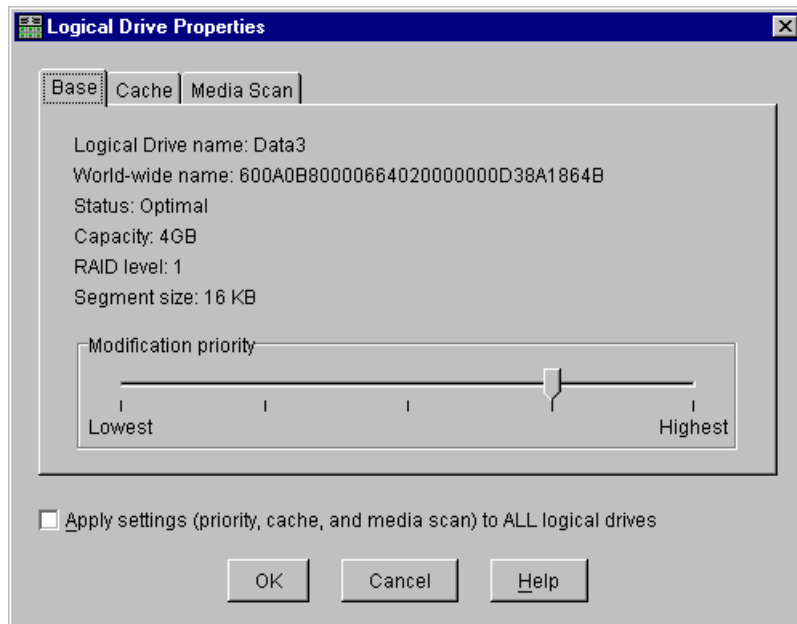


Figure 129. Modification priority for a logical drive

#### 7.6.1.4 Fibre Channel disk cache settings

Another important feature of the new Netfinity FAStT500 RAID Controller is the cache manageability.

Enabling read and write caching makes the biggest impact on performance when the server workload consists of predominantly sequential I/O accesses. Even for servers with mainly random access I/O, however, it can significantly improve throughput because control may return to the application much sooner, especially when write caching is enabled.

The write cache with mirroring feature mirrors the two independent cache memories of both controllers, providing an extra level of redundancy in case of a single controller failure. As the controller unit comes with built-in batteries for cache protection, we do not recommend that you select **Enable write caching without battery support**. A power supply failure or power outage could mean that the data in the cache is lost and could not be written to the disks. The dialog window for these features is shown in Figure 130:

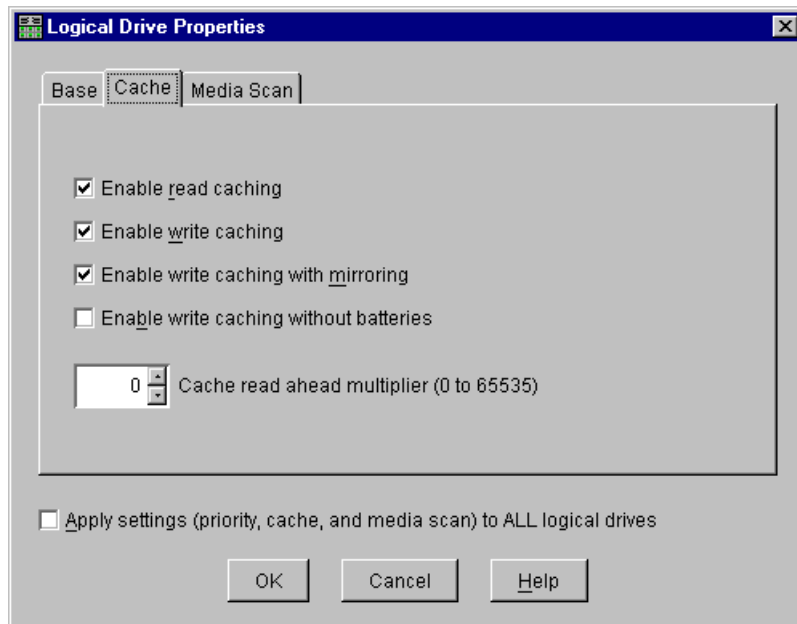


Figure 130. Setting the cache properties of a logical drive

This dialog also allows you to specify the cache read-ahead multiplier. You should use this parameter to optimize the performance of the controllers in your specific environment. Increasing this parameter boosts performance when your data tends to be accessed more sequentially rather than randomly.

It makes most sense to use this parameter in conjunction with the cache flush settings of the controller, shown in Figure 131 on page 218. The cache flush settings give you the ability to specify at what level of cache utilization (Start flushing) data will be written to disk and when it will stop doing so (Stop flushing).

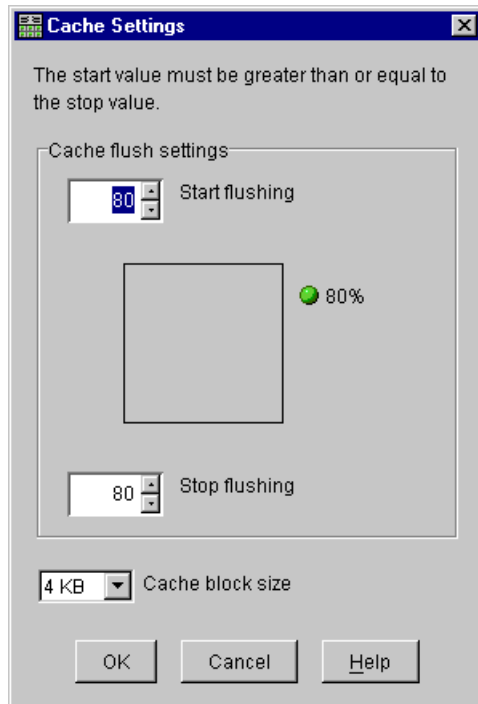


Figure 131. Cache settings for a controller

#### 7.6.1.5 Scripting

As mentioned previously, the Fibre Channel controller firmware incorporates a built-in scripting engine for automating large, time-consuming, or tedious tasks. The Enterprise Management window provides a script editor and an execution interface. Scripts can also be validated before you execute them on the controller. The commands available allow you to perform the following tasks:

- Controller cache configuration.
- Logical drive and array configuration.
- Drive configuration.
- Logical drive ownership.
- Change controller mode.
- Download NVSRAM and firmware to controllers.
- Dump controller unit profile information.
- Battery management.



- Housekeeping tasks, such as reset configuration to default, labeling, health checking, set time-of-day, clear event log and set the media scan rate.

In the current scripting language implementation you cannot program iteration, conditional commands, functions, variables, or parallel threads.

The online help of the Enterprise Management window provides information and syntax for all of the available scripting commands and parameters. It also provides guidance on usage of the script editor.

### **7.6.2 Storage partitioning**

Storage partitioning allows you to share your storage subsystem among multiple hosts. A storage partition is a logical entity used by the controller firmware to allow hosts to share or use a portion of the available storage within the storage subsystem. A storage partition is created when you define a collection of hosts, called a host group, or a single host and their associated host ports and then define one or more logical drive-to-LUN mappings. The mappings allow you to define which host groups or hosts will have access to particular logical drives in your storage subsystem. A partition can contain one or more logical drives.

Host ports are physically represented by the PCI Fibre Channel adapter. Each time an adapter joins a Fibre Channel fabric, it is assigned a worldwide name, which is associated with a specific host as you will see below.

We will describe storage partitioning by going through an example.

The first step in creating storage partitions is to define a topology. This entails the assignment of specific host ports to hosts that in turn belong to a host group. As you can see in Figure 132 on page 220, we have assigned a single host port with the name “zorro hba” to the host “Host zorro”. Host “Host zorro” belongs to a host group “Zorro Group” which in turn is under the default host group:

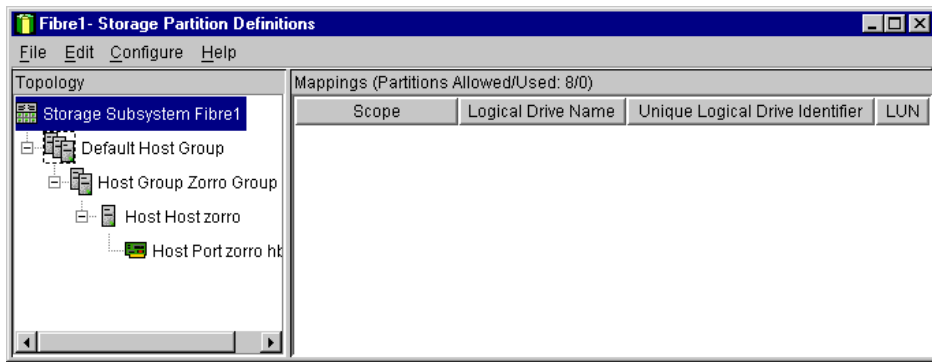


Figure 132. Storage partition with one host port

Default Host Group cannot be deleted and initially contains all discovered host ports. In addition, when new host ports join the fabric, they will be assigned to the default host group. They can later be assigned to other host groups or hosts or left in the default host group. You can leave a discovered host port logically undefined in the default host group if it is used to access tape drives or other devices that are not part of the storage subsystem.

Expanding our example, we now define a new host group named "MagicBox" containing the host "Server02" as in Figure 133.

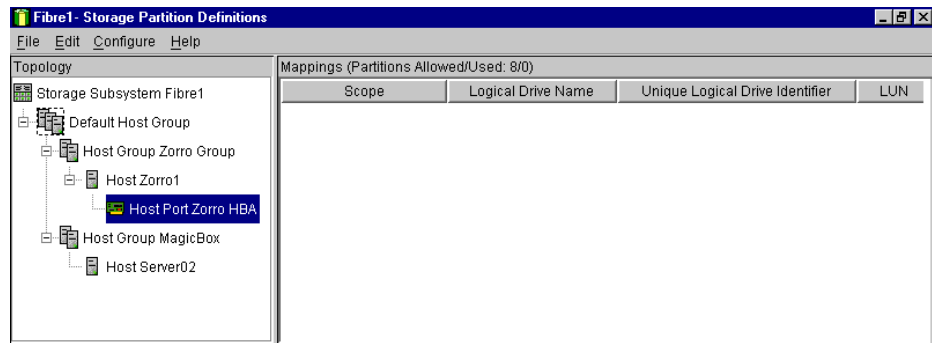


Figure 133. Adding a new host group and host

The next step in partitioning storage lies in assigning specific logical drives to hosts or host groups that can access them. This process is called logical drive-to-LUN mapping. When creating a logical drive as in Figure 134, you can select either the default mapping or defer it until later.

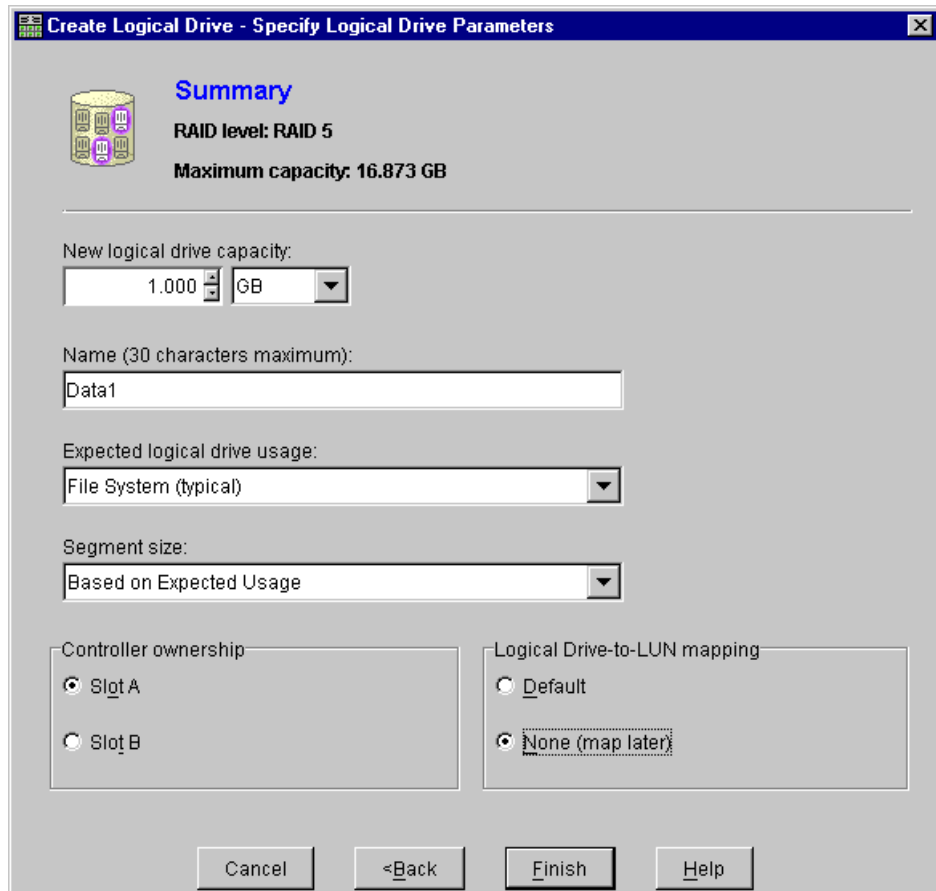


Figure 134. Creating a logical drive

In choosing to map the logical drives to the default host groups, this logical drive can be accessed by any host or host group that lies under the default host group in the topology view. When choosing to defer it later, the logical drive cannot be accessed by any host initially.

For our example, we have created four logical drives, which are initially assigned to the default host group, meaning they are all accessible by both defined host groups and hosts. You can regard the default host group as a parent container that presents all of its logical drives to its children.

Figure 135 on page 222 shows the dialog window where you specify the association between the logical drive and the LUN in the operating system. We mapped logical drive “Data4” to LUN 0 on a Windows NT 4.0 host.

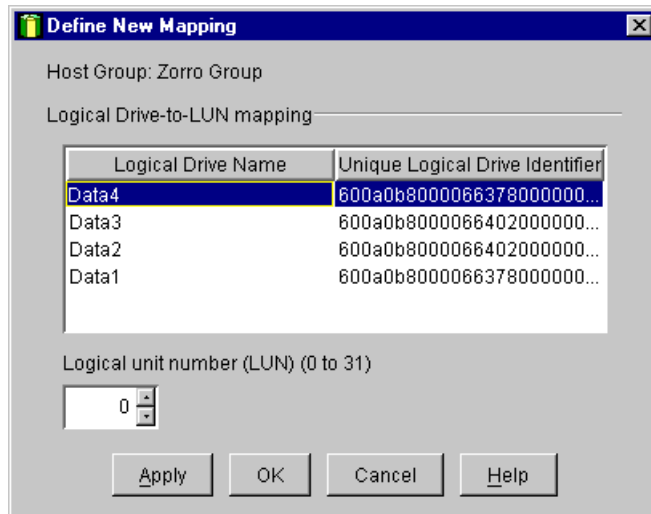


Figure 135. Mapping a logical drive to a LUN

Instead of leaving all logical drives assigned to the default host group, we assigned them to the host group “Zorro Group”, which now has two hosts, as you can see in Figure 136. Both hosts in this group can access all four logical drives. This illustrates the use of a partition as it would be used in a two-host clustering scenario. One of the logical drives can act as a quorum drive in Microsoft Cluster Server and the remaining three drives could be data drives. The lock managing function, which regulates which host gets access to which logical drive at which point in time, is the responsibility of MSCS. Note the change in the hierarchy of host groups when performing this step, and compare Figure 133 on page 220 with Figure 136.

You can see that Host Group "Zorro Group" is now no longer under the Default Host Group but is on its own, meaning it is also directly under the Storage Subsystem object. All currently defined storage partitions in "Zorro Group" can no longer be accessed from new hosts in the default host. They can only be accessed from hosts in "Zorro Group". Both "Zorro Group" and the Default Group are now independent parent containers. The Host Group "MagicBox" was deleted and is not relevant to this example.

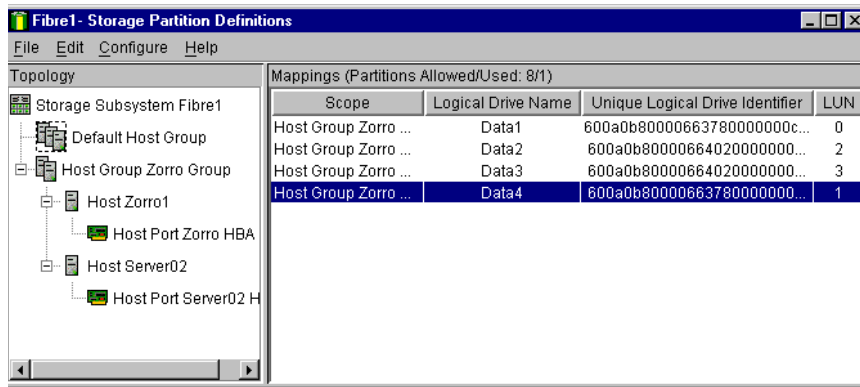


Figure 136. All logical drives are assigned to a host group

We now change the assignment and distribute the logical drives between the two hosts as shown in Figure 137 on page 224. This represents the scenario of two separate file servers that are in the same storage partition. Each host has its specifically assigned logical drives, which are only accessible by itself. In the figure, you can see that host “Zorro1” owns logical drives “Data1” and “Data2” (top) and host “Server02” owns logical drives “Data3” and “Data4”. No logical drives are assigned to the host groups.

You can create a mix of both scenarios by assigning shared logical drives to the corresponding host group and other logical drives to specific hosts for granting exclusive access.

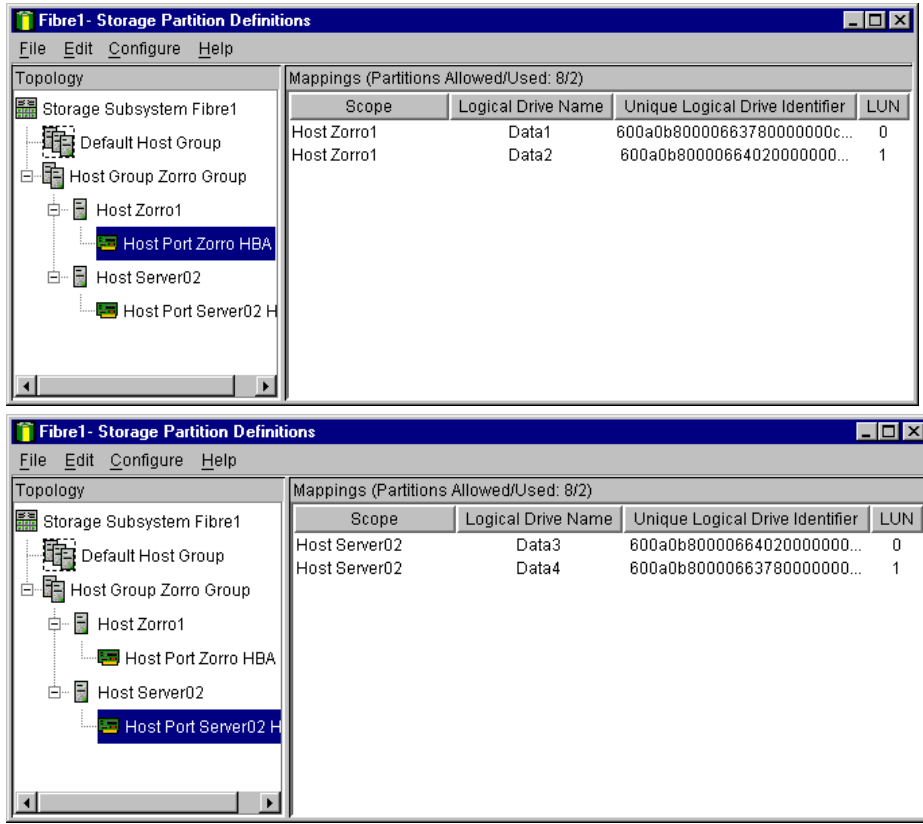


Figure 137. Assigning logical drives to servers







---

## Chapter 8. Introduction to serial storage architecture (SSA)

This chapter gives you an introduction to SSA technology. In it, we explain the underlying protocol and technology, and guide you from there to the currently available SSA adapters and enclosures available from IBM.

We also cover briefly the attributes of older SSA hardware that is still in use and which customers may wish to migrate to the newer versions.

---

### 8.1 The SSA protocol: IBM's implementation

Serial storage architecture (SSA) is a high-performance serial interface designed to connect I/O devices to host adapters. It is a two-way signal connection (transmit and receive), providing full-duplex communication between host and devices over a non-arbitrated loop.

Some people have the perception that SSA is a proprietary IBM interface. Yes, the first SSA interface was developed by IBM, but in 1991 this technology was made available as one of the serial storage interface options for the SCSI-III standard.

SSA started out under the control of the Serial Storage Architecture - User Industry Group (SSA-UIG), which was made up of many leading computer companies. Since 1994, however, SSA standardization and documentation has been under the control of the ANSI X3T10.1 committee, which is a daughter committee to the ANSI SCSI-III X3T10.

This has made the SSA interface an international standard, open to any interested company.

Understanding the complete SSA solution is made easier by having knowledge of the topology of SSA and the inner workings of its protocol.

Before delving into these areas, however, we define some important SSA terms in Table 18:

Table 18. Definition of important SSA terms

Term	Definition
Port	A port is the physical connector on a device, such as an adapter or a disk. SSA adapters available for Intel-based servers are quad-ported, which means that the adapter has four physical SSA connectors. A port consists of one transmit path and one receive path.
Adapter	This is the physical SSA adapter that is located in the PCI bus of the host server.
Initiator	An initiator is a node that issues commands to other nodes in the loop. Adapters are examples of initiators.
Target	A target is a node that receives and responds to commands issued by initiators. Disk drives are examples of target nodes.
Node	In an SSA network, a node is an element that implements one or more SSA ports, through which it may send and receive frames containing dateless node will usually implement some function (such as data storage, as for a disk drive) that may originate or receive frames. Frames received by the node but for which it is not the target node are passed on around the loop. Disk drives and adapters are examples of SSA nodes.
Loop	An SSA network in which a closed path exists. That is, a path around which a message may travel and end up at the node from which it originated. Loops are made up of nodes each containing two ports. Netfinity SSA adapters have four ports and can manage two independent loops.
SIC	A serial interface chip (SIC) controls a loop and resides on a node. Since the Netfinity SSA adapters have two loops, there are two SICs on each adapter. Each SSA disk also has a SIC.
UID	Unique Identifier. A unique factory-assigned number given to every SSA node. A UID consists of 15 characters. Also commonly used is the short UID, which consists of eight characters from the full UID.
JBOD	<i>Just a Bunch Of Disks</i> . This term is used for a non-RAID disk resource, that is a collection of single SSA disks.

### 8.1.1 SSA topology

There are various configurations that can be implemented using SSA, including loop, string and switch topologies. Netfinity disk subsystems do not currently implement the switch topology.

The topology offering the best redundancy and performance and implemented by IBM for Netfinity servers is the SSA loop. This is illustrated in Figure 138:

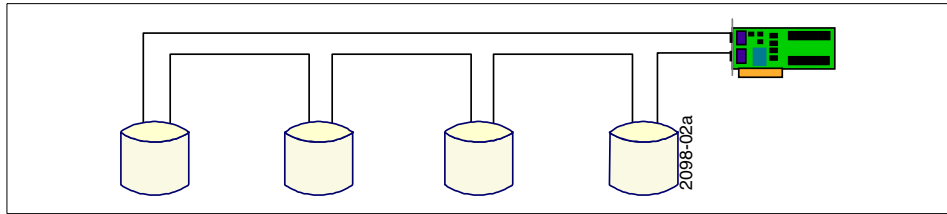


Figure 138. SSA loop topology

A loop contains only dual-ported nodes, and SSA can support a maximum of 127 nodes in one loop (this is in many cases further restricted by the capabilities of the adapter itself). If a break occurs in the loop, such as a cable being damaged, for example, each device in the loop adjusts its routing methods, under direction from the master initiator, so that frames are automatically rerouted to circumvent the break. This allows devices to be removed from or added to the loop while the subsystem continues to operate without interruption.

The loop topology also allows you to achieve the maximum bandwidth from your adapter. Today's SSA adapters for Intel-based servers provide simultaneous full-duplex 40 MBps connections. That means data can be both sent and received at 40 MBps at the same time on each port, or an aggregate data rate of 80 MBps per port, transferred over a single cable.

Because the adapter sits on a loop, two ports are used, which means communication may take place using both ports simultaneously. This results in a total maximum transfer rate of 160 MBps per adapter, controlled by a single SIC.

**Device speed**

Confusion can arise due to the different ways of discussing speed. In this document we will refer to 40 MBps devices when we mean a device that is capable of transferring 40 MBps over a single cable, in a single direction, to provide the aggregate speed of 160 MBps as described in the preceding paragraph.

Figure 139 illustrates the maximum data transfer rate per SIC. The IBM SSA adapters for Netfinity each have two SICs, giving two independent loops.

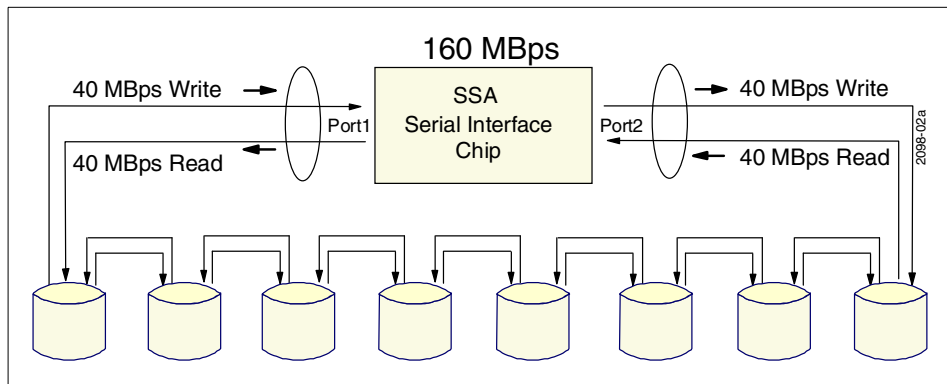


Figure 139. 160 MBps aggregate transfer rate

If you remove one end of the loop you are creating a string topology. It only provides half the maximum data throughput to the adapter in comparison with a loop, and does not offer a loop's redundancy features. Normally you would not implement this topology; we mention it only to assure you that a "broken" loop still works and poses no threat to your data.

#### 8.1.1.1 SSA nodes

Devices in an SSA network are called nodes. From a communication point of view, a node can be either an *initiator* or a *target*. An initiator issues commands, and targets respond with data and status. Typically, adapters provide initiator nodes and disk drives are target nodes. Each SSA node is given a unique address (UID) at the time of its manufacture, which allows the initiators in the loop to determine what SSA nodes are connected in a specific loop.

The SSA architecture allows more than one initiator to be present in a loop. In that case, commands and data from multiple initiators can be directed to the same or different targets and intermixed freely. This has an obvious application in clustering environments.

In an SSA loop, one initiator must be configured as a master node. By default this is the initiator with the highest UID. If a new initiator is added to the network with a higher UID than those currently present, it takes over the master responsibilities for that loop. Similarly, if a master initiator is removed from the loop, the initiator with the next highest UID takes over the role of master node. This handover of responsibility occurs automatically, requiring no user intervention.

The SSA specification supports the use of a mixture of 40 MBps and slower 20 MBps devices in the same loop. Any two 40 MBps devices that are installed next to each other in a loop can communicate at an aggregate maximum throughput of 80 MBps. When a 40 MBps SSA device is installed next to a 20 MBps SSA device, data will be transferred between the two devices at the slower speed.

### 8.1.2 The SSA frame

The basic unit of data transferred between SSA nodes is a *frame* (see Figure 140), which contains one character (10 bits) describing the frame type, four characters of CRC, up to six characters of addressing, and up to 128 characters of data. The data block contains either an SSA message or the user data.

SSA was specifically designed with storage solutions in mind. Because of this, much attention has been paid to the way in which SSA handles fault isolation and performs error recovery. When a fault is detected, error recovery occurs at the lowest level of the protocol, on a frame-by-frame basis. SSA uses cyclic redundancy checks (CRC) to protect its data. The CRC covers the complete frame content, that is, data, address, and control information. The value in this is that the SSA hardware and microcode handle error recovery procedures automatically between neighboring ports, isolating any problem down to a specific frame.

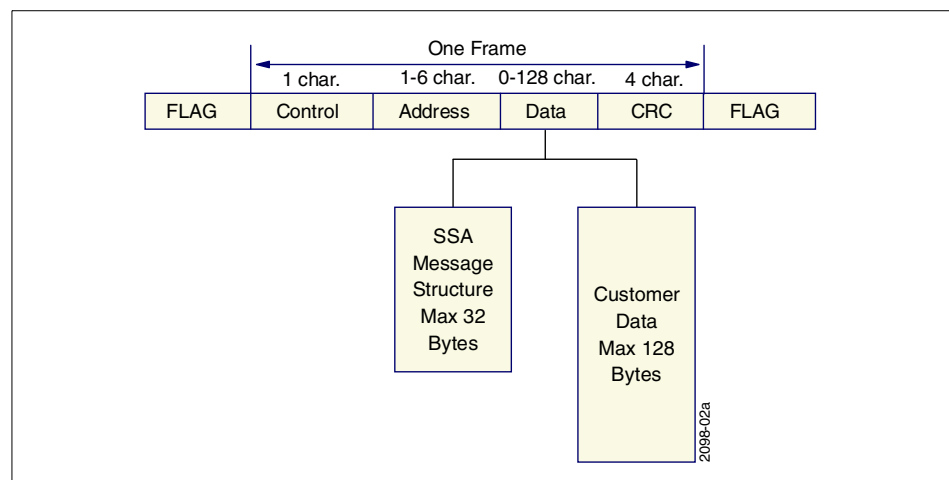


Figure 140. The SSA frame

Error recording in an SSA network is coordinated by the *master initiator* or node, to which we referred earlier. The master initiator is responsible for

receiving notification of, and coordinating the recovery from, high-severity errors and conditions. These include an invalid message being received by a node, and excessive numbers of low-severity errors. Lower-severity errors, such as noise corrupting a transmission between two nodes, are handled at the node level.

### 8.1.3 SSA performance

Two techniques of the SSA architecture provide a significant enhancement for the performance in SSA networks:

- Cut-through routing
- Spatial reuse

We describe these in more detail in the following sections.

#### 8.1.3.1 Cut-through routing

It might appear that, with all the error checks and flow control between links, a frame that has to traverse several nodes as it travels from its source to its destination would have a very sluggish journey, with it being held up at each node while its credentials were checked. However, the expected error rate across SSA links is so low that such delays can be avoided by a method known as *cut-through routing*.

Cut-through routing (sometimes called worm-hole routing) allows a node to forward a frame character-by-character as it is received; it does not have to wait to confirm that the frame passes its CRC check. Using this method, the delay can be as little as 5-10 characters or 0.5 microseconds at 20 MB per second, and half this figure at 40 MBps.

If an error is detected on an inbound frame after a router has already started to forward the frame, then it sends an ABORT character (followed by a FLAG) to the receiver. The receiver will send an ABORT if it, too, has already begun forwarding. An ABORT character tells a receiver to discard the frame in which it occurs. In this situation, error recovery processes take over to ensure that the data is correctly received or that the host is notified of the error.

#### 8.1.3.2 Spatial reuse

One of the characteristics that distinguishes an SSA loop (where each link between nodes is a separate connection) from a bus (where each node connects to the same piece of wire) is that an SSA loop allows the possibility of *spatial reuse*. This is the technique whereby links that are not involved in a particular data transaction are available for use in another transaction.

A data transfer in an SSA loop only involves the source, target, and the nodes between them. Other nodes may concurrently be involved in other transactions. This is not possible on a bus because devices not involved in a transaction are not allowed access to the bus.

Because of buffering that happens on the frame level, data can be travelling on all links at once. Each link essentially behaves as an independent point-to-point path.

An example of spatial reuse is shown in Figure 141. The single host adapter can transmit data X to or from drives A, B or C at the same time as it is transmitting data Y to or from drives D, E or F.

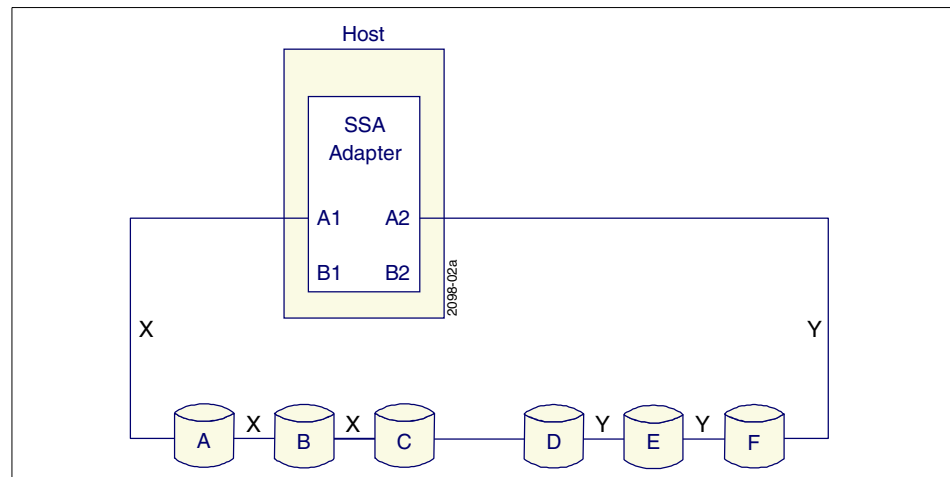


Figure 141. SSA spatial reuse

SSA determines which route to use to transfer data between two nodes based on the hop-count, that is, it uses the path that contains the smallest number of intervening nodes between source and target.

Of the three disk technologies considered in this book, SSA is the only one that allows aggregate data transfer rates that can be higher than the transfer rate of the physical medium. This is achieved because of spatial reuse.

You can find out more about the SSA standard by visiting the X3T10 committee's Web site at:

<http://www.t10.org>.

---

## 8.2 Netfinity SSA hardware

This section examines the principal hardware items used to implement an SSA disk subsystem: cables, adapters, and the disk enclosures and disks.

### 8.2.1 SSA cabling

SSA introduces a whole new cabling technology. Unlike fragile and difficult-to-manage SCSI cabling, SSA copper cabling (see Figure 142) is simpler in its structure and far more robust. SSA uses shielded cables with two differential pairs of wires for the signal transport: one pair for incoming (received) signals and one for outgoing (transmitted) signals. The external connectors are 9-way micro-mini D-connectors with two integrated screws allowing the connector to be securely attached.



*Figure 142. SSA copper cable and connector*

SSA cables come in different lengths. Copper cables may be up to 25 m in length but it is good practice to use cabling of the appropriate length for your installation. This makes cable management easier and minimizes the potential for electrical interference.



There are two types of copper cables used to create SSA loops:

- Standard SSA cables for up to 20 MBps (40 MBps full duplex) data transfer rate. These cables are black in color.
- Advanced SSA cables for up to 40 MBps (80 MBps full duplex) data transfer rate per connection. These cables are blue in color.
- You can use both black and blue cables in the same loop, but bear in mind that lower-speed black cables can cause performance degradation between two 40 MBps devices.

Table 19 summarizes the currently available SSA copper cables:

*Table 19. Feature codes of SSA copper cables*

Length	Feature Code
1.0 m (3.3 ft)	8801
2.5 m (8.2 ft)	8802
5.0 m (16.4 f)	8805
10.0 m (32.8 ft)	8810
25.0 m (82.0 ft)	8825

#### **8.2.1.1 Fiber optic cables**

As we have stated, the SSA interface allows nodes to be separated by up to 25 meters using copper cables. You can achieve longer distances, from 2.4 km to 10 km, with fiber optic extenders. The fiber optic cables can be connected to SSA adapters or to a storage enclosure such as the 7133.

Figure 143 shows the fiber optic extender, which is sold in pairs.

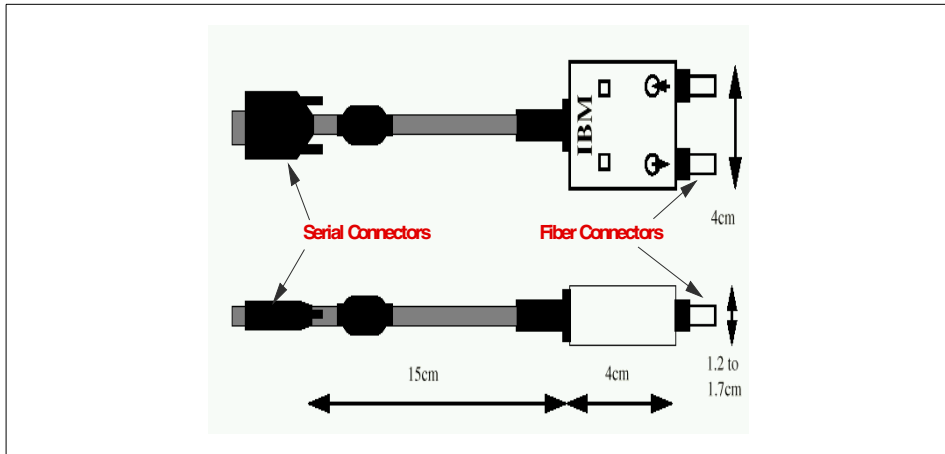


Figure 143. Fiber optic extender (side and top view)

The current generation of fiber optic extenders (Feature Code 8851) support single-mode fiber cables whereas the older generation (Feature Code 5500) only supported multi-mode cabling. As a reference we summarize the different cabling scenarios in Table 20:

Table 20. Compatibility overview between old (5500) and new (8851) fiber optic extenders

Characteristic		SSA fiber optic extender	SSA advanced fiber optic extender
Feature code		5500	8851
Link speed	20 MBps	Yes	Yes
	40 MBps	Not supported	Yes (only with advanced SerialRAID/X adapter FC6225 & 7133-D40/T40)
Fiber cable type	Multi-mode	Either 50 or 62.5 microns	Either 50 or 62.5 microns (Mode Conditioning Patch Cords FC8852 or FC8853 must be installed)
	Single-mode	Not supported	9 microns

Characteristic		SSA fiber optic extender	SSA advanced fiber optic extender
Maximum distance for fiber cable type	Multi-mode 500 MHz	Up to 2.4 km	Up to 2.4 km
	Multi-mode 800 MHz	Up to 2.4 km	Up to 3 km with FC6225, up to 2.4 km with all other adapter types
	Single-mode	Not supported	Up to 10 km (only with advanced SerialRAID/X adapter FC6225 & 7133-D40/T40)
Supported storage enclosures		All models (010/500, 020/600, D40/T40)	All models (010/500, 020/600, D40/T40)
Supported adapters		All	All

The advanced SSA multi-mode conditioners (Feature Code 8852 for 50 microns and Feature Code 8853 for 62.5 microns cables) are available to improve signal quality. In environments where old and new generations of Fibre Optic Extenders are used, the following rules should be observed:

- FC5500 is not supported for use between two 40 MBps capable devices (7133 D40/T40 and FC6225 adapters), since this would cause the link speed to drop to 20 MBps and also log link-speed errors.
- You are not allowed to have FC 5500 at one end of a link and FC 8851 at the other end. You can have FC 5500 and FC 8851 in the same loop, however, as long as they are connected between devices that support them. (It is also possible to have FC5500 on the A-loop of an SSA adapter and FC8851 on the B-loop of the same adapter.)
- When FC8851 is used on multi-mode fibre cables, then FC 8852 or FC 8853 (depending on the diameter of the cable) must be installed.
- It is recommended that the loop be “symmetrical” in terms of disk drives on both sides of your loop. It is also suggested that long lengths of fiber should also be balanced as well. If you consider the extreme of 10 km of fiber then this will introduce 100 microseconds of delay and bring the maximum data rate down to around 1 MBps from 40 MBps. If on the other side of loop you have a copper connection, then you will have the full 40 MBps and hardly any delay.

## 8.2.2 How to identify SSA adapters and their features

At the time of writing, the advanced SerialRAID/X SSA adapter is the only available SSA adapter for Netfinity Servers. However, for reference, we provide information about older adapters because much confusion has been caused by different SSA adapters.

There are in total four SSA adapters that are or have been supported in Netfinity servers:

- The IBM SSA RAID Adapter (1MB and 512 KB versions, see below)
- The IBM SSA RAID Cluster Adapter
- The IBM SerialRAID Adapter
- The IBM Advanced SerialRAID/X Adapter

There are two versions of the IBM SSA RAID Adapter: one fitted with 512 KB SRAM and the other with 1 MB SRAM. This difference is important because the adapters utilize different firmware.

It is vital to know which adapter you are dealing with, particularly when technical support is required. Since it is not easy to distinguish the adapters from each other by looking at them, here are two ways to determine the type of adapter:

- Determine the loadable microcode level of the adapter by using the RSM utility.
- Identify the field replaceable unit (FRU) number of the adapter.

**Note:** The microcode level is unique and remains unchanged, regardless of the actual loaded microcode version.

Table 21 lists the SSA adapters, their microcode load level, option number, and FRU numbers.

*Table 21. Properties of SSA adapters*

<b>Adapter</b>	<b>Microcode load level</b>	<b>Option number</b>	<b>FRU number</b>
SSA RAID Adapter - 512 KB	LL03	32H3811	32H1607 32H1610 32H3825 32H1611 97H0419 32H1612
SSA RAID Adapter - 1 MB	LL11	32H3811	32H1613 32H1614 02L7668 25L5806 09L2066
SSA RAID Cluster Adapter	LL10	96H9835	96H9848 02L7671 25L5809 09L2070
SerialRAID Adapter	LL04	09L2084	09L5544
Advanced SerialRAID/X Adapter	LL05	09L2123	09L2124

Please remember that only the IBM Advanced SerialRAID/X Adapter is now available for purchase.

The two variants of the SSA RAID adapter can also physically be distinguished from each other, as you can see in Figure 144:

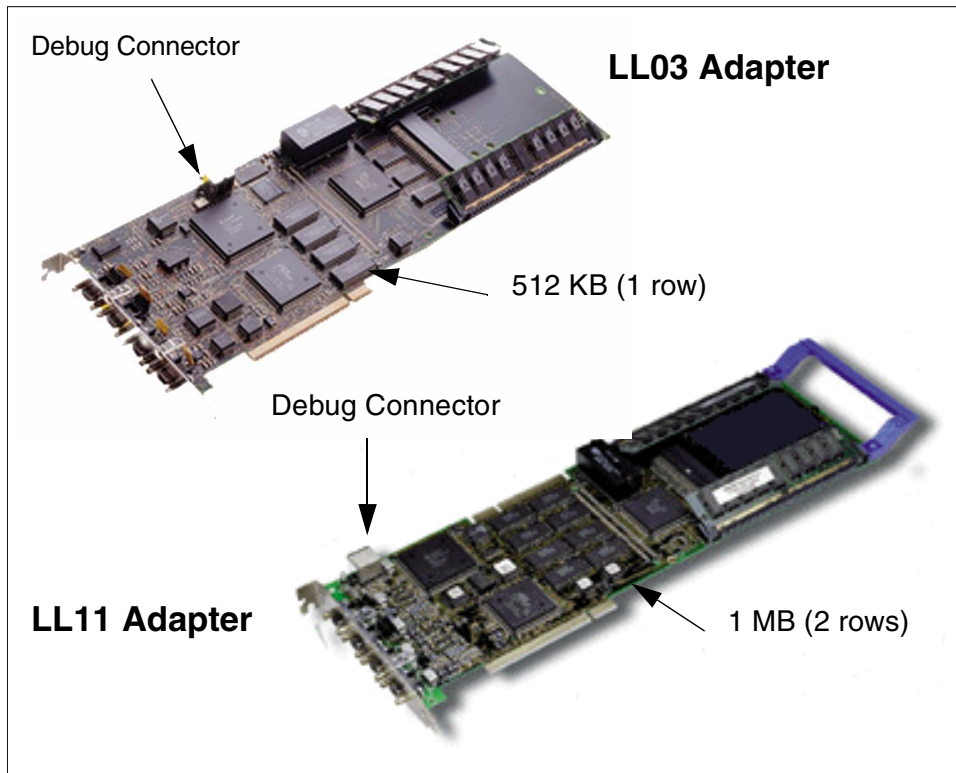


Figure 144. The SSA RAID Adapters

A summary of specifications and features supported by these adapters is provided in Table 22:

Table 22. Feature comparison of SSA adapters

Feature	SSA RAID LL03/LL11	SSA RAID Cluster LL10	SerialRAID LL04	SerialRAID/X LL05
Adapter bandwidth (base/aggregate)	20/80 MBps	20/80 MBps	20/80 MBps	40/160 MBps
Max adapters per server	3	3	3	6
Configurations	JBOD and RAID-0, 1, 5	JBOD and RAID-1	JBOD and RAID-5	JBOD and RAID-0, 1, 10, 5
Two-way cluster	No	Yes	Yes	Yes

Feature	SSA RAID LL03/LL11	SSA RAID Cluster LL10	SerialRAID LL04	SerialRAID/X LL05
Bootable drives	JBOD, 0, 1, 5	JBOD (local)	JBOD, 5 (both local)	JBOD, 0, 1, 10, 5
Array span loops	Yes	Yes	No	No
Hot spares	Global to Adapter	Global to Adapter	Global to Loop	Global to Loop
Operating systems	NT, OS/2, NetWare (4.2 and 5.0)	NT	NT	NT, NetWare (4.2 and 5.0)
Support for 7133-D40/T40	No	No	Yes	Yes

If you need information about RAID levels, these have been covered in the SCSI section of this book (see 4.2.2, “RAID levels supported by ServeRAID adapters” on page 48). RAID levels 0 through 5 are standard terms, used universally within the disk controller manufacturer community, and are independent of the underlying disk technology. Indeed, RAID techniques have been used for other storage media, such as tape.

Other RAID levels, such as RAID-10, can mean different things depending on the manufacturer, and even the particular adapter from a single manufacturer, and the SerialRAID/X adapter is a case in point.

#### **RAID-10**

In the discussion of RAID-10 for ServeRAID (4.2.2.8, “RAID-10” on page 55), we showed that this comprises data striped (RAID-0) across a number of RAID-1 arrays. In contrast, RAID-10 for the advanced SerialRAID/X adapter consists of mirroring (RAID-1) a striped (RAID-0) array.

This distinction, while somewhat subtle, is important when you need to understand how your data is being distributed across the disks in the array.

### 8.2.3 The IBM Advanced SerialRAID/X Adapter

The IBM Advanced SerialRAID/X Adapter (shown in Figure 145 on page 243) has the following features:

- The adapter supports high-performance serial disk storage with an aggregate 160 MBps bandwidth.
- One adapter supports up to two loops with up to 48 disks in each loop, that is, a maximum of 96 disks.
- The adapter supports a number of configurations: non-RAID (JBOD), and RAID levels 0, 1, 10 and 5.
- All RAID levels support a maximum of 16 disks per array (including any mirror or parity disks), except for RAID-1, which consists of two disks.
- All RAID levels except RAID-0 can be used in two-way (cluster) configurations. RAID-0 is supported only in one-way configurations.
- The adapter enables disk sharing between two servers and provides host failover support when used with MSCS or Novell Cluster Services.
- Up to 32 RAID-5 arrays can be shared per adapter pair.
- The adapter provides excellent performance with up to 16,000 I/Os-per-second throughput for clustered server non-RAID environments (70:30 read/write workload).
- Disks can be up to 10 km away from the host adapter when used with the IBM Advanced SSA Optical Extender and the IBM 7133 Serial Disk System Advanced Models.
- The adapter offers fast-write cache (32 MB) and read cache (64 MB) for dual or single-host attachment. This comes as standard.
- The 7133-D40/T40 storage enclosure is supported at 160 MBps. All other enclosures are supported at 80 MBps. Mixed-speed operation is possible.
- The adapter provides access to data on RAID-5 arrays created with the SSA adapter LL03 and LL11. This is intended as a means of migrating data only. Please note that RAID-1 drives that were created with the SSA RAID cluster adapter cannot be migrated to the advanced SerialRAID/X adapter.



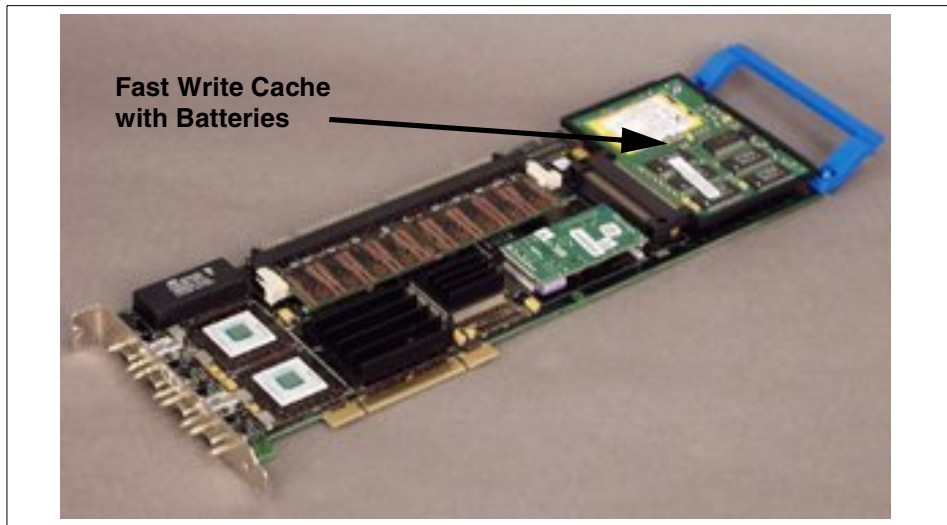


Figure 145. The advanced SerialRAID/X adapter

Table 23 summarizes the configuration details of the RAID levels supported by the advanced SerialRAID/X adapter:

Table 23. RAID level features

RAID level	Member disks per array	Stripe sizes
RAID-0	2-16	4-256 KB in 4 KB increments
RAID-1	2	n/a
RAID-10	4-16	16, 32 or 64 KB
RAID-5	3-16	16, 32 or 64 KB

Table 24 on page 244 lists all currently supported Intel-based server platforms, from IBM and other vendors, for the advanced SerialRAID/X adapter. For up-to-date support information on SSA, check:

<http://www.storage.ibm.com/hardsoft/products/ssa/pcserver/index.html>

Table 24. Supported Intel server platforms for the SerialRAID/X Adapter

Server Vendor	Models	Operating System
IBM Netfinity Servers	5000 5500-M10 5500-M20 5600 7000 7000-M10 8500R	NT 4.0 Server and Enterprise Edition Novell NetWare 4.2 and 5.0
Compaq ProLiant Servers	5500 6500-XEON 7000-XEON	NT 4.0 Server and Enterprise Edition Novell NetWare 4.2 and 5.0
Dell PowerEdge Servers	4300 6350	NT 4.0 Server and Enterprise Edition Novell NetWare 4.2 and 5.0
HP NetServer Models	LH2	NT 4.0 Server and Enterprise Edition Novell NetWare 4.2 and 5.0

#### 8.2.4 The write cache

The non-volatile write cache makes redundant copies of write data to ensure that there is no single point of failure. The advanced SerialRAID/X adapter makes a non-volatile copy in the fast-write cache along with a volatile copy in SDRAM. In a clustered environment, the other adapter makes a further copy in its SDRAM.

In case of an adapter failure, the cache module can be transferred to a new adapter. The outstanding transactions will then be written to the disks. The battery protects data in the cache for up to 10 years.

The cache is now supported in both one-way and two-way configurations on all RAID levels.

Table 25. Supported Netfinity Servers and slot considerations

Netfinity server model	PCI slot supported	Maximum number of adapters supported
5000 5500 5500-M10	All	4
5500-M20	All, except in slot 5. Slot 5 is only half-length.	4
5600	Not in slot 4 and 5 at the same time	2
7000	Not in slot 5 if more than one adapter is used. If you use slot 1 and 6 for SSA Adapters, then other non-SSA high-performance adapters should be placed on the other PCI bus. When using only NetWare 4.11 Do <i>not</i> install the SSA Adapter in slot 5 or 6.	3
7000-M10	All	6 (4 with Novell HAM driver)
8500R	All	6 (4 with Novell HAM driver)

### 8.2.5 The 7133 disk storage enclosures

The storage enclosure Model 7133 is available in two variants:

- Model D40, a drawer model (4U high) that can be installed in a standard 19-inch rack.
- Model T40, a stand-alone deskside tower unit.



Figure 146. Several 7133-D40s in a rack solution and a 7133-T40

The 7133-D40 and T40 enclosures have the following features:

- They provide outstanding performance, supporting the 40 MBps data rate offered by the advanced SerialRAID/X adapter.
- Each enclosure has 16 disk drive slots, and supports 36.4 GB, 18.2 GB, 9.1 GB and 4.5 GB disk drives.
- The enclosures provide storage capacity of up to 582 GB per tower or drawer. The number of daisy-chained enclosures is only limited by the maximum supported number of drives on the loop, which is 48. This means that up to three fully-populated 7133-D40/T40 enclosures can be supported on a single loop, and up to six by one adapter. Using the 36.4 GB drives, a total disk capacity of up to 3.5 TB can be supported by a single adapter.
- High availability solutions are easy to implement by using redundant data paths, redundant cooling units, and two power supplies.
- Remote mirroring at up to 10 km is supported by the enclosures, using the Advanced SSA Optical Extender.
- These enclosures can be intermixed in the same loop with other models of the 7133 (010, 020, 500, and 600), the 7131-405 and the 3527. However,

the operational speed of the loop will then be limited by the other enclosures to 20 MBps per link.

- Four pairs of ports, allowing attachment of up to four independent loops for configuration flexibility are provided on each enclosure.
- Each enclosure contains a controller card that provides SCSI-3 Enclosure Services information on FRUs, part numbers for components of the disk subsystem, drive location and enclosure ID, temperature information, cooling management, and status of bypass cards.
- Four bypass cards are included in each enclosure. These cards provide the connection between the SSA cables and the disk drive modules. Each bypass card has two external connectors.

Figure 147 on page 248 shows a detailed parts diagram of a 7133-D40. We will describe the exact functionality of the bypass cards in 9.1, “Configuration of bypass cards” on page 251.

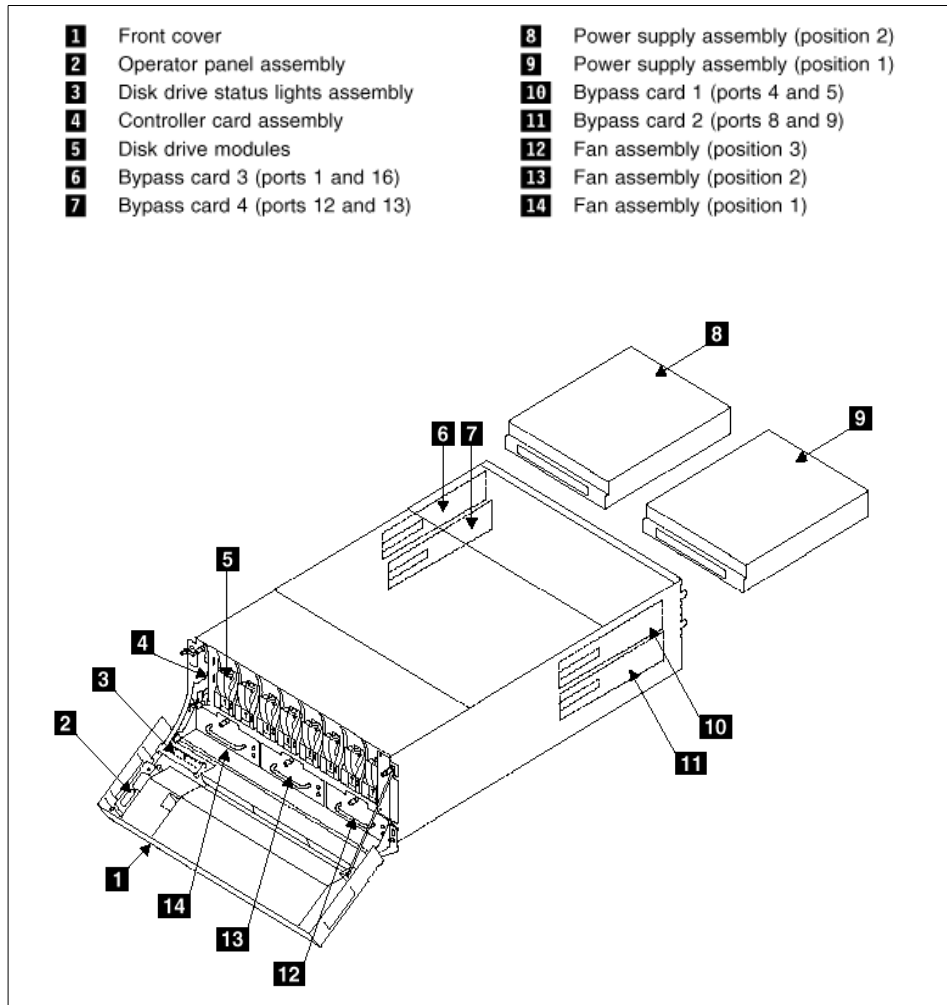


Figure 147. Parts diagram of 7133-D40

### 8.2.6 The IBM SAN Data Gateway for SSA

The IBM SAN Data Gateway 2108-S20 provides the attachment of the IBM SSA Disk Subsystems 7133, 7131 and 3527 through Ultra SCSI or SCSI host bus adapters.



Figure 148. The IBM SAN Gateway 2108-S20 for SSA (top)

The Gateway S20 has the following features:

- UltraSCSI host attachment for up to 64 Serial Disk drives (over 2 TB) on a single loop.
- Up to eight Gateway S20s on a single SSA loop.
- RAID-1 data protection and triple mirroring for improved availability and performance.
- Disk concatenation for larger volumes.
- Global hot spare disk assignment.
- Attachment to SSA Fibre Optic Extender for distances of up to 2.4 km.
- Intermix of Gateway S20 supported host platforms on a single SSA loop.
- StorWatch Specialist for configuration setup, monitoring, and management.
- Availability in stand-alone or rack-mounted enclosures.

The Gateway S20 is managed through the IBM StorWatch S20 Specialist management software and is currently only supported under NT 4.0 Server or Enterprise with ServicePack 4 or later.

Currently the Netfinity platform supports the Adaptec SCSI Adapter AHA-2944UW. It is also supported on Compaq ProLiant Servers under Windows NT 4.0 with SP 4.

For the latest support information, especially for non-Intel-based servers, visit:

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





---

## Chapter 9. Implementing SSA disk subsystems

This chapter covers the implementation of disk subsystems using the advanced SerialRAID/X adapter with the latest firmware currently available and the storage enclosures 7133-D40 and -T40.

We also describe a disaster recovery configuration that can be implemented with SSA in clustered environments, such as Novell Cluster Services or MSCS.

We also give a brief introduction into the available management tools for SSA subsystems.

---

### 9.1 Configuration of bypass cards

Before discussing specific loop configuration details (9.2, “Building SSA loops” on page 254), we need to explain how you should prepare the bypass cards of a 7133-D40. In Figure 149, you can see the internal wiring of a 7133-D40 enclosure. The four groups of four disk drives are wired to the ports on the four bypass cards.

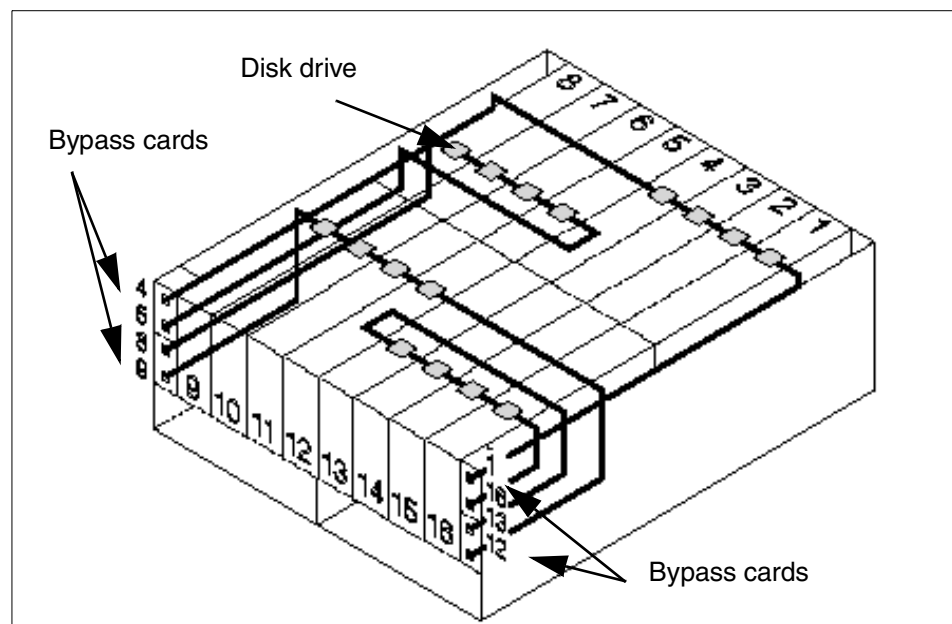


Figure 149. Internal view of the 7133-D40

This diagram can be simplified to a logical schematic as shown in Figure 150. We will use this view for future configuration diagrams. You will find the port numbers from the diagrams printed on the enclosure exterior. The bypass cards act as configurable switches that allow the four groups of disk drives to be grouped different ways.

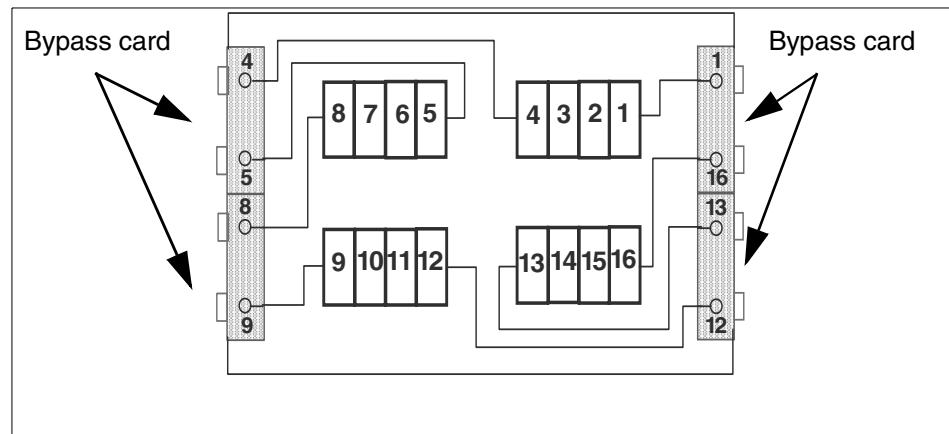


Figure 150. Logical view of the 7133-D40

### 9.1.1 Bypass card modes

A bypass card (see Figure 151) has a jumper for each port, which allows the ports to be set to one of two modes:

- Automatic mode
- Forced Inline mode

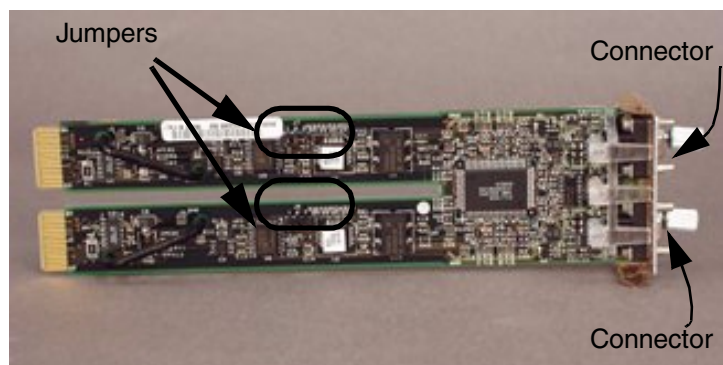


Figure 151. Bypass card and jumper locations

When a bypass card is set to Automatic mode, the card can be set into one of three additional modes with the Remote System Management tool. These modes are:

- Forced Inline mode
- Forced Bypass mode
- Forced Open mode

#### **Setting jumpers**

The four bypass cards in a 7133 can be configured independently of each other. Note, however, that both ports on a single bypass card must be set to the same state. Using RSM to set the bypass card state affects both ports.

#### ***Automatic mode***

When a bypass card is jumpered to operate in Automatic mode, it monitors both of its external connectors. If it detects that at least one of its connectors is connected to a powered-on SSA attachment or device, it switches the port to the Inline state; that is, it connects the internal SSA links to the external connector.

If the bypass card detects that neither of its connectors is connected to a powered-on SSA attachment or device, it switches into the Bypass state; that is, it connects the internal strings together, disconnecting them from the external connectors. (As a consequence, a 7133 that has all its bypass cards jumpered for Automatic mode, and is not connected externally, connects all 16 disk drive module slots in one internal SSA loop.)

#### ***Forced Inline mode***

When a port is set to operate in Forced Inline mode, either by using the jumper or with RSM, its switching ability is disabled and the internal SSA links are connected to the external connector.

If Forced Inline mode is set by the jumpers, the mode cannot be changed by the subsystem service aids or by the command line tools. Only Automatic mode allows configuration of the bypass cards using software tools. Automatic mode is the preferred operating mode unless one of the two above-mentioned reasons dictates otherwise.

#### ***Forced Bypass mode***

Using RSM, a bypass card that is jumpered for Automatic mode can be switched to Forced Bypass mode, which forces the links to break the connection to the external connector and to link the two internal SSA strings.

This ensures that any external SSA connections cannot break the connection between two internal disks that are on either side of the bypass card.

### ***Forced Open mode***

When the bypass card is set to Forced Open mode, no connections through the bypass card are active. This could be used to ensure that the interface is isolated. RSM allows you to set a port to Forced Open mode.

Note, however, that once in this state, RSM would no longer be able to communicate with the card, and you would have to reset the card from the 7133 operator panel.

There is no practical reason to use Forced Open mode.

#### **Bypass setting recommendation**

We recommend that you should leave the bypass jumper settings to Automatic mode. In this mode you should select Forced Inline (through RSM) if you have disks in a 7133 that are on different SSA loops and you wish to ensure that any fault condition will not join the loops together. In all other cases you should leave the bypass cards programmed to Automatic mode.

---

## **9.2 Building SSA loops**

This section examines the typical ways to configure cabling between a 7133-D40/T40 storage enclosure and the IBM Advanced SerialRAID/X Adapter.

We will use the diagram we introduced in Figure 150 on page 252 for the representation of the enclosure, and the diagram in Figure 152 as a schematic drawing for the adapter connectors. As previously discussed, an adapter has four full-duplex connectors, allowing two loops to be configured. Connectors designated with the same letter form part of the same single loop.

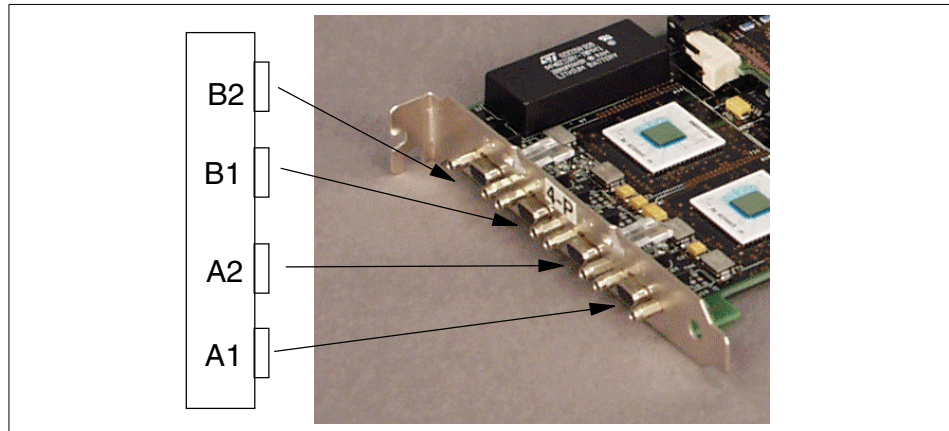


Figure 152. The connectors on the Advanced SerialRAID/X adapter

## 9.2.1 Single-host configurations

Configurations connecting disk subsystems to a single host are the simplest and we examine some of these initially. Following this, we will discuss some configurations utilizing two hosts.

### 9.2.1.1 Single host, two loops

We start with the example shown in Figure 153 on page 256, in which we have a single host with two independent loops. Each loop provides a data path that starts at one connector of the SSA adapter, passes through a link to the devices and then returns to the second connector on the SSA adapter. Loop A consists of disks 1 to 8, and Loop B consists of disks 9 to 16.

The bypass cards in 4-5 and 12-13 are set to automatic and since no external cable is connected to them they are effectively bridged (bypass mode). Ports 1-16 and 8-9, however, should be set to Forced Inline mode by jumpers ensuring that the two loops A and B do not interfere.

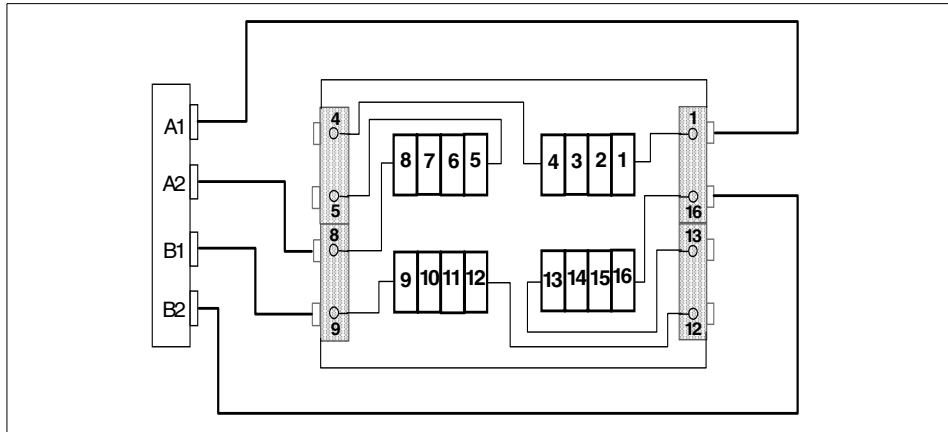


Figure 153. Single host with two loops

### 9.2.1.2 Single host, two enclosures

The next example (see Figure 154) shows one host connected to two 7133 enclosures. Each enclosure is configured in its own separate loop. All bypass cards should be set to Automatic mode in this configuration. You have 16 drives in each loop. Keep in mind that you cannot span arrays across loops.

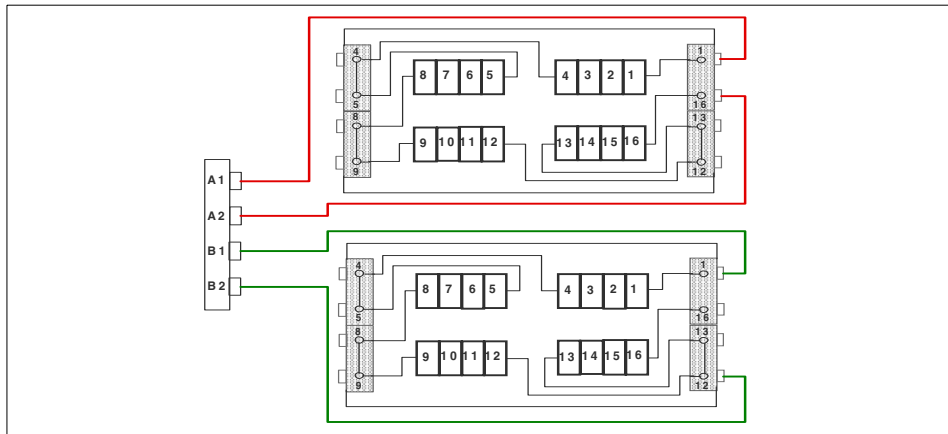


Figure 154. Single host with two 7133 enclosures

### 9.2.1.3 Maximum configuration

The maximum number of enclosures allowed on a single adapter is six, as shown in Figure 155. Each loop encompasses three enclosures which, if fully populated, provide the maximum configuration of 48 disk drives per loop. The

bypass cards jumpered to Forced Inline mode, as we suggested in 9.1, “Configuration of bypass cards” on page 251, are marked with two asterisks.

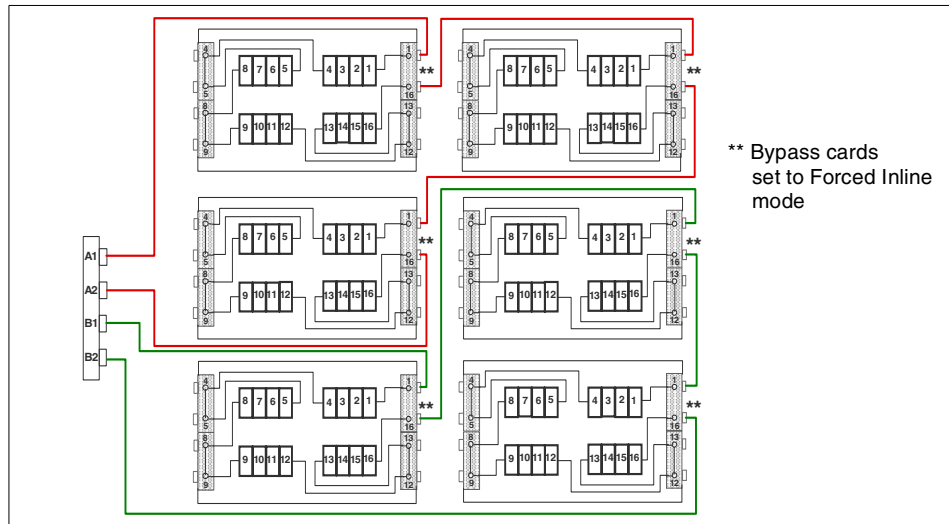


Figure 155. Single host with six 7133 enclosures

## 9.2.2 Dual-host configurations

Now we examine dual-host configurations as used for clustered configurations using MSCS or Novell Cluster Services.

### 9.2.2.1 Clustering with a single 7133 enclosure

The example in Figure 156 on page 258 demonstrates a dual host configuration with two loops. This is a valid configuration for MSCS, using one 7133 enclosure. Bypass cards 1-16 and 8-9 are jumpered to Forced Inline mode to prevent possible loop mixing. Each loop is shared by both hosts (or initiators). We can afford to have two cable link failures in each loop and still have access to the data.

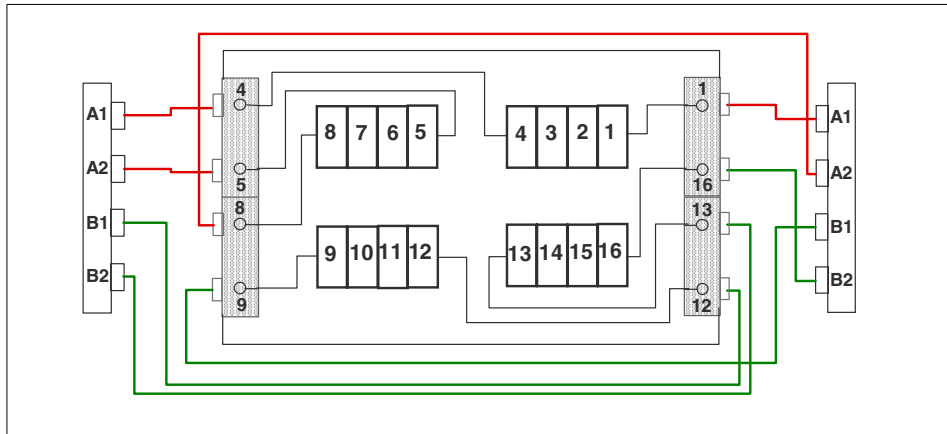


Figure 156. Dual-host configuration with two loops (MSCS)

Data on disks in this configuration can potentially be accessed by both hosts using Windows NT. The lock mechanism preventing this is provided by MSCS.

### 9.2.2.2 Clustering with two 7133 enclosures

Going a step further, we can use two 7133 enclosures with two hosts as shown in Figure 157. In this configuration, there is a dedicated enclosure per loop, removing the need to jumper the open bypass cards to Forced Inline mode. The bypass cards are switched automatically to Bypass mode when they are jumpered in Automatic mode.

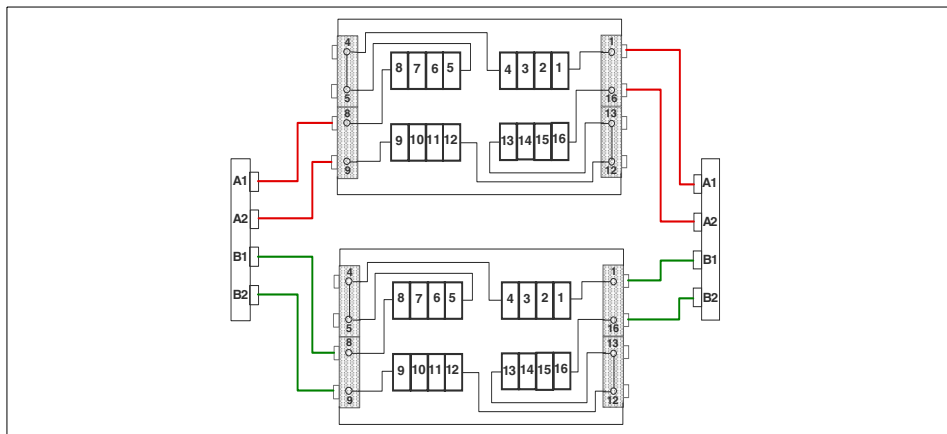


Figure 157. Dual-host configuration with two 7133 enclosures



### 9.2.2.3 Maximum configuration

Concluding this section, we show the maximum dual host configuration, using six 7133 enclosures, as in Figure 158:

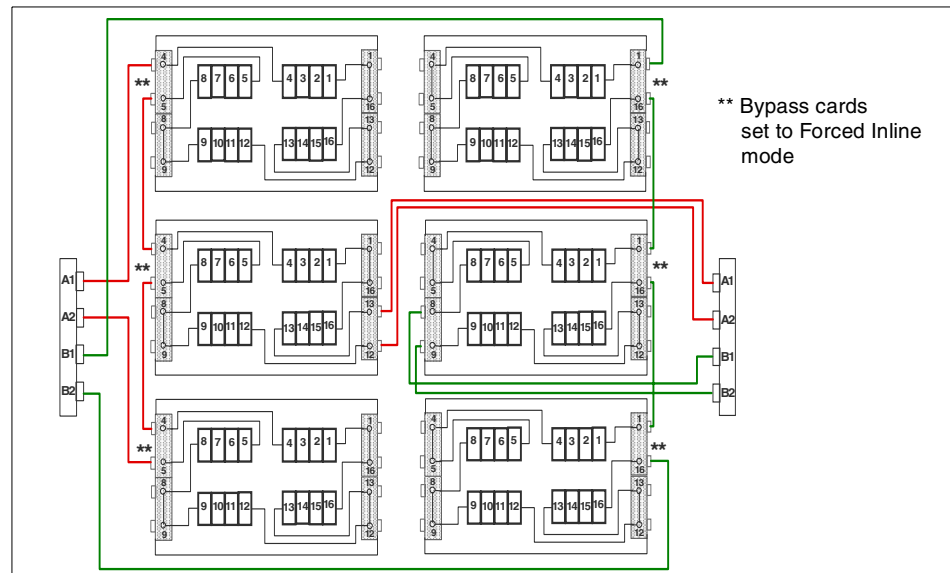


Figure 158. Dual host configuration with six 7133 enclosures

### 9.2.3 Cabling summary

To summarize, these are the rules you need to observe when building loops in an SSA environment:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (either connectors A1 and A2 or connectors B1 and B2).
- Only one pair of adapter connectors can be connected in a particular loop.
- Only one adapter can be connected in a particular loop if any drives in that loop are configured in RAID-0 arrays. This means that two-node clustering solutions cannot be implemented using RAID-0 arrays. There are no limitations on the other RAID levels.
- A maximum of two adapters can be connected in one loop. Each adapter must be in a separate Netfinity Server. This means that you cannot build clusters with more than two hosts even when supported by the clustering software (Novell Cluster Services, for example).
- All disk drives belonging to a specific array must reside on the same loop. Arrays cannot span across loops.

- A maximum of 48 disks can be connected in a single SSA loop, meaning you can have a maximum of three 7133-D40 or -T40 enclosures per loop.
- When an SSA adapter is connected to two loops, and each loop is connected to other adapters, all adapters must be connected to both loops. This rule means that it is not possible to daisy-chain or cross-connect loops among SSA adapters. If you did that you would violate one or more of the above rules.

---

### 9.3 Disaster recovery configuration

Advanced SerialRAID/X adapter firmware Release 6806 (released in the fourth quarter of 1999) or later enables technology allowing you to implement a solution offering full disaster recovery.

This provides a higher degree of availability for critical data and applications through the use of geographically separated servers and SSA disk subsystems (also known as a multiple domain operation). With the use of suitable clustering software, such as Microsoft's Cluster Server (MSCS), it is possible for users to have access to their applications and data even in the event of one of the sites going offline (for example, due to a power outage or a building fire).

In order to take full advantage of these facilities, it is necessary to understand the operation of the IBM Advanced SerialRAID/X Adapter in various failure situations.

#### 9.3.1 RAID-10 and RAID-1 explained

To explain the operation of the disaster recovery configurations, we first of all provide a brief explanation of RAID-10 and RAID-1 arrays. For basic information about RAID levels, we refer you to 4.2.2, "RAID levels supported by ServeRAID adapters" on page 48. Although the descriptions there are discussing RAID levels supported by ServeRAID adapters, RAID levels 0 through 5 are standard for all technologies. RAID-10, as we shall see, is implemented slightly differently between ServeRAID and SerialRAID adapters.

RAID-1, also known as disk mirroring, takes two disks and writes the same data to both of them. If one of the disks fails, your data is still available from the other disk. This means that the usable capacity of your array is limited to the capacity of a single disk. To overcome this limitation, and thus provide larger storage capacity to the operating system being run on the server, the

SerialRAID adapters provide RAID level 10. Before we can tell you about RAID-10, however, we need to explain RAID-0.

In RAID-0, a collection of disks, an array, is operated as if they were one large disk, and the data to be written is spread (or striped) across the members of the array. This configuration provides increased capacity and improved performance in comparison with the use of a single disk. The performance improvement is due to a number of data accesses being performed in parallel, as separate chunks of data are written to or read from the member disks in the array. In the case of a single disk, all operations are completed as a number of sequential operations to the one drive. The problem with RAID-0 is that there is, in fact, no redundancy; if one of the disks in the array fails, then the data is lost. There is no protection for disk failure.

RAID-10 combines RAID-0 and RAID-1, extending RAID-1 by replacing each of the two disks in the mirror with a RAID-0 array. Now you have improved performance (RAID-0), with high availability in the event of a disk error (RAID-1). In fact, you can tolerate multiple disk failures, provided the other half of the mirror (the disk which holds the same data held as the disk that failed) is still operational. The two RAID-0 arrays in the RAID-10 array are known as the primary and the secondary domains - the secondary domain mirrors all the data held in the primary domain.

In the event of a disk failure, operation will continue without user intervention. You can also define extra disk drives to act as hot-spares, that is disks that are powered up and will automatically replace disks that fail by recreating the data from the mirror disk that is still operational. There is some loss of performance while the new disk is being rebuilt, and the user must decide the balance between operational performance and how quickly the new disk is rebuilt (this is known as rebuild priority) and fault tolerance is restored.

### **9.3.2 Split-site operation with RAID-10 and RAID-1**

In the discussion that follows, reference will be made to the RAID-10 configuration shown in Figure 159 on page 262. This could equally well be a RAID-1 configuration, the only difference being the number of disks involved; the following remarks apply equally to both configurations.

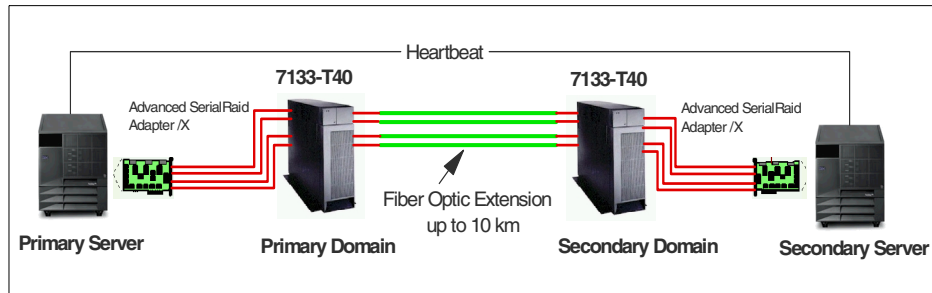


Figure 159. 2-way RAID-10 cluster

Figure 159 shows a 2-way cluster, which consists of two Netfinity servers, each containing an SerialRAID/X adapter. The adapters drive the two SSA loops running between the machines. By using suitable optical fiber connections, the two servers may be up to 10 km apart. Also connected into the SSA loops are two disk drive subsystems (using 7133-D40 disk drawers that can each hold up to 16 SSA disk drives). The disk subsystem at one site holds the primary domain of the RAID-10 array and is located locally to the primary server; likewise for the secondary domain.

In SSA a RAID-10 array must consist of a minimum of four disks, and may contain up to 16 disks (for example using 18 GB disk drives, the maximum usable capacity for a RAID-10 array is  $8 \times 18 = 144$  GB). The SerialRAID/X firmware maintains a small amount of information, known as “metadata”, on the SSA disk drives, which is hidden from the user. Part of this metadata is a flag, known as “Split-Resolution”, which can be controlled by the user through the SSA configuration tool, RSM. The use of this flag will become clear as we consider different failures.

In normal operation the split-resolution flag is set to “primary”. As data is written to the primary domain, it is replicated on the secondary domain. At the operating system level, (Windows NT with MSCS for example), only one of the two servers controls the disks; the other server does not update the array. The names primary and secondary are deliberate; it is expected that the primary server will normally be running the applications, with the secondary server used to fail over the applications in the event of a failure to the primary server.

Now let us consider what happens in each of the following error situations:

1. Failure of the primary server

If the primary server has some sort of failure, which means that it can no longer operate, then the cluster software should automatically switch

operation of those applications to the other server. This could include a server being taken down for upgrading or maintenance. In this case the primary domain of disks is still accessible from the secondary server (even if the SSA loop is broken at the SerialRAID/X adapter in the primary server), and operation will continue normally.

2. Failure of the secondary server

If the secondary server fails, that is the server which is not currently running the applications, then the primary server continues normal operation. The secondary domain is still operational and mirroring the data in the primary domain.

3. Failure of the primary domain

This failure is meant to imply the complete failure of the primary disk subsystem (for example a power outage), not individual disk failures that can be handled through the hot-spare mechanism. It also covers the case where the primary domain is inaccessible due to multiple faults in the SSA loop (SSA can tolerate a single loop failure). In this situation the secondary domain and the primary server are still available and will continue in normal operation.

When the primary domain comes back online, then the SerialRAID/X adapter will automatically start to rebuild the data on the primary domain from the secondary domain.

4. Failure of the secondary domain

This failure is handled in the same way as for a failure of the primary domain (in 3 above), except that when the secondary domain comes online again, it has its data rebuilt from the primary domain.

5. Failure of the secondary site

This failure occurs when both the secondary domain of disks and the secondary server become unavailable (for example due to a power outage that affects only the secondary site). Operation will continue normally, with data being updated on the primary domain. As soon as the secondary domain is available again, its data is rebuilt from the primary domain.

6. Failure of the primary site

This is similar to situation number 5, except that, now, the primary server and the primary array of disks have failed. The cluster software will attempt to fail over the applications to the secondary server; however, the SerialRAID/X adapter in the secondary server will not allow the server to access the array in the secondary domain. To understand why this behavior has been built into the adapter consider the following:

From the viewpoint of the secondary site the primary site has had a complete failure. However, a similar situation would also arise if the SSA loop had been disrupted (perhaps a utility company has cut through the cables while digging a trench), so that each server can no longer “see” the disk drive domain at the remote site. This situation is called a “cluster partition”. The cluster software in both servers would also believe that the other server had failed, and would try to start running the applications on both servers using their local copies of the applications and data held on their local domains. If both servers were allowed to update their own local data, then the two domains would rapidly diverge and it would be impossible to sort out what information was valid once communications had been restored.

This is where the split-resolution flag comes into action. Remember that by default it is set to primary, and it is this setting that prevents the secondary server from accessing the secondary domain when the primary site has failed. The secondary server cannot update information on the disks, and it cannot read information from those disks because the data can't be guaranteed to be valid any more, since the same information might have been changed on the primary domain. At this point an operator has to make the decision about whether to change the flag's setting.

If the primary site is still operational, then the operator would leave the split-resolution flag set to primary, since he would wish the primary site to continue running and the secondary site to remain offline. When the cluster partition is resolved, the secondary domain is automatically rebuilt.

On the other hand, if the primary site had really failed, then the operator would set the split-resolution flag to Secondary, to allow the secondary server to have access to the secondary domain and continue operations.

Once the primary site is running again, its data would be rebuilt from the secondary domain. When the data rebuild has completed the adapter automatically resets the split-resolution flag to primary.

### **9.3.3 Operation of hot-spares**

To restore an array to a high level of availability following an individual disk failure, the user can create, using RSM, pools of hot-spare drives. It is then possible to associate a hot-spare pool with the disks in the primary domain, and a second pool with the secondary domain disks. For a split-site operation, this is important to ensure that any hot-spares that are automatically introduced into the domain through disk failures are local to that domain.

Unless care is taken, the situation might arise that a hot-spare at the secondary site is used as a replacement for a drive in the primary site. Then, if there was a communication failure between the sites (a cluster partition), the primary domain would fail because one of its disks would be inaccessible.

In the event of a cluster partition, with the effective loss of one domain, how should the SerialRAID/X adapter behave? Should it replace all the drives in the missing half of the mirror with local hot-spares? This is an implementation decision that you must make. The SerialRAID/X hardware allows this flexibility in configuration by being able to have multiple hot-spare pools and by being able to control hot-spare replacements when a cluster partition occurs.

#### **9.3.4 Performance**

When planning a split-site operation, it is necessary to consider the effect of long distances on performance. SSA operates by transmitting a packet of information (commands and data) from one node to the next, sequentially, around the SSA loop. A node is a disk drive or an SSA adapter. To maintain reliability, when a node sends a packet it expects a response from the next node in the loop to say that the packet was successfully received before sending another packet.

When the distance between nodes is increased dramatically through the use of optical fibers, the delays associated with sending packets and waiting for responses start to become significant. As the delay is increased, the throughput across the optical connection decreases. The effective maximum data speed of 40 MBps starts to drop, until, at the maximum distance using optical fibre extenders, 10 km, the throughput is only 1 MBps per on each port in each direction.

Of course, arrays that are local to an adapter do not suffer this penalty. For example, a RAID-5 array whose disks are local to, say, the primary site retains a 40 MBps throughput in each direction when being accessed by the primary server.

When using split-site configurations with optical fibre, care should be taken in using the fast-write cache on the advanced SerialRAID/X adapter. The fast-write cache is a battery-backed memory that is used to hold data being written to the disks. When it is used, the adapter will signal to the operating system that a write request has been completed as soon as the data is safely stored in the cache. The adapter subsequently writes the data to the disk subsystem in its own time. This enhances write performance, especially when used in conjunction with RAID-5.

In a cluster configuration, however, the data must be written to the fast-write cache on the adapter that received the write request, and then sent over the SSA loop to the other adapter in the cluster to be stored in its cache, before the operating system is told that the write request has been completed. When large distances between the two sites are involved, the overhead for writing the data to the remote cache will be significant, and may mean that better throughput is achieved by not enabling the fast-write cache.

### **9.3.5 Disk placement**

An SSA adapter communicates with the disks attached to it through its four ports (A1 and A2 on the A-loop, and B1 and B2 on the B-loop). Each port can communicate in full duplex mode (that is, it can receive and send data at the same time). When the adapter wants to communicate with a disk on one of the loops, it uses the port that has the lowest “hop count” between the port and the disk. The hop count is determined by calculating how many disks or other SSA adapter nodes exist between the adapter port and the disk with which it wants to communicate. The packets of data will always be sent along the path with the lowest hop count.

For this reason, arrays and single disks on an SSA loop should always be spread between the two ports on the adapter. This becomes even more important in a split-site arrangement where there is a large distance between the primary and secondary sites. Consider the following two arrangements of a RAID-10 array, made up of eight member disks (four disks are local: X1, X2, X3, X4 and four disks are remote in the secondary domain: Y1, Y2, Y3, Y4):



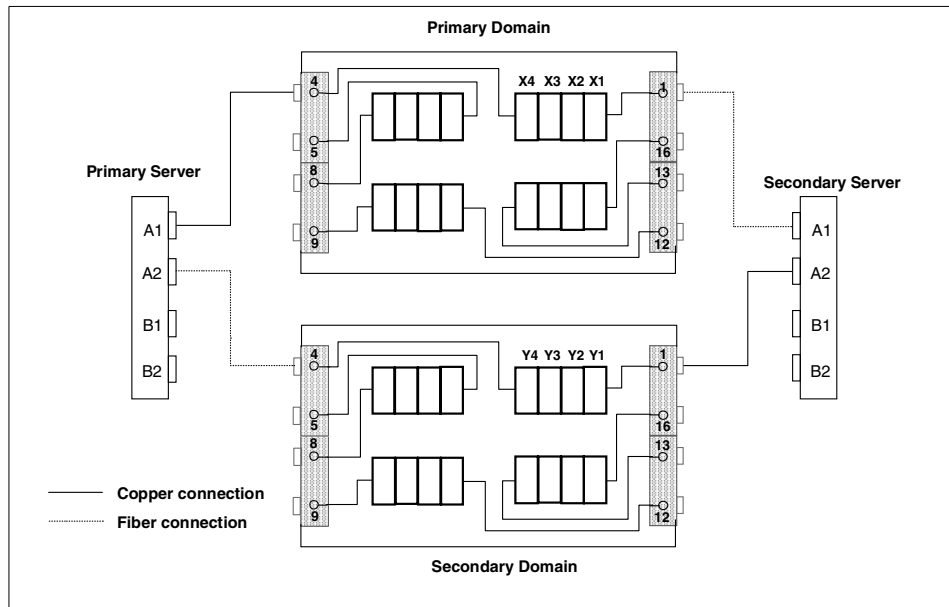


Figure 160. 2-way RAID-10 cluster, non-distributed arrangement

In Figure 160, the disks are arranged in their enclosures so that all the primary disks are on one half of the loop, and all the secondary disks on the other half. When the primary SSA adapter wants to communicate with the disks in the secondary domain it will *only* use the bottom half of the loop - because the hop count to the 'Y' disks is lower on the bottom half than through the top half. When the sites are widely separated, and optical fiber is being used to connect them, then the available bandwidth across the fibers is already constrained, for the reasons explained in the previous section. With this arrangement of disks, only half of that limited bandwidth will be used.

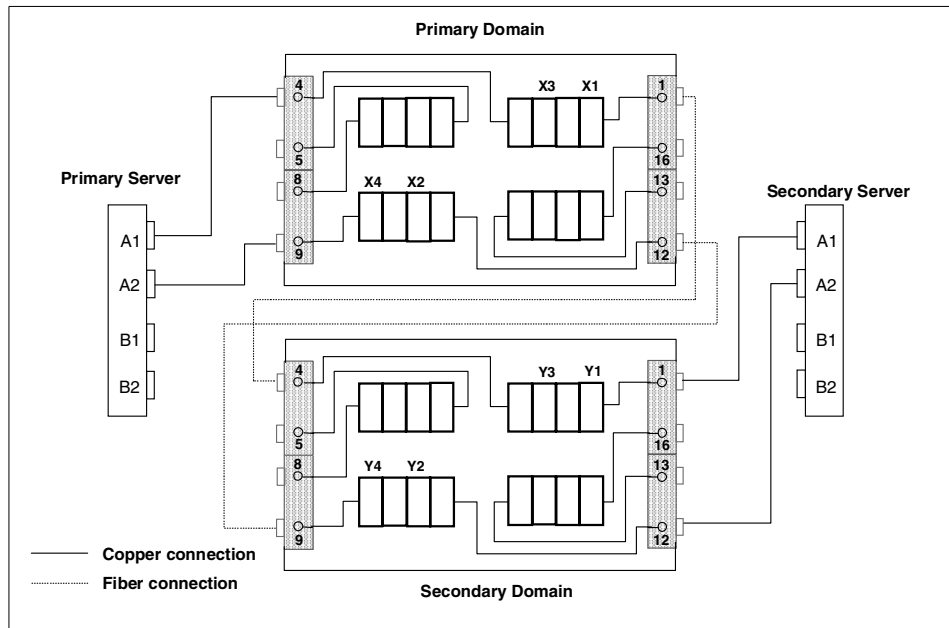


Figure 161. 2-way RAID-10 cluster, distributed arrangement

Consider now the arrangement shown in Figure 161. Here, the disks in the 7133 enclosures have been cabled so that half the disks are in the top half of the loop, and the other half are in the bottom (we are referring here to the same disks as above, we have just located them in different slots in the 7133 and cabled them differently). Now when the primary SSA adapter talks to the 'Y' disks, it will use the top half of the loop to talk to the disks Y3 and Y1, and the bottom half (the A2 link) to talk to Y4 and Y2. This maximizes the use of the available bandwidth between the two sites.

Finally, we give you in Figure 162 a complete cabling diagram of a 2-way cluster configuration with fully populated storage enclosures:

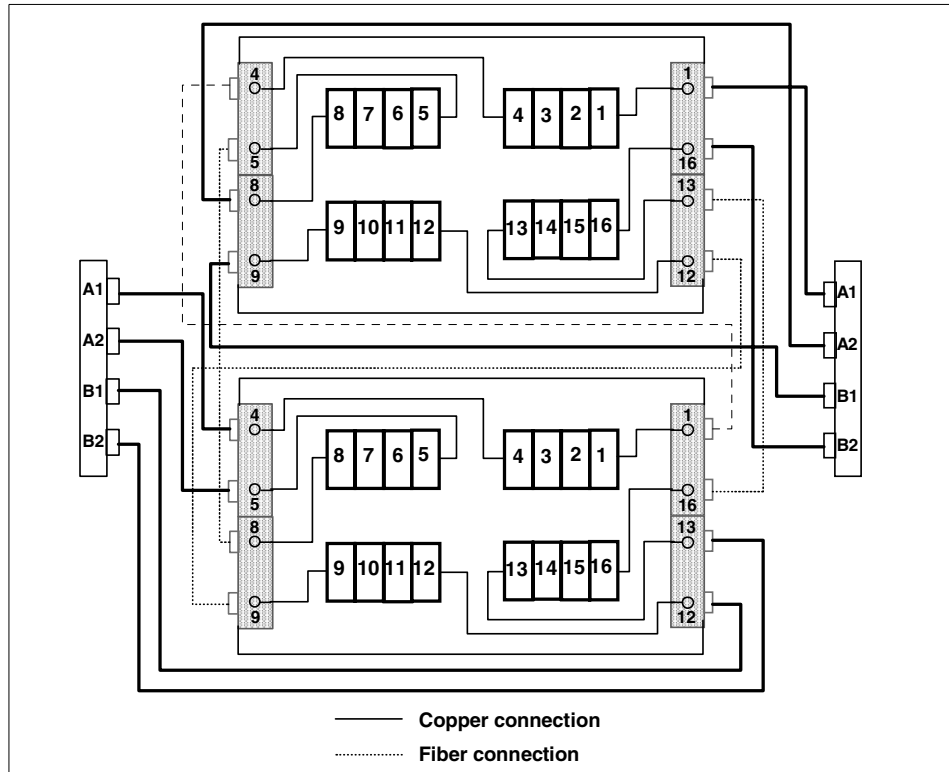


Figure 162. 2-way cluster with maximum disk configuration in two 7133 enclosures

The cabling in this diagram looks quite complicated, but if you trace it out, you will see that it follows the guidelines already discussed.

### 9.3.6 Summary

RAID-10 and RAID-1 are important features of the IBM Advanced SerialRAID/X Adapter, coming, as it does, at a time when many IBM customers are eager to achieve the highest levels of availability. The burgeoning Web economy is placing heavier demands on businesses to provide 24x7 operation all year round. IBM now has an SSA solution that provides this level of reliability together with outstanding levels of performance.

## 9.4 Managing the advanced SerialRAID/X adapter

A number of tools are provided to enable you to effectively manage your advanced SerialRAID/X adapter installation. There may have been updates

since you first obtained your adapter, so to obtain the latest software tools, documentation and microcode, we strongly recommend that you check:

<http://www.storage.ibm.com/hardsoft/products/ssa/pcserver/index.html>

or

<http://www.hursley.ibm.com/ssa>

To begin this section, we will examine the various tools, described below, for flashing the different components of a complete SSA solution. These are:

- The Advanced SerialRAID/X Adapter
- The 7133-D40/T40 enclosures (updates to the enclosure controller card)
- The hard disk drives

We will then describe the Web-based Remote System Management (RSM) tool in detail in 9.4.2, "Remote System Management" on page 272.

#### **9.4.1 Operating system-specific tools**

DOS utilities are provided for configuring and installation purposes only. There is no supported DOS driver for SSA available. When downloading the DOS tools diskette (which boots to a command prompt) you are offered the following DOS command-line tools:

- ISSACFG.EXE - this is the text-based configuration utility with which you can configure SSA disk drives. It allows configuration of your subsystem, as does the Web-based RSM tool but without some additional functions provided by RSM. The DOS tool should only be used by experienced support people and not for normal administration purposes. The DOSRME.TXT file on the diskette gives instructions on how to use this utility.
- ISSAADLD.EXE - this program allows you to download the microcode for the advanced SerialRAID/X adapter.
- ISSADDLD.EXE - this program allows you to download the microcode for the IBM SSA Ultrastar disk drives (XP, 2XP, 3XP).
- ISSAEDLD.EXE - this program allows you to download the enclosure controller microcode for all IBM 7113 Model 40 enclosures. This tool can also be downloaded separately as Enclosure Microcode.

The Windows NT 4.0 utilities either come in a complete package, which is currently LL05ALL.EXE, or they can be downloaded as separate components. The components are:

- All command-line utilities for flashing the microcode of the adapter, the enclosure and the disk drives. These have the same names as the above-mentioned DOS utilities. However, the executables are different from the DOS executables!
- The Windows NT driver, IBMSSA.SYS. This is installed through the SCSI adapter icon in the control panel. There is also a special driver to allow you to load Windows NT as a bootable OS, installed on SSA disks. This special driver has to be copied onto the Windows NT installation diskettes.
- The RSM Web-based management tool. This is the premier tool for administration of SSA disk subsystems. It comes both as a standalone utility and integrated into Netfinity Manager. We describe RSM in detail in the next section.
- The SSA Event Logger, which is normally packaged with the RSM tool. This Windows NT service logs SSA events and controls their forwarding to Netfinity Alert Manager. Note that this service is needed even if you are not using RSM with Netfinity Manager.

The Advanced SerialRAID/X adapter is supported under Novell NetWare Versions 4.2 and 5.0, including Novell Cluster Services. The tools available are:

- Bootable DOS diskettes for flashing the microcode of the adapter, the disk drives and the 7133-D40/T40 enclosures. These are the same diskettes as mentioned above for DOS.
- The host device driver diskette. This diskette also contains the NetWare loadable module (NLM). The NetWare driver is a \*.HAM driver. The following modules should be loaded on a Novell server:  
ISSAV4E0.HAM - the device driver  
ISSAINJ.NLM - the SSA Injector Service  
ISSAELOG.NLM - the SSA Event Logger
- The configuration tool (ISSACFG.NLM) is available as an NLM making it possible to control disk resources directly from the Novell console.
- The RSM administration tool is used through a Web browser from a client machine. A background HTTP service (ISSARSM.NLM) must be running on the Novell server.

For technical reasons related to the installation process for Windows NT, it is much easier to install NetWare as a bootable operating system on SSA disks than is Windows NT. You first configure the SSA disks with the above mentioned DOS tools. Next, you install DOS and start the NetWare

installation program from CD-ROM. The NetWare driver for SSA is then easily configured during the NetWare installation process, making the disks readily available to the operating system thereafter.

### 9.4.2 Remote System Management

Remote System Management (RSM) is a JavaScript-based tool that is accessed through an HTTP service, using a Web browser.

The RSM tool requires the SSA device driver to be installed before installation. The event logger included in the RSM package is also required under Windows NT, but is installed automatically during the RSM installation process.

Table 26 summarizes the versions of RSM and their prerequisites. Note that you do not download different versions of the software for stand-alone operation and for use as a Netfinity Manager extension. There is only one download package. You choose which version you want at install time (and in fact it is possible, but not recommended, to have both versions running at the same time, the only difference being through which HTTP port they are accessed).

Table 26. RSM versions for Windows NT and NetWare

RSM Version	Prerequisites
Windows NT Stand-alone Version V1.51	<ul style="list-style-type: none"> <li>• Windows NT Server 4.0 or EE with TCP/IP installed.</li> <li>• Windows NT SSA Event logger V1.39</li> <li>• Netscape V4.07 or Internet Explorer V4.72 with SP2</li> </ul>
Windows NT Web Extension for Netfinity Manager V5.2	<ul style="list-style-type: none"> <li>• Windows NT Server 4.0 or EE with TCP/IP installed</li> <li>• Windows Netfinity Manager V5.2 with Web Manager Services installed and enabled</li> <li>• Windows NT SSA Event Logger V1.39</li> <li>• Netscape V4.07 or Internet Explorer V4.72 with SP2</li> </ul>
NetWare RSM Stand Alone Service V1.51	<ul style="list-style-type: none"> <li>• NetWare 4.2 or 5.0</li> <li>• NetWare TCP/IP installed and configured</li> <li>• NetWare SSA Event Logger V1.39</li> <li>• NetWare SSA Injector</li> <li>• Netscape 4.5x or Internet Explorer 4.72 with SP2</li> </ul>

Using a Web browser, the RSM start page for the stand-alone versions are accessed at:

http://<machine-name>:511/ssaindex.htm#SSAExplorer

whereas the Netfinity Manager version can be accessed at:

http://<machine-name>:411//ssaindex.htm#SSAExplorer

The RSM start page is shown in Figure 163. When trying to access the SSA Configurator you will be asked for a user name and password. These are both set to SSA by default and can, and should, be changed to protect your disk subsystem from unwanted tampering.

Going into the configurator lets you perform all necessary tasks required to manage your SSA disk subsystem.

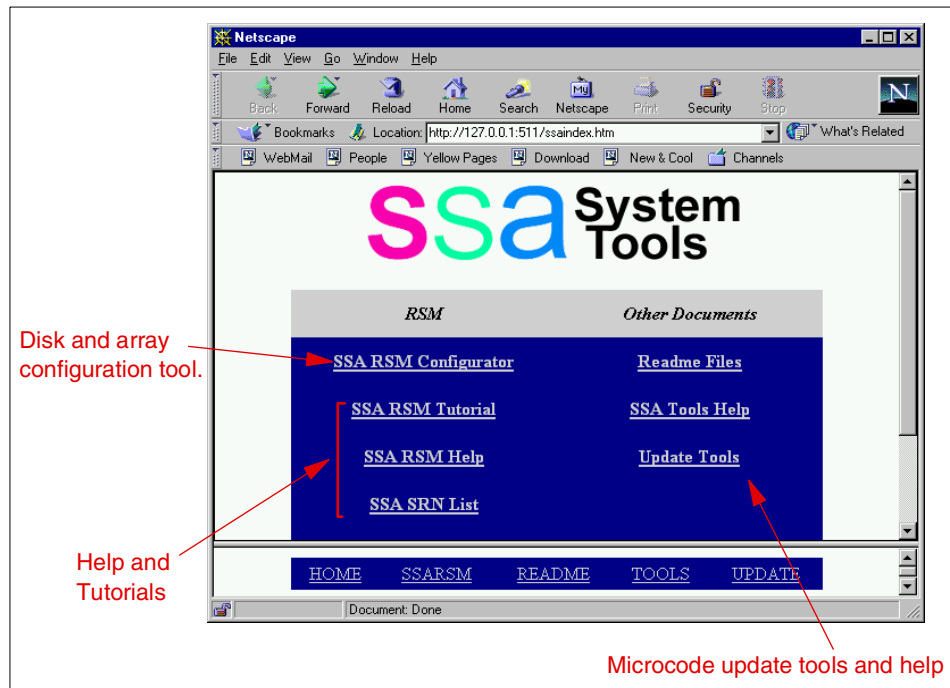


Figure 163. The RSM start page

RSM lets you perform additional management tasks beyond those offered by the DOS-based configurator. The additional functions are:

- Event logging and analysis, checking the error log, help function on SRNs, configuring the Event Logger control file.

- Check level function: Downloads microcode through the Web if access is provided.
- Context-sensitive help.
- RAID-10 and Hot-Spare Pools (**Note:** the latest DOS-based configurator does support RAID-10).
- Remote configuration with RSM.
- Security and Access Control of RSM of the stand-alone version. The Netfinity Manager version uses Netfinity Security Manager.
- SSA event forwarding: The event logger can be configured to route events to the Netfinity Alert Manager.
- Specialized views: Logical, Physical and Enclosure.

### 9.4.3 Managing disk resources with RSM

SSA handles disk resources differently from the ServeRAID or Fibre Channel solutions also discussed in this book. The reasons for this are primarily due to SSA's long and proven heritage in the IBM RS/6000 world. Because of this, some of the concepts used in SSA may be unfamiliar to users of Intel-based servers. You should not be distracted by this; SSA offers a very high degree of flexibility and scalability as a storage solution.

In SSA terminology, a resource is anything with which an SSA adapter can exchange I/O operations. At the system level, resources include a RAID drive or a single disk. The adapter maps such resources seamlessly to the operating system as a physical disk resource.

**New Resource:** A New Resource is normally one that is new from the factory. It can also be a resource that has been used in a non-PC environment (such as AIX). Also, resources that are members of array types not supported by the current adapter are placed under the New Resource classification, with the word "Preconfigured" displayed next to them. Finally, if two system resources with the same disk number are accidentally connected to an adapter, one will be listed as a system resource, and the other will be listed as a New Resource. Deleting a New Resource converts it to a "Free Resource".

**Free Resource:** A Free Resource is one that is not currently used for any purpose. It can be assigned for a particular use, or used to construct other types of resource (such as RAID array resources and fast-write resources).



**System Resource:** A System Resource is one that the operating system may use. Every System Resource has a unique resource number that the operating system uses to distinguish between resources. System Resources are presented to the system in ascending order of resource number.

**Hot Spare Resource:** A Hot Spare Resource is one which is currently not in any active use. If a (suitably configured) array resource loses one of its component resources, the Hot Spare Resource will be taken and used to replace the failed/missing component.

**Rejected Resource:** A Rejected Resource is one which used to be a member of an array resource, but is no longer. Arrays reject member resources if they fail or are found to contain corrupt data. If an array loses a component, replaces it with a Hot Spare, and then the original component returns, it will be Rejected.

**RAID Resources:** RAID-0, RAID-1, RAID-5 or RAID-10 resources are available for configuration.

**Fast Write Resource:** Fast Write can provide a performance boost for certain workloads by completing a write command as soon as the data enters the adapters non-volatile memory, without having to wait for it to be transferred from the memory onto the disk (or array) resource below it.

Disk drives that are connected to the advanced SerialRAID/X adapter do not need to be configured to be components of an array. The adapter handles such disk drives in the same way as a non-RAID SSA adapter does. It transfers data directly between the disk drives and the system, and uses no RAID functions.



---

## Part 5. Storage area networks (SANs)



---

## Chapter 10. Introduction to SAN

For any organization using information technology to help run its business, data rapidly becomes an important asset. It is an underlying resource on which the organization's computing processes are based. As such, storage technology is critical to the success of data-intensive computer environments. With the ever-growing importance of computers to business operations, organizations are faced with the problem of managing growing data needs, data protection, and availability.

In a host-based computing system, the management of storage is centralized in that storage devices are connected directly to the host and managed directly from the host. This is the traditional data-processing environment in which it is relatively easy to manage storage.

With the advent of distributed computing, client/server computing and internet technology, together with downsizing and rightsizing trends, today's business environment has been transformed into a highly competitive, expanding, and unpredictable networked environment.

While the flexibility of distributed systems is attractive, managing a set of servers spread across a geographically dispersed network can present problems. Because of this, there is growing focus on server and storage consolidation. As a way to address this issue, storage area network (SAN) approaches to storage have been developed to help businesses manage, track, and more easily share the complex and ever-increasing volume of data created by the Internet and e-business applications, and the emergence of data-intensive technologies such as multimedia and data warehousing.

In this chapter, we will look at the shortfalls of traditional storage architectures, then move on to define storage area networks and the way they address some of the problems. We examine the building blocks of SAN, and how solutions can be designed in a SAN environment, discussing the benefits and challenges faced when implementing SAN technology.

**Note**

Much of the discussion in this chapter is generic and conceptual in nature. Configurations depicted here should not be taken as examples of functional SAN implementations. We explore Netfinity-specific products in Chapter 11, "SANs and Netfinity servers" on page 293.

## 10.1 Traditional storage architectures

Many businesses have implemented standard Intel server-based systems to provide computing resources where they are required. Figure 164 shows a simple example of the typical network environment in which these servers often operate. The figure shows only three servers, but the network environment for a large enterprise may have hundreds or even thousands of servers distributed over a number of sites. The hard disks attached to all these servers can add up to a huge amount of storage space, terabytes (TB) or petabytes (PB) of data capacity, containing vital business information that is spread across the entire network.

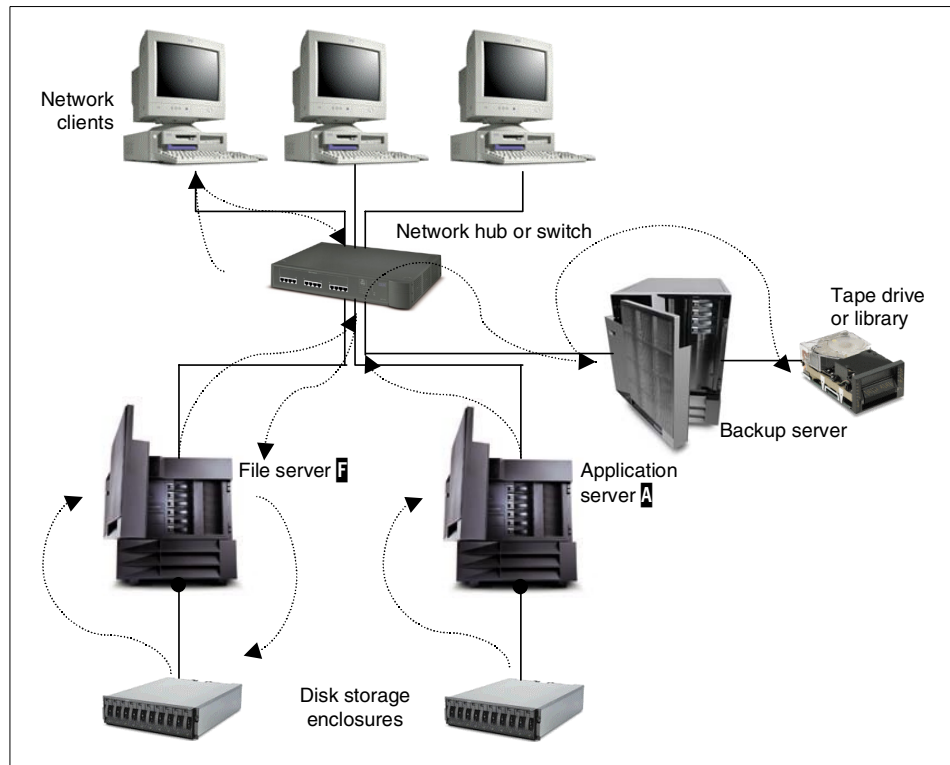


Figure 164. Typical disk and tape storage network resources

#### n-bytes definition

- 1 byte (**B**) is equivalent to 8 bits (**b**)
- 1 kilobyte (**KB**) is equivalent to 1024 B
- 1 megabyte (**MB**) is equivalent to 1024 KB
- 1 gigabyte (**GB**) is equivalent to 1024 MB
- 1 terabyte (**TB**) is equivalent to 1024 GB
- 1 petabyte (**PB**) is equivalent to 1024 TB
- 1 exabyte (**EB**) is equivalent to 1024 PB

### 10.1.1 Problems arising with server-attached storage

Server-attached storage has been the standard method for providing storage resources for a network for some time, at least in the Intel-based server marketplace. There are a number of potential issues with this approach, however, which become more pressing as system complexity increases.

#### **Data availability**

In an environment such as that shown in Figure 164, the servers may, for example, host applications or provide file services and are connected directly to disk storage devices that hold the data used by the clients. That is, we have server-attached storage in which data access is dependent on the server. If a server fails, all disk storage devices connected to that system will be unavailable. We can term this a *data availability* problem. SAN techniques isolate the storage from the servers to prevent this.

#### **Network performance**

If a client PC is running an application from the application server (A in Figure 164) that requires data from the storage connected to the file server (F), server F will fetch the required data and pass it to server A over the local area network, and, perhaps, subsequently to the client PC. Typically, servers A and F manage other server functions as well, such as user authentication, file and print, and additional application processing. In the event server F is busy with another function, server A must wait until it is available to service its request. This can create a *network performance* problem as the number of servers and clients grow. Having a separate network dedicated to storage traffic, as in a SAN, removes this traffic from the user network.

#### **Backup overhead**

During data backup operations, the network performance may be more severely affected as the backup server transfers data from other servers to back up data on the disks attached to them. The network impact can be eliminated by attaching a backup device directly to each server, but as the number of servers grows this will increase the overall hardware and

management costs. This can be called the *backup overhead* problem, which is addressed by implementing tape pooling in a SAN environment.

### **Connectivity issues**

SCSI-attached disks are widely used in current storage technology, thanks to their reliability and relatively low cost. However, they do impose configuration limitations. At best, you are limited to a maximum of 25 meters from the server's SCSI adapter to the attached disk storage devices and, depending on the exact form of SCSI used, this can be reduced to just a meter or two. This is a problem of *connection distance* between storage and server that can make configuration of large capacity disk subsystems difficult. Storage area networks are designed to allow storage to be accessed over large distances.

### **Storage scalability**

Additionally, a single SCSI bus can only connect a maximum of 15 devices and thus, while today's adapters support several channels, large configurations may require more adapters to be installed in the server. This will increase the cost of managing the server as the configuration is more complex, and there is a limit to the number of adapters that can be installed in a system. We can call this a *scalability* problem. SANs can typically support much larger disk configurations than is possible with more traditional approaches.

### **Storage consolidation**

As the requirement for storage grows, more disks may be added to the server but when the limit is reached, additional servers are required to accommodate the growth in storage needs even though the available server processing power may still be adequate. If servers and disk storage devices are acquired from various computer vendors, compatibility may be an issue. These issues can be described as *server and storage consolidation* problems. Generally, SANs often offer connectivity to a single storage repository by multiple hosts, with data for each server being partitioned for use by the owning system, to address this type of problem.

### **Heterogeneous systems**

In a typical organization, there may be more than one operating environment in use, based on the applications used. For example, you may have your e-mail and file and print servers running on either Windows NT or Windows 2000, your business applications running on OS/400, and Internet servers running on UNIX, and so on. Each type of system has its own operating environment and its own disk storage devices. Currently, data held on a single disk storage subsystem may be shared in a homogeneous environment but not by different operating environments, since the underlying operating technologies are proprietary. We term this the *heterogeneous data sharing*



problem. The partitioning offered by SANs can allow support of multiple operating systems.

### **Storage management**

As the number of servers in a network grows, the effort required to manage them will increase non-linearly as more network components (such as hubs, switches, tape drives and libraries, cabling and so on) are required. As the number of nodes and devices increase, the cost of managing them will increase. This is the *manageability* problem. Server and storage consolidation, enabled by SAN approaches to storage, can alleviate the management load.

In the following section, we look at what storage area networks are and at what is driving the move towards SANs. We also expand on the way in which SANs can help to address and overcome the problems we have defined in the preceding paragraphs.

---

## **10.2 What is a storage area network?**

Storage area networks (SANs) have been developed to address the problems described in the previous section. A simple description of an idealized SAN is a dedicated, centrally managed, high performance storage infrastructure, providing any-to-any interconnection between storage devices and a portfolio of heterogeneous servers.

SANs provide storage devices and services in a highly available, scalable and manageable fashion. However, let us state clearly that a SAN, as such, is not a product that is readily available. Instead, a SAN could be defined as an evolution of multiple products, designed to address various customer requirements. The goal is that, ultimately, all the problems described above, and perhaps others, will be addressed by SANs using a collection of hardware, software and specialized interconnect technologies. Current SAN implementations offer some, but not yet all, of their promised benefits.

SANs provide a method of attaching storage that revolutionizes the network because of the improvement in availability and performance. They are currently being used in several different ways:

- To provide shared storage arrays for multiple hosts.
- To provide a shared storage solution for high availability clusters.
- They have been used for several years for distributed mainframe storage and tape attachment.
- To provide high performance and highly available storage.

In essence, SANs are another network technology, similar to a subnet, but constructed from storage interfaces. SANs enable storage to be physically and logically separated from the server and, in doing so they allow storage to be shared among multiple host servers without impacting either system performance or the primary network. A SAN is sometimes referred to as the “network behind the server”.

**SAN building blocks**

The building blocks of a SAN-based network comprise four main elements, namely servers utilizing SAN-based resources, SAN software, SAN interconnects, and SAN storage as illustrated in Figure 165:

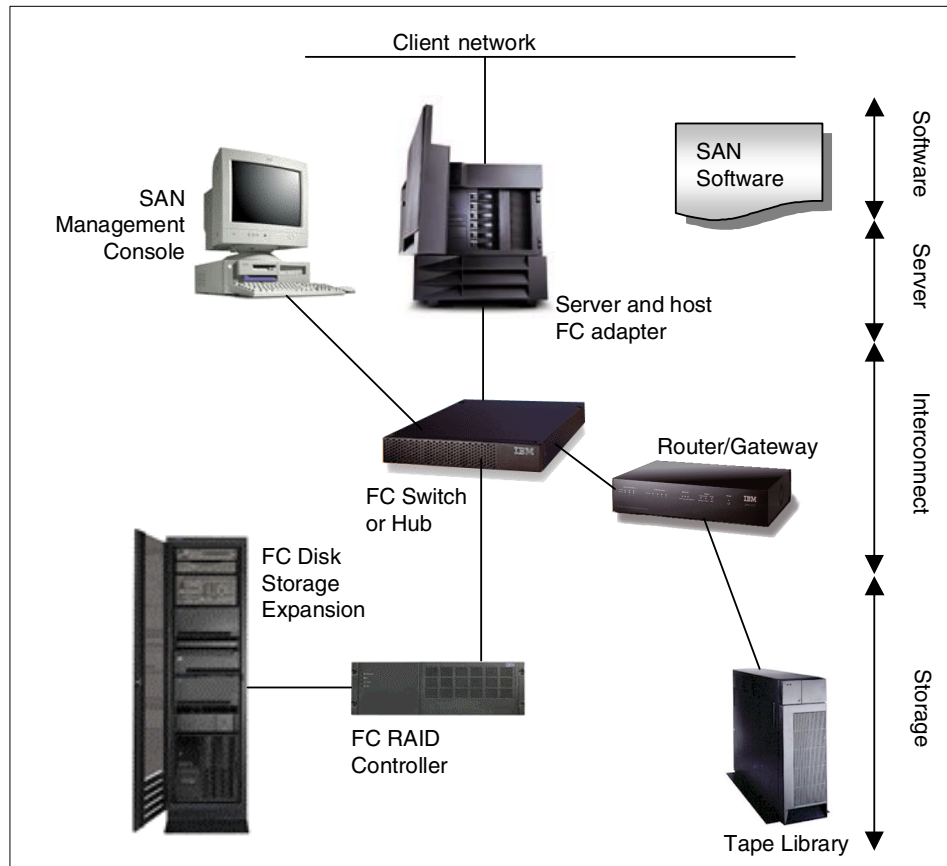


Figure 165. SAN building blocks

Each of these elements is important in creating a storage area network.

- SAN servers are the processing resources for all user requests and data access requirements in SAN-based computing environments, just as they are for any network. They will serve as file and print, mail, Internet, multimedia, and application and database processing servers in the primary LAN, and interact with the SAN-based storage for access to the required information.
- SAN software is an enabler of intelligent storage and plays a significant role in driving the adoption of SAN technology.
- SAN interconnects are another key element in SAN architecture. They comprise network components such as extenders, multiplexers, hubs, gateways, routers, switches and others that are already in use in local and wide area networks today.
- SAN storage is a shared repository, attached to SAN interconnects using interfaces such as SCSI, SSA or FC-AL to form a separate network specifically for storage access. Storage components that may be used in a SAN environment include disks and disk subsystems, tape drives and libraries, intelligent storage servers, and others.

### ***SAN data transfer***

A SAN bypasses traditional network bottlenecks and supports direct, high-speed data transfers between servers and storage systems in three different ways, namely *server-to-storage*, *server-to-server* and *storage-to-storage*. These data paths are shown in Figure 166:

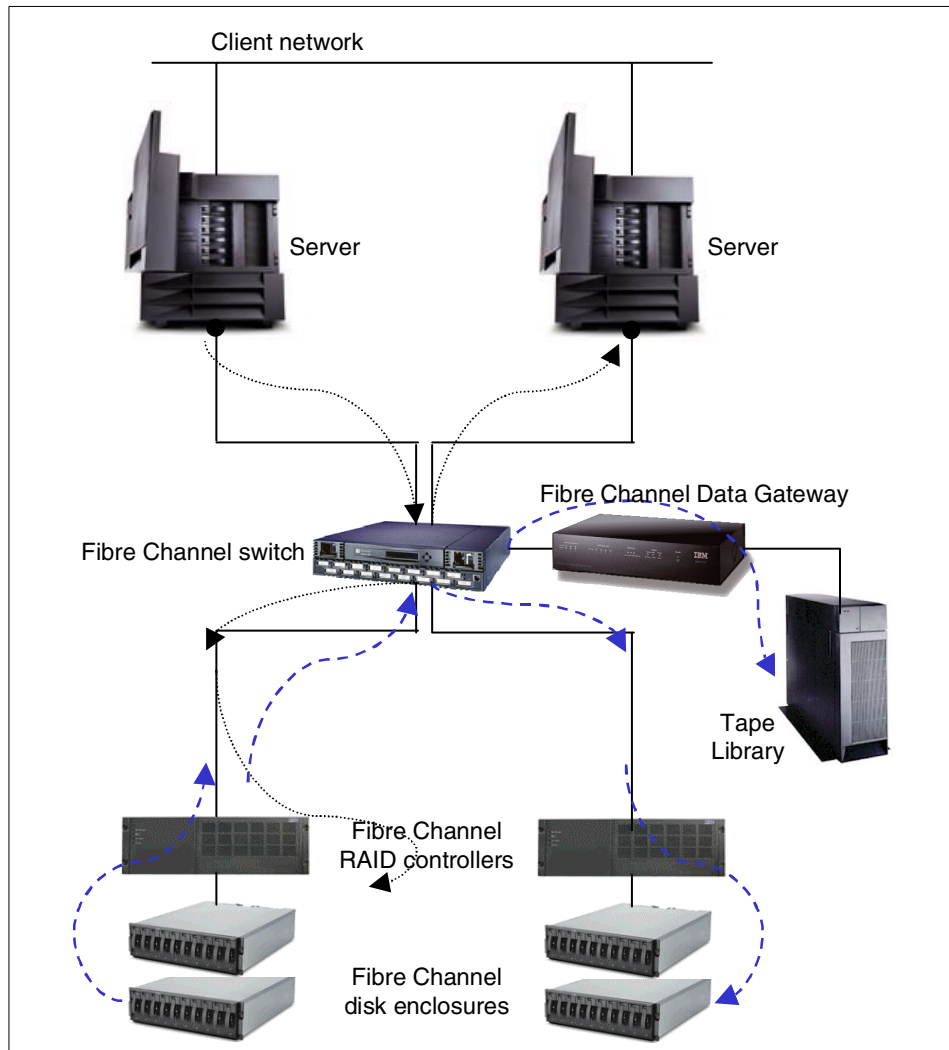


Figure 166. SAN, the network behind the server

The diagram shows a simple storage area network attached to two servers.

- In a server-to-storage access environment, storage devices (the tape library and disk enclosures) are connected to the SAN fabric via the Fibre Channel switch. They are accessed by the two servers through the SAN fabric (a term commonly used for the set of interconnects). This allows all the storage devices to be shared and accessed by multiple servers connected to the SAN without affecting the public LAN.

- In a server-to-server access environment, the two servers communicate through the SAN fabric by way of the Fibre Channel switch. The SAN fabric may be used as a high performance interconnect for servers, high performance workstations, or thin-client servers. As SANs are evolving, protocols such as IP are now beginning to be supported, opening up a wide variety of potential applications for this type of environment.
- In a storage-to-storage access environment, the SAN-attached storage devices (the tape library and disk enclosures) may interact with each other through the SAN fabric. These are serverless transactions that enable data to be moved from one storage device to another without intervention from the servers. This is particularly relevant for tape backup, which will require minimal server involvement beyond initiating a backup process and acknowledging its completion. Storage-to-storage data transfers allow high performance data mirroring, backup and recovery.

### ***SAN storage attachment***

Storage resources in a network environment can be broadly placed in one of three classes:

- Server-attached storage
- Network-attached storage
- SAN-attached storage

Figure 167 on page 288 shows a server having access to storage available in each of these ways.

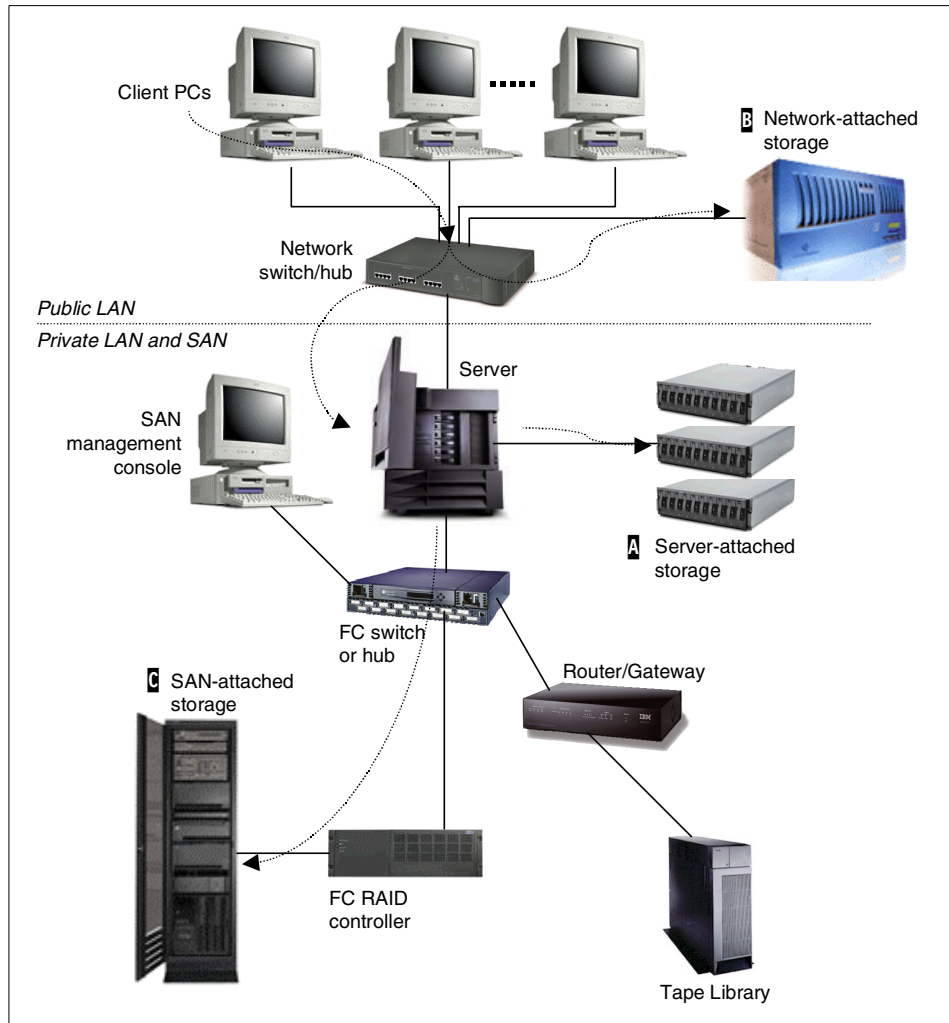


Figure 167. Types of storage attachment

In a traditional client/server network, storage devices are directly connected to the server, typically using the SCSI interface. This type of attachment is commonly known as server-attached storage (A) in the figure). In this environment, the availability and performance of the storage devices is determined by the hosting server's capability and loading.

Network-attached storage (B) is connected directly to the public LAN, and communicates with requesting devices using network protocols such as Ethernet or token-ring. Typical implementations include a processor residing

in the storage devices which responds to requests for data. The storage is separate from the network server but is connected to the same network and functions as a thin server, processing requests for data. The network-attached storage may be accessed serially or concurrently by multiple servers through the public LAN. In this environment, the availability and performance of the storage subsystem are dependent on the network-attached storage itself. To optimize these factors, the dedicated thin server can be fine-tuned for the specific single task it has to perform and the storage can include redundancy features.

Since network-attached storage is connected directly to the public LAN, it will have an impact on the public LAN performance during data movement, but offloads data processing from the server.

SAN technology offers the advantages of network-attached storage (NAS), but also avoids moving storage traffic over the public LAN. Storage devices are no longer connected to the server directly as in server-attached storage, but through shared hubs or switches. This creates a SAN-centric environment, in which data is stored, shared, and accessed from a storage subsystem by one or multiple servers.

SAN-attached storage (C in Figure 167) allows data transfer between storage devices, such as from one disk to another, and backing up or restoring data from a tape device without server intervention. The data transfer will take place through the SAN fabric comprised of Fibre Channel hubs, switches, and so on. This offloads both the public LAN and servers for high performance back-end solutions such as disaster protection, remote mirroring, serverless backup, and storage consolidation.

In a nutshell, a SAN is a data-centric environment where storage is accessed by way of an independent, high performance and highly available network. It is also highly scalable and centralized. Since the storage is in its own private network, traffic in the primary network is no longer an issue. SAN technology is becoming critical to the success of data-intensive computer environments. To build a flexible and secure storage environment takes a good understanding of your own storage needs and how to leverage current technology to meet those requirements.

---

### 10.3 SAN challenges

The advent of Internet technology in the last few years has transformed today's business environment to cope with a highly competitive, expanding, unpredictable, networked economy. This leads to the need to adapt your

computing environment so that information may be exchanged freely and immediately.

The Internet explosion and the emergence of other data-intensive and data-centric applications such as multimedia, enterprise resource planning and business intelligence has created a phenomenal growth in storage capacity requirements. In parallel with this, the need to manage these resources virtually anywhere, anytime, and to be able to share information across storage networks regardless of systems and applications suppliers has come to the fore.

Data storage systems have become mission-critical. The demands of e-business lead to the need for the resources to be available 24 hours a day, seven days a week in multiple time zones. The growing need for storage capacity is challenging traditional storage technologies. The move to SAN has been driven by the need to manage the dramatically increasing volume of business data, and to mitigate its effect on network performance.

In today's networks, the growth in storage capacity is coupled with an increasing number of users accessing data through the network, and limited growth of network bandwidth. Data access, backup, and recovery slow down the network tremendously and affect end users across the enterprise network.

This transformation requires organizations to have ability to continually adapt, while immediately accessing and processing information to drive successful business decisions. To gain competitive edge by exchanging information immediately, storage network needs to be flexible, responsive and reliable.

---

#### **10.4 SAN benefits**

SANs allow applications that move data to perform better by sending data directly from source to target without server intervention. They also enable new network architectures, where multiple hosts can access multiple storage systems connected to the same storage area network.

Client network loading is reduced as data traffic for backup processes is removed from the network, thus giving IT managers a strategic way to improve system performance and application availability.

By using Fibre Channel connections, SANs provide high-speed network communication and attachment at the distances needed for remote workstations and servers to easily access the shared data storage pools. These same features allow you to centralize storage systems and consolidate



backups, thus increasing overall system efficiency. The increased distance provided by Fibre Channel technology makes it easier to deploy remote disaster recovery sites. Fibre Channel and switched fabric technology can eliminate single points of failure on the network.

With a SAN, virtually unlimited expansion is possible using hubs and switches. Nodes can be removed or added with minimal disruption to the network.

In summary, using SAN technology can potentially offer the following benefits:

- Improved application availability as storage becomes separated from processors, independent from applications, and accessible through alternate data paths as found in clustered systems.
- Improved application performance as storage processing is offloaded from servers and moved onto a separate network.
- Centralized and consolidated storage for simpler management, scalability, flexibility, and availability.
- Data transfer and vaulting to remote sites as remote copying of data is enabled for disaster protection.

---

## 10.5 SAN evolution

SANs for Intel-based servers are in the early stages of evolution. Fibre Channel technology provides a foundation for today's SAN solutions, since it offers the key features of performance, distance, reliability, and scalability, but you can expect to see continuing innovation in SAN capabilities in the future.

It is expected that the features richness of SANs will continue to grow with the advent of intelligent disk subsystems, the increasing capability of SAN software, and the growing acceptance of Fibre Channel technology.



---

## Chapter 11. SANs and Netfinity servers

In this chapter, we will look at Netfinity's initiatives toward storage area networks (SANs), its challenges and building blocks in terms of SAN software, SAN servers, SAN interconnects, and SAN storage.

Based on the Netfinity SAN solutions, we will also look at where Netfinity SAN is on the SAN evolutionary scale in comparison with the SAN features and functions defined in the earlier chapter.

### Supported configurations

Configurations described in this chapter are for example only. The flexibility of Fibre Channel subsystems allows many different configurations to be implemented. You should check the IBM Netfinity Fibre Channel Web site for current information about supported configurations:

<http://www.pc.ibm.com/ww/netfinity/fibrechannel/>

---

### 11.1 Netfinity server SAN challenges

IBM has recently added a number of products to augment its family of Netfinity servers that allow SAN approaches to providing storage resources. This initiative is intended to help businesses manage, track and more easily share an ever-increasing volume of data largely created by the combined impact of the Internet, e-business, multimedia, enterprise resource planning and business intelligence applications.

The Netfinity server SAN products provide the next step in the pathway towards centrally managed, open software and hardware solutions designed to help companies get the most value out of their business information and IT infrastructures. The initiative is based on the following objectives:

- Open network architecture for deploying data access and data sharing capabilities across the enterprise.
- Consolidation of servers and storage.
- Increased data availability.
- Centralized storage management.
- The ability to back up and migrate data without affecting enterprise network performance.
- Increased reliability offered by cluster technology.
- Security and protection of data in the event of disaster.

Netfinity servers are poised to deliver on the vision of SAN technology by combining the best of IBM and industry-standard technologies using IBM's experience from decades of mainframe, UNIX and Intel processor-based computing in cooperation with strategic industry partners.

---

## 11.2 Netfinity SAN components

As illustrated in Figure 168 on page 295, the major components of a storage area network are software, servers, interconnects, and storage. The building blocks of Netfinity storage are networks are as follows:

- Netfinity SAN software is an enabler that consolidates SAN storage, LAN-free backup and restore operations, and so on by utilizing available management software such as Tivoli Storage Manager, Tivoli SANergy, Netfinity Fibre Channel Storage Manager, HP SAN Manager LM, and others. SAN software continues to develop and holds the promise of a high performance, highly scalable and highly available storage subsystem, offering features such as serverless data movement, heterogeneous data sharing and more.
- Netfinity servers are high performance, scalable systems that offer features, such as redundant options and clustering, that make them ideal for use in a SAN environment.
- Netfinity SAN interconnects provide the connectivity and performance necessary to implement powerful SAN-based systems. Products such as the IBM SAN Fibre Channel switch, IBM SAN Fibre Channel Managed Hub, Netfinity Fibre Channel hub, IBM SAN Data Gateway Router, and their associated cabling systems let you implement a high performance, reliable SAN infrastructure.
- Netfinity SAN storage is a shared storage repository comprising redundant Fibre Channel RAID controllers, Fibre Channel and SCSI disk enclosures, and tape drives and libraries.

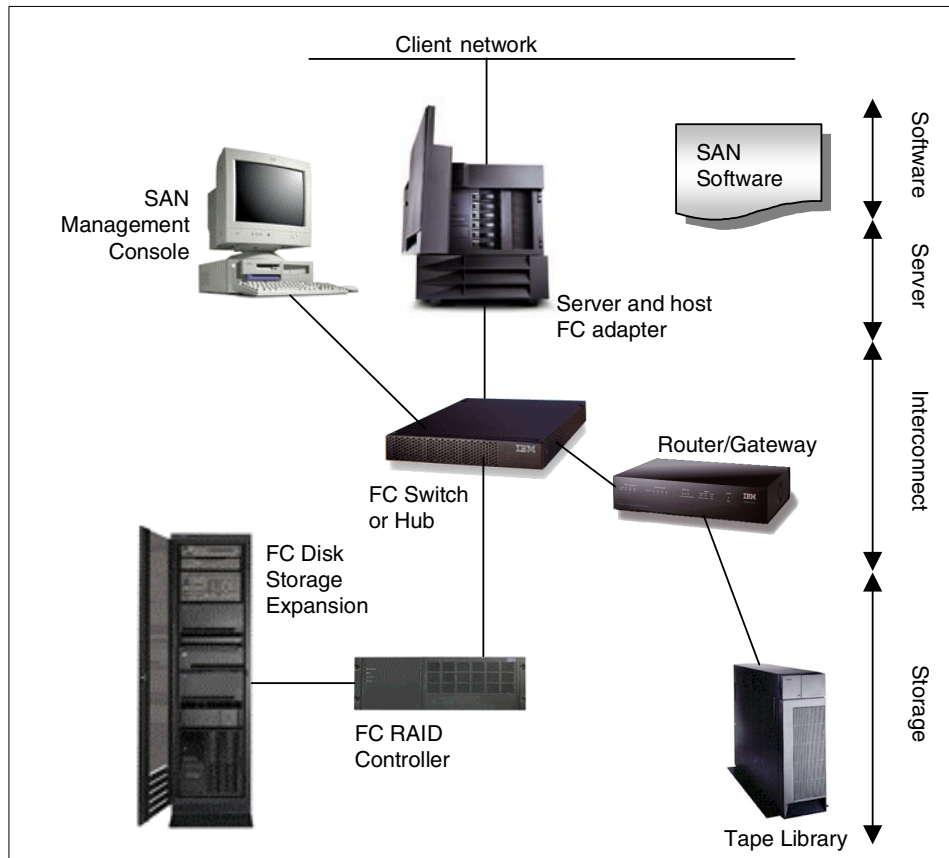


Figure 168. Netfinity SAN components

You will find a discussion of the major components that enable you to implement SANs for Netfinity servers in Part 3, “Fibre Channel subsystems” on page 165. In the following section we examine some of the software that will help you to implement a SAN.

### 11.2.1 SAN software for Netfinity servers

SAN software enables intelligent storage by moving away from the traditional architecture that directly attaches storage devices to their host servers. SAN-based solutions that can be implemented include disk pooling, tape pooling, incremental storage growth, and LAN-free data transfer for data backup, archiving and recovery.

SAN-based solutions and their associated software for Netfinity servers are as follows.

- The combination of Computer Associates (CA) ARCserveIT and CA ARCserveIT Enterprise Library Option enables tape library sharing.
- The combination of Hewlett Packard SAN Manager LM, CA ARCserveIT, and CA ARCserveIT Enterprise Library Option enables disk subsystem and tape library sharing.
- The combination of Legato Networker and Legato SmartMedia enables tape library sharing.
- The combination of Tivoli SANergy, VERITAS BackupExec and VERITAS BackupExec Shared Storage Option enables file sharing and tape library sharing.
- The combination of HP SAN Manager LM and Tivoli Storage Manager enables disk subsystem sharing and tape library sharing.
- The combination of Tivoli SANergy and Tivoli Storage Manager enables file sharing and tape library sharing.

This list is not exhaustive and new products may be added over time. For the latest news, you should visit the IBM SAN Web site at:

<http://www.storage.ibm.com/ibmsan/>

Netfinity server-specific SAN-related information can be found at:

<http://www.pc.ibm.com/us/compat/san/index.html>

A brief description of the software tested and planned for Netfinity SAN solutions is given in Table 27:

Table 27. Netfinity SAN software components

Netfinity server with FC-attached tape storage	Netfinity MS Cluster with tape pooling	Netfinity server consolidation with tape pooling	Netfinity server disk and tape pooling	IBM Availability Extensions for MSCS	Netfinity Advanced Cluster Enabler for OPS	Netfinity server and storage consolidation	
Tested or Planned as of December 1999							
T	T	T	T	T	P	P	Windows NT 4
T	T	T	T	T			Windows NT 4 Service Pack 5
T			T				CA ARCserveIT 6.61
			T				CA ARCserveIT Enterprise Library Option 6.61
			T				HP SAN Manager LM 1
				T			IBM Availability Extension for MSCS
				T			IBM DB/2 Universal Database (EEE) 5.2
				T			IBM Lotus Domino 5
							IBM Lotus Notes 5
							IBM StorWatch ESS Expert
						P	IBM StorWatch ESS Specialist
T	T		T			P	IBM StorWatch SAN Data Gateway Specialist 1.3
	T	T	T	T	P	P	IBM StorWatch SAN FC Switch Specialist
T			P				Legato Networker 5.51
			P				Legato SmartMedia
			T				Tivoli SANergy 1.6
					P		Oracle Parallel Server 1.5
T	T	T	T	T	P		Netfinity FC RAID Support for Windows NT 06.22.25
P	P	P	P	P	P		Netfinity FC Storage Manager 7
T	T	T	T			P	Tivoli Storage Manager 3.7
T	T		T				VERITAS BackupExec 7.3
	T		T				VERITAS BackupExec Shared Storage Option 7.3
T							VERITAS NetBackup 3.2

#### **11.2.1.1 ARCserveIT and ARCserveIT Enterprise Library Option**

ARCserveIT automates every traditional storage management task, lowers operating expense, reduces errors, and enables administrators to easily perform essential data protection routines. The backup and restore operations can be managed from one central location, virtually anywhere on the network, to provide fast and efficient operation.

Computer Associates (CA) has added SAN management capability to the Unicenter TNG Framework, which is an integral part of all of CA's IT products, including ARCserveIT, and delivers an open platform for SAN device management, from server to storage device. The ARCserveIT Enterprise Library Option enables users to exploit Fibre Channel technology, the key enabler of SAN solutions.

For more information on ARCserveIT, visit the CA Web site at:

<http://www.cai.com>

#### **11.2.1.2 Hewlett Packard SAN Manager LM**

Hewlett Packard (HP) SAN Manager LM manages Fibre Channel attached storage in a highly efficient manner as an administrative cluster. The product allows heterogeneous or homogeneous systems to utilize a common pool of storage devices on the SAN. Administrators can assign storage from the pool when and where it is needed in a matter of seconds through a simple drag-and-drop interface.

Using HP SAN Manager LM, storage can be added to the pool or assigned to nodes (servers) without reboots. Storage is available for immediate use because it is mounted automatically when assigned. These features eliminate the costly downtime normally associated with adding and moving storage.

For more information, visit the HP SAN Web site at:

<http://hpsanmanager.com>

#### **11.2.1.3 Legato Networker and Legato SmartMedia**

Legato Networker manages distributed data, offering heterogeneous platform support, automated media handling, interoperable tape format, data stream parallelism, and remote tape management.

Legato SmartMedia is an open media management application that provides standard interfaces for applications, robotic library control, drive control, and administration, thereby enabling enterprises to better manage their growing base of removable media and devices distributed across a heterogeneous storage environment.



For more information about Legato products, visit their Web site at:

<http://www.legato.com>

#### **11.2.1.4 IBM Netfinity Availability Extensions for MSCS**

IBM Netfinity Availability Extensions (NAE) for Microsoft Cluster Server (MSCS) complements and extends the capability of MSCS from the standard two-node configuration to an up to eight-node, any-to-any failover solution.

At present, NAE supports IBM's DB2 Universal Database (UDB), Lotus Domino Server and Microsoft Windows NT File and Print services. It is planned to add support for more applications over time.

For more information about NAE, visit IBM Web site at:

<http://www.pc.ibm.com/us/solutions/netfinity/index.html>

#### **11.2.1.5 IBM StorWatch**

IBM offers a number of products that are components of its Enterprise Storage Server suite. Briefly, their names and functions are:

- IBM StorWatch Enterprise Storage Server (ESS) Expert manages the performance and the total assigned and free capacity of disk storage within an enterprise. It also identifies which volumes in the storage are allocated to or shared by SCSI-attached servers.
- IBM StorWatch ESS Specialist is a storage management tool that enables storage administrators to centrally monitor and manage the ESS. Using commonly available Web browsers, you may manage from the IBM StorWatch ESS Specialist from anywhere at work, home, or on the road via a secure network connection.
- IBM StorWatch SAN Data Gateway Specialist provides tools to manage any-to-any access between Fibre Channel ports and SCSI ports by defining data gateway zoning, virtual private SAN, and LUN-masking, and controls which host systems have access to specific storage devices.
- IBM StorWatch SAN Fibre Channel Switch Specialist provides a comprehensive set of management tools that support a Web browser interface for flexible, easy-to-use integration into existing enterprise storage management structures. It also provides security and data integrity by zoning host system attachment to specific storage systems and devices.
- IBM StorWatch Fibre Channel RAID Specialist is a network-based integrated storage management tool that helps storage administrators configure, monitor, dynamically change, and manage multiple Fibre

Channel RAID subsystems. High availability and full redundancy are provided with the host-specific Fibre Channel Storage Manager software, which resides on the host system and provides automatic I/O path failover when a Netfinity Fibre Channel Host Bus Adapter, IBM SAN Fibre Channel switch, Netfinity Fibre Channel hub, or Netfinity Fibre Channel RAID Controller unit fails.

For more information about these products, visit the IBM Web site at:

<http://ssdweb01.storage.ibm.com/software/storwatch/swovis.htm>

#### **11.2.1.6 Tivoli SANergy**

Tivoli SANergy software allows you to dynamically share files on SAN-based storage using standard network and file systems at 100 MBps simultaneously using Fibre Channel, SCSI or SSA. It supports multiple Windows NT, Macintosh and UNIX systems concurrently sharing SAN-based storage.

This product was formerly known as Mercury SANergy. On December 14, 1999, IBM announced the purchase of Mercury's Shared Storage Business Unit (which owned SANergy) and this is now part of IBM's Tivoli Systems.

For more information about SANergy, visit this Web site:

<http://www.sanergy.com>

#### **11.2.1.7 Tivoli Storage Manager**

Tivoli Storage Manager provides the only application-centric approach to information management by delivering true end-to-end solutions spanning the entire enterprise. Tivoli Storage Manager is an enterprise-wide solution integrating automated network backup, restore and archive, storage management and disaster recovery functions for PC servers to midrange and mainframe host servers.

Tivoli Storage Manager exploits SAN infrastructures by providing LAN-free data transfer over traditional IP and Fibre Channel networks for backup and recovery, and tape resource sharing.

Tivoli has outlined the Tivoli SAN management solutions listed below. Those not available now are due to be delivered through 2001:

1. Tape resource sharing and LAN-free data transfer through Tivoli Storage Manager on September 20, 1999.
2. Serverless backup is targeted for second quarter 2000.

3. SAN management tools based on Tivoli NetView management software scheduled for third quarter 2000.
4. Tivoli SAN Disk management, in which heterogeneous servers address the same disk device, is targeted for fourth quarter 2000.
5. Tivoli SAN Data management is scheduled for 2001.

Tivoli Systems is an IBM company providing the industry's leading open, highly scalable and cross-platform IT management solutions that span networks, systems, applications and e-business to business solutions.

For more information about Tivoli products, visit their Web site at:

[http://www.tivoli.com/products/index/storage\\_mgr](http://www.tivoli.com/products/index/storage_mgr)

#### **11.2.1.8 VERITAS**

VERITAS BackupExec provides high performance, reliable Windows NT data backup and restore. It incorporates data protection and storage management for complex messaging and transactional intensive databases, and storage system management and server recovery from a centralized backup server.

VERITAS BackupExec Shared Storage Option allows multiple distributed backup servers to share a common, centralized storage device connected over a Fibre Channel network. It balances the processing load across multiple backup servers, increases performance and backup speed, and provides centralized management.

VERITAS NetBackup automates enterprise backup operations for thousands of users across multiple servers and consolidates management of all storage devices. It supports heterogeneous environments including Windows NT, UNIX, Novell, Macintosh and others.

For more information about VERITAS products, visit their Web site at:

<http://support.veritas.com>

### **11.2.2 Netfinity servers**

IBM Netfinity X-architecture is a blueprint that plans the migration of features such as management capabilities from larger IBM systems such as RS/6000, AS/400 or S/390 and adapts them to the Netfinity server environment.

Capabilities that were once only in the domain of these more costly systems are now being demanded by customers using Intel-based systems. Examples include:

- S/390 mainframes for proven high availability of 99.999%.
- RS/6000 midrange servers offer scalability of 512- to 1024-node clusters of 8 and, soon, 12-way SMP systems.
- AS/400 midrange servers for solution partnerships and self-maintaining capabilities such as auto-tuning and auto-configuration.

A SAN is intended to provide high performance, high availability, and scalability, so the recommended Netfinity servers for a SAN environment are the midrange Netfinity 5000 and larger machines. These servers adopt the Netfinity X-architecture, which includes powerful processors, reliable and highly available memory systems, scalable I/O, advanced caching software and world-class silicon and module technology.

For example, the top-of-the-range Netfinity 8500 supports up to 8-way SMP with the power of Intel's fastest processors, 2 MB of level 2 cache and up to 16 GB SDRAM, and comes in a rack or tower form factor. Most Netfinity servers also provide hot-pluggable hard disks, PCI adapters, fans and power supplies; features that keep your server running even when components fail. Future X-architecture features planned include hot-pluggable core components, such as memory and processors. By adding Fibre Channel-attached storage options for scalable, highly available, cluster-enabled storage, improved security and disaster protection, Netfinity servers become an excellent platform for implementing SAN solutions such as server consolidation, clustering, e-business or enterprise resource planning.

Table 28 shows those Netfinity servers tested and planned as the basis of a number of SAN solutions. This list is not exhaustive and new products may be added over time.

For latest news about SAN solutions, visit the IBM Web site at:

<http://www.storage.ibm.com/ibmsan/>

Table 28. Netfinity SAN servers

Netfinity Server with FC Attached Tape Storage	Netfinity MS Cluster with Tape Pooling	Netfinity Server Consolidation with Tape Pooling	Netfinity Server Disk and Tape Pooling	IBM Availability Extensions for MSCS	Netfinity Advanced Cluster Enabler for OPS	Netfinity Server and Storage Consolidation	
Tested or Planned as of December 1999							
T	T	T	T			P	IBM Netfinity 5000
T	T	T	T			P	IBM Netfinity 5500
T	T	T	T			P	IBM Netfinity 5500M10
T	T	T	T			P	IBM Netfinity 5500M20
P	P	P	P			P	IBM Netfinity 5600
T	T	T	T	T	P	P	IBM Netfinity 7000M10
T	T	T	T			P	IBM Netfinity 8500R

### 11.2.3 Netfinity SAN interconnects

Interconnection components such as cables and connectors, converters, adapters, extenders, multiplexors, hubs, routers, bridges, gateways, switches and directors are used for local area network (LAN) or wide area network (WAN) implementations. SAN, like LAN or WAN, interconnects storage interfaces together in many network configurations and over increasingly greater distances.

In this section, we will list the Netfinity SAN Fibre Channel interconnection components that handle the connection of storage devices to Netfinity servers. SAN fabric hardware includes the IBM SAN Fibre Channel switch, the IBM SAN Fibre Channel Managed Hub, the Netfinity Fibre Channel Hub, the IBM SAN Data Gateway Router and the GBICs that allow fiber cabling to connect to some of the hardware.

Hardware descriptions can be found in 6.2, "Netfinity Fibre Channel hardware" on page 171.

Table 29 shows those Netfinity SAN interconnects tested and planned for Netfinity SAN solutions. This list is not exhaustive and new products may be added over time.

For the latest information, you may visit the IBM Web site at:

<http://www.storage.ibm.com/ibmsan/>

Table 29. Netfinity SAN interconnects

Netfinity Server with FC Attached Tape Storage	Netfinity MS Cluster with Tape Pooling	Netfinity Server Consolidation with Tape Pooling	Netfinity Server Disk and Tape Pooling	IBM Availability Extensions for MSCS	Netfinity Advanced Cluster Enabler for OPS	Netfinity Server and Storage Consolidation	
Tested or Planned as of December 1999							
T	T			T	P		Netfinity Fibre Channel Hub, 3523-1RU
	T	T	T	T	P	P	IBM SAN Fibre Channel Switch, 2109-S16
	T	T	T	T	P	P	IBM SAN Fibre Channel Switch, 2109-S08
T	T	T	T				IBM SAN Data Gateway Router, 2108-R3S
	P	P	P	P			IBM SAN Data Gateway Router, 2108-R3D
						P	IBM SAN Data Gateway, 2108-G07
T	T	T	T	T	P	P	Netfinity Fibre Channel Short Wave GBIC
T	T	T	T	T	P	P	Netfinity Fibre Channel Long Wave GBIC

### 11.2.4 Netfinity SAN storage

SAN-ready disk and tape storage devices include the Netfinity FAStT500 RAID Controller, the Netfinity FAStT EXP500 Storage Expansion unit, and the IBM Magstar Tape Library 3570-C22.

Details of the Fibre Channel disk subsystems can be found in Chapter 6, “Introduction to Fibre Channel” on page 167. For information about Magstar products, visit the IBM Web site at:

<http://www.storage.ibm.com/ibmsan/>

Netfinity servers may also be connected to IBM's Enterprise Storage Server (ESS), which is the third generation of the Seascape architecture for disk systems. This system has many autonomous capabilities that offload key functions from the host servers it supports, making it ideal for SAN solutions. ESS provides extensive heterogeneous server connectivity, high availability, and large capacity, features that make it ideal for consolidation purposes and to support the growing needs of e-business. For more information on ESS products, see:

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

---

## 11.3 Netfinity SAN solutions

Netfinity SAN solutions are still at an early stage in their evolution. As enterprise storage requirements continue to increase, especially as more companies are moving toward a “zero latency” business environment, Netfinity server SAN products will continue to enhance their performance, capacity and reliability as Fibre Channel technology matures.

At this stage, Fibre Channel products are able to enhance Netfinity SAN by improving storage scalability, by offering server and storage consolidation, and by increasing data transfer and backup performance. As technologies evolved over time, more solutions and intelligent devices will be developed and thus the full promise of SANs will be realized.

In this section, we briefly describe several currently available Netfinity SAN solutions.

### 11.3.1 Storage consolidation

In a typical network environment today, large organizations may have hundreds or thousands of servers distributed over a number of sites across the nation or internationally. Data resides on the server's internal disk storage and, if greater capacity is required, external storage expansion units may also

be required. The disks will usually be attached to a SCSI RAID controller, a maximum distance of 25 meters from the server to the expansion units being imposed, thanks to the limitations of the SCSI bus.

A representation of this type of network is shown in Figure 169:

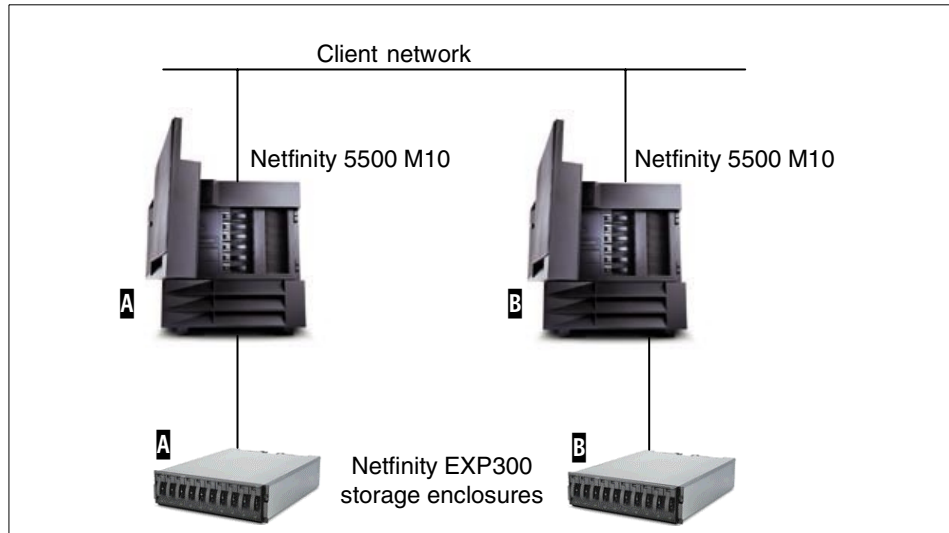


Figure 169. Traditional storage architecture

As the requirement for data storage grows, additional storage can be added to a server. However, the task of managing the process of growth becomes significant as the number of servers increases, particularly when servers are spread over a wide geographic area.

Having storage distributed around the network like this can also cause performance problems if backup data has to be transferred over the client network.

In a Netfinity server SAN storage consolidation configuration (see Figure 170), multiple servers utilize a common pool of SAN-attached disk devices. Storage resources are pooled within a disk subsystem or across multiple disk subsystems, and the data capacity is assigned to independent file systems supported by the operating systems on servers. In both Figure 169 and Figure 170, each server has its own dedicated storage as indicated by the reference letters **A** and **B**.



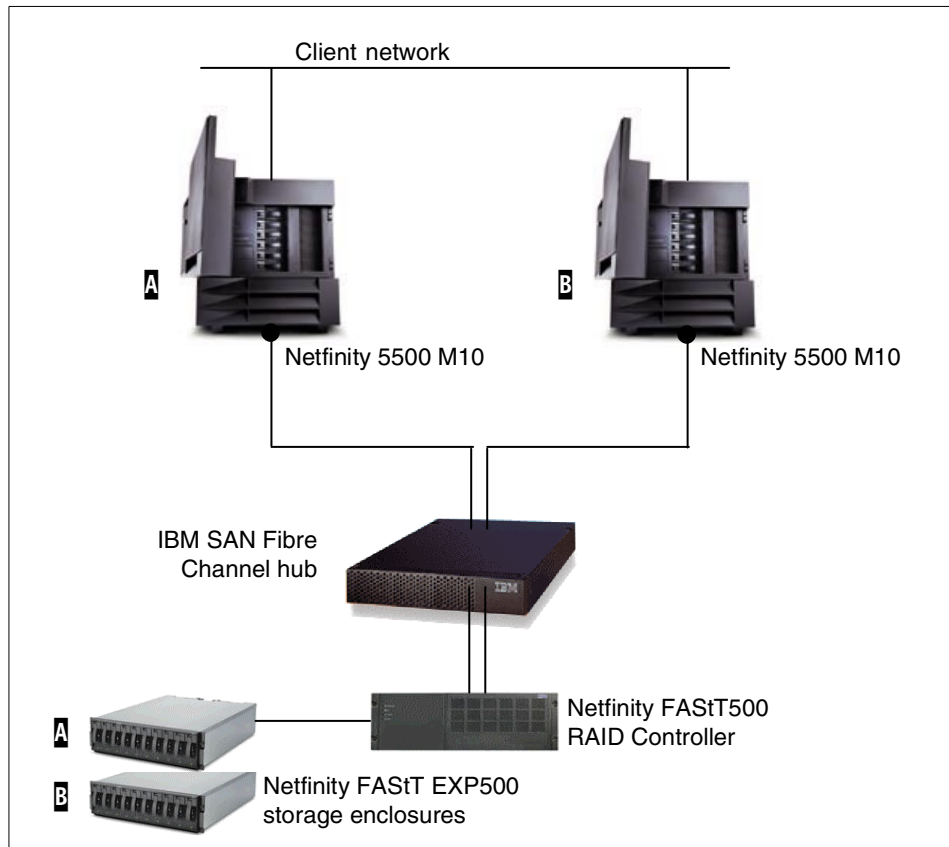


Figure 170. Netfinity SAN storage consolidation

As the need for additional capacity grows, up to 220 disks can be attached to a single FASt500 RAID controller providing ample storage. The amount of storage assigned to an individual server can be managed from a central location by using the Netfinity FASt Storage Manager software.

For large, complex storage consolidation exercises, particularly those for heterogeneous environments, consideration should be given to IBM's Enterprise Storage Server, discussed briefly in 11.2.4, "Netfinity SAN storage" on page 305.

### 11.3.2 Netfinity server consolidation

At the departmental level, a manager may be able to authorize a new server purchase without higher approval. As a result, many organizations have allowed the number of servers attached to their network grow as the needs of

the business have demanded. In time, the task of managing these servers becomes so unwieldy that no one person may have a clear view of the total server population on the network. In this situation, essential management processes, such as backing up critical data, may begin to fail.

Even a relatively modest server with, say, a single Pentium II processor will spend a lot of its time waiting for data to be transferred to or from disk. An examination of the configurations used to get top benchmark scores (TPC-C is a good example) shows that these figures are only achieved by configuring servers with huge arrays of disks feeding the processors with sufficient data to keep them busy. Replacing multiple small servers with one, or a few, more powerful servers connected to a Fibre Channel disk subsystem can reduce overall hardware costs, better utilize server hardware, and simplify the task of managing the network resources. A consolidated environment is shown in Figure 171:

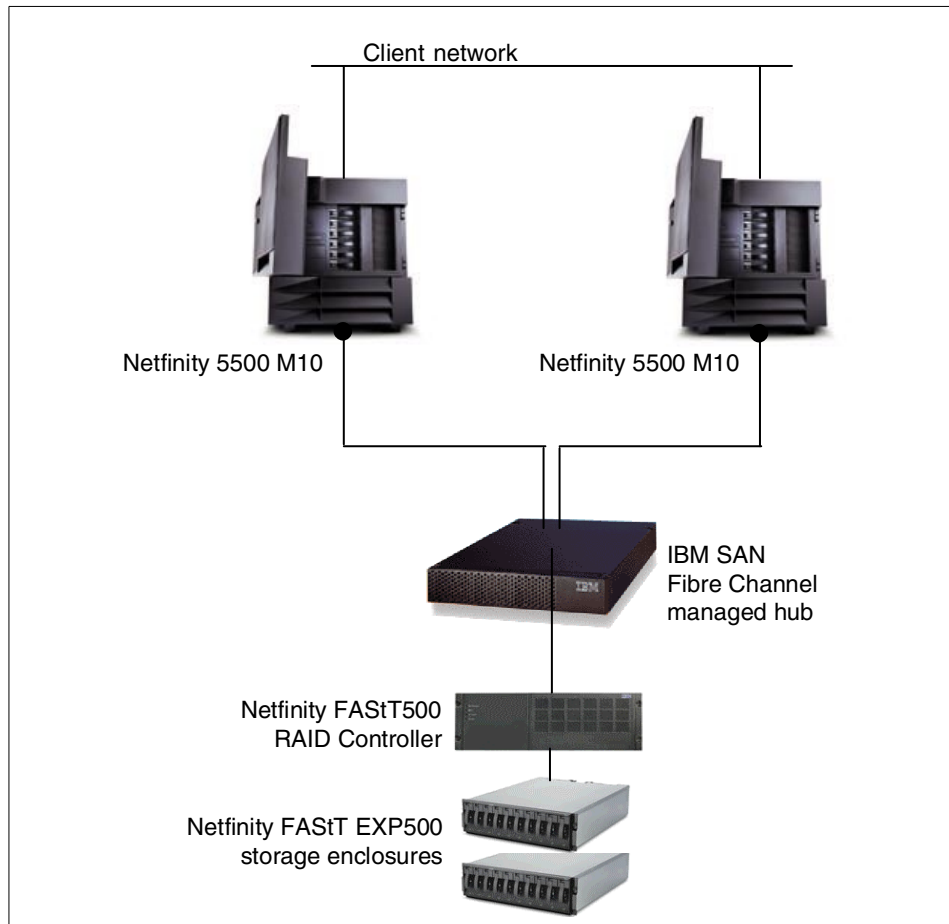


Figure 171. Netfinity SAN server consolidation

In the above example, growing storage capacity simply means adding disks to the subsystem, the management software can then allocate storage as required. As more processing power is required over time, servers may also be added incrementally.

Additional examples of SAN configurations, such as tape pooling, remote backup and high availability solutions can be found in Chapter 3, “Sample disk configurations” on page 19.

## 11.4 Netfinity SAN at present and beyond

SAN is the next generation of enterprise storage architecture. Fibre Channel technology provides the foundation for the current crop of SAN solutions, since it provides immediate benefits in terms of performance, distance and device connectivity and thus, satisfies significant requirements for SAN implementation.

Netfinity server SAN solutions are currently limited to LAN-free backup and recovery, storage consolidation, and limited disaster protection and recovery. A Netfinity server SAN allows sharing of devices as an alternative to expensive investment in additional equipment. The high speed and performance available with Fibre Channel technology help to eliminate any bottleneck between servers and storage.

To drive toward a more complete SAN solution, there is a need to increase the functionality in storage devices, that is, the development of intelligent disk subsystems such as those found in larger systems. Continuing development of SAN management software is also important in enabling SAN features and functions such as serverless data transfer, remote mirroring, fault management, incremental storage growth, and continuous availability.

The following table indicates the current Netfinity SAN against the SAN features and functions.

Table 30. Netfinity SAN features

	Netfinity SAN features and functions
<b>Supported, Partially supported, Not supported yet</b>	
Heterogeneous and homogeneous data sharing	P
Serverless data transfer operations in data mirroring, backup and recovery, and similar operations	N
Fault tolerance	P
Incremental storage growth or storage scalability	P
Any-to-any access	P
Server and storage consolidation	P
Remote data transfer for disaster protection and recovery	N
High-speed data transfer rate (100 MBps to 1 GBps)	P

	Netfinity SAN features and functions
<b>Supported, Partially supported, Not supported yet</b>	
Long connection distance support	S
Continuous availability	P
Single point of SAN management	P









---

## Appendix A. Network operating system support

This appendix provides information about the network operating systems under which the major features of the disk subsystems discussed in this book are supported.

---

### A.1 IBM ServeRAID adapters

Table 31 shows which ServeRAID adapter utilities and functions are supported in various operating systems.

Table 31. *ServeRAID operating system support*

Operating System	ServeRAID Manager	Active PCI	Fault tolerant pair	Command Line Utilities
Windows NT	Yes	Yes	Yes	Yes
Windows 2000	Yes	Yes	Yes	Yes
Novell NetWare	Yes	Yes	No	Yes
IBM OS/2	Yes	No	No	Yes
OpenServer 5.0.5	Yes	No	No	Yes
UnixWare 7.1	Yes	No	No	Yes
Red Hat Linux	Yes	No	No	Yes

**Note:** The table is valid for Version 4.0 of ServeRAID Manager and command-line utilities. ServeRAID Manager Version 3.60 does not support Windows 2000, SCO OpenServer and Linux operating systems.

You can find detailed information on ServeRAID adapters in Part 2, "ServeRAID SCSI subsystems" on page 43.

---

## A.2 Netfinity Fibre Channel and FAST RAID Controllers

Operating system support of these products is shown in Table 32.

Table 32. Netfinity Fibre Channel RAID Controller operating system support

Operating System	Fibre Channel RAID Controller (3526)	Netfinity FAST RAID Controller (3560)	RDAC functionality
Windows NT	Yes	Yes	Yes
Windows 2000 <sup>1</sup>	Yes <sup>1</sup>	Yes <sup>1</sup>	Yes <sup>1</sup>
Novell NetWare	Yes	No <sup>2</sup>	No
SCO UnixWare	Yes	No <sup>2</sup>	No

### Notes:

1. At the time of writing, Windows 2000 support is not available, however it is planned for 2Q2000.
2. Novell NetWare and SCO UnixWare operating systems are likely to be supported at a later date.

Netfinity Fibre Channel products are discussed in Part 3, "Fibre Channel subsystems" on page 165.

---

## A.3 IBM Advanced SerialRAID/X Adapter

This adapter is supported on the following operating systems:

- Windows NT
- NetWare 4.2 and 5

You can use it in the following clustering environments:

- Microsoft Cluster Server
- NetWare Cluster Services

The Advanced SerialRAID/X Adapter does not support a redundant pair of adapters in the same server.

More detailed information on this adapter and other SSA products is available in Part 4, "SSA subsystems" on page 225.

---

## Appendix B. Troubleshooting

This appendix contains information to help you to implement ServeRAID, Fibre Channel, or SSA solutions. We cover areas that may cause problems and provide hints and tips to resolve issues as quickly as possible. You should check IBM's support Web site for current information if you continue to have difficulties:

<http://www.ibm.com/pc/support>

In particular, the Hints and Tips section of the IBM support site lists known problems and any resolutions or workarounds to overcome them.

We also describe the new e-Gatherer tool, which will greatly enhance IBM's support services for Netfinity servers.

---

### B.1 Troubleshooting ServeRAID solutions

The ServeRAID family of adapters are described in Part 2, "ServeRAID SCSI subsystems" on page 43. This section provides information that should save you time when working with ServeRAID-based disk subsystems.

#### B.1.1 Windows NT Server

##### ***Upgrading ServeRAID Manager***

If ServeRAID Manager 3.50 is installed on your system, you should not remove this version before upgrading to 3.60. If you do remove 3.50 first, you will lose all of the ServeRAID Manager's customization files (for example, managed system tree nodes, and the Notification List).

After upgrading the ServeRAID Manager to Version 3.60, remove Version 3.50 using the following steps:

1. From the Start menu, select **Settings -> Control Panel**.
2. From the Control Panel, double-click **Add/Remove Programs**.
3. Remove **ServeRAID Manager 3.50**.

##### ***Fault-tolerant pair - recovering from a power failure during failover***

If a power failure occurs during a failover, it is possible that the two controllers in the active-passive pair might be in a state where some logical drives are configured on one controller and others are configured on the second controller. It is also possible that there might be a logical drive that does not show up as belonging to either controller. To recover from this problem, run IPSSSEND MERGE once for every merge ID that you configured in the pair on the

controller that you want to become active. Then, run `IPSSSEND UNMERGE` once for every merge ID that you configured in the pair on the passive controller. Finally, reboot Windows NT to pair the controllers again.

### B.1.2 Novell NetWare

The following table lists potential problems when starting the ServeRAID Manager software in a Novell NetWare environment, and their possible causes and solutions:

Table 33. Problems and solutions with ServeRAID Manager under NetWare

Problem	Possible solution
The ServeRAID Manager hangs on the splash screen.	You might be using an old version of the ServeRAID device driver. Update the ServeRAID device driver.
When launching ServeRAID Manager an error message. Unable to find load file RAIDMAN is displayed.	The ServeRAID Manager was not installed to the root directory of the SYS volume. There should be a directory called RAIDMAN under the root directory of the SYS volume if the installation was completed properly.
When launching ServeRAID Manager an error message -autounload is an invalid parameter is displayed.	You are using an old version of the Novell Java Virtual Machine. Download and install the latest JVM from Novell. You can download the latest JVM from the following Web site: <a href="http://developer.novell.com/ndk/download.htm">http://developer.novell.com/ndk/download.htm</a>
When launching ServeRAID Manager an error message ERROR: Unable to find Java is displayed.	The Novell Java Virtual Machine is not installed on the NetWare machine. Install the Novell Java Virtual Machine available from: <a href="http://developer.novell.com/ndk/download.htm">http://developer.novell.com/ndk/download.htm</a>

### B.1.3 Linux

Linux is a fast-growing operating system that is supported on IBM Netfinity servers. This section gives some useful tips for this environment.

#### **Linux Device Driver Version 3.60**

The Linux device driver bundled with Caldera OpenLinux V2.3, Red Hat Linux V6.1, SuSE Linux V6.2, and Pacific HiTech TurboLinux V4.0 is not compatible with Version 3.60 of the ServeRAID firmware. To use ServeRAID 3.60 firmware you must update the ServeRAID Linux device driver to Version 3.60 (or newer) *before* updating the firmware on the controller.

The updated Linux device driver is available from:

- IBM ServeRAID Support CD (Version 3.60 or higher)
- IBM ServeRAID Device Driver diskette (Version 3.60 or higher)
- IBM ServeRAID for Linux Web site

<http://www.developer.ibm.com/welcome/netfinity/serveraid.html>

- IBM Netfinity Support Web site

<http://www.pc.ibm.com/support>

#### ***ServeRAID adapter with BIOS/Firmware Version 3.60***

If you received a ServeRAID-3H or ServeRAID-3HB controller with Version 3.60 of the ServeRAID firmware (the firmware version number is displayed during the boot process) you will need to downgrade the firmware on your controller to Version 3.50 before installing Caldera OpenLinux 2.3, Red Hat Linux 6.1, SuSE Linux 6.2, or Pacific HiTech TurboLinux 4.0.

You can obtain Version 3.50 of the ServeRAID BIOS and firmware diskette from the IBM Netfinity Support Web site:

<http://www.pc.ibm.com/support>

Normally the ServeRAID BIOS and Firmware Diskette will not update the BIOS and firmware on your controller if you already have a newer version. To force the update, press Ctrl+F at the update screen.

After installing Linux you must update the ServeRAID Linux device driver (see section above). Then, you can update the firmware on the controller back to Version 3.60.

### **B.1.4 SCO UnixWare**

#### ***Installing ServeRAID Manager***

To use the ServeRAID Manager program with UnixWare, you must have either the Java runtime environment (JRE) or the Java development kit (JDK) installed on your server.

Version 3.60 of the ServeRAID Manager program supports up to eight ServeRAID controllers when using UnixWare.

If ServeRAID Manager 3.50 is installed on your system, you are required to remove this version before upgrading to 3.60. All customization files (for example, managed system tree nodes or Notification list) are saved and used in Version 3.60.

To remove the ServeRAID Manager program from a UnixWare system, run the following command:

```
pkgrm RaidMan
```

### ***The ServeRAID Manager hangs when starting***

There is an infrequent, intermittent timing window that might cause the ServeRAID Manager to hang when the ServeRAID Manager is starting. If you experience this problem, simply stop the Java process that is running the ServeRAID Manager and start the program again.

To stop the ServeRAID Manager on a SCO UnixWare platform, find the process ID using the `ps -ef` command, then issue the `kill` command, specifying the process ID.

Unfortunately, this problem does not lie within the ServeRAID Manager program logic, so we are unable to address this issue at this time.

### ***Miscellaneous problems***

There is an intermittent problem when using the Back button in the online help. In certain cases, you will need to press the Back button two times to return to the previous page.

SCO UnixWare 7.01 shows the time in the event viewer as PST.

Images (such as .gif) may not display when using the DISPLAY environment variable to display the application on a remote X platform.

## **B.1.5 OS/2**

Double-clicking may cause the ServeRAID Manager to freeze and stop responding to mouse input.

To unlock the ServeRAID Manager, press Ctrl+Esc to open the OS/2 Window List. Using the mouse or arrow keys, select one of the ServeRAID Manager windows in the Window List and press the Enter key.

---

## **B.2 Troubleshooting Netfinity Fibre Channel solutions**

This section suggests ways to avoid problems you may encounter with Fibre Channel-based systems.

## **B.2.1 Replacing controllers in a controller unit**

When replacing controllers in a controller unit, the following procedure should be observed when this is done with the subsystem online:

- The controller that remains in the controller unit determines the firmware level of the new controller to be inserted into the unit.
- The new controller (that joins the controller already present in the unit) is flashed to the firmware level of the already present controller. If the present controller has firmware level 3.x and you are re-inserting a controller with 4.x firmware, this controller will be downgraded to 3.x. This is also true in the reverse order.
- When receiving new controllers from stock, make sure they have the correct firmware level before inserting them into a controller unit that is used in a live environment.

## **B.2.2 The Major Event Log (MEL)**

The Major Event Log (MEL) provides a single place for finding information relating to significant actions performed on any component of the controller unit or storage enclosure, and modifications to the controller configuration and controller state. It can be accessed through Netfinity FAST Storage Manager 7. As you can see in Figure 172 on page 322, the log is structured to make it simple to read. Additional detailed information on any listed event can be accessed by double-clicking on the relevant entry in the log. It may also be saved to disk for later analysis.

The MEL is stored in the NVRAM region of the controllers and is therefore independent of the host or management station accessing it. A detailed list of the types of event you can find in the MEL is supplied in the controller unit's documentation.

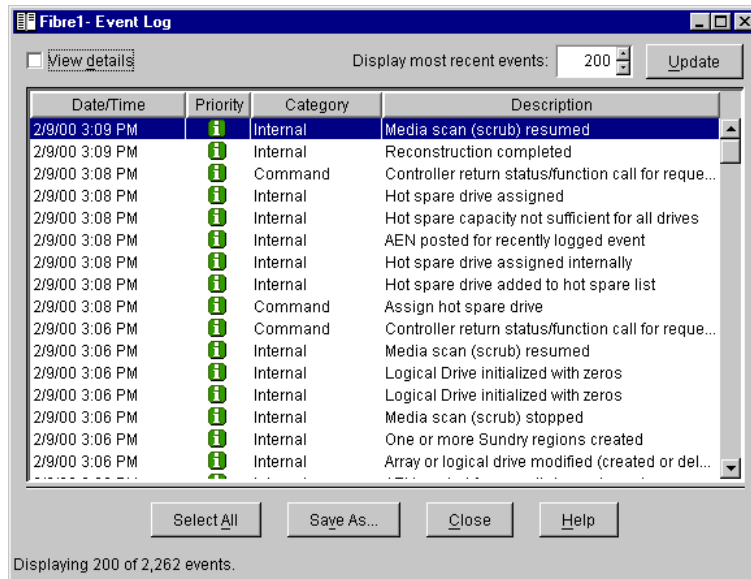


Figure 172. Major Event Log (MEL)

### B.2.3 Power on/off sequence

Vital configuration information is stored on the disk drives and the RAID controller units of a Fibre Channel disk subsystem. To avoid corruption of this information, it is wise to ensure the equipment is powered on and off in a controlled way. We strongly advise the procedure shown in Figure 173 on page 323 to protect against data loss.

It is not necessary to shut down all of these components each time a server is powered off. For example, you should not power down a disk drive enclosure in a clustered environment while any cluster node is still running. The suggested procedure should only be followed when shutting down all nodes in this situation.

We want to emphasize the order in which you power down/up the equipment. In particular, if you power off the disk enclosure before the RAID controller unit, it can lead to unpredictable results, since data may still be in the controller's cache.



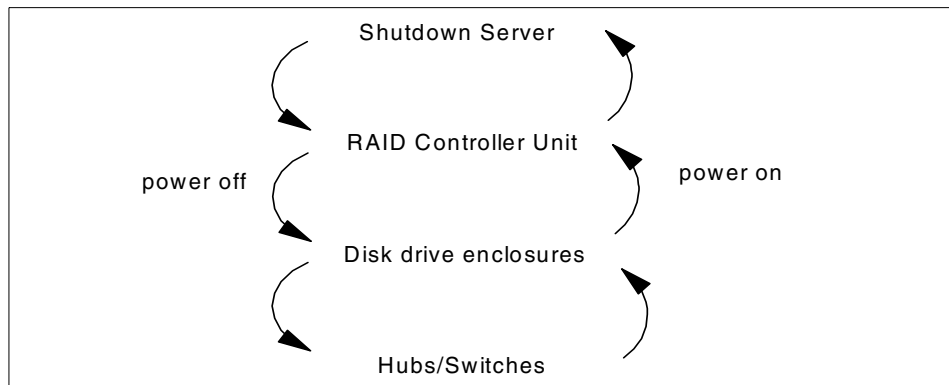


Figure 173. Power on/off sequence for Fibre Channel equipment

## B.3 Troubleshooting SSA disk subsystems

This section looks at problems you may encounter with SSA-based systems.

### B.3.1 Using the SSA service aids

The SSA service aid functions are provided in the RSM configurator:

- Set service mode
- Identify
- Format disk
- Certify disk
- The Event/Error Log utility

The Set service mode function lets you isolate a specific disk drive from a loop without affecting I/O in the background. Before you can do this, the disk drive has to be made a free resource if it was a system resource beforehand. This prevents the operating system from using the drive. If you then activate the service mode on a certain disk drive through RSM, the following actions will happen:

- The Check light of the disk drive comes on for identification.
- All SSA loop activity through the disk drive stops.
- The disk drive motor stops.
- The Check light (if present) of the enclosure that contains the selected disk drive comes on.

- The SSA loop is broken and no communication to the disk drive is possible. However, access to other disk drives is still possible through the open ends of the loop.

You can then safely remove and replace the affected disk drive. In order to make the disk drive (or array) re-usable by the operating system, you have to attach the disk drive.

The Identify function enables you to determine the location of a particular disk drive that you want to identify, but do not want to remove.

The Format disk utility performs a low-level format on a hard disk, and the Certify Disk function verifies the media of the hard disk.

The adapter software includes an event/error logger that runs as a background service on the host system. It collects information about SSA errors, and tells you when an error occurs that needs a service action. All errors are reported and logged in the event/error log files of cluster members.

The format in which the event/error logger reports errors is controlled by the file EVNCTRLF.TXT, which can be changed through the RSM tool.

Before any actions should be taken on the contents of the event log, you need to analyze the log. You can do this using the RSM tool, which parses and summarizes the event log and produces an output file. This output file indicates service actions by service request numbers (SRNs). The online help provides an excellent reference to all SRNs and the recommended procedures to perform as a response to them. Figure 174 shows an example of an entry in the event/error log.

```
Date and time: 22-09-1999 20:39:18
SSA unique ID: S8765231
Error type: Adapter
Template: SSA_DETECTED_ERROR
SRN: 301C0
Flag: Current
          09 23 F5 00 11 22 23 87 82
          90 87 22 22 99 12 87 23 56
```

Figure 174. Event/error log entry

### **B.3.2 Service Request Number list**

When analyzing the SSA Event Logger within RSM, if any problems exist, a Service Request Number (SRN) will be shown. Click the link to open the SRN window, where instructions will be given to help you correct the problem.

Some browsers will not refresh the window when subsequent SRN links are selected. Either use the Find function, or close the window and re-click the link.

### **B.3.3 RSM Proxy configuration**

Some problems may be experienced when accessing RSM via a proxy server. It is recommended that you use an automatic proxy configuration script (if available). Alternatively, if you are accessing SSA RSM from behind a firewall, and SSA RSM is itself behind the firewall, ensure you set your browser so that the proxy is not used for sites that are inside the firewall.

Some problems have also been found using Netscape when an array contains a failed resource and no hot spare is available. At this point, the failed resource is replaced by a virtual disk, a "blank reserved". When using a proxy server, some versions of Netscape fail to process the "blank reserved" resource.

With the enhanced IP Address-based security, if you access RSM through a proxy, you will not be granted access unless you add the IP address of the proxy server to RSM. However, this will nullify the IP security, since it will allow all users entering through the proxy server to access RSM.

For local (behind a firewall) access you should ensure you are using an automatic proxy configuration script.

### **B.3.4 Service port**

The NetWare RSM service will default to port 511. This can be changed by modifying the autoexec.ncr file. Change the line containing:

```
load SYS:SYSTEM\issarsm.nlm
```

to:

```
load SYS:SYSTEM\issarsm.nlm (port number)
```

where (port number) is any valid TCP/IP port number.

### B.3.5 7133 bypass card settings

The bypass cards in a 7133-D40/T40 enclosure should be set to Automatic, which is also the default setting. Some technical personnel have the misconception that jumpering bypass cards to Forced Inline will give a performance improvement. This is false and you should set jumpers to Inline mode only when guided to do so by the documentation or by IBM support.

You can also set the bypass cards into Forced Inline mode used by the RSM too. This is used to disable the switching ability of bypass cards. You may wish to set the Forced Inline mode if the disk drive modules in one 7133 are not all connected to the same SSA loop. In this type of environment, Forced Inline mode removes the risk that a fault condition might cause the disk drive modules of different loops to be connected to each other.

You may also have heard a recommendation to set Forced Inline mode when two 713's are connected to each other. In this type of configuration, if one of the boxes fails you will have lost disks off the loop, and the problem needs to get fixed anyway. In addition, there is not much advantage one way or the other whether the loop wraps in the remaining 7133 or remains forced inline.

### B.3.6 Documentation

Documentation for the SSA subsystem components can be downloaded as PDF files from <http://www.hursley.ibm.com/ssa/pcserver>. These documents include:

Documentation for the Advanced SerialRAID/X Adapter:

- *Advanced SerialRAID/X Adapter Installation Guide*
- *Advanced SerialRAID Adapter Technical Reference*
- *Advanced SerialRAID/X Adapter User Guide and Maintenance Information*

Documentation for the 7133-D40/T40 enclosures:

- *7133-D40 Installation Guide*
- *7133-T40 Installation Guide*
- *7133-D40/T40 Operators Guide*
- *7133-D40/T40 Service Guide*
- *7133-D40/T40 Technical Information*
- *7133 T40/D40 Safety Notices*

---

## B.4 e-Gatherer

e-Gatherer is a tool used by IBM support personnel to quickly resolve a customer's problem. It is a call assistance tool that helps to reduce the resources involved in a support case once a customer has called IBM with a problem.

If a problem cannot be solved immediately, the IBM support representative needs information about the customer's machine and its environment in order to be able to initiate problem determination and, perhaps, to replicate the problem. In the past, the customer was asked to fill in a so-called *escalation template*. This template was essentially a questionnaire asking the relevant questions about the customer's system.

Sometimes it could take significant technical expertise on the part of the customer to know where to find certain information (for example, "What is the stripe size of your RAID-5 logical drive?") and it might take one or two hours for someone with that expertise to fill in all the information. Another issue arises when information from the BIOS, requiring a reboot, is needed. This can be a problem, especially on production servers.

With e-Gatherer that has all changed. The customer is asked to execute a utility program that automatically collects all the information necessary for the IBM support personnel to begin to resolve the customer's problem.

This way, no expert is required to fill in the template, and the machine can virtually be diagnosed remotely. The customer does not need to install any sophisticated management software on his already troubled machine, he does not need to reboot his machine, and a single mouse click provides IBM with a comprehensive and accurate report.

The following operating systems are currently supported:

- Windows NT and 2000, 95/98
- Linux

There are also plans to support OS/2, NetWare 5.x and SCO UNIX.



---

## Appendix C. Special notices

This publication is intended to help technical staff within IBM, our business partners, and our customers to understand the characteristics of available Netfinity server disk subsystem technologies. In addition, it provides guidance in selecting the most appropriate technology for a specific server installation. The information in this publication is not intended as the specification of any programming interfaces that are provided by Netfinity servers. See the PUBLICATIONS section of the IBM Programming Announcement for Netfinity servers for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

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

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The information about non-IBM ("vendor") products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. The use of this information or the implementation of any of these techniques is a customer

responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

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

AIX	AS/400
DB2	DB2 Universal Database
Domino	Enterprise Storage Server
ESCON	EtherJet
FICON	IBM
Lotus	Lotus Notes
Magstar	Micro Channel
Netfinity	Netfinity Manager
NetView	Notes
OS/2	OS/400
Power PC 603	Predictive Failure Analysis
RS/6000	S/390
ServeRAID	StorWatch
System/390	ThinkPad
Tivoli	Ultrastar
Versatile Storage Server	Wizard

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

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

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

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.



ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

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

SET and the SET logo are trademarks owned by SET Secure Electronic Transaction LLC.

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



---

## Appendix D. Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

---

### D.1 IBM Redbooks

For information on ordering these publications see “How to get IBM Redbooks” on page 335.

- *Netfinity Clustering Planning Guide*, SG24-5845
- *Netfinity Director - Integration and Tools*, SG24-5389
- *Tuning Netfinity Servers for Performance - Getting the most out of Windows 2000 and Windows NT 4.0*, ISBN 0130406120

---

### D.2 IBM Redbooks collections

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at [ibm.com/redbooks](http://ibm.com/redbooks) for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	Collection Kit Number
IBM System/390 Redbooks Collection	SK2T-2177
IBM Networking Redbooks Collection	SK2T-6022
IBM Transaction Processing and Data Management Redbooks Collection	SK2T-8038
IBM Lotus Redbooks Collection	SK2T-8039
IBM Tivoli Redbooks Collection	SK2T-8044
IBM AS/400 Redbooks Collection	SK2T-2849
IBM Netfinity Hardware and Software Redbooks Collection	SK2T-8046
IBM RS/6000 Redbooks Collection (PDF Format)	SK2T-8043
IBM Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

---

### D.3 Other resources

These publications are also relevant as further information sources:

- *ServeRAID-3HB, ServeRAID-3H and ServeRAID-3L Ultra2 SCSI Controllers Installation and User's Guide*, shipped with the product
- *IBM Netfinity EXP500 - Installation and User's Handbook*, shipped with the product

- *IBM Netfinity Fibre Channel Storage Manager for Windows NT Installation and Support Guide*, shipped with the product
- *Netfinity Manager User's Guide*, shipped with the product

---

#### D.4 Referenced Web sites

These Web sites are also relevant as further information sources:

- <http://www.t11.org>
- <http://www.t10.org/scsi-3.htm>
- <http://www.pc.ibm.com/support>
- [ftp://ftp.pc.ibm.com/pub/pccbbs/pc\\_servers/00n9126.iso](ftp://ftp.pc.ibm.com/pub/pccbbs/pc_servers/00n9126.iso)
- <http://www.pc.ibm.com/us/compat/hotplug>
- <http://www.pc.ibm.com/us/solutions/netfinity/index.html>
- <http://www.pc.ibm.com/ww/netfinity/fibrechannel/>
- <http://www.pc.ibm.com/us/netfinity/clustering.html>
- [http://www.pc.ibm.com/ww/netfinity/systems\\_management/nfdir.html](http://www.pc.ibm.com/ww/netfinity/systems_management/nfdir.html)
- <http://www.pc.ibm.com/us/compat/san/index.html>
- <http://www.ibm.com/storage>
- <http://www.storage.ibm.com/hardsoft/products/ess/ess.htm>
- <http://www.storage.ibm.com/hardsoft/products/ssa/pcserver/index.html>
- <http://www.cai.com>
- <http://hpsanmanager.com>
- <http://www.legato.com>
- <http://ssdweb01.storage.ibm.com/software/storwatch/swovis.htm>
- <http://www.sanergy.com>
- [http://www.tivoli.com/products/index/storage\\_mgr](http://www.tivoli.com/products/index/storage_mgr)
- <http://support.veritas.com>
- <http://www.developer.ibm.com/welcome/netfinity/serveraid.html>
- <http://developer.novell.com/ndk/download.htm>

---

## How to get IBM Redbooks

This section explains how both customers and IBM employees can find out about IBM Redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** [ibm.com/redbooks](http://ibm.com/redbooks)

Search for, view, download, or order hardcopy/CD-ROM Redbooks from the Redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

Send orders by e-mail including information from the IBM Redbooks fax order form to:

	<b>e-mail address</b>
In United States or Canada	pubscan@us.ibm.com
Outside North America	Contact information is in the "How to Order" section at this site: <a href="http://www.elink.ibm.com/pbl/pbl">http://www.elink.ibm.com/pbl/pbl</a>

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the "How to Order" section at this site: <a href="http://www.elink.ibm.com/pbl/pbl">http://www.elink.ibm.com/pbl/pbl</a>

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the "How to Order" section at this site: <a href="http://www.elink.ibm.com/pbl/pbl">http://www.elink.ibm.com/pbl/pbl</a>

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the Redbooks Web site.

### IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and Redbooks by accessing the IBM Intranet Web site at <http://w3.itso.ibm.com/> and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access MyNews at <http://w3.ibm.com/> for redbook, residency, and workshop announcements.

---

## IBM Redbooks fax order form

Please send me the following:

Title	Order Number	Quantity

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

Invoice to customer number \_\_\_\_\_

Credit card number \_\_\_\_\_

Credit card expiration date Card issued to Signature

**We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.**

---

## Abbreviations and acronyms

<b>ANSI</b>	American National Standards Institute	<b>FICON</b>	fiber distributed data interface
<b>API</b>	application programming interface	<b>FRU</b>	field replaceable unit
<b>ATA</b>	advanced technology attachment	<b>GB</b>	gigabytes
<b>BIOS</b>	basic input/output system	<b>GBIC</b>	gigabit interface converter
<b>CA</b>	Computer Associates	<b>GUI</b>	graphical user interface
<b>CAD/CAM</b>	computer-aided design/computer-aided manufacturing	<b>HIPPI</b>	high performance parallel interface
<b>CD-ROM</b>	compact disk read-only memory	<b>HTTP</b>	Hypertext Transfer Protocol
<b>CPU</b>	central processing unit	<b>HVD</b>	high voltage differential
<b>CRC</b>	cyclic redundancy check	<b>I/O</b>	input/output
<b>DHCP</b>	Dynamic Host Configuration Protocol	<b>IBM</b>	International Business Machines Corporation
<b>DLT</b>	digital linear tape	<b>IDE</b>	integrated drive electronics
<b>DMA</b>	direct memory access	<b>IP</b>	Internet Protocol
<b>DMI</b>	Desktop Management Interface	<b>IPI</b>	intelligent peripheral interface
<b>DOS</b>	disk operating system	<b>ISO</b>	International Standards Organization
<b>ECC</b>	error checking and correcting	<b>ISP</b>	Internet Service Provider
<b>EDO</b>	extended data out	<b>ITSO</b>	International Technical Support Organization
<b>EEPROM</b>	electrically erasable programmable read-only memory	<b>JBOD</b>	“just a bunch of disks” - a set of disks attached to a controller
<b>EIDE</b>	enhanced integrated drive electronics	<b>JDK</b>	Java development kit
<b>ESCON</b>	enterprise systems connection	<b>JRE</b>	Java runtime environment
<b>ESM</b>	Enclosure Services Monitor	<b>JVM</b>	Java virtual machine
<b>ESS</b>	Enterprise Storage Server	<b>KB</b>	kilobytes
<b>FC</b>	Fibre Channel	<b>LAN</b>	local area network
<b>FC-AL</b>	Fibre Channel Arbitrated Loop	<b>LCD</b>	liquid crystal display
<b>FCP</b>	Fibre Channel Protocol	<b>LCT</b>	Life Cycle Tools
		<b>LDM</b>	logical drive migration
		<b>LED</b>	light-emitting diode
		<b>LUN</b>	logical unit number
		<b>LVDS</b>	low-voltage differential SCSI

<b>MAC</b>	medium access control	<b>SCO</b>	Santa Cruz Operation, Inc.
<b>MB</b>	megabytes	<b>SCSI</b>	small computer system interface
<b>Mbps</b>	megabits per second	<b>SDRAM</b>	static dynamic random access memory
<b>MBps</b>	megabytes per second	<b>SIC</b>	serial interface chip
<b>MEL</b>	Major Event Log	<b>SMP</b>	symmetric multiprocessing
<b>MIA</b>	media interface adapter	<b>SNMP</b>	simple network management protocol
<b>MIB</b>	management information base	<b>SQL</b>	structured query language
<b>MIME</b>	multipurpose internet mail extensions	<b>SRAM</b>	static random access memory
<b>MSCS</b>	Microsoft Cluster Server	<b>SRN</b>	service request number
<b>NAE</b>	Netfinity Availability Extensions for MSCS	<b>SSA</b>	serial storage architecture
<b>NAS</b>	network attached storage	<b>TB</b>	terabytes
<b>NLM</b>	NetWare loadable module	<b>TCP/IP</b>	Transmission Control Protocol/Internet Protocol
<b>NVRAM</b>	non-volatile random access memory	<b>TPC-C</b>	Transaction Processing Council - C benchmark
<b>PCI</b>	peripheral component interconnect	<b>UID</b>	unique identifier
<b>PDF</b>	portable document format	<b>ULP</b>	Upper Layer Protocols
<b>PFA</b>	predictive failure analysis	<b>URL</b>	Uniform Resource Locator
<b>POST</b>	power on self test	<b>VHDCI</b>	very high density connector interface
<b>RAID</b>	redundant array of independent disks	<b>WAN</b>	wide area network
<b>RAM</b>	random access memory		
<b>RDAC</b>	Redundant Disk Array Controller		
<b>ROM</b>	read only memory		
<b>RPM</b>	revolutions per minute		
<b>RPO</b>	rotational positioning optimization		
<b>RSM</b>	Remote System Management		
<b>SAN</b>	storage area network		
<b>SAP</b>	Systeme, Anwendungen und Programme in der Datenverarbeitung (Systems, Products, and Programs in Data Processing)		



---

## Index

### Numerics

64-bit PCI data path 64  
8-way SMP 302

### A

Active PCI 67, 106, 107  
adapters per server  
    ServeRAID 24, 48  
    SSA 245  
arbitrated loops 168  
arrays  
    adding drives improves performance 148  
    logical drive configuration 152  
    performance 146  
    spanned 47, 54  
arrays per adapter 48  
ATA interface 8  
autosync 69  
availability, technologies compared 15

### B

BIOS 75, 91

### C

cabling  
    redundancy for Fibre Channel 188  
    rules for SSA 259  
    SSA 266  
cache  
    battery-backup 60, 65  
    effectiveness under load 158  
    Fibre Channel 216  
    overheads 160  
    ServeRAID 64  
    SSA 244  
    write-back 65  
    write-back versus write-through 158  
cache operation 157  
changing RAID levels 95, 96  
channel subsystem 8  
choosing the storage technology to use 16  
clustering  
    Fibre Channel 28  
    IBM Redbook 127

Novell Cluster Services (NCS) 29  
    ServeRAID 124  
    SSA 257, 262  
    Windows 2000 Advanced Server 37, 41  
    Windows 2000 Datacenter Server 41  
coexistence installation 203  
command overhead 150  
comparing hardware and software RAID 48  
configuration data, ServeRAID 69  
configuring fault-tolerant adapters 114, 118  
consolidation, server 32  
cyclic redundancy check 59

### D

data mirroring 50  
data scrubbing 68  
data transfer time 151  
definitions, Fibre Channel 199  
Desktop Management Interface 107  
development, storage 7  
device drivers 161  
disaster recovery 35, 260  
disk and tape pooling 32  
disk configuration changes 127  
disk performance 150  
disks per RAID array  
    Fibre Channel 180  
    ServeRAID 48  
    SSA 243  
distance, technologies compared 13  
distributed hot spare 53  
DMI (Desktop Management Interface) 107

### E

e-Gatherer 327  
Enclosure Services Monitor (ESM) boards 186  
enhanced IDE 8  
Enterprise Storage Server 299  
ESCON 8  
evolution of storage 7  
EXP200 Storage Expansion Unit 78  
EXP300 Storage Expansion Unit 79  
external storage enclosures 78

### F

fabric, Fibre Channel 171

- failover of fault-tolerant adapter 115
- failure of disk drives, recovery from 99
- FAST products
  - Fibre Array Storage Technology*
  - See Fibre Channel
- fault tolerance, technologies compared 14
- fault-tolerant adapter pair 67, 107, 114
- fiber optical cabling 172, 235
- Fibre Array Storage Technology
  - See Fibre Channel
- Fibre Channel
  - adding drives 213
  - administration tasks 210
  - arbitrated loops 168
  - basic configuration 25
  - cabling 172, 187
  - cache settings 216
  - changing NVSRAM settings 209
  - cluster configurations 28
  - coexistence installation 203
  - communication protocols 168
  - components 172
  - connectors 178
  - controller blades 176
  - Controller Unit 3526 176
  - Data Gateway 196
  - data scrubbing 214
  - definitions 199
  - direct management 204
  - disk drives 185
  - dual-loop wiring 26
  - duplex-SC connector 172
  - duplicate controllers 210
  - Enclosure Services Monitor (ESM) boards 186
  - fabric 171
  - Failsafe RAID Controller 176
  - FAST EXP500 Storage Expansion Unit 185
  - FAST200 RAID/Storage Unit 183
  - FAST500 connectors 178, 188
  - FAST500 RAID Controller 174, 177
  - fiber cable 172
  - firmware 203
  - frame structure 170
  - GBIC 173
  - gigabit interface converter 173
  - high-capacity subsystem 29
  - host group 200, 219
  - host-agent management 206
  - hub 174
  - Immediate Availability Feature 213
  - intermediate firmware level 204
  - logical drive migration 180
  - long-wave 172
  - Major Event Log (MEL) 214, 321
  - major features 167
  - managed hub 196
  - maximum capacity 179
  - maximum configuration 190
  - maximum number of ports per loop 168
  - media interface adapters 173
  - media scan 213
  - mini-hubs 178
  - multi-mode cables 172
  - multiple hosts 192, 219
  - multiple server configuration 26
  - Netfinity hardware 171
  - Novell Cluster Services (NCS) 29
  - NVSRAM 208
  - operating system support 316
  - overview 11, 167
  - point-to-point connections 168
  - power on/off sequence 322
  - protocol layers 169
  - RAID levels 179
  - RAID-0 181
  - RAID-1 182
  - RAID-3 182
  - RAID-5 183
  - RDAC 201
  - read-ahead multiplier 217
  - redundancy, considerations 195
  - Redundant Disk Array Controller 201
  - replacing controllers 321
  - SAN switches 193
  - scripting 209, 218
  - segment size 211
  - server attachment 191
  - setting the tray ID 187
  - short-wave 172
  - single-mode cables 172
  - SM7agent 201, 206
  - SM7client 200, 207
  - Storage Manager 7 (SM7) 200
  - storage partitioning 29, 191, 219
  - switched connections 168
  - terminology 199
  - topology 168
  - troubleshooting 320

- upgrading firmware 204
- FICON 8
- firmware
  - Fibre Channel 203, 208
  - ServeRAID 75, 162
  - SSA 260
- FlashCopy 73
- force failover, ServeRAID 122
- frame structure, Fibre Channel 170
- free capacity 199

## G

- gigabit interface converter (GBIC) 173

## H

- hardware RAID versus software RAID 48
- high-availability solutions 37
- high-capacity SCSI subsystem 22
- host group 200
- hot-add
  - no need to reboot Windows NT 111
  - ServeRAID adapter 111
- hot-plug tools
  - hardware wizard 110
  - PCI hot-plug controls 109
  - rules and guidelines 111
  - system tray pop-up menu 108
- hot-remove not supported under Windows NT 111
- hot-replace an adapter 112
- hot-spare 86, 99
- hot-swap
  - Fault Tolerant Management Interface 117
  - rebuild 67

## I

- I<sub>2</sub>O (Intelligent Input/Output) 66
- IBM Advanced SerialRAID/X Adapter 242
- IBM Netfinity Availability Extensions for MSCS 299
- IBM StorWatch 299
- IBM StorWatch Switch Specialist 195
- integrated drive electronics 8
- Intelligent Input/Output (I<sub>2</sub>O) 66
- interconnects, SAN 303
- interface data rate 151
- IPSEND command 131

## L

- Linux, using ServeRAID 318
- logical drive 199
- logical drive configuration 152
- logical drive migration 72, 94, 180
- logical drives 46, 48
- logical unit number 200
- long-wave fiber optic cable 172
- low voltage differential signalling 63
- LUN (logical unit number) 200
- LVDS 63

## M

- Major Event Log (MEL) 214
- management, remote 104
- media data rate 151
- Microsoft Cluster Server
  - See MSCS
- Mini-Configuration Utility, ServeRAID 89
- mini-hubs 178
- mirrored data 50
- monitoring performance 149
- monitoring physical disk activity 156
- MSCS
  - Microsoft Cluster Server
  - /localquorum switch 127
  - Fibre Channel clustering 29
  - high-availability solutions 37
  - quorum resource 126, 127
  - ServeRAID clustering 127
  - SSA clustering 260
- multi-mode cables 172
- multiple failed disk drives 101

## N

- Netfinity
  - 8-way SMP 302
  - disk and tape pooling 32
  - EXP200 Storage Expansion Unit 78, 185
  - EXP300 Storage Expansion Unit 22, 79, 185
  - FAStT EXP500 Storage Expansion Unit 185
  - FAStT Host Adapter 175
  - FAStT200 RAID/Storage Unit 183
  - FAStT500 RAID Controller 28, 177
  - Fibre Channel definitions 199
  - Fibre Channel hardware 171
  - Fibre Channel Storage Manager 7 (SM7) 200
  - Fibre Channel subsystem configurations 25

- Fibre Channel terminology 199
- remote backup, archiving and recovery 35
- SAN components 294
- SAN interconnects 303
- SAN servers 303
- SAN software 295
- SAN solutions 305
- SANs 293
- SCSI disk subsystem configuration 20
- servers 301
- SSA hardware 234
- X-architecture 301
- Netfinity Director 132
- Netfinity Manager 132
- network operating system support 315
- network-attached storage 9, 287
- Novell Cluster Services 29

## O

- operating system support 315
- optimal speed (SCSI) 64

## P

- partitioning storage 29
- PCI
  - peripheral component interconnect*
  - 64-bit bus 64
  - Active PCI 67, 106
  - fault-tolerant adapter pair 67
  - hot-swap 67
- performance
  - adding drives, effect of 148
  - cache operation, ServeRAID 157
  - device drivers, effect of 161
  - disk failure for RAID-5, effect of 52
  - disk technologies compared 14
  - firmware, effect of 162
  - individual disk performance 150
  - monitoring 149
  - RAID level, effect of 146
  - rule of thumb for adding drives 150
  - SCSI bus speeds, effect of 164
  - ServeRAID 145
  - SSA 232, 265
  - stripe size recommendations 156
  - stripe size, effect of 154
  - write-back versus write-through cache 158
- physical layers, Fibre Channel protocols 170

- planning RAID subsystems 146
- point-to-point connections 168
- Power PC 750 processor 60
- protocol layers, Fibre Channel 169

## Q

- quorum disk drive 126

## R

### RAID

- redundant array of independent disks*

- avoiding data inconsistencies 68
- distributed hot spare 53
- hardware versus software 48
- hot-spare disk 53
- levels (spanned arrays) 47
- redundancy 9
- subsystem planning 146
- synchronizing arrays 68
- Windows NT Stripe Sets 50

### RAID levels

- changing 95, 180
- Fibre Channel 179
- firmware 162
- performance differences 146
- RAID-0 49, 181, 261
- RAID-00 54
- RAID-1 50, 182, 260
- RAID-1 Enhanced (RAID-1E) 51
- RAID-10 55, 261
- RAID-1E0 56
- RAID-3 182
- RAID-5 51, 183, 243
- RAID-5 Enhanced (RAID-5E) 53
- RAID-50 57
- software RAID-5 53
- SSA 240

### RDAC

- See under* Fibre Channel
- recommended cabling method, EXP500 188
- redundant array of independent disks
  - See* RAID
- remote backup, archiving and recovery 35
- remote system management 104
- Remote System Management (RSM) 272
- requirements, storage 7
- rotational latency 150
- rotational positioning optimization 151

rules and guidelines for hot-plug PCI 111

## S

### SAN

- storage area network*
- benefits 290
- building blocks 284
- concepts 10
- Data Gateway 196
- Data Gateway for SSA 248
- data transfer 285
- defined 283
- disk and tape pooling 32
- disk subsystem sharing 296
- enabled by Fibre Channel 10
- Enterprise Storage Server 299
- evolution 291
- Fibre Channel managed hub 196
- Fibre Channel switches 193
- IBM StorWatch 299
- Netfinity components 294
- Netfinity interconnects 303
- Netfinity servers 303
- Netfinity solutions 305
- overview 279
- problems addressed by 281
- remote backup, archiving and recovery 35
- software for Netfinity servers 295
- storage attachment 287
- storage consolidation 305
- supported configurations 293
- tape attachment 196
- tape library sharing 296
- Tivoli SANergy 300
- Tivoli Storage Manager 300

SAN-attached storage 287

scalability, technologies compared 16

SCO UnixWare, using ServeRAID 319

scripting language 209

### SCSI

- small computer systems interface*
- bus organization 156
- bus transfer rate 164
- connection distances 11
- devices 11
- domain validation 59
- fault-tolerant configurations 21
- high-capacity subsystem 22

- low voltage differential signalling (LVDS) 63
- optimal speed 64
- overview 11
- redundant adapters 22
- sample configurations 19
- ServeRAID 11
- technology 11
- Ultra3 11
- Ultra3 160/m 58

Security Manager, of ServeRAID Manager 104

seek time 150

serial storage architecture

- See SSA

server consolidation 32

server roles 9

### ServeRAID 60

- Active PCI support 106
- active/passive pair 115
- adapter cache 64
- adapters per server 24, 48
- alert handling 132, 144
- arrays 46
- arrays per adapter 48
- autosync 69
- BIOS 75, 91
- boot-time messages 127
- cache operation 157
- cache size 159
- changing RAID levels 95, 96
- clustering 124
- clustering installation procedures 127
- command-line utilities 72, 130
- configuration data 69
- Configuration Program 82
- Configuration Utility 70
- configuring a fault-tolerant pair 114, 118
- creating arrays 85, 93
- creating logical drives 86, 93
- data scrubbing 68
- device drivers 161
- disk configuration changes 127
- disk failure 99
- disks per RAID array 48
- DMI Component Service 107
- drives per channel 47
- failover of fault-tolerant adapter 115
- fault-tolerant adapter pair 67, 107, 317
- fault-tolerant configurations 21
- feature comparison by adapter 76

- features and options 46
- firmware 75, 162
- FlashCopy 73
- force failover 122
- hardware 45
- hot-add a new adapter 111
- hot-plug tools 108
- hot-replacement 112
- hot-spare 86
- hot-spare disk 53
- hot-swap rebuild 67
- increasing logical drive size 98
- installing fault-tolerant adapters 116
- IPSEND command 131
- limitations of spanned arrays 60
- Linux tips 318
- logical drive configuration 152
- logical drive migration 72, 94
- logical drives 46, 48
- managing the subsystem 132
- Mini-Configuration Utility 70, 89
- multiple adapters 24
- multiple failed disk drives 101
- Netfinity Director integration 132
- Netfinity Manager integration 137
- network operating system support 315
- number of disks in a spanned array 48
- older adapters 62
- performance 145
- potential problems with NetWare 318
- power-on self test 127
- quorum disk drive 126
- RAID adapter information 141
- RAID levels supported 48
- RAID Manager 138
- RAID-0 49
- RAID-00 54
- RAID-1 50
- RAID-10 55
- RAID-1E 51
- RAID-1E0 56
- RAID-5 51
- RAID-50 57
- RAID-5E 53
- redundant adapters 22
- remote system management 104
- running Manager as an agent 104
- SCO UnixWare 319
- SCSI bus organization 156
- Security Manager 104
- ServeRAID Manager 70, 92, 134
- ServeRAID-3 adapter family 46, 60
- ServeRAID-4 adapter family 45, 58
- spanned arrays 22, 47
- stripe unit size 153
- subsystem planning 146
- troubleshooting 317
- Ultra3 160/m 58
- upgrading ServeRAID manager 317
- utilities 70, 81
- working with fault-tolerant adapters 121
- write-back versus write-through cache 158
- server-attached storage 287
- short-wave fiber optic cable 172
- single-mode cables 172
- small computer system interface
  - See SCSI
- software RAID versus hardware RAID 48
- spanned arrays
  - benefits of 47
  - limitations 60
  - number of disks 48
  - RAID-00 54
  - RAID-10 55
  - RAID-1E0 56
  - RAID-50 57
  - ServeRAID-4 47
  - sub-logical drives 54
- spatial reuse 232
- split-site operation 261
- SSA
  - 7133 storage enclosure 245
  - accessing RSM using a proxy 325
  - adapter connectors 254
  - adapter feature comparison 240
  - ANSI standard 227
  - Automatic mode 253
  - bandwidth 229, 267
  - bypass card settings 326
  - bypass cards 251
  - cabling 234, 259, 266
  - cache 244, 265
  - clustering 257, 262
  - cut-through routing 232
  - cyclic redundancy checks 231
  - data rate 237
  - device speed 229
  - disaster recovery configuration 260

- disk placement 266
- disk storage enclosures 245
- distance, effect on performance 265
- documentation 326
- DOS utilities 270
- dual-host configurations 257
- error log 324
- error recovery 231
- event logging 273
- failure scenarios 262
- Fast Write Resource 275
- fault isolation 231
- fiber optic cables 235
- fiber optic extender 236
- firmware 260
- Forced Bypass mode 253
- Forced Inline mode 253
- Forced Open mode 254
- frame 231
- Free Resource 274
- hardware 234
- hop count, significance of 266
- Hot Spare Resource 275
- hot-spares 264
- IBM Advanced SerialRAID/X Adapter 242
- identifying adapters 238
- implementation 251
- initiator 230
- jumpers for bypass cards 252
- Just a Bunch Of Disks (JBOD) 228
- loop configuration 254
- loop topology 228
- management 269
- managing disk resources 274
- managing with a Web browser 273
- master node 230
- maximum configuration 256, 259
- maximum usable capacity 262
- metadata 262
- modes, bypass cards 252
- MSCS 260
- Netfinity server PCI slots supported 245
- New Resource 274
- nodes 230
- nodes per loop 229
- Novell NetWare tools 271
- operating system support 316
- operating system-specific tools 270
- overview 12
- performance 232, 265
- RAID levels 240
- RAID Resources 275
- RAID-0 261
- RAID-1 260
- RAID-10 241, 261
- RAID-5 243
- recommended settings for bypass cards 254
- Rejected Resource 275
- Remote System Management (RSM) 272
- RSM Proxy configuration 325
- rules for cabling 259
- SAN Data Gateway 248
- serial interface chip 228
- SerialRAID/X adapter 265
- service aids 323
- service port 325
- Service Request Number 325
- single-host configurations 255
- spatial reuse 232
- Split-Resolution flag 262
- split-site operation 261
- string topology 230
- stripe sizes 243
- supported servers 244
- System Resource 275
- target 230
- terminology 228, 274
- throughput 265
- topology 228
- troubleshooting 323
- two enclosures, single host 256
- Windows NT 4.0 utilities 270
- storage
  - ATA interface 8
  - availability comparison 15
  - choosing the appropriate technology 16
  - comparing solutions 12
  - development 7
  - distance comparison 13
  - enclosures 78, 245
  - evolution of 7
  - fault tolerance comparison 14
  - Fibre Channel 11
  - first generation 7
  - network operating system support 315
  - network-attached 287
  - partition, defined 200
  - partitioning 29, 191, 219

- performance comparison 14
- requirements 7
- SAN-attached 287
- scalability comparison 16
- server-attached 280, 287
- subsystem planning 146
- system-attached 7
- technologies 10
- traditional architectures 280
- troubleshooting 317
- storage area network
  - See SAN
- storage consolidation 305
- Storage Manager 7 (SM7) 200
- stripe size recommendations 156
- stripe unit size 153, 243
- sub-logical drives 54
- switched connections 168
- SYMPlicity Storage Manager 203
- synchronizing RAID arrays 68
- System/390 8
- system-attached storage 7

## T

- tape, attached to a SAN 196
- technologies, storage 10
- terminology
  - Fibre Channel 199
  - SSA 228, 274
- Tivoli SANergy 300
- Tivoli Storage Manager 300
- topology
  - Fibre Channel 168, 205
  - SCSI 156
  - SSA 228
- traditional storage architectures 280

## U

- Ultra3 160/m SCSI 58
- UltraATA 8
- unconfigured capacity 199
- UnixWare, using ServeRAID 319
- upper layers, Fibre Channel protocols 170

## W

- Web site
  - ANSI Fibre Channel 171

- ANSI SCSI 58
- ANSI SSA 233
- hot-plug, supported adapters 108
- IBM PC technical support 82
- IBM SANs 296
- IBM storage 249
- Netfinity Fibre Channel 25
- Netfinity SANs 296
- SSA 243, 270
- Windows 2000
  - Advanced Server 37, 41
  - Datacenter Server 41
- write-back cache 65

## X

- X-architecture 301



---

## IBM Redbooks review

Your feedback is valued by the Redbook authors. In particular we are interested in situations where a Redbook "made the difference" in a task or problem you encountered. Using one of the following methods, **please review the Redbook, addressing value, subject matter, structure, depth and quality as appropriate.**

- Use the online **Contact us** review redbook form found at [ibm.com/redbooks](http://ibm.com/redbooks)
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to [redbook@us.ibm.com](mailto:redbook@us.ibm.com)

<b>Document Number</b>	SG24-2098-03
<b>Redbook Title</b>	Netfinity Server Disk Subsystems
<b>Review</b>	          
<b>What other subjects would you like to see IBM Redbooks address?</b>	   
<b>Please rate your overall satisfaction:</b>	<input type="radio"/> Very Good <input type="radio"/> Good <input type="radio"/> Average <input type="radio"/> Poor
<b>Please identify yourself as belonging to one of the following groups:</b>	<input type="radio"/> Customer <input type="radio"/> Business Partner <input type="radio"/> Solution Developer <input type="radio"/> IBM, Lotus or Tivoli Employee <input type="radio"/> None of the above
<b>Your email address:</b> The data you provide here may be used to provide you with information from IBM or our business partners about our products, services or activities.	<input type="checkbox"/> Please do not use the information collected here for future marketing or promotional contacts or other communications beyond the scope of this transaction.
<b>Questions about IBM's privacy policy?</b>	The following link explains how we protect your personal information. <a href="http://ibm.com/privacy/yourprivacy/">ibm.com/privacy/yourprivacy/</a>





**Redbooks**

**Netfinity Server Disk Subsystems**

(0.5" spine)

0.475" <-> 0.875"

250 <-> 459 pages







# Netfinity Server Disk Subsystems



**Covers the major disk subsystem technologies used in Netfinity servers**

**Helps you to select the best storage solution for your server**

**Updated with the latest on Netfinity SANs**

This IBM Redbook is the definitive guide to IBM Netfinity disk subsystems. It is aimed at technical staff within IBM, customers, and business partners who wish to understand the range of available storage options for IBM's Netfinity family of servers. Reading it will provide you with sufficient information to be able to make informed decisions when selecting disk subsystems for Netfinity servers. It will also prove invaluable to anyone involved in the purchase, support, sale, and use of these leading-edge storage solutions.

The storage subsystems covered in the book are those based on the ServeRAID family of SCSI-based RAID controllers, Netfinity Fibre Channel products, and serial storage architecture (SSA) products. All currently available Netfinity products are discussed. Comparisons between the technologies are made, to provide you with guidance when selecting a storage subsystem for your applications.

With the rapid increase in interest in storage area network (SAN) approaches to data storage, and a growing number of Netfinity SAN products, we have included a new section to address this topic in this edition. After an introduction to SAN concepts, we discuss current Netfinity SAN products and solutions, and outline the direction we expect Netfinity SANs to take in the future.

## **INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION**

### **BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE**

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:**  
[ibm.com/redbooks](http://ibm.com/redbooks)