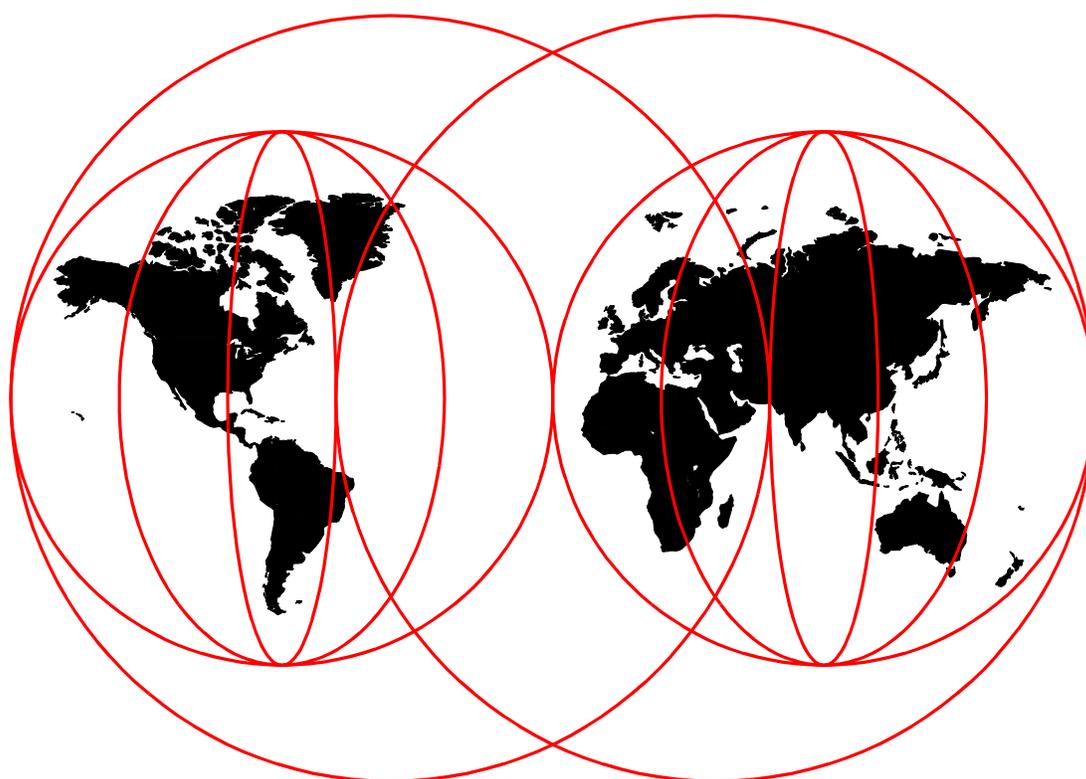


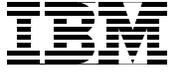
Monitoring and Managing IBM SSA Disk Subsystems

Mark Blunden, Bernd Albrecht, Jean-Paul Bardinet, Barry Mellish



International Technical Support Organization

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



International Technical Support Organization

SG24-5251-00

Monitoring and Managing IBM SSA Disk Subsystems

August 1998

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix H, "Special Notices" on page 183.

First Edition (August 1998)

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

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.

Contents

Figures	vii
Tables	ix
Preface	xi
The Team That Wrote This Redbook	xi
Comments Welcome	xii
Chapter 1. Overview of Serial Storage Architecture (SSA)	1
1.1 Rules for SSA Loops	3
Chapter 2. Major SSA Hardware Components	7
2.1 SSA adapters	7
2.2 SSA Disks	7
2.3 SSA Subsystems	8
2.3.1 3527 SSA Storage Subsystem for PC Servers	8
2.3.2 7131 SSA Multistorage Tower Model 405	9
2.3.3 7133 Serial Storage Architecture (SSA) Disk Subsystem	10
2.3.4 IBM 7190-100 SCSI Host to SSA Loop Attachment	10
2.3.5 IBM 7190-200 Ultra-SCSI Host-to-SSA Loop Attachment	14
2.3.6 Distances between SSA nodes	18
Chapter 3. Tools for Managing SSA Storage	19
3.1 Managing SSA Disks in an AIX Environment	19
3.1.1 Concepts	21
3.1.2 Types of Control	23
3.1.3 Other Tools	38
3.2 Managing SSA Disks in a Sun Solaris Environment	43
3.2.1 Other Tools	44
3.3 Managing SSA Disks in an HP-UX Environment	44
3.4 Managing SSA Disks in a Windows NT Environment	45
3.5 Managing SSA Disks in Other Environments	49
Chapter 4. Introducing IBM StorWatch Serial Storage Expert	51
4.1 StorX Introduction	51
4.1.1 Functional Overview	52
4.2 What is a Host?	53
4.3 What is a Storage Network?	53
4.4 What is a Management Set?	54
4.5 What is on the StorX Display?	54
4.6 Identifying Missing Connections	57
4.7 Device States	57
4.8 Working with the StorX Planner	58
4.9 Working with the StorX Live Viewer	59
4.10 How Do I Discover a Storage Network?	59
4.11 Performance	60
4.12 Event Monitor	60
4.13 Installing StorX	61
4.13.1 System Requirements	61
4.13.2 Licenses	61
4.13.3 Installing the IBM StorWatch Serial Storage Expert from the Web	61
4.13.4 Installing the StorWatch Serial Storage Expert from the Product CD62	61

4.13.5	Operating Requirements	62
4.13.6	Installing the IBM SSA Network Agent for Windows NT 4.0	63
4.13.7	Starting StorX	64
4.14	Problem Analysis	64
4.15	Unsupported Devices	66
4.16	Performance impacts of StorX	66
4.16.1	Impact on Disk Performance	66
4.16.2	Impact on CPU Performance	66
4.16.3	Impact on Network performance	68
4.16.4	Summary	69
4.17	Printing	69
4.17.1	Printing the Canvas	69
4.17.2	Printing Cable Labels	69
4.17.3	Printing Enclosures	70
4.17.4	Displaying Page Boundaries	71
4.17.5	Printing a Large Management Set	71
4.18	Updating StorX	71
4.19	Uninstalling the StorX	71
4.19.1	Planner and Live Viewer	71
4.19.2	Removing StorX Network Agents from an AIX Client	72
Chapter 5. Monitoring and Managing SSA Environment with StorX		73
5.1	Customizing a StorX Live View	73
5.2	Event Monitor	76
5.3	Monitoring and Managing a Shared Configuration	78
5.4	Monitoring and Managing a Complex Configuration	80
5.5	Using StorX for Subsystem Troubleshooting	83
5.6	Monitoring and Managing an hdisk Failure	84
5.6.1	Recovery Procedure	86
5.7	Monitoring and Managing an SSA Open Loop Problem	88
5.7.1	Initial Configuration	88
5.7.2	Recovery Procedure	90
5.8	Monitoring and Managing an SSA Adapter Failure	90
5.8.1	Two SSA Loops Connected to One SSA Adapter Card	90
5.8.2	One SSA Loop Shared by Two Systems	91
5.9	Monitoring and Managing an SSA RAID Array Creation	92
5.9.1	Step1: Hot Spare Disk Definition	93
5.9.2	Step 2: Array Candidate Disk Definitions	94
5.9.3	Step 3: SSA RAID Array Creation	95
5.9.4	Step 4: RAID5VG Volume Group Creation	96
5.10	Monitoring and Managing Disk Problems in a RAID 5 Configuration	97
5.10.1	Pdisk 1 Crash	97
5.10.2	Recovery Procedure	98
5.10.3	New Hot Spare Disk Definition	99
5.10.4	Complete Recovery Procedure	100
5.11	Monitoring, Managing, and Diagnosing with StorX	102
5.11.1	Disk Properties	102
5.11.2	Adapter Properties Window	108
Chapter 6. Availability, Mirroring, and Clustering with SSA		109
6.1	General	109
6.1.1	Definitions	109
6.1.2	SSA Availability Features	111

6.2 AIX	112
6.2.1 High Availability	113
6.2.2 RAID Arrays and Disk Mirroring	113
6.3 Windows NT and Other PC Operating Systems	116
6.3.1 Standard Systems	116
6.3.2 PC Clustering	118
6.4 Other Operating Systems	119
6.4.1 RAID and Disk Mirroring	121
6.4.2 Clustering and High Availability Support	121
Chapter 7. Managing Backup and Recovery	123
7.1 Generalities	123
7.1.1 Basic Backup Strategy	123
7.2 System Backup and Restore Using Standard Utilities	124
7.2.1 Backing Up the AIX Operating System	124
7.2.2 Backing Up a Customer Volume Group: DATAVG	126
7.2.3 Restoring the AIX Operating System: ROOTVG	127
7.2.4 Restoring a Customer Volume Group: DATAVG	127
7.3 Managing Backup and Recovery with ADSM	128
7.3.1 How ADSM Store Client Data?	130
7.3.2 How ADSM Controls Backup, Archive, and Space Management	131
7.3.3 ADSM Database	131
7.3.4 Optional ADSM Tools	132
7.3.5 ADSM Device Support	135
Chapter 8. Performance and Tuning	137
8.1 Introduction	137
8.1.1 Performance at Disk Level	137
8.2 Performance at Adapter Level	139
8.2.1 Number of Disks in the Subsystem	139
8.2.2 Basic Configuration Rules	141
8.2.3 Number of Adapters	142
8.2.4 Adapter Types	146
8.3 Performance on OS Level	149
8.3.1 Guidelines for improving I/O Performance	151
Appendix A. How to Change an AIX Mirrored Disk	153
8.3.2 Fixed Disk - Removing and Replacing a Fixed Disk	153
Appendix B. Booting from SSA Disks	159
Appendix C. Replacing a Mirrored Disk - Documentation	161
C.1 SSA spare tool Scripts	161
C.1.1 Working with the SSA spare tool Scripts	161
C.1.2 Operating Requirements	162
C.1.3 SSA spare tool Scripts	162
C.1.4 Installation Requirements	163
C.1.5 Installing the SSA spare tool	163
C.1.6 Setting Up the Daemons on the Client	163
C.1.7 Setting Up the Daemons on the Server	164
C.1.8 Handling Problems	164

Appendix D. Tuning RAID 5 Arrays and Using Fastwrite Cache	167
Appendix E. Disk Striping and Sizing	169
E.1 Stripe Size	169
E.2 Disk Sizing	170
Appendix F. Laying Out an Oracle Database on AIX Disks.	171
Appendix G. Backing Up and Restoring the Operating System in AIX 4.2 175	
G.1 MKSYSB	175
G.1.1 MKSYSB Tape Images	175
G.1.2 Writing to a Tape Drive	176
G.1.3 Creating a MKSYSB	176
G.1.4 Writing to a File	176
G.1.5 Verifying.	177
G.1.6 Boot Verification	177
G.1.7 Restoring a MKSYSB	178
G.1.8 Restore Menus	180
G.1.9 Restoring Individual Files from a MKSYSB Tape	181
Appendix H. Special Notices	183
Appendix I. Related Publications	185
I.1 International Technical Support Organization Publications	185
I.2 Redbooks on CD-ROMs	185
I.3 Other Publications	185
How to Get ITSO Redbooks	187
How IBM Employees Can Get ITSO Redbooks	187
How Customers Can Get ITSO Redbooks	188
IBM Redbook Order Form	189
LIST OF ABBREVIATIONS	191
Index	193
ITSO Redbook Evaluation	195

Figures

1. Basic SSA Configuration	1
2. Illustration of Spatial Reuse	2
3. Two 7190-100s in One System Sharing One SSA Loop	12
4. Host Logical View of Disks	13
5. 7190-200 Highlights and Compatibilities	16
6. 7190-200 Drawer Highlights.	17
7. 7190-200 SCSI Connected to System and SSA Loop with 7133-600	18
8. SMIT Devices Menu.	23
9. SSA Adapter Menu	24
10. SSA Disks Menu	24
11. Logical Disk Menu	25
12. SSA Physical Disk Menu	25
13. SSA RAID Arrays Menu	26
14. Output of Command maymap - d ssa2 -p	28
15. StorX View of Network Shown in Figure 14 on page 28	29
16. Typical PTX 3D Screen View	43
17. SSA RSM Initial Screen View	46
18. Detailed View of Adapter Product Information.	46
19. Physical View from Adapter	47
20. VPD of Individual Disk	47
21. Logical View from Adapter: Part 1	48
22. Logical View from Adapter: Part 2	48
23. View of Event Logger.	49
24. Sample StorX conceptual view	53
25. StorX Palette	54
26. Event Monitor	60
27. Structure of StorX Functioning	65
28. Discovery of a RS/6000-G40	67
29. Refresh of an RS/6000-G40.	68
30. Uncustomized StorX Live View	73
31. Management Set Properties	74
32. Customized Storx Viewer.	75
33. Customized StorX Viewer with Annotations.	76
34. Event Monitor Display	77
35. Event Monitor Showing the Associated Device	77
36. StorX Live View from RS/6000-1: Logical Disk Option	78
37. StorX Live View from RS/6000-2: Logical Disk Option	79
38. StorX Live View from RS/6000-1.Pdisk Name Option	79
39. StorX Live View from RS/6000-2. Pdisk Name Option	80
40. SSA Planning without Using StorX	81
41. Live View of Uncustomised StorX	82
42. Live View of Customized StorX from System-1, Pdisk Name Option	83
43. Initial Configuration.-Logical Disk Name Option.	84
44. Initial Configuration. -Physical Disk Name Option	84
45. StorX Live View of an hdisk Problem.	85
46. Hdisk Replacement StorX Live View	86
47. Configuration of the New SSA Disk.-Pdisk Name Option	87
48. Recovery Procedure Completed	88
49. Initial Configuration of Two SSA Loops Connected to One SSA Adapter	89
50. StorX Live View of an SSA Open Loop	89

51. Initial Configuration with Two Loops - StorX Live View	90
52. SSA Adapter Problem Affecting Two Loops - StorX Live View	91
53. Initial Configuration with One Loop and Two Adapters - StorX Live View	91
54. SSA Adapter Problem on System1 - StorX Live View	92
55. Initial Configuration before SSA RAID Array Creation - StorX Live View	93
56. Hot Spare Definition StorX Live View	94
57. Hot Spare and Array Candidate Disks - StorX Live View	94
58. SSA Raid Array - StorX Live View	95
59. SSA Raid Array, Pdisk Option - StorX Live View	96
60. Volume Group RAID5VG - StorX Live View	97
61. Crash of Pdisk 1 in a SSA RAID Array	98
62. Replacement of Pdisk 1 - StorX Live View	99
63. Pdisk 8 Definition - StorX Live View	100
64. Figure 59. New SSA RAID 5 Environment. - Pdisk Option	101
65. New SSA RAID5 Environment. - Logical Disk Option	101
66. Two Loops Connected to an SSA Adapter with RAID5 Array as one Loop	102
67. General Page of Disk Properties Window	103
68. Description Page of Disk Properties Windows	104
69. System Page of Disk Properties Window	105
70. Advanced Page of Disk Properties Window	106
71. General Page of Disk Properties Window	107
72. General Page of Disk Properties Windows	107
73. Adapter Properties Window	108
74. Basic SSA Loop Topology	111
75. Features of the 7133 SSA Subsystem	112
76. Octopus and Vinca PC Clusters	117
77. SSA Cluster Adapter for PC Servers	118
78. Microsoft Cluster Server Configuration: Part 1	119
79. Microsoft Cluster Server: Part 2	119
80. Back Up the System SMIT Screen	126
81. Back Up a Volume Group SMIT Screen	127
82. Restore Files in a Volume Group SMIT Screen	128
83. Platforms Supported by ADSM	129
84. How ADSM Stores Client Data	130
85. How ADSM Controls Backup, Archive, and Space Management	131
86. Hierarchical Storage Management	133
87. Disaster Recovery Manager Overview	135
88. Performance Chart 1 - SSA versus SCSI	140
89. Performance Sample 1 - Single Loop	142
90. Performance Sample 2 - Two Loop	142
91. Performance Sample 3 - Four Loops and Two Adapters	144
92. Performance Sample 4 - Two Loops on Two Adapters	145
93. Performance Sample 5 - Four Loops on Three Adapters	146

Tables

1. SSA Adapters	7
2. SSA Disks	7
3. 3527 SSA Storage Subsystem for PC server	8
4. 7131-Model 405 SSA Multi-Storage Tower Specifications	9
5. 7133 Models 10, 20, 500, and 600 Specifications	10
6. 7190-100 Host to SSA Loop Attachment:	14
7. 7190-200 Ultra-SCSI Host to SSA Loop Attachment Specifications	17
8. Maximum Distance Between Two SSA Nodes	18
9. SSA Adapter Features and Compatibility	20
10. Vicom and IBM Product Comparison	50
11. StorX Toolbar Icons	55
12. StorX Parts Icons	56
13. StorX Device States	57
14. IBM 7190 and Vicom Product Details	120
15. RS/6000 Models that Support Booting from SSA disks	125
16. SSA Disk Parameter	137
17. Comparison of Adapters	143
18. Overview RAID Hardware	147
19. RS6000 Models that Support Booting from SSA Disks	159

Preface

This redbook describes how to design, configure, monitor, and control SSA disk subsystems that are connected to IBM and non-IBM operating system servers.

The material demonstrates how to manage SSA loop configurations using various storage management facilities such as AIX Logical Volume Manager, StorX, and the latest Web user-interface technology.

Platform-specific storage management offerings are discussed, and customers are provided with a complete and comprehensive guide showing how to perform flexible, low-risk, remote-site, open storage subsystem design, configuration, and management in a variety of SSA disk subsystem environments.

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 (location) Center.

Mark Blunden is the project leader for Open Systems Storage at the International Technical Support Organization, (location) Center. He has coauthored four previous redbooks and teaches IBM classes worldwide on all areas of Storage. Mark has worked for IBM for 18 years in many areas of the IT business. Before joining the ITSO in 1998, Mark worked in Sydney, Australia, as an Advisory Storage Specialist

Bernd Albrecht is a Systems Engineer in Germany. He joined the IBM in 1991 as Systems engineer in the Telecommunications Business unit. He moved into the RISC Sysytem/6000 business unit when it was set up in 1993. He worked in various fields such as RS/6000 SP, and product support groups for networking, AIX and performance. His current position is in the Product Support Group for Storage at the Technical Presales Support Germany/EMEA Central Region. This is Bernhard's first ITSO redbook.

Jean-Paul Bardinet is a Support Center Customer Engineer in France. He joined IBM in 1984 in the manufacturing division. He joined Operational Services in 1990 where he worked for 3 years on large systems. In 1992 he moved to the RISC Sysytem/6000 unit. He has 5 years experience in the RS/6000 field. His current areas of expertise include RS/6000, SP2, Magstar, and the SSA storage environment. This is Jean-Paul's first ITSO redbook.

Barry Mellish is a Senior Storage Specialist in the UK. He has worked for IBM for the last 15 years. Barry joined IBM as a Property Services Engineer responsible for IBM premises in Central London. He moved into system engineering 11 years ago, initially working on mid range systems, he started specializing in the IBM 6150, the forerunner of today's RS/6000. He joined the AIX Business Unit when it was set up following the launch of the RS/6000 in 1990. Barry has worked extensively with Business Partners and Systems Integrators, providing technical support with systems design. Over the last three years he has specialized in storage and storage systems, joining SSD EMEA when it was set up in January 1997. His current role is as a member of the UK open systems storage team within MSS. Barry is also the coauthor of the IBM Versatile Storage Server redbook.

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

Dave McAuley, ITSO San Jose Center

Dean Underwood, IBM Tucson

Earl Timmons, IBM Rochester

Gary Axberg, IBM Rochester

Dan Morcom, IBM Rochester

Fay Weamer, IBM Rochester

Jon Schmidt, IBM Rochester

Mike Griese, IBM Rochester

Dan Braden, IBM Dallas

David Sinclair, IBM UK

Ron Case, IBM San Jose

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 "ITSO Redbook Evaluation" on page 195 to the fax number shown on the form.
- Use the electronic evaluation form found on the Redbooks Web sites:

For Internet users <http://www.redbooks.ibm.com>

For IBM Intranet users <http://w3.itso.ibm.com>

- Send us a note at the following address:

redbook@us.ibm.com

Chapter 1. Overview of Serial Storage Architecture (SSA)

The following is a brief description of SSA and the basic rules to follow when designing SSA networks. For a full description of SSA and its functionality, please read *A Practical Guide to Serial Storage Architecture for AIX, SG24-4599-00*.

SSA is a high performance, serial interconnect technology used to connect disk devices and host adapters. SSA is an open standard, and SSA specifications have been approved by the SSA Industry Association and are approved as an ANSI standard through the ANSI X3T10.1 subcommittee.

SSA subsystems are built up of loops of adapters and disks. A simple example is shown in Figure 1.

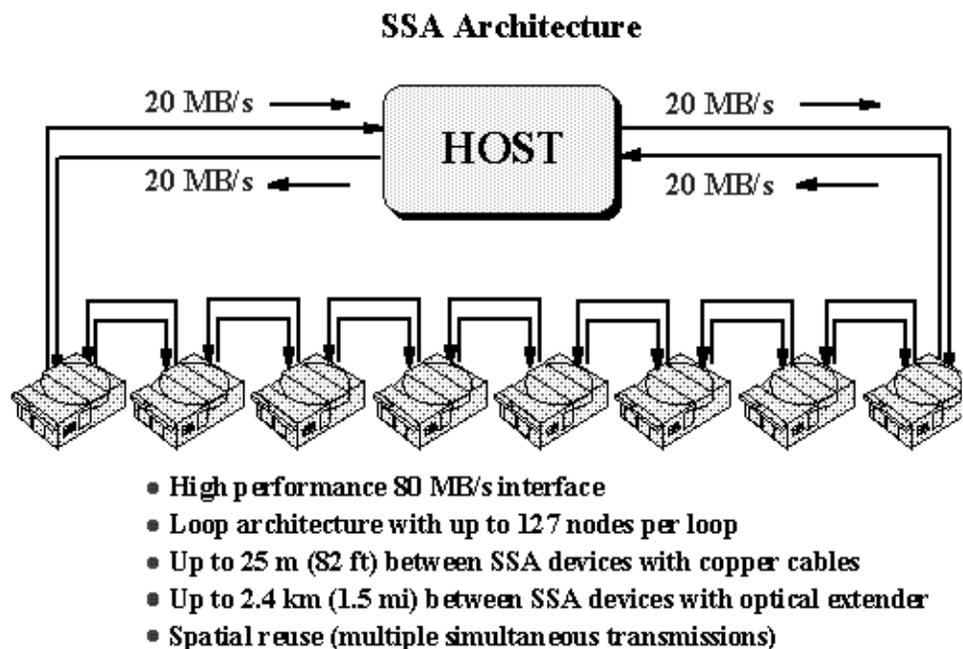


Figure 1. Basic SSA Configuration

Here, a single adapter controls one SSA loop of eight disks. Data can be transferred around the loop, in either direction, at 20 MB/s, consequently, the peak transfer rate of the adapter is 80 MB/s. The adapter contains two SSA nodes and can support two SSA loops. Each disk drive also contains a single SSA node. A node can be either an initiator or a target. An *initiator* issues commands, while a *target* responds with data and status. The SSA nodes in the adapter are therefore initiators while the SSA nodes in the disk drives are targets. Each SSA node is given a unique address at manufacturing time, known as the *UID*. This allows the initiators in the loop to determine what other SSA nodes have been attached to the loop and to understand the SSA loop topology.

SSA is a point-to-point architecture. This means that when two adjacent SSA nodes are communicating, the rest of the SSA network is free to allow other node-to-node communications to take place. This concept is known as *spatial reuse*; see Figure 2 on page 2. 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. The SSA network is managed by one particular initiator, known as the *master initiator*. This is the initiator with the highest UID. If a new initiator is added to the network with a higher UID than those already present, then it will take over the master responsibilities for that network.

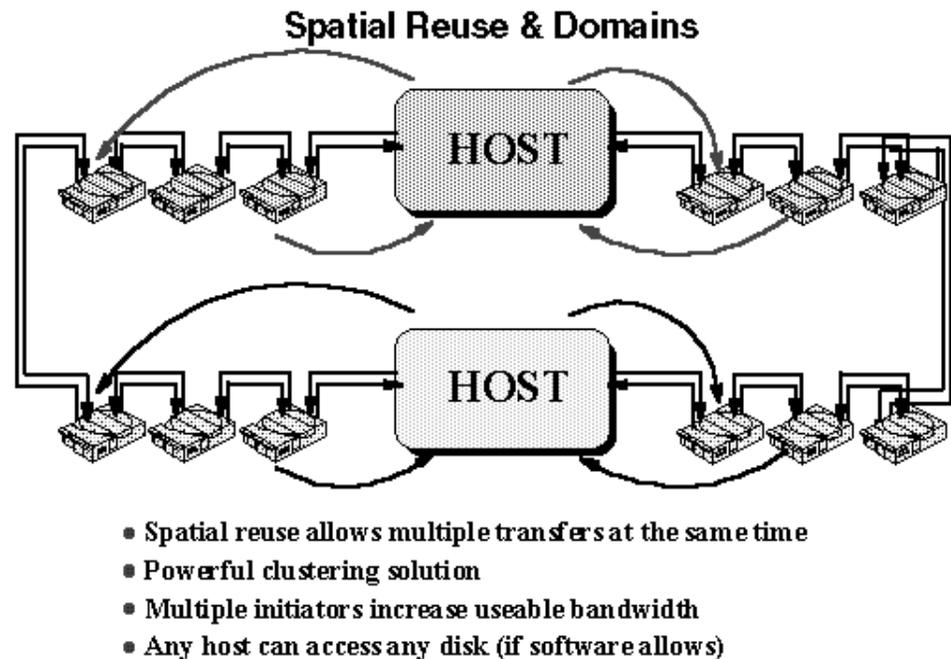


Figure 2. Illustration of Spatial Reuse

Similarly, if a master initiator is removed from the SSA network, then the initiator with the next highest UID takes over master responsibility. This master takeover and usurpation occurs automatically without any user intervention. Unlike SCSI and fiber channel arbitrated loop, SSA does not use bus phases. Also, SSA does not use arbitration to contend for the bus. Instead, it makes use of frame multiplexing. All data is transported as frames, which are the basic unit of data transferred between SSA nodes. A frame can contain up to 128 data bytes. The loop network has several advantageous characteristics:

- Full duplex support is provided on each link, so that traffic can be present in both directions on the loop simultaneously.
- The loop also supports spatial reuse; that is, different frames may be in between different devices on the loop concurrently. For instance, a frame could be moving from Disk 1 to Disk 2 at the same time as the adapter is sending a frame to Disk 1 (since each disk is dual ported).
- Since a loop topology is used, should a break occur in the loop (for example a cable is disconnected), each adapter on the loop adjusts its routing algorithms, under direction from the master initiator, so that frames are automatically rerouted to avoid the break. This allows devices to be removed or added to the loop while the subsystem continues to operate without interruption.

Because SSA allows SCSI-2 mapping, all functions associated with initiators, targets, and logical units are translatable. Therefore, SSA can use the same command descriptor blocks, status codes, command queuing, and all other aspects of current SCSI systems. The effect of this is to make the type of disk subsystem transparent to the application. No porting of applications is required to move from traditional SCSI I/O subsystems to high performance SSA. SSA and SCSI I/O systems can coexist on the same host running the same applications.

The advantages of SSA are summarized as follows:

- Dual paths to devices
- Simplified cabling - cheaper, smaller cables and connectors, no separate terminators
- Faster interconnect technology
- Not an arbitrated system
- Full duplex, frame multiplexed serial links
- 40 MB/s total per port, resulting in 80 MB/s total per node, and 160 MB/s total per adapter
- Spatial reuse allows concurrent activity in different portions of the same loop
- Hot pluggable cables and disks
- Very high capacity per adapter - up to 127 devices per loop, although most adapter implementations limit this. For example, current IBM SSA adapters provide 96 disks per Micro Channel or PCI slot)
- Distance between devices of up to 25 m (82 ft) with copper cables, 2.4 km (1.5 mi.) with optical links.
- Auto-configuring - no manual address allocation
- SSA is an open standard
- SSA switches can be introduced to produce even greater fan-out and more complex topologies

1.1 Rules for SSA Loops

The following rules must be followed when configuring and connecting SSA loops:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2).
- Only one of the two pairs of connectors on an adapter card can be connected in a particular SSA loop.
- A maximum of 48 devices can be connected in a particular SSA loop
- A maximum of two pairs of adapter connectors can be connected in a particular loop if one adapter is an SSA 4-Port adapter, Feature 6214
- A maximum of eight pairs of adapter connectors can be connected in a particular loop if all the adapters are Enhanced SSA 4-Port Adapters, Feature 6216
- A maximum of two SSA adapters, both connected in the same SSA loop, can be installed in the same host using system.

For SSA loops that include an SSA Four-Port RAID adapter, Feature 6217, or a PCI SSA Four-Port RAID adapter (Feature 6218), the following rules apply:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2)
- A maximum of 48 devices can be connected in a particular SSA loop
- Only one pair of adapter connectors can be connected in a particular SSA loop
- Member disk drives of an array can be on either SSA loop

For SSA loops that include a Micro Channel Enhanced SSA Multi-initiator/RAID EL adapter, Feature 6215 or a PCI SSA Multi-initiator/RAID EL adapter, Feature 6219, the following rules apply:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2)
- A maximum of two adapters can be connected in a particular loop if none of the disk drives in the loops are array disk drives, configured for fast-write operations. The adapters can be two Micro Channel Enhanced SSA Multi-initiator/RAID EL Adapters, two PCI Multi-initiator/RAID EL Adapters, or one adapter of each type.
- Only one Micro Channel Enhanced SSA Multi-initiator/RAID EL Adapter or PCI SSA Multi-initiator/RAID EL Adapter can be connected in a particular loop if any disk drives in the loops are members of a RAID-5 array, or are configured for fast-write operations
- All member disk drives of an array must be on the same SSA loop
- A maximum of 48 devices can be connected in a particular SSA loop
- Only one pair of adapter connectors can be connected in a particular loop
- When an SSA adapter is connected to two SSA loops, and each loop is connected to a second adapter, both adapters must be connected to both loops.

For the IBM 7190-100 SCSI to SSA converter, the following rules apply:

- Up to 48 disk drives per loop
- Up to four IBM 7190-100 can be attached to any one SSA loop

For the Sun SBus adapter the following rules apply:

- Up to 48 disk drives per loop
- Up to eight adapters in any one loop
- Up to two adapters in any single Sun host

For the PC RAID adapter, the following rules apply:

- One adapter per SSA loop
- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2).
- Only one of the two pairs of connectors on an adapter card can be connected in a particular SSA loop.

For the PC RAID cluster adapter, the following rules apply:

- Up to 22 SSA RAID-1 arrays can be supported per cluster.
- Two disks per RAID-1 array
- Up to three SSA adapters for each server in the cluster.

Chapter 2. Major SSA Hardware Components

In this chapter we discuss SSA adapters, SSA disks, and SSA subsystems. SSA performance discussed in Chapter 8, "Performance and Tuning" on page 137.

2.1 SSA adapters

Table 1 lists the different SSA adapters and presents an overview of their characteristics.

There are two loops per adapter (except for Feature 4003).

Each loop per adapter can manage 48 disks except Feature 4012.

Table 1. SSA Adapters

Feature code	Bus	Operating system	Adapter description	Adapters per system	Diagnostic and maintenance	Hardware Raid types
6214	MCA	AIX	Classic	4	Diag Smit	n/a
6215	PCI	AIX	Enhanced	4	Diag SMIT	5
6216	MCA	AIX	Enhanced	4	Diag SMIT	5
6217	MCA	AIX	RAID-5	4	Diag SMIT	5
6218	PCI	AIX	RAID-5	4	Diag SMIT	5
6219	MCA	AIX	Enhanced	4	Diag SMIT	5
4003	SBus	Solaris 2.4 2.5.1	Classic	2	SSAU SSACF	n/a
4011	PCI	Windows NT, OS/2, Novell NetWare, DOS		2	SSA RSM	0, 1, 5
4012	PCI	Windows NT		3	SSA RSM	0, 1, 5
RPQ: 07H9604	PCI	Windows NT, OS/2, Novell Netware	RAID	3		0, 1, 5

2.2 SSA Disks

Table 2 on page 7, lists the different SSA disks, and provides an overview of their characteristics

Table 2. SSA Disks

Name	Capacities (MB)	Buffer size (KB)	Transfer rate (max)
Starfire 1100	1.1	0	20

Name	Capacities (MB)	Buffer size (KB)	Transfer rate (max)
Starfire 2200	2.2	0	20
Starfire 4320	4.5	512	20
Scorpion 4500	4.5	512	80
Scorpion 9100	9.1	512	160
Sailfin 9100	9.1	1024	160
Thresher 9100	9.1	1024	160

2.3 SSA Subsystems

The five SSA subsystems are discussed in this section:

1. 3527 SSA Storage Subsystem for PC Servers
2. 7131 SSA Multi-Storage Tower Model 405
3. 7133 Serial Storage Architecture (SSA) Disk Subsystem
4. IBM 7190-100 SCSI Host to SSA Loop Attachment
5. IBM 7190-200 Ultra-SCSI Host to SSA Loop Attachment

2.3.1 3527 SSA Storage Subsystem for PC Servers

The IBM 3527 SSA Storage Subsystem for PC Servers uses SSA technology to deliver outstanding performance, highly scalable capacity, and enhanced availability.

The 3527 SSA Storage Subsystem uses the SSA RAID Adapter for PC Servers (refer to item RPQ 07H9604 in Table 1 on page 7,) to run under a variety of operating systems such as:

- IBM Windows NT 3.5.1 or 4.0, Novell Netware 4.1 or 4.1 SMP, OS/2 Warp Server or Warp Server SMP, and OS/2 2.11 SMP
- Compaq Windows NT 4.0
- HP Windows NT 4.0

The 3527 SSA Storage Subsystem uses the SSA RAID Adapter for PC Servers and supports various PC Servers such as:

- IBM PC Server families PC320, PC325, PC330, PC520, PC704 and PC720
- Compaq ProLiant 1500, 2500, 5000 and 6000
- HP NetServer LH Pro and LX Pro

Table 3 lists the 3527 specifications.

Table 3. 3527 SSA Storage Subsystem for PC server

Item	Specification
Transfer Rate SSA Interfaces	80 MB/s drive interface bandwidth
Configuration	2 to 5 disks (4.5 GB or 9.1 GB) per subsystem

Item	Specification
Configuration range	9.0 GB-22.5 GB (with 4.5 GB disks) 18 GB-45.5 GB (with 9.1 GB disks)
Supported RAID level	0, 1, and 5
Hot swappable disks	Optional

2.3.2 7131 SSA Multistorage Tower Model 405

i7131 SSA Multistorage Tower Model 405 provides a solution for those who require the high performance of SSA and for applications that exceed the capacity of direct SCSI-attached storage. The 7131 Model 405 delivers outstanding performance, high capacity, and enhanced availability.

The 7131 SSA Multistorage Tower Model 405 runs under various operating systems such as:

- IBM - AIX 4.1.4 (and above) and AIX 4.2, Windows NT 3.51 or 4.0, Novell NetWare 4.1 or 4.1 SMP, OS/2 Warp Server or Warp Server SMP, and OS/2 2.11 SMP
- HP - HP-UX 10.01, 10.10, or 10.20
- Sun - Solaris 2.4, 2.5, or 2.51

7131 SSA Multistorage Tower Model 405 is supported on various systems such as:

- IBM - Supported on RS/6000 and RS/6000 SP systems with adapters Feature 6214, 6216, 6217, 6218, and PC Server systems with the SSA RAID Adapter for PC Server (refer to Table 1, RPQ 07H9604)
- HP - HP 9000 Series 800 with the IBM 7190 SCSI Host to SSA Loop Attachment
- Sun - Supported on selected SPARCstation, SPARCserver, SPARCcenter, and Ultra Enterprise models with the IBM 7190 SCSI Host to SSA Loop Attachment

Table 4 lists the 7131 Model 405 specifications.

Table 4. 7131-Model 405 SSA Multi-Storage Tower Specifications

Item	Specification
Transfer rate SSA interface	80 MB
Configuration	2 to 5 disk drives (4.5 GB or 9.1 GB) per subsystem
Configuration range	Up to 4.4 GB to 11 GB (with 2.2 GB disk drives) Up to 9.0 GB to 22.5 GB (With 4.5 GB disk drives) Up 18.2 GB to 45.5 GB (With 9.1 GB disk drives)
Supported RAIDS	0, 1, and 5
Hot swap disks	Yes

2.3.3 7133 Serial Storage Architecture (SSA) Disk Subsystem

The IBM 7133 SSA disk subsystem uses SSA technology to deliver outstanding performance, highly scalable capacity, and enhanced availability.

The 7133 models 10 and 500 were the first SSA products announced in 1995 with the revolutionary new Serial Storage Architecture. Some IBM customers still use the Models 10 and 500, but these have been replaced, respectively by 7133 Model 20, and 7133 Model 600.

7133 Models 10, 20, 500 and 600, have redundant power and cooling, which is hot swappable.

The 7133 runs under various operating systems, such as:

- IBM - AIX 3.2.5 (with PTFs), AIX 4.2, Windows NT 3.51 or 4.0, Novell NetWare 4.1 or 4.1 SMP, OS/2 Warp Server or Warp Server SMP, and OS/2 2.11 SMP
- HP - HP-UX 10.01, 10.10, or 10.20
- Sun - Solaris 2.4, 2.5, or 2.5.1

The 7133 supports various systems such as:

- IBM - RS/6000 and RS/6000 SP systems with SSA adapters FC 6214, FC 6216, FC 6217, and FC 6218 and on PC Server systems with the SSA RAID Adapter for PC Server (Refer to Table 1, RPQ 07H9604)
- HP - HP 9000 series 800 with the IBM 7190 SCSI Host to SSA Loop Attachment
- Sun - Supported on selected SPARCstation, SPARCcenter, and Ultra Enterprise models with the IBM 7190 SCSI Host to SSA Loop Attachment

Table 5 lists the 7133 specifications.

Table 5. 7133 Models 10, 20, 500, and 600 Specifications

Item	Specification
Transfer rate SSA interface	80 MB/s
Configuration	4 to 16 disks (1.1 GB, 2.2 GB, 4.5 GB, for Models 10, 20, 500, and 600, and 9.1 GB for Models 20 and 600 only. With 1.1 GB disk drives you must have 8 to 16 disks.
Configuration range	8.8 to 17.6 (with 1.1 GB disks) 8.8 to 35.2 GB (with 2.2 GB disks) 18 to 72 GB (with 4.5 GB disks) 36.4 to 145.6 GB (with 9.1 GB disks)
Supported RAID level	5
Hot swappable disk	Yes (and hot swappable redundant power and cooling)

2.3.4 IBM 7190-100 SCSI Host to SSA Loop Attachment

The 7190-100 can be considered a gateway that enables other systems that support SCSI Fast/Wide attachment to work with all IBM SSA subsystems. So with the 7190-100, you can improve the performance of your Sun, HP and DEC Fast Wide SCSI adapters and device drivers.

2.3.4.1 Powerful SCSI Host to SSA Loop Attachment

As the number of storage devices increases on a SCSI bus, performance decreases dramatically.

The IBM 7190-100 SCSI Host to SSA Loop Attachment is an external device that enables attachment of high-performance SSA subsystems, such as IBM 7131 Model 405 and the 7133, to existing SCSI-based systems. Key features are these:

- The 7190-100 overcomes key SCSI limitations by providing increased bandwidth, reliability, number of devices per channel, and cabling flexibility of the SSA.
- The 7190-100 provides maximum SCSI data throughput of as much as 18 MB/s, with up to 1,900 I/O per second.
- The 7190-100 supports up to 48 SSA disks and 437 GB on a single loop (but only four 7190-100 units per loop). You can run any combination of the above SSA disks of 2.2 GB, 4.5 GB or 9.1 GB disks.
- The 7190-100 provides outstanding scalability and performance. It maintains the SCSI features such as tagged command queuing, scatter/gather, disconnect/reconnect, and synchronous as well as asynchronous data transfers. It provides the SSA technology and offers up to 80 MB/s of bandwidth for faster and more efficient data transfers, it supports the spatial reuse characteristics, and allows the attached adapter and host to transfer data simultaneously to and from the loop SSA disks. In addition, SSA's unique loop architecture and auto-addressing capability enables disks to be added without powering down the systems.

2.3.4.2 High Reliability for Improved Availability

SSA's node architecture provides high reliability for critical storage applications by eliminating single points of path failure on the loop. Should an SSA loop be broken, the 7190 automatically reconfigures the loop into two strings - transparently to the host - with no loss of data. When the defective SSA cable is replaced, the device reconfigures the disk into a loop once again. Additionally, the SSA disk drives are hot swappable, allowing users to add or remove SSA devices while the disk subsystem is powered on. Applications running on the host system continue to have access to existing disks on the SSA loop. Further, the 7190-100 attachment uses parity to ensure that the highest data integrity is maintained. This parity information is designed to protect the data transmitted between the SCSI host adapter and the IBM 7190. Data on the serial SSA loop is protected by cyclical redundancy check bits.

Figure 3 on page 12 shows two 7190-100 controllers in one system and sharing an SSA loop.

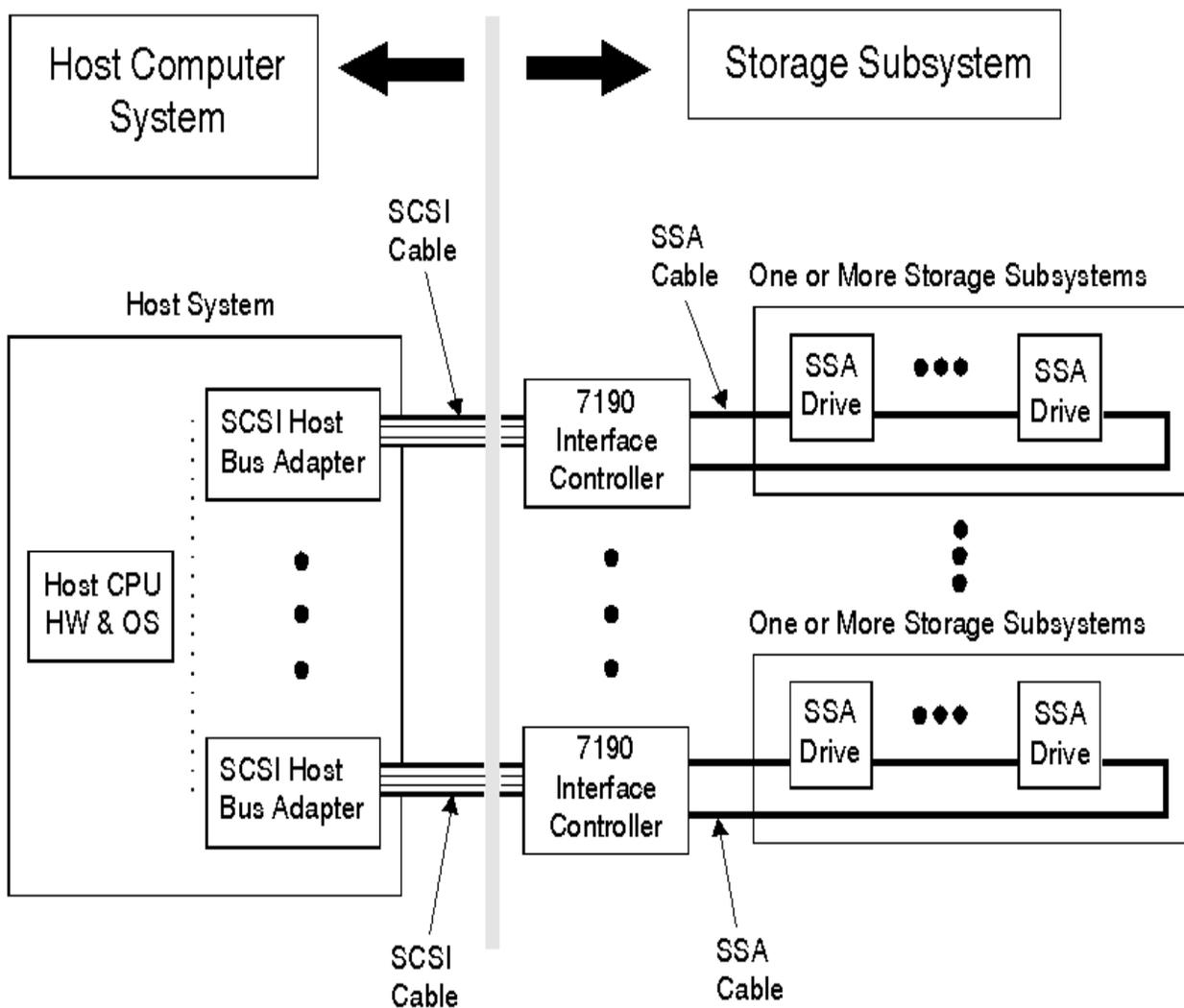


Figure 3. Two 7190-100s in One System Sharing One SSA Loop

Figure 3 also indicates where the SCSI and SSA interfaces begin and end.

The host operating system sees the system differently. The 7190-100 is transparent to the system. The operating system therefore sees the SSA drives as a group of SCSI drives, as indicated in Figure 4 on page 13.

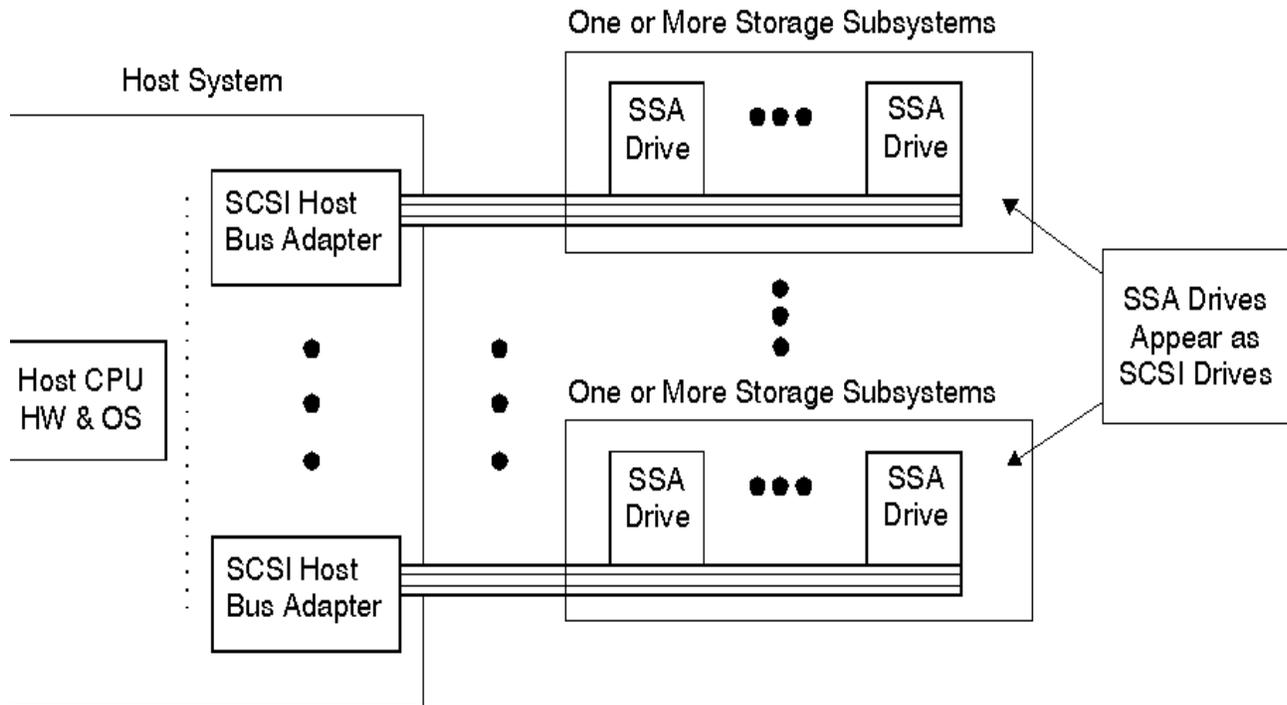


Figure 4. Host Logical View of Disks

Online Service Monitoring

Service features built into the 7190 allow Sun and HP systems to efficiently monitor all attached storage subsystems. These online capabilities include:

- Displaying vital product data (VPD)
- Showing the SSA loop topology
- Showing the SSA-to-SCSI ID-to-LUN mapping
- Setting or changing modes for drive identification and diagnostics
- Downloading microcode to the 7190-100 and SSA disk drives as required
- Providing a detailed system wide error log
- Alerting the system to errors in the redundant power and cooling systems and loop topology changes

The IBM 7190-100 SCSI Host to SSA Loop Attachment runs under various operating systems, such as:

- Sun solaris 2.4, 2.5 or 2.51 for Sun servers
- HP-UX 10.01, 10.10, or 10.20 for HP Servers
- Windows NT 3.5 through 4.0
- Digital UNIX 3.2B through 3.2G and 4.0 through 4.0C for digital servers

The IBM 7190-100 SCSI Host to SSA Loop Attachment runs under various systems, such as:

- Sun SPARCstation 10 and 20, SPARCserver Models 1000 and 1000E, SPARCcenter Models 2000 and 2000E, and Ultra Enterprise Models 2, 150, 3000, 4000, 5000, and 6000.
- HP 9000 series 8000
- Digital AlphaServer 300, 800, 1000, 1000A, 2000, 2100, 2100A, 4000, 4100, 8200, and 8400 series

The IBM 7190-100 SCSI Host to SSA Loop Attachment can be attached to 7133 SSA Storage Subsystems, and to the 7131 Model 405.

Table 6 on page 14 lists the 7190-100 specifications.

Table 6. 7190-100 Host to SSA Loop Attachment:

Item	Specification
Transfer Rate SSA Interface	80 MB
Configuration	2 to 48 disks (2.2 GB, 4.5 GB, and 9.1 GB)
Configuration range	4.4 GB to 105.6 GB (with 2.2 GB disks) 9.0 GB to 216 GB (with 4.5 GB disks) 18.2 GB to 437 GB (with 9.1 GB disks)
Support RAID level	5
Hot swappable disks	Yes

2.3.5 IBM 7190-200 Ultra-SCSI Host-to-SSA Loop Attachment

The 7190-200 Ultra-SCSI Host Loop Attachment announced on April 21 1998, provides HP, Sun, and Digital products with the flexibility and performance of serial disk storage.

SSA Management is the same as in 7190-100, but greatly improved. The 7190-200 offers the following features:

- Enables SCSI-based servers to benefit from the performance improvements of non arbitrated serial disks. Each Ultra SCSI adapter performs at a sustained speed up to 2600 I/O per second and 29 MB/s depending on the server and SCSI adapter used. Up to four 7190-200 interfaces can be connected to the SSA loop. The potential sustained performance is increased by 2600 I/O per second and 29 MB/s for each additional 7190-200.
- Supports advanced SCSI features which enhance the performance of HP, Sun, and Digital adapters and device drivers. SCSI tagged command queuing, scatter/gather, and disconnect/ reconnect benefit from the 7190-200 inter-face's ability to perform multiple simultaneous reads and writes to multiple non-arbitrated SSA disks.
- The SCSI adapter sees the 7190-200 as a single SCSI target and the SSA disks as parallel SCSI logical unit numbers (LUNs).The 7190-200 automatically groups up to 48 SSA disks into 15 balanced SCSI LUNs. The LUNs can be configured through administrative utilities and can be used by list-based RAID managers.

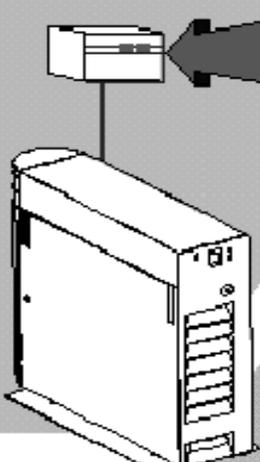
- Enables highly scalable configurations for HP, Sun, and Digital servers. Each server may have multiple SCSI adapter connections to a single loop of an SSA disk, providing redundant connections to data. A pair of servers with redundant SCSI adapter connections provides the highest level of availability and facilitates remounting a file system from a failed host to a backup host. Additional software may be required. Fiber-optic extenders enable 7133 can be geographically separated by up to 2.4 km (1.5 mi.) while maintaining the high transaction rates of the 7190-200.
- Supports up to 48 disks and 437 GB on a single SSA loop. Disk drives and servers may be added or removed while the entire configuration remains online and operational. Each SSA component, down to the individual power supply, disk drive, and interface adapter can be remotely configured and diagnosed by using the server's existing remote console facility.
- Includes a drawer option for integrating up to four 7190-200s into a single rack-mounted drawer. The drawer of 7190-200 may be connected to over 1.7 TB of 7133 SSA disk drawers delivering up to 10.400 I/O per second and 116 MB/s across full Ultra SCSI ports.

Figure 5 on page 16 shows the 7190-200 highlights and compatibilities.

The IBM 7190-200

**Improved Data Throughput for Open Systems Hosts
The High Performance of the 7133 Now Available to
UltraSCSI Users**

Entry Level



7190-200 Ultra-SCSI

- ✓ Up to 40 MB/s Bandwidth
- ✓ Up to 29 MB/s
- ✓ Up to 2,600 IO/s
- ✓ Up to 2.4km with Fiber Optic Extender

7133-600 SSA Disk Towers*

- 18 GB to 435 GB
- 1 to 3 towers
- 4 to 48 disks
- One SSA-80 loop

*or 7131-405 for lower cost entry

Tested, supported, set-up, and installed by IBM on:

- HP with HP-UX
- DEC with UNIX
- DEC with NT
- Sun with Solaris

Figure 5. 7190-200 Highlights and Compatibilities

Figure 6 on page 17 shows the 7190-200 drawer highlights.

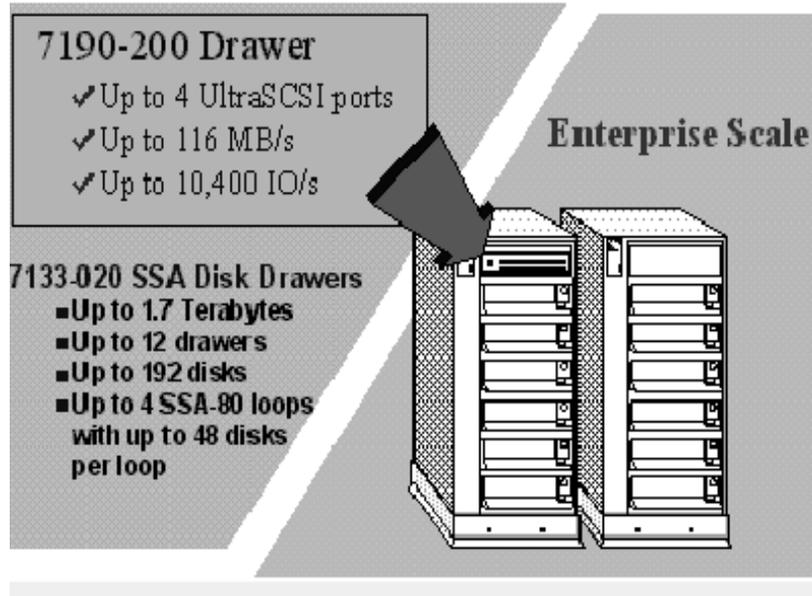


Figure 6. 7190-200 Drawer Highlights

Table 7 shows the 7190-200 specifications

Table 7. 7190-200 Ultra-SCSI Host to SSA Loop Attachment Specifications

Item	Specification
Transfer Rate SSA Interface	160 MB
Configuration	2 to 48 disks (2.2 GB, 4.5 GB, and 9.1 GB)
Configuration range	4.4 GB-105.6 GB (with 2.2 GB disks) 9.0 GB-216 GB (with 4.5 GB disks) 18.2 GB-437 GB (With 9.1 GB disks)
Support RAID level	5
Hot swappable disks	Yes

The 7190-200 systems, operating systems, and storage subsystems are the same as the 7090-100.

Figure 7 on page 18 represents an example of 7190-200 connected with a SCSI attachment to an HP, SUN, or Digital system, and connected with SSA attachment to 7133-600 with fiber-optic extenders

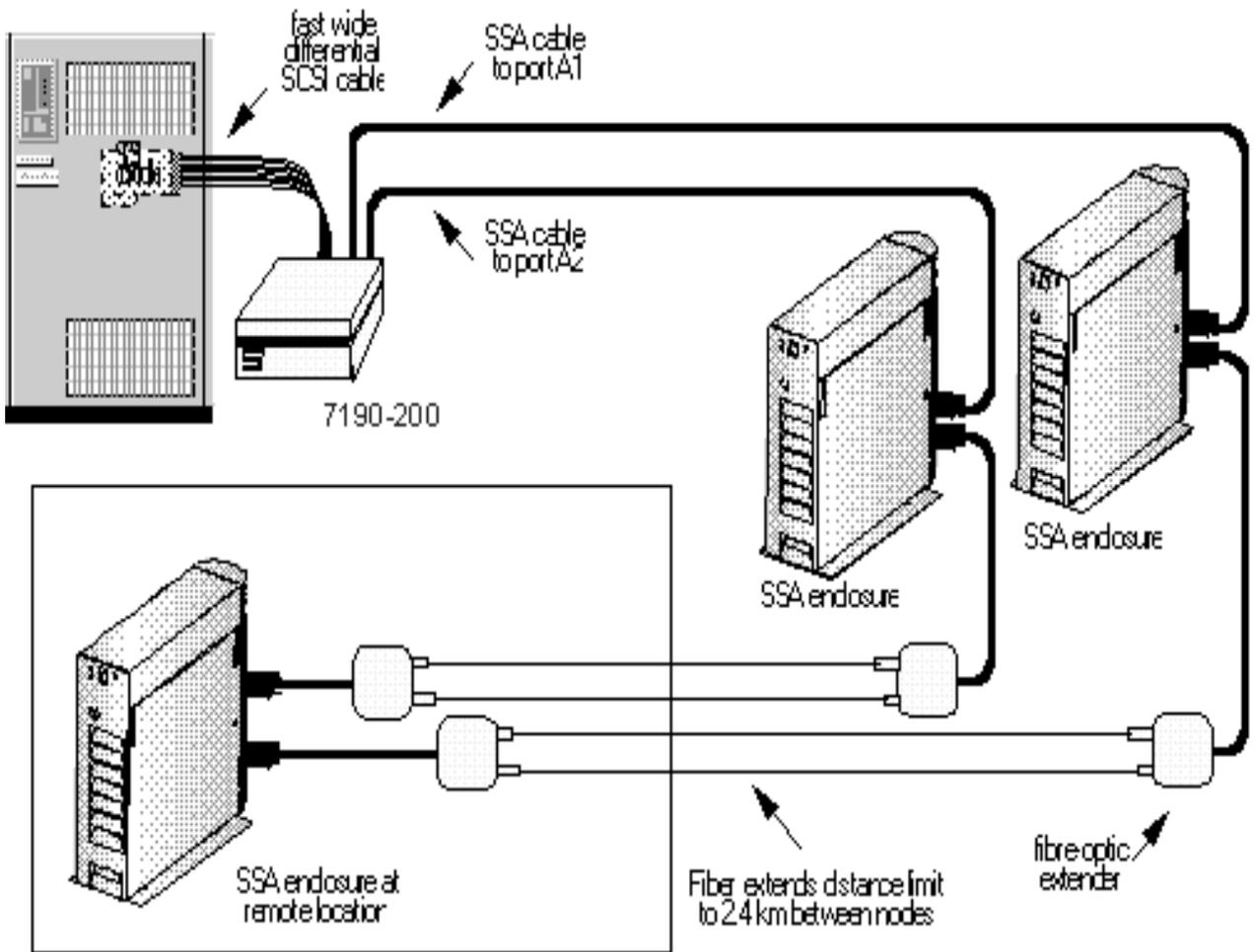


Figure 7. 7190-200 SCSI Connected to System and SSA Loop with 7133-600

2.3.6 Distances between SSA nodes

Table 8 on page 18 shows the maximum distances between two SSA nodes.

Table 8. Maximum Distance Between Two SSA Nodes

Name	Distance (Max)
3527 SSA Storage Subsystem for PC Server	0.5m to 25m (1.6ft to 82ft)
7131 SSA Multi-Storage Tower Model 405	0.5m to 25m (1.6ft to 82ft)
7133 Serial Storage Architecture (SSA) Disk Subsystem	0.5m to 25m (1.6ft to 82ft) with copper cables up to 2400 m (1036 ft) with the fiber-optic extender
IBM 7190-100 SCSI Host to SSA Loop Attachment	Up to 25m (82 ft) with copper cable
IBM 7190-200 Ultra-SCSI Host to SSA Loop Attachment	0.5m to 25m (1.6ft to 82 ft) with copper cables, and up to 2400m (1.4mi.) with fibre-optic if attached to 7133 Mod 20 or 600

Chapter 3. Tools for Managing SSA Storage

In this chapter, we discuss the tools that are available for managing SSA disk subsystems. We consider five different scenarios:

- AIX
- Sun Solaris
- HP-UX
- Windows NT
- Other

Each operating system usually has some form of storage management capability. In addition to this, tools and utilities are provided with the SSA adapter, the IBM 7190 or the Vicom SLIC.

Care must be taken to ensure that all the adapters or SCSI to SSA converters in the same loop are compatible with one another. The basic guidelines are:

1. All Vicom products and the IBM 7190 are compatible with one another, regardless of the host to which they are attached. You can have up to 16 Vicom adapters in a loop, but only four IBM 7190s.
2. The Sun SBus adapter cannot be put in a loop with any other type of adapter. You can have up to eight SBus adapters per loop.
3. The IBM 6214 and 6216 can coexist in the same loop. You can have two adapters per loop if one of them is the 6214; if they are all 6216s, then you can have up to eight adapters per loop.
4. The IBM 6215 and 6219 can coexist in the same loop. The maximum is two adapters per loop, but only if the disks are Just a Bunch Of Disks (JBOD). If a RAID array is configured, then there can be only one adapter in the loop.
5. The IBM 6217 and 6218 are single host attach only (one adapter per loop).

For full details, please refer to Table on page 20.

3.1 Managing SSA Disks in an AIX Environment

The SSA disk subsystem was designed and developed on the RS/6000 platform running the AIX operating system. A range of tools built to help the development process are available either on Web sites or you can obtain them from your IBM representative. The IBM StorWatch Serial Storage Expert (StorX) product is

discussed in detail in Chapter 4, "Introducing IBM StorWatch Serial Storage Expert" on page 51

Table 9. SSA Adapter Features and Compatibility

Feature Code	6214	6215	6216	6217	6218	6219
Adapter Description	Classic	Enhanced	Enhanced	RAID-5	RAID-5	Enhanced
Bus	MCA	PCI	MCA	MCA	PCI	MCA
Architecture	SSA	SSA-EL	SSA	SSA	SSA	SSA-EL
Loops:Disks Supp	2:96	2:96	2:96	2:96	2:96	2:96
Max JBOD Multiattach	2	8	8	1	1	8
Max RAID5 Multiattach	n/a	2	n/a	1	1	2
Max Fast Write-Attach	n/a	1	n/a	n/a	n/a	1
Fast Write Cache (Opt)	n/a	4MB	n/a	n/a	n/a	4MB
Fast Write Feature Code	n/a	6222	n/a	n/a	n/a	6222
Read Cache	n/a	32MB	n/a	8MB	8MB	32MB
Features that can be mixed in the same loop	6216 6214	6219 6215	6214 6216	n/a	n/a	6215 6219
HA/CMP Qualified	Yes	Yes	Yes	No	No	Yes
Target Mode Support	No	Yes	No	No	No	Yes
IOP/second (Non-RAID)	3000	3000	3000	3000	3000	3000
IOP/second (RAID-5 **)	n/a	3000	n/a	3000	3000	3000
IOP/second (RAID-5 ***)	n/a	1000	n/a	1000	1000	1000
MB/s (Non-RAID)	35	35	35	35	35	35
MB/s (RAID-5)	n/a	29reads 13 writes	n/a	29 reads 7writes	29 reads 7 writes	29 reads 13 writes

Notes: JBOD = Just a Bunch of Disks
IOP = I/O Operations
** = Cache hits 100% reads
*** = Non cache hits (70% reads 30% writes)
Attach = denotes the max number of initiators/loop

3.1.1 Concepts

3.1.1.1 Device Drivers

In order to effectively manage SSA disks it is helpful to have an understanding of how the device driver works and how it helps the management process.

SSA disk drives are represented in AIX as SSA logical disks (hdisk0, hdisk1,...,hdiskN) and SSA physical disks (pdisk0, pdisk1,...,pdiskN). SSA RAID arrays are represented as SSA logical disks (hdisk0, hdisk1,...,hdiskN). SSA logical disks represent the logical properties of the disk drive or array, and can have volume groups and file systems mounted on them. SSA physical disks represent the physical properties of the disk drive. By default one pdisk is always configured for each physical disk drive. One hdisk is configured for each disk drive that is connected to the using system, or for each array. By default, all disk drives are configured as system (AIX) disk drives. The array management software can be used to change the disks from hdisks to array candidate disks or hot spares.

SSA logical disks:

- Are configured as hdisk0, hdisk1,...,hdiskN.
- Support a character special file (/dev/rhdisk0, /dev/rhdisk1,...,/dev/rhdiskN).
- Support a block special file (/dev/hdisk0, /dev/hdisk1,...,/dev/hdiskN).
- Support the I/O Control (IOCTL) subroutine call for non service and diagnostic functions only.
- Accept the read and write subroutine calls to the special files.
- Can be members of volume groups and have file systems mounted on them.

SSA physical disks:

- Are configured as pdisk0, pdisk1,...,pdiskN.
- Have errors logged against them in the system error log.
- Support a character special file (/dev/pdisk0, /dev/pdisk1,...,/dev/p.diskN).
- Support the IOCTLI subroutine for servicing and diagnostic functions
- Do not accept read or write subroutine calls for the character special file.

3.1.1.2 Twin Adapters in a Loop, Single Host

Some SSA subsystems allow a disk drive to be controlled by up to two adapters in a particular host system. The disk drive has therefore two paths to the host system, and the SSA subsystem can continue to function if one adapter fails. If an adapter fails or the disk drive becomes inaccessible from the original adapter, the SSA disk device driver switches to the alternative adapter without returning an error to any working application. Once a disk drive has been successfully opened, takeover by the alternative adapter does not occur simply because a drive becomes reserved or fenced out. However, during an opening of an SSA logical disk, the device driver attempts to access the disk drive through the alternative adapter if the path through the original adapter experiences reservation conflict or fenced-out status.

Takeover does not occur because of a medium error on the disk drive. Takeover occurs only after extensive error-recovery activity within the adapter and several retries by the device driver. Intermittent errors that last for approximately 1

second usually do not cause adapter takeover. Once takeover has successfully occurred and the device driver has accessed the disk drive through the alternative adapter, the original adapter becomes the standby adapter. Takeover can, therefore, occur repeatedly from one adapter to another so long as one takeover event is completed before the next one starts. A takeover event is considered to have completed when the device driver successfully accesses the disk drive through the alternative adapter. Once takeover has occurred, the device driver continues to use the alternative adapter to access the disk drive until either the system is rebooted, or takeover occurs back to the original adapter.

Each time the SSA disks are configured, the SSA disk device driver is told which path or paths are available to each disk drive, and which adapter is to be used as the primary path. By default, primary paths to disk drives are shared equally among the adapters to balance the load. This static load balancing is performed once, when the devices are configured for the first time. You can use the **chdev** command to modify the primary path.

Because of the dynamic nature of the relationship between SSA adapters and disk drives, SSA pdisks and hdisks are not children of an adapter but of an SSA router. This router is called *ssar*. It does not represent any actual hardware, but exists only to be the parent device for the SSA logical disks and SSA physical disks.

Note: When the SSA disk device driver switches from using one adapter to using the other adapter to communicate with a disk, it issues a command that breaks any SSA-SCSI reserve condition that might exist on that disk. The reservation break is performed only if this host had successfully reserved the disk drive through the original adapter. This check is to prevent adapter takeover from breaking reservations held by other using systems.

3.1.1.3 Disk Fencing

If multiple host systems are connected to the SSA disks, SSA-SCSI reserve should not be used as the only method for controlling access to the SSA disks. SSA disks support a fence operation that is used by HACMP to control access to a particular disk within a group of host machines. A disk may either be *fenced in*, in which case only nominated hosts can access that disk, or the disk may be *fenced out*, in which case nominated hosts are specifically excluded from accessing it.

3.1.1.4 Enhanced Loop Adapters

PCI SSA Multi-Initiator/RAID EL adapters and Micro Channel Enhanced SSA Multi-Initiator/RAID EL adapters are capable of reserving to a node number rather than reserving to an adapter. We highly recommend that you make use of this capability by setting the SSA router `node_number` attribute if multiple adapters are to be configured.

3.1.1.5 Configuring Drives

Usually, all the disk drives connected to the system are configured automatically by the system boot process and the user need not take any action to configure them. Because SSA devices can be connected to the SSA network while the system is running without taking the system offline, it may be necessary to

configure SSA disks after the boot process has completed. In this case, configure the devices by running the configuration manager with the **cfgmgr** command.

3.1.2 Types of Control

There are four main methods of controlling SSA disks on AIX:

- AIX SMIT interface
- AIX diagnostics
- Special tools
- Other Tools

3.1.2.1 AIX SMIT Interface

The fast path SMIT devices can be used to access the SMIT menu system. Use the last three options on the SMIT devices menu, Figure 8 to control:

- SSA Adapters
- SSA Disks
- SSA RAID Arrays.



Figure 8. SMIT Devices Menu

Selecting **SSA Adapters** takes you to the SSA Adapters menu, Figure 9 on page 24, which provides an additional six options.

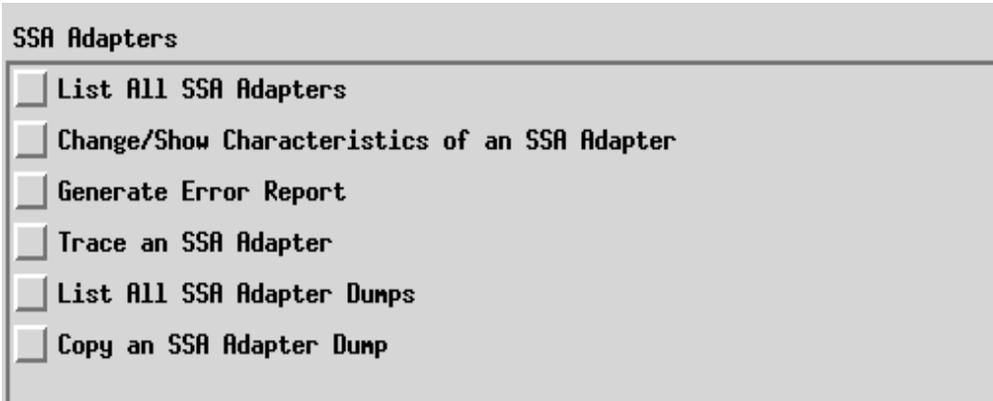


Figure 9. SSA Adapter Menu

Selecting **SSA Disks** on the SMIT Devices menu takes you to the SSA Disks menu, shown in Figure 10, which provides two options.



Figure 10. SSA Disks Menu

Selecting **SSA Logical Disks** on the SSA Disks menu takes you to the SSA Logical Disks menu. As can be seen on Figure 11 on page 25, this gives you a further 14 options.

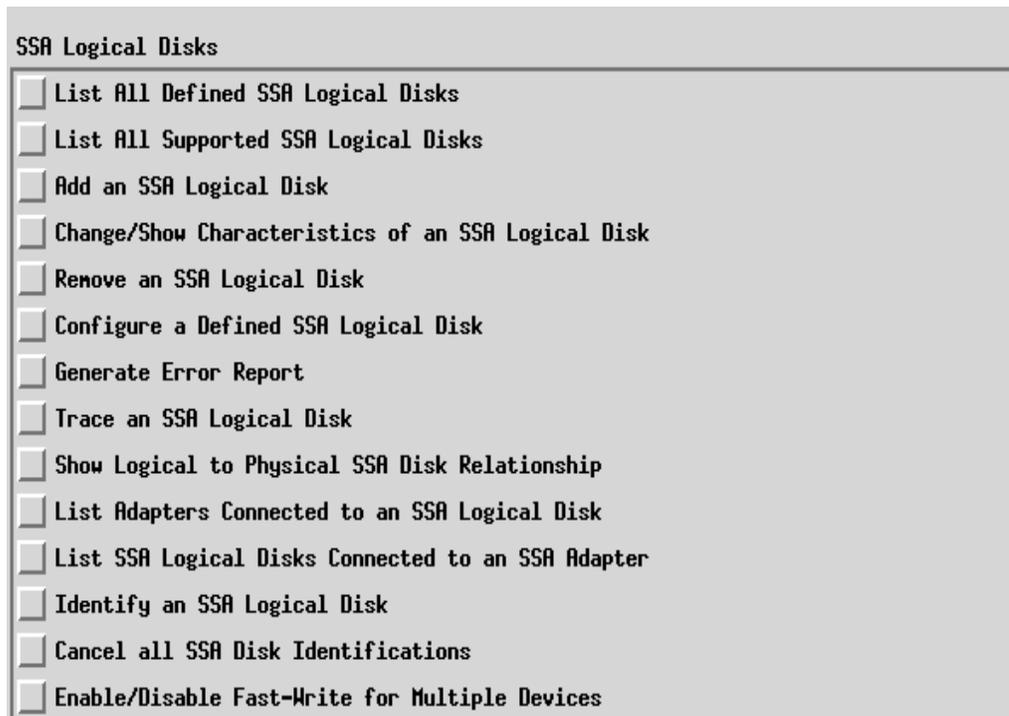


Figure 11. Logical Disk Menu

Selecting **SSA Physical Disks** on the SSA Disks menu takes you to the SSA Physical Disks menu, Figure 12 on page 25, which gives you a further 14 options.

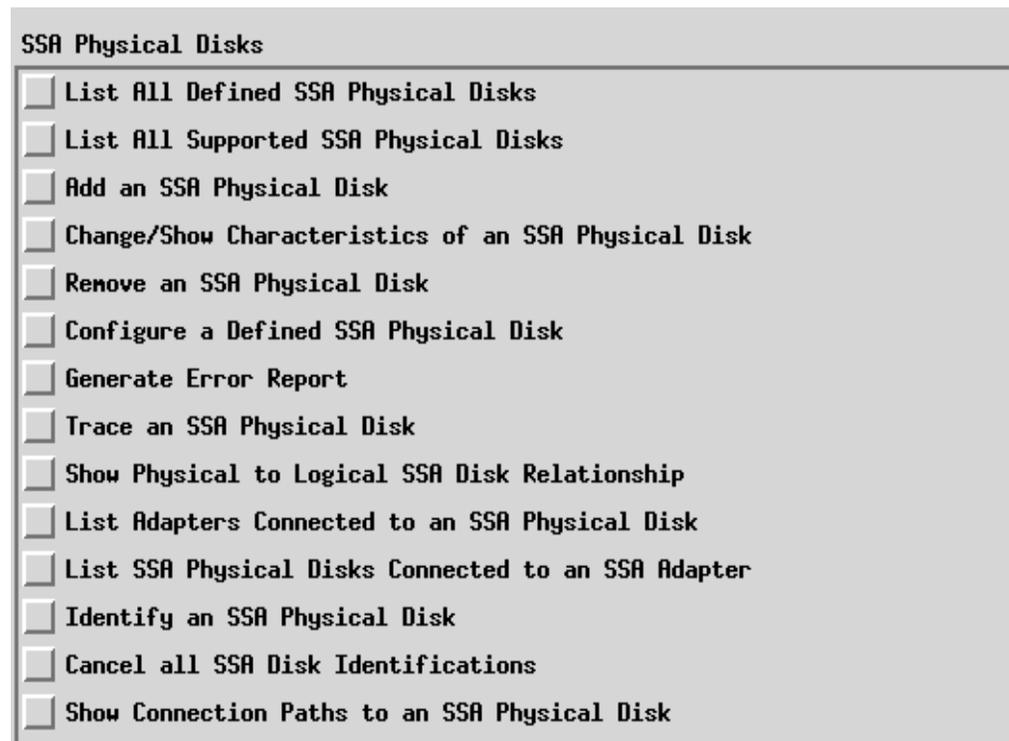


Figure 12. SSA Physical Disk Menu

Selecting **SSA RAID Arrays** on the SMIT Devices menu takes you to the SSA RAID Arrays menu, Figure 13 on page 26 which gives you a further 12 options.

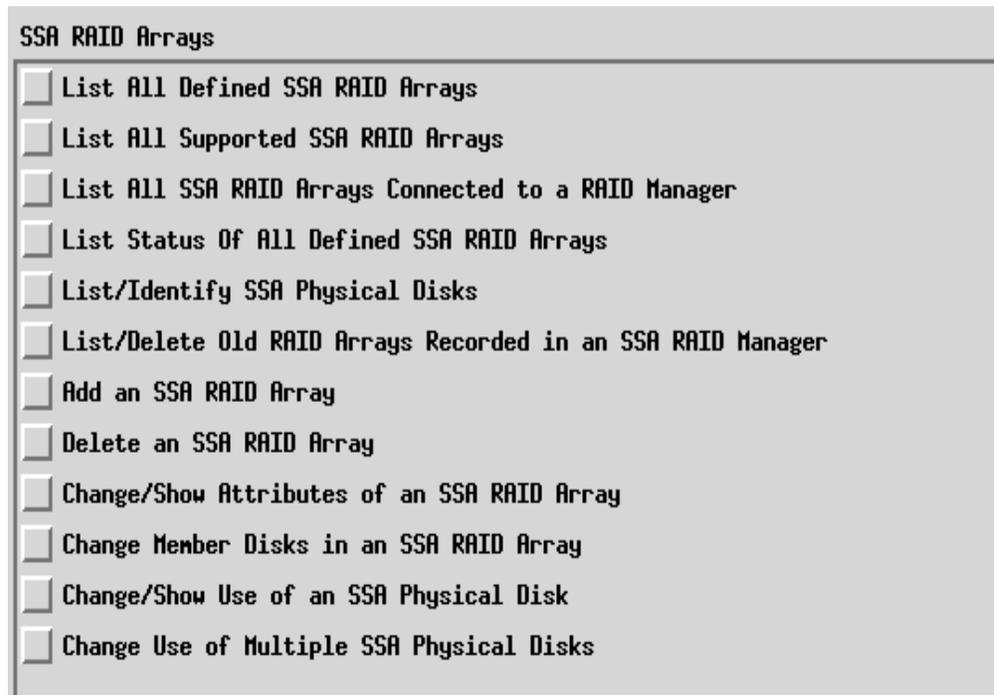


Figure 13. SSA RAID Arrays Menu

3.1.2.2 Diagnostics

The SSA AIX diagnostic routines are fully documented in *A Practical Guide to SSA for AIX, SG24-4599*. A brief overview of them is included below for the sake of completeness.

The SSA service aids are accessed from the standard AIX diagnostics menu. This menu is invoked with the **diag** command and list these aids:

- Set Service Mode
This option enables you to determine the location of a specific SSA disk drive within a loop and to remove the drive from the configuration, if required.
- Link Verification
This option enables you to determine the operational status of a link
- Configuration Verification
This option enables you to display the relationships between physical (pdisks) and logical (hdisks) disks.
- Format Disk
This option enables you to format SSA disk drives.
- Certify Disk
This option enables you to test whether data on an SSA disk drive can be read correctly.
- Identify

This option is accessible from any of the previous options and enables you to determine the physical location of SSA disk drives within a subsystem.

NOTE: When an SSA loop is attached to multiple host systems, do not invoke the diagnostic routines from more than one host simultaneously, to avoid unpredictable results that may result in data corruption.

3.1.2.3 Special Tools

A variety of tools have been developed as "one off" tools to resolve a specific situation. These have become more generally available and the main ones are described here.

Maymap

Maymap was developed as an aid to the SSA adapter development process. The functionality of maymap is largely incorporated into StorX. Maymap is a command-line driven program that will map out the disks in an SSA loop and report on details such as pdisk and hdisk numbers.

A good starting point is:

\$maymap -p -a

Maymap is compatible with all revisions of AIX Versions 3.2.5 and later. Here is the structure of the command set.

usage: /maymap [[-d device]][-a] -n [node] [-4] [-b B] [-t] [-i x] [-p] [-h] [-y] {[-l] [-m|u|w|s|v|o|f]} [-z]

where:

- d [device] is the adapter device driver name; the default is ssa0
- n [node] IPN node number override for the above, for support of old levels of host software
- a Display for all adapters; do not use with -d
- b [B] Display up to B boxes per row; the default is 4 (Range 2 through 16)
- 3 Use nice test characters for boxes. Forced if AIX 3.2.5. Can be used to override -4 option
- 4 Use plain test characters for boxes. (This is an interim fix for AIX 4.1 and is forced if 4.1).
- t Display DASD map without titles
- l Switches Maymap into netlist mode
- i Time out period in seconds for ioctl() calls. The default is 20 (0=Off)
- p Display with pdisk numbers, also work with -l (list) option.
- h Display with hdisk numbers, also work with -l (list) option.
- y Highlights whether the devices yellow light is on or off.

Note: In a multiway environment, an adapter cannot see if a disk is flashing, if the disk was told to flash by any other adapter, other than the adapter you are viewing.

The following apply only when -l is specified:

- u List with SSA uniqueID (ISALMgr Serial number)

- m List with block count and block size (max.LBA+1)
- f List with drive type
- w List with shrew level (Starfires only)
- s List only VPD serial numbers (to pipe into scripts)
- v List with entire VPD
- o List with OMT type
- z Enable debug and transaction error messages (default: off)

The following may be used together:

- lmuphfy Block counts, UID with pdisk/hdisk numbers/drive type
- luophfy UID, OMT with pdisk/hdisk numbers, drive type and flash
- lww Shrew level with entire VPD

You can remove options from the above, for example -luo or -lmu

-s may be not used in conjunction with any other option.

Maymap requires root privileges.

When we executed the command **maymap -d ssa2 -p** we received the output shown in Figure 14 on page 28.

```

*****
MAYMAP Version v1.98n 6-Feb-98 18:00 : Colonial SSA Network Map Utility
Written by Lee Sanders - Copyright (C)IBM 1994-1998 - IBM Internal Use Only
*****
SSA Adapter (2-way) Card: ssa2 [Node 0x82] [Bus: 0] [Slot: 5] found.
Type: adapter/mca/ssa Firmware: 2401
Time Out Period set to 120
*****
Device ssa2 [Node 0x82] - 16 DriverPhysical resources reported
Allicat/SSA = 0664 SHH ... Starfire/Scorfire SSA = DFHC (Model Varies)
              Scorpion/SSA = DFHC (Model Varies)
*****
-----
|DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B|
|Node 82|---|AC9D4178|---|AC50DED3|---|AC50DF16|---|AC9D1C9A|---|AC903ADC|---|AC9D4182|---
|Port A1| |lpdisk26| |lpdisk16| |lpdisk18| |lpdisk22| |lpdisk23| |lpdisk25| |
-----
|DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B|
|AC50DF41|---|AC50DED5|---|AC9D42C8|---|5AE83CF4|---|AC9D418B|---|AC50DF44|---
|lpdisk19| |lpdisk17| |lpdisk29| |lpdisk31| |lpdisk27| |lpdisk20| |
-----
|DFHC C4B| |DFHC C4B| |DFHC C4B| |DFHC C4B|
|AC50DF49|---|AC9D471B|---|AC9D41B7|---|AC9D40FA|---
|lpdisk21| |lpdisk30| |lpdisk28| |lpdisk24| |
-----
|Integr |
|Node 82|
|Port #2|
-----

```

Figure 14. Output of Command **maymap -d ssa2 -p**

The same network appears to StorX as shown in Figure 15 on page 29.

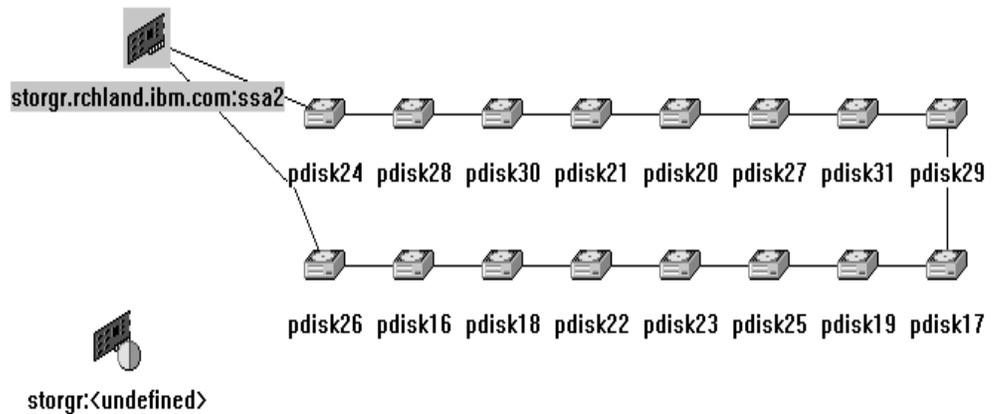


Figure 15. StorX View of Network Shown in Figure 14 on page 28

Maymap not only displays the pdisk number but also the disk serial number, type of disk, and connection ports to the adapter. Currently, StorX cannot display multiple types of data simultaneously. Maymap, however, does not have any code that runs continuously to enable it to monitor changes in the SSSA network as StorX has. Later releases of StorX will enable it to display multiple types of data about each disk simultaneously.

SSATools

The SSATools package is available on the internal tools disk, AIXTOOLS, and contains six command line programs and three shell scripts. The commands enable you to access some of the functions available in the SSA service aids from the command line. The commands are simple and they are intended to be used primarily from within shell scripts. They do not have extensive error checking and error messages. The SSA service aids should be used if you require extensive error checking and error messages.

In general, the command prints a usage string if the syntax is incorrect but will not print a message if the command fails. If the command executes without error, the return code is 0. If there is an error, the return code is a value other than 0.

The shell scripts contain no error checking. They enable you to assign SSA pdisk and hdisk names to meet your requirements and perform error log analysis on all SSA devices from a single command-line entry. These scripts can be used alone or in combination with other shell scripts.

ssaxlate command

Purpose: Translates between SSA logical disks (hdisks) and SSA physical disks (pdisks)

Syntax:

```
ssaxlate -l LogicalDiskName
ssaxlate -l PhysicalDiskName
```

Description: If the parameter is an SSA logical disk, the output will be a list of SSA physical disk names supporting that logical disk. If the parameter is an SSA

physical disk, the output will be a list of SSA logical disk names that depend on that physical disk.

Flags:

-l DiskName - Specifies the SSA logical or physical disk

ssaadap command

Purpose: Lists the adapters to which an SSA logical disk or physical disk is connected.

Syntax:

```
ssaadap -l LogicalDiskName  
ssaadap -l PhysicalDiskName
```

Description: The output is the list of SSA adapters to which the SSA logical or physical disk is connected. If there is more than one adapter, the first adapter listed is the primary adapter.

Flags:

-l DiskName - Specifies the SSA logical or physical disk

ssaidentify command

Purpose: Sets or clears identify mode for an SSA physical disk

Syntax:

```
ssaidentify -l PhysicalDiskName -y  
ssaidentify -l PhysicalDiskName -n
```

Description: If the -y flag is specified the disk is set to identify mode and flashes an amber light at approximately 1 second intervals. The -n flag turns off identify mode.

Flags:

-l PhysicalDiskName Specifies the device to place in identify mode.

-y Turns on identify mode

-n Turns off identify mode

ssaconn command

Purpose: Displays the SSA connection details for the physical disk

Syntax:

```
ssaconn -l PhysicalDiskName -a AdapterName
```

Description: The **ssaconn** command performs a function similar to the SSA Link Verification service aid.

The output is the PhysicalDiskName followed by the adapter name followed by four hop counts. The four hop counts represent the number of SSA devices between the adapter and the A1 A2 B1 B2 ports of the adapter, respectively. If the disk is not connected to a particular adapter port, the hop count is replaced by a hyphen character.

Flags:

- l PhysicalDiskName. Specifies the SSA physical disk whose connection details should be listed
- a AdapterName. Specifies the adapter to list the connection details relative to

ssacand command

Purpose: Displays the unused connection locations for an SSA adapter

Syntax:

```
ssacand -a AdapterName -P -L
```

Description: The **ssacand** command is used to list out the available connection locations on an SSA adapter. These locations correspond to devices that are connected to the adapter, but for which AIX devices are not configured.

Flags:

- a AdapterName. Specifies the adapter for which to list the connection locations
- P Produces a list of possible connection locations for SSA physical disks
- L Produces a list of possible connection locations for SSA logical disks

ssadisk command

Purpose: Displays the names of SSA disks connected to an SSA adapter

Syntax:

```
ssadisk -a AdapterName -P -L
```

Description: The **ssadisk** command is used to list out the names of SSA disks connected to an SSA adapter. These names correspond to devices that are in the customized device database, and have the SSA adapter as their adapter_a or adapter_b attribute.

Flags:

- a AdapterName. Specifies the adapter to which the disks are connected
- P Produces a list of SSA physical disks
- L Produces a list of SSA logical disks

change_ssahdisk_name shell script

Purpose: Sometimes during the installation or maintenance of an SSA subsystem, the assigned hdisk names do not meet customer requirements. This shell script makes it possible to change the assigned hdisk name to any other unused hdisk name.

Syntax:

```
change_ssahdisk_name <old hdisk name> <new hdisk name>
```

Description: This script can only be used to change the names of SSA hdisks. The old hdisk name is first removed (rmdev) and the new hdisk name is created (mkdev) using the same physical volume.

Example: change_ssahdisk_name hdisk1 hdisk12, removes hdisk1 and creates hdisk12

change_pdisk_name shell script

Purpose: Sometimes during the installation or maintenance of an SSA subsystem the assigned pdisk names may not meet customer requirements. This shell script makes it possible to change an assigned pdisk name to any other unused pdisk name.

Syntax:

```
change_pdisk_name <old pdisk name> <new pdisk name>
```

Description: This script can only be used to change the names of SSA pdisks. The old pdisk name is first removed (rmdev) and the new pdisk name is created (mkdev) using the same physical disk drive.

Example: change_pdisk_name pdisk0 hdisk5 removes pdisk0 and creates pdisk5

ssa_log_analysis shell script

Purpose: This script can be used to perform error log analysis on all configured SSA adapters and disk drives.

Syntax:

```
ssa_log_analysis
```

Description: This shell script may be executed from any RS/6000 system but is primarily designed to support SP2 installations where it is necessary to obtain a summary of SSA errors that require service action through a single command line entry.

On SP2 systems, the ssa_log_analysis shell script should be installed on all nodes that contain SSA devices. The directory selected for the installation should be accessible from a Distributed Shell (dsh) command. Installation in /usr/sbin is recommended.

You can now execute error log analysis on all nodes through the following command line entry:-

```
dsh "ssa_log_analysis"
```

Most of these commands are an integral part of the latest releases of AIX and of the SMIT panels. The script files are useful and the descriptions of how to use the commands in the tools README file produce a good understanding of exactly what is happening. If there are situations where the operating system is still at AIX V3 level, it is essential to have this SSAtools package.

SSA Cable

This is not a tool for controlling SSA networks. Instead, the program is designed to help you cable an SSA subsystem composed of multiple loops and multiple hosts. All you need to specify is which hosts go into which loops, and how many disks you intend to assign to each host in each loop. You may also choose between the default 7133-010 and 7133-020.

The program will optimize the order of nodes in the loop. It does this by optimizing to the nearest group of disks, for example to the nearest singleton or pair or quad. It will always try to deliver the minimum number of unassigned disks or gaps in the disk drawers.

The grouping factor may be specified differently for each loop, and is the first numeric field in the loop:' line. Additionally, if you have a line beginning *clus:* it will start a fresh disk tray for every host, and fill these up as best it can, trying to keep drawers associated with hosts.

A node will have the disks that it owns in a loop attached equally to both ports of the connector, as far as this is consistent with the grouping factor. The grouping factor can be any integer greater than zero, and controls how disks assigned to a node in a loop are distributed on each of the adapter ports - for example B1 and B2. A different grouping factor may be specified for each loop.

For example, a node with six disks in a loop with a grouping factor of 1 will have three disks on port A1 and three on port A2. A grouping factor of 2 will give four disks and two disks, as will a factor of 4. A node with eight disks in a loop and a grouping factor of 1 or 2 or 4 will split its disks four and four, but a grouping factor of 3 would cause it to split them five and three. The disks are always packaged in quad slots, and the program will display any vacant slots.

Example file *ssainxmp4c20*

```
clus:

type: 020

loop: 4 4 8 12 -2

loop: 4 12 8 4 -2

name: Matthew

name: Andrew

name: Nicholas
```

This input file gives the following output:

```
*****
*
*  SSACABLE V2.1  - example file ssainxmp4c20  *
*
*                J S Watts IBM UK Petroleum    *
*
*****

Node Matthew   needs 1 adaptors and 16 disks
Node Andrew    needs 1 adaptors and 16 disks
Node Nicholas  needs 1 adaptors and 16 disks
```

This configuration needs:
3 enhanced SSA adaptors, 48 disks in 3 trays and 16 cables

The cabling reflects the use of 7133-020
The disks are grouped within the rack, where possible, by hostname

```
**** Program limits (changeable on request) ****

Number of nodes: 48
Number of loops: 112
```

Number of trays: 200
Number of disks/loop: 96

1

```
*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* Placement of disks within trays, by host ID
*
*****

      1 ====Quad=== 4 5 ====Quad=== 8 9 ====Quad===12 13 ====Quad===16

Tray 2:  2 2 2 2      2 2 2 2      2 2 2 2      2 2 2 2
Tray 1:  1 1 1 1      1 1 1 1      1 1 1 1      1 1 1 1
Tray 0:  0 0 0 0      0 0 0 0      0 0 0 0      0 0 0 0

      1 ====Quad=== 4 5 ====Quad=== 8 9 ====Quad===12 13 ====Quad===16
```

Legend: Numbers represent hosts, as follows:

0 is Matthew
1 is Andrew
2 is Nicholas

1

```
*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* Placement of disks within trays, by loop ownership*
*
*****

      1 ====Quad=== 4 5 ====Quad=== 8 9 ====Quad===12 13 ====Quad===16

Tray 2:  0 0 0 0      0 0 0 0      0 0 0 0      1 1 1 1
Tray 1:  0 0 0 0      0 0 0 0      1 1 1 1      1 1 1 1
Tray 0:  0 0 0 0      1 1 1 1      1 1 1 1      1 1 1 1

      1 ====Quad=== 4 5 ====Quad=== 8 9 ====Quad===12 13 ====Quad===16
```

Legend: Numbers represent loops

```
1*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* HOST/DISK LAYOUT FOR LOOP 0
*
* (optimized to nearest 4 disks)
*
```

```

*
*****

```

```

*****
*           port A1 *>-----+
Matthew  * adaptor 0      *
(4 disks) *           port A2 *>-----+
*****
*           |
*           *****
*           * 0 *
*           8 disks in 8 slots: *====*
*           * 8 *
*           *****
*****
*           |
*           port A1 *>-----+
Nicholas * adaptor 0      *
(12 disks) *           port A2 *>-----+
*****
*           |
*           *****
*           * 4 *
*           8 disks in 8 slots: *====*
*           * 4 *
*           *****
*****
*           |
*           port A1 *>-----+
Andrew  * adaptor 0      *
(8 disks) *           port A2 *>-----+
*****
*           |
*           *****
*           * 4 *
*           8 disks in 8 slots: *====*
*           * 4 *
*           *****
*           |
*           +-----+

```

```

1*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
*           HOST/DISK LAYOUT FOR LOOP 1
*
*           (optimized to nearest 4 disks)
*
*****

```

```

*****
*           port B1 *>-----+
Matthew  * adaptor 0      *
(12 disks) *           port B2 *>-----+
*****
*           |
*           *****
*           * 4 *
*           8 disks in 8 slots: *====*
*           * 4 *

```

```

*****
*****
*          port B1 *>-----+
Andrew      * adaptor 0      *
(8 disks)  *          port B2 *>-----+
*****
*****
* 4 *
8 disks in 8 slots: *====*
* 4 *
*****
*****
*          port B1 *>-----+
Nicholas    * adaptor 0      *
(4 disks)  *          port B2 *>-----+
*****
*****
* 0 *
8 disks in 8 slots: *====*
* 8 *
*****
|
+-----+

```

```

1*****
*
* SSACABLE V2.1 - example file ssainxmp4c20 *
*
*          CABLING SCHEDULE FOR LOOP 0      *
*
* (optimized to nearest 4 disks)           *
*
*****

```

```

Cable 1: Matthew adapter 0 port A2 to tray 2 port 1
                                     tray 2 port 4 bypass to port 5
Cable 2: Nicholas adapter 0 port A1 to tray 2 port 8
Cable 3: Nicholas adapter 0 port A2 to tray 2 port 9
Cable 4:.....tray 2 port 12 to tray 1 port 1
Cable 5: Andrew adapter 0 port A1 to tray 1 port 4
Cable 6: Andrew adapter 0 port A2 to tray 1 port 5
Cable 7:.....tray 1 port 8 to tray 0 port 1
Cable 8: Matthew adapter 0 port A1 to tray 0 port 4

```

```

1*****
*
* SSACABLE V2.1 - example file ssainxmp4c20 *
*
*          CABLING SCHEDULE FOR LOOP 1      *
*
* (optimized to nearest 4 disks)           *
*
*****

```

```

Cable 9: Matthew adapter 0 port B2 to tray 0 port 5
Cable 10:.....tray 0 port 8 to tray 1 port 9

```

```

Cable 11: Andrew      adapter 0 port B1 to tray 1 port 12
Cable 12: Andrew      adapter 0 port B2 to tray 1 port 13
Cable 13:.....tray 1 port 16 to tray 2 port 13
Cable 14: Nicholas    adapter 0 port B1 to tray 2 port 16
Cable 15: Nicholas    adapter 0 port B2 to tray 0 port 9
                                     tray 0 port 12 bypass to port 13
Cable 16: Matthew     adapter 0 port B1 to tray 0 port 16

```

1

```

*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* Cables which interconnect disk ports
*
*****

```

***** Loop 0 *****

```

Cable 4:.....tray 2 port 12 to tray 1 port 1
Cable 7:.....tray 1 port 8 to tray 0 port 1

```

***** Loop 1 *****

```

Cable 10:.....tray 0 port 8 to tray 1 port 9
Cable 13:.....tray 1 port 16 to tray 2 port 13

```

1

```

*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* Cables which connect processors to disks
*
*****

```

***** Loop 0 *****

```

Cable 1: Matthew     adapter 0 port A2 to tray 2 port 1
Cable 2: Nicholas    adapter 0 port A1 to tray 2 port 8
Cable 3: Nicholas    adapter 0 port A2 to tray 2 port 9
Cable 5: Andrew      adapter 0 port A1 to tray 1 port 4
Cable 6: Andrew      adapter 0 port A2 to tray 1 port 5
Cable 8: Matthew     adapter 0 port A1 to tray 0 port 4

```

***** Loop 1 *****

```

Cable 9: Matthew     adapter 0 port B2 to tray 0 port 5
Cable 11: Andrew     adapter 0 port B1 to tray 1 port 12
Cable 12: Andrew     adapter 0 port B2 to tray 1 port 13
Cable 14: Nicholas    adapter 0 port B1 to tray 2 port 16
Cable 15: Nicholas    adapter 0 port B2 to tray 0 port 9
Cable 16: Matthew     adapter 0 port B1 to tray 0 port 16

```

1

```

*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* Cables which join hosts to hosts
*
*****

***** Loop 0 *****

***** Loop 1 *****

```

1

```

*****
*
* SSACABLE V2.1 - example file ssainxmp4c20
*
* Cables sorted by hostname
*
*****

Cable 5: Andrew adapter 0 port A1 to tray 1 port 4
Cable 6: Andrew adapter 0 port A2 to tray 1 port 5
Cable 11: Andrew adapter 0 port B1 to tray 1 port 12
Cable 12: Andrew adapter 0 port B2 to tray 1 port 13
Cable 8: Matthew adapter 0 port A1 to tray 0 port 4
Cable 1: Matthew adapter 0 port A2 to tray 2 port 1
Cable 16: Matthew adapter 0 port B1 to tray 0 port 16
Cable 9: Matthew adapter 0 port B2 to tray 0 port 5
Cable 2: Nicholas adapter 0 port A1 to tray 2 port 8
Cable 3: Nicholas adapter 0 port A2 to tray 2 port 9
Cable 14: Nicholas adapter 0 port B1 to tray 2 port 16
Cable 15: Nicholas adapter 0 port B2 to tray 0 port 9

```

3.1.3 Other Tools

3.1.3.1 AIX Commands

AIX incorporates some standard UNIX commands that are very useful for measuring disk performance. These commands and their usage are fully described in *Aix Performance Monitoring & Tuning Guide SC23-2365*. For completeness, a brief description of the main commands is given here.

iostat

The ***iostat*** command is used for monitoring system input/output device loading by observing the time the physical disks are active in relation to their average transfer rates. The ***iostat*** command generates reports that can be used to change system configuration to better balance the input/output load between physical disks.

The first report generated by the **iotstat** command is the tty and CPU Utilization Report. For multiprocessor systems, the CPU values are global averages among all processors. Also, the I/O wait state is defined systemwide and not for each processor. The report has the following format:

tin: Shows the total number of characters read by the system for all ttys.

tout: Shows the total number of characters written by the system to all ttys.

% user: Shows the percentage of CPU utilization that occurred while executing at the user level (application).

% sys: Shows the percentage of CPU utilization that occurred while executing at the system level (kernel).

% idle: Shows the percentage of time that the CPU or CPUs were idle and the system did not have an outstanding disk I/O request.

% iowait: Shows the percentage of time that the CPU or CPUs were idle during which the system had an outstanding disk I/O request. This value may be slightly inflated if several processors are idling at the same time, an unusual occurrence.

This information is updated at regular intervals by the kernel (typically 60 times a second). The tty report provides a collective account of characters per second received from all terminals on the system as well as the collective count of characters output per second to all terminals on the system.

The second report generated by the **iotstat** command is the Disk Utilization Report. The disk report provides statistics for each physical disk. The report has a format similar to the following:

% tm_act: Indicates the percentage of time the physical disk was active (bandwidth utilization for the drive).

Kbps: Indicates the amount of data transferred (read or written) to the drive in KB per second.

tps: Indicates the number of transfers per second that were issued to the physical disk. A transfer is an I/O request to the physical disk. Multiple logical requests can be combined into a single I/O request to the disk. A transfer is of indeterminate size.

Kb_read: The total number of KB read.

Kb_wrtn: The total number of KB written.

vmstat

The **vmstat** command reports statistics about kernel threads, virtual memory, disks, traps and CPU activity. Reports generated by the **vmstat** command can be used to balance system load activity. These systemwide statistics (among all processors) are calculated as averages for values expressed as percentages, and as sums otherwise.

If the **vmstat** command is invoked without flags, the report contains a summary of the virtual memory activity since system startup. If the -f flag is specified, the **vmstat** command reports the number of forks since system startup. The PhysicalVolume parameter specifies the name of the physical volume. The Interval parameter specifies the amount of time in seconds between each report. The first report contains statistics for the time since system startup. Subsequent reports contain statistics collected during the interval since the previous report. If the Interval parameter is not specified, the **vmstat** command generates a single report and then exits. The Count parameter can only be specified with the Interval parameter. If the Count parameter is specified, its value determines the number of reports generated and the number of seconds apart. If the Interval parameter is specified without the Count parameter, reports are continuously generated. A Count parameter of 0 is not allowed.

The kernel maintains statistics for kernel threads, paging, and interrupt activity, which the **vmstat** command accesses through the use of the knlist subroutine and the /dev/kmem pseudo-device driver. The disk input/output statistics are maintained by device drivers. For disks, the average transfer rate is determined by using the active time and number of transfers information. The percent active time is computed from the amount of time the drive is busy during the report.

The vmstat command generates a report that contains the following column headings, described here:

kthr: kernel thread state changes per second over the sampling interval.

r: Number of kernel threads placed in run queue.

b: Number of kernel threads placed in wait queue (awaiting resource, awaiting input/output).

Memory: information about the usage of virtual and real memory. Virtual pages are considered active if they are allocated. A page is 4096 bytes.

vm: Active virtual pages.

fre: Size of the free list.

Note: A large portion of real memory is used as a cache for file system data. It is not unusual for the size of the free list to remain small.

Page: information about page faults and paging activity. These are averaged over the interval and given in units per second.

re: Pager input/output list.

pi: Pages paged in from paging space.

po: Pages paged out to paging space.

fr: Pages freed (page replacement).

sr: Pages scanned by page-replacement algorithm.

cy: Clock cycles by page-replacement algorithm.

Faults: trap and interrupt rate averages per second over the sampling interval.

in: Device interrupts.

sy: System calls.

cs: Kernel thread context switches.

Cpu: breakdown of percentage usage of CPU time.

us: User time.

sy: System time.

id: CPU idle time.

wa: CPU cycles to determine that the current process is wait and there is pending disk input/output.

Disk: Provides the number of transfers per second to the specified physical volumes that occurred in the sample interval. The PhysicalVolume parameter can be used to specify one to four names. Transfer statistics are given for each specified drive in the order specified. This count represents requests to the physical device. It does not imply an amount of data that was read or written. Several logical requests can be combined into one physical request.

filemon

The **filemon** command monitors a trace of file system and I/O system events, and reports on the file and I/O access performance during that period. In its normal mode, the **filemon** command runs in the background while one or more application programs or system commands are being executed and monitored. The **filemon** command automatically starts and monitors a trace of the program's file system and I/O events in real time. By default, the trace is started immediately; optionally, tracing may be deferred until the user issues a **trcon** command. The user can issue **trcoff** and **trcon** commands while the **filemon** command is running in order to turn monitoring off and on, as desired. When tracing is stopped by a **trcstop** command, the **filemon** command generates an I/O activity report and exits.

In AIX Version 4, the **filemon** command is packaged as part of the Performance Toolbox for AIX. To determine whether **filemon** is available, use:

lspp -ll perfagent.tools

If this package has been installed, **filemon** is available.

3.1.3.2 Performance Toolbox

Performance Toolbox is a Motif-based application that contains performance management tools in a toolbox framework to help you monitor, analyze, and tune AIX, Hewlett Packard HP-UX, Sun Microsystems SunOS, or Sun Solaris 2.4 systems for optimal local (AIX) or client/server performance (network environment). Underlying Performance Toolbox are some of the standard UNIX commands such as **iostat**, **vmstat** and **filemon**. With performance management tools you can:

- Monitoring of local or remote system performance statistics

- Analyze performance statistics
- Tune system performance parameters to balance the utilization of fixed system resources

Performance Toolbox is a client/server type of application, the server or manager code can run only on an RS/6000 running AIX. The client or agent code can run on host systems running the AIX, Sun Solaris, or HP-UX operating systems. In addition to monitoring general disk statistics, such as those collected in **iostat** reports, a filter can be set so that alarms are triggered if the set thresholds are exceeded. The filter is set by using the AIX Filter and Alert daemon (filtld). The filtld program allows raw statistics, as supplied by xmservd, to be processed and fed back to the xmservd data pool as new statistics. These new statistics are then available to users and service personnel. The primary purpose of filtld is that of data reduction. With filtld you can also establish thresholds defined by arithmetic and Boolean expressions using statistics from the xmservd data pool, including new statistics defined through data reduction. For example, you can define a threshold as being passed when any one of the disks in a system is more than 90% busy, or when the average disk-busy percentage exceeds 70%. Each threshold defined must specify one or more actions to be initiated when the threshold is exceeded for a certain length of time and with no more than a certain frequency. You can specify any combination of the following actions:

- Invocation of system commands or shell scripts
- Sending of exception messages to an exception monitor (exmon)
- Sending of an SNMP trap (for AIX agents only).

Data reduction, alarms, and thresholds are specified in an ASCII filter configuration file. You do not have to have programming experience to modify the configuration file.

Although Performance Toolbox is not specifically designed for use with SSA Storage Networks, it is an invaluable tool for the storage administrator.

A typical view of the 3D output from Performance Toolbox is shown in Figure 16 on page 43. The left side of the grid lists the statistics you chose to include in the configuration set. The right side of the grid shows the parts of the system on which you are taking the measurements. In this example, it is a set of disks from different host systems. The third dimension is represented by the actual statistics values as received from the data supplier daemon. The values are plotted as rectangular pillars placed on the fields of the chessboard, each filling its field except for a user-modifiable spacing between the pillars. The actual value is displayed at the top of each pillar.

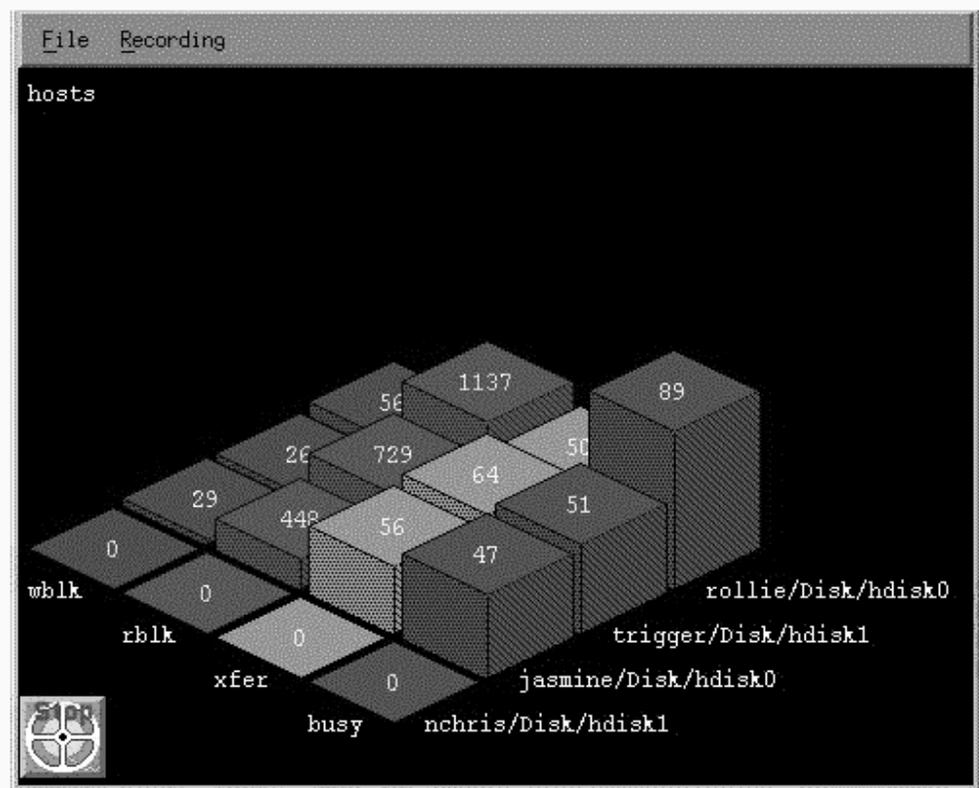


Figure 16. Typical PTX 3D Screen View

3.2 Managing SSA Disks in a Sun Solaris Environment

There are two methods of attaching SSA disks to a SUN host system. For systems that use the Sun SBus, there is a native adapter card, which supports one SSA loop of up to 48 disks. You can have two adapter cards per host, which means that up to 96 SSA disk drives can be attached to a single host. For Sun systems that do not use the SBus you can use the IBM 7190 SCSI to SSA converter. This external attachable device does not require internal installation or configuration of the host system.

The disk attachment method determines how the disks are managed. With the SSA SBus adapter, you can invoke the SSA Service Aid Utility through the command **ssau**. The utility is automatically installed as part of the adapter installation process. As the Sun operating system does not understand the concept of SSA disks, you must map the SSA IDs to SCSI IDs. Use the SSA Configuration Utility which you invoke through the **ssacf** command.

If you use the IBM 7190 to attach the SSA disks then install the SSA Service Functions from the CD-ROM that is supplied with the IBM 7190. The command **SSA_SERV** invokes the following functions:

1. Download microcode
2. Display VPD
3. Identify a disk drive module

4. Set/reset service mode
5. Disk drive diagnostics
6. 7190 model 100 diagnostics
7. Error log analysis
8. Clear Check mode
9. Activities monitor
10. Show SSA topology
11. Force SSA master
12. Force Web reset
13. Write cache option
14. Synchronous write
15. Quit

The above functions are fully described in the *Exploiting SSA Disk Subsystems in Sun Solaris Platform Environments (SG24-5083-00)*.

3.2.1 Other Tools

Other tools that are commonly used in the Sun Solaris environment are DiskSuite and Veritas. These tools are described in the *Exploiting SSA Disk Subsystems in Sun Solaris Platform Environments (SG24-5083-00)*. In addition, a Performance Toolbox agent runs in the Sun Solaris environment; see 3.1.3.2, "Performance Toolbox" on page 41.

3.3 Managing SSA Disks in an HP-UX Environment

The way of connecting SSA disks to a host running the HP-UX operating system is to use a SCSI to SSA converter such as the IBM 7190 or the Vicom SLIC.

If the SSA disks are attached using the IBM 7190 then the SSA Service Functions should be installed from the CD-ROM that is supplied with the IBM 7190. To invoke the following service functions, use the command SSA_SERV:

1. Download microcode.
2. Display VPD.
3. Identify a disk drive module.
4. Set/reset service mode.
5. Disk drive diagnostics.
6. 7190 Model 100 diagnostics
7. Error log analysis
8. Clear check mode
9. Activities monitor
10. Show SSA topology.
11. Force SSA master.
12. Force Web reset.

13. Write cache option.
14. Synchronous write
15. Quit.

The above functions are fully described in *Exploiting SSA Disk Subsystems in Sun Solaris Platform Environments (SG24-5083-00)*.

In addition a Performance Toolbox agent runs in the HP-UX environment see 3.1.3.2, "Performance Toolbox" on page 41.

3.4 Managing SSA Disks in a Windows NT Environment

The main tool used to manage SSA disks in this environment is SSA Remote Systems Management (SSA RSM). SSA RSM, formerly known as WebSSA, is a new Web-based SSA device configurator and service aids utility. It is designed as a replacement for the text-based SSA configurator that ships as standard with the SSA PC adapter. Two versions are available, a Netfinity plug-in module for IBM PC servers running TME10 Netfinity Version 5.xx, and a Windows NT standalone device. Both versions have the same look, feel, and operability.

With SSA RSM you can easily manage SSA-based storage systems locally and remotely. (Currently a beta version of SSA RSM is available on the Internet). SSA RSM, is a web browser-based tool, that provides for RAID Array configuration and service activities of the IBM SSA disk storage system. If the Netfinity Manager, is used it forms part of the WebGUI panel. The Netfinity Alert Manager already notifies you of critical alerts and thresholds warnings. Now all SSA alerts will be forwarded to the Netfinity Alert Manager as well, thus simplifying the management of error conditions.

SSA RSM enables you to:

- View the SSA adapter configuration information, including the physical location.
- Show all of the components attached to a particular SSA adapter.
- Create, attach and detach RAID arrays.
- Assign hot spares to RAID arrays .
- Identify individual disks by flashing the disk LEDs so that you can locate specific disks.
- Check and download the microcode levels of SSA adapters and disk drives.

Figure 17 on page 46 shows the initial screen view when SSA RSM is started.

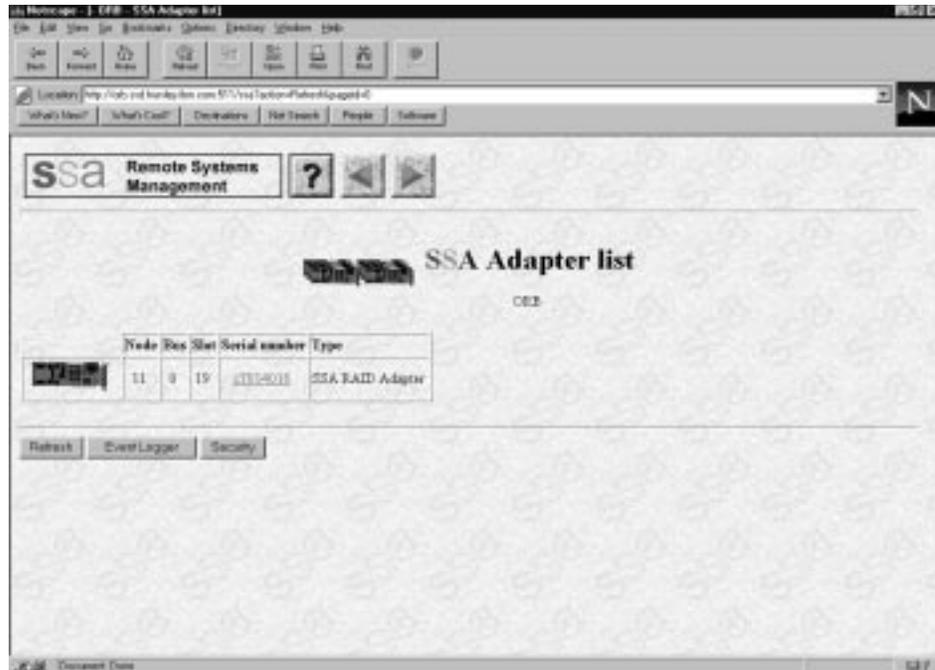


Figure 17. SSA RSM Initial Screen View

In this example, there is only one SSA adapter in the host PC. By clicking on the serial number of the adapter, you bring up a detailed view of the product information, as shown in Figure 18 on page 46.



Figure 18. Detailed View of Adapter Product Information.

In the SSA Adapter view, clicking on the **Physical view** of the disks brings up the screen shown in Figure 19 on page 47. You can see that the four ports on the

adapter are listed, with the disks shown in the order as they connect to the adapter.



Figure 19. Physical View from Adapter

Click on one of the disks to display the disk VPD, as shown in Figure 20 on page 47.



Figure 20. VPD of Individual Disk

When you select the **Logical view** in the SSA Adapter view window (Figure 18 on page 46) you get to the windows shown in Figure 21 and Figure 22.



Figure 21. Logical View from Adapter: Part 1



Figure 22. Logical View from Adapter: Part 2

If in SSA Adapter list window (Figure 17 on page 46), you select the **Event logger**, you get a detailed view of all recorded SSA events (see Figure 23 on page 49).

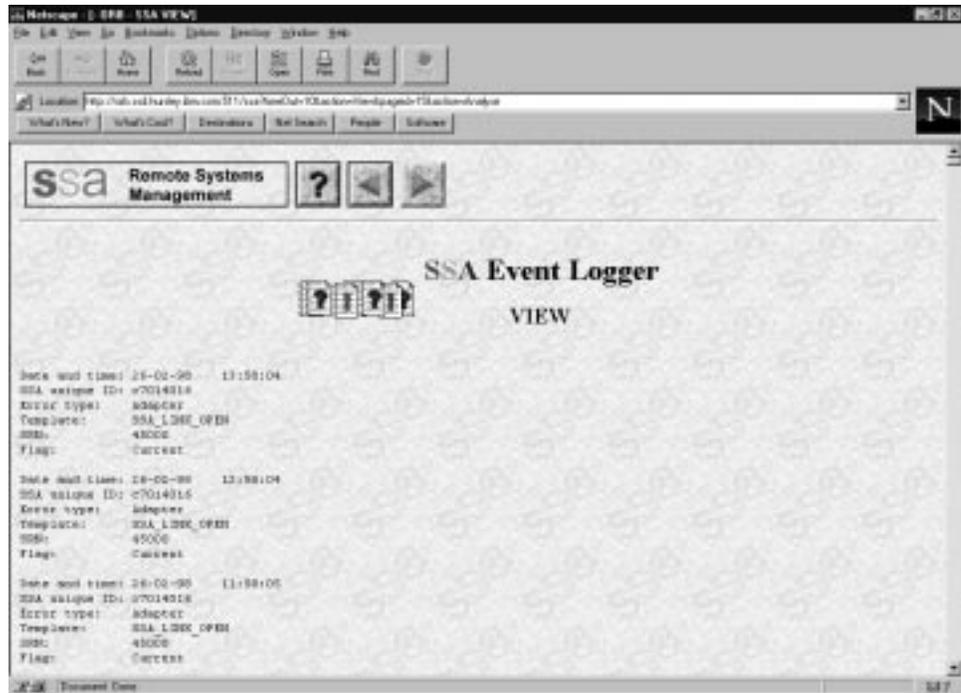


Figure 23. View of Event Logger

You can refresh and analyze this log.

3.5 Managing SSA Disks in Other Environments.

In order to attach SSA disks to other operating environments, you have to use non-IBM adapters such as those made by Pathlight, or SCSI-to-SSA convertors, such as the Vicom 2000.

Pathlight Adapters

Pathlight makes a range of adapters that can attach SSA disks to Windows NT machines, Apple Macs, and Sun systems. We recommend using an IBM adapter unless the Pathlight adapter has additional functionality that is required by the customer.

The Pathlight Toolbox is a utility for updating the Image80 adapter firmware and SSA disk drive firmware on a Macintosh PowerPC computer. It checks your SSA cables, tells you whether your network topology is functioning properly, and shows you a map of your devices. Apple Guide help makes it easy for you to learn how to use the Pathlight Toolbox.

Vicom Product Range

Vicom produces a range of SCSI-to-SSA convertors as well as manufacturing the IBM 7190. A product comparison is given in Table 10 on page 50.

Table 10. Vicom and IBM Product Comparison

	IBM 7190	SLIC 1355	UltraLink 1000	UltraLink 2000
Platform tested	SUN, HP, DEC	UNIX, Windows NT	UNIX, Windows NT	UNIX, Windows NT
HW RAID-1, RAID-0	No	No	No	Yes
Multihost support	1 - 4 hosts	1 - 16 hosts	1 - 16 hosts	1 - 16 hosts
Fiber extender support	No	No	Yes	Yes
19 inch rack mount option	No	No	Yes	Yes
Burst data rate	20 MB/s	20 MB/s	40 MB/s	40 MB/s
Sustained data rate	18 MB/s	18 MB/s	35 MB/s	35 MB/s
I/O throughput	1900 IO/second	1900 IO/second	3000 IO/second	3000 IO/second
I/O throughput with 2 UltraLinks	N/A	N/A	5100 IO/second	5100 IO/second
Capacity per SCSI channel	1 - 48 drives	1 - 64 drives	1 - 64 drives	1 - 64 drives
Capacity per SCSI channel with multiple UltraLinks	N/A	N/A	1 - 120 drives	1 - 512 drives

The tools available to help manage the disk subsystem are those utilities provided with the adapter and any tools provided with the operating system. The utilities provided with the Vicom products enable you to:

1. Download microcode.
2. Display VPD.
3. Identify a disk drive module.
4. Set/reset service mode.
5. Run disk drive diagnostics.
6. Run Vicom self test diagnostics.
7. Analyze error log .
8. Clear check mode.
9. Monitor activities .
10. Show SSA topology.
11. Force SSA master.
12. Force Web reset.
13. Write cache option.
14. Synchronous Write.
15. Quit.

All of the above functions are fully described in the Redbook, *Exploiting SSA Disk Subsystems in Sun Solaris Platform Environments (SG24-5083-00)*.

Chapter 4. Introducing IBM StorWatch Serial Storage Expert

Storage management is becoming a more complex, costly, and time-consuming task for computer systems managers and users. The advent of serial storage interfaces has brought with it the ability to create networks of storage devices. These networks can be distributed among multiple systems. Thus a need exists for a simple way to plan, manage, and monitor these types of storage networks.

The IBM StorWatch Serial Storage Expert product has been designed to fill this need.

In this chapter we discuss the IBM StorWatch Serial Storage Expert (StorX) product, and the functions it provides for storage management in the SSA environment.

The industry standard SSA interface is the first target implementation platform for StorX. This being the case, many of the examples and requirements for StorX are driven from the IBM implementation of SSA. The reader should keep this in mind when using this document.

The first section provides an overview of StorX. This is not an exhaustive description of all the StorX components. It is included here to familiarize you with the terminology and components discussed in the later sections. For a more detailed description of StorX, please refer to the *IBM StorWatch Serial Storage Expert (StorX) User's Guide (SC26-7267)*.

In the later sections we present some specific examples and scenarios of storage management and show what part StorX plays in management and monitoring of your SSA environment. Items we look at include these:

- StorX overview
- Understanding the topology of your SAA subsystem
- Monitoring event changes in your environment
- Installing StorX's Agent and client software
- Problem analysis
- Performance impacts of StorX
- Deinstalling StorX

4.1 StorX Introduction

StorX is a graphical tool for managing SSA storage networks. StorX functions with the family of SSA adapters, SSA disk units and SSA enclosures that are developed by the IBM Storage System Division. StorX provides:

- A Windows 95/NT graphical user interface that enables you to plan new SSA storage networks, create plans from existing SSA storage networks, or monitor existing storage networks.
- Discovery and display of the SSA topology, capacity, and availability of storage resources.
- Monitoring for adapter and disk errors, topology changes, and Redundant Array of Independent Disks (RAID) events.

- A physical view of disk storage in a storage network.
- Support for SSA enclosures (7133, 7131, and 3527).
- The ability to simplify the graphical presentation of your storage networks by collapsing storage networks into a single storage network icon.
- Support for SSA adapters and disks that are attached to RS/6000 servers and SP2 systems that are running AIX.
- Support for SSA adapters and disks that are attached to Intel platform servers that are running Windows NT 4.0.
- The ability to extract detailed information about the SSA devices in your storage networks through a report function.
- The ability to automatically locate a device associated with an event that has occurred in your storage network.
- The ability to print cable labels.

You can use the StorX planner to visually plan your SSA storage network. StorX provides an SSA parts palette for you to add or remove adapters, disks, connections, and text and box annotations. Editing tools enable you to select, move and change attributes for any devices in the storage network configuration you are planning.

Discovery of SSA storage networks on AIX or Windows NT 4.0 hosts is possible using the Live Viewer. StorX automatically discovers topology and attributes and displays them on any network-attached PC that is running Windows 95 or Windows NT 4.0. Monitoring for events allows adapter errors, disk errors, topology changes, and RAID events to be reported, alerting storage network managers to items that may need attention.

4.1.1 Functional Overview

When you are using the live viewer, StorX uses a:

- Manager that is installed on a Windows 95/NT 4.0 system. The manager contains the interface that displays the graphical information that represents the storage networks.
- SSA network agent installed on each system, AIX or Windows NT 4.0, that contains devices in the storage network.
- TCP/IP protocol to communicate between the SSA network agent and the StorX manager.

The manager polls the SSA network agent to gather information about the storage network. The SSA network agent receives the manager's polling request and reports the status of the devices in the storage network.

The manager uses remote procedure calls (RPC) over a TCP/IP communications network to communicate with the network agent on the host system. Because of this, RPC and TCP/IP must be enabled and active when you are using StorX.

Figure 24 on page 53 shows a conceptual view of the StorX manager and the SSA network agents on the host systems.

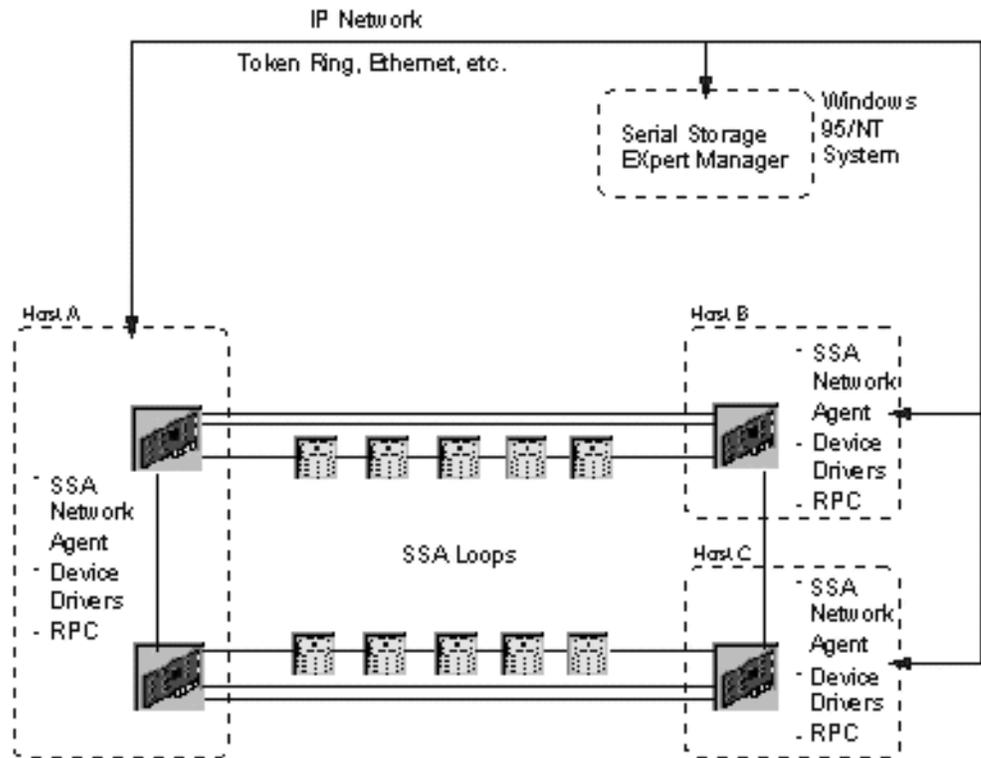


Figure 24. Sample StorX conceptual view

4.2 What is a Host?

Hosts are computers that communicate with StorX. A host must have a StorX-compatible SSA network agent installed to enable StorX to communicate with an adapter in the host. The SSA network agent software is usually installed with the adapter.

4.3 What is a Storage Network?

A storage network is a configuration of devices (adapters, disks, cables) that enable a host computer system or set of host computer systems to access data.

The connections between its devices determine the boundary of an individual storage network. StorX considers two devices to be in the same storage network if the devices are connected together. StorX considers that a cable is in the same storage network as the devices it connects.

You can use StorX to manage several storage networks by placing them in a StorX management set. StorX determines the topology of a storage network by requesting information from agent programs installed on each of the hosts in the host list.

When you use the live viewer, StorX queries local agents that are installed on host systems by using TCP/IP. The local agents return information that is related to the devices (disk units, cables, adapters) that are attached to the host. StorX

then uses this information to create a graphical representation of the storage network which StorX calls a *management set*.

Note: StorX is able to discover storage networks that are configured on AIX and Windows NT systems. However, SSA rules do not allow AIX and Windows NT host systems to coexist in the same storage network.

4.4 What is a Management Set?

A management set is the description of all objects (for example, host computers, disk units, cables, adapters, and storage networks) known to StorX. There are two types of management sets, those that you create with the Planner, which are used to plan to design the configuration of a new storage network, and those that are dynamically created by the Live Viewer, which are used to determine the topology of a storage network and monitor the states of the devices in the storage network.

You can use the StorX Planner to merge management sets. You cannot merge any management sets while using the live viewer.

4.5 What is on the StorX Display?

When you start StorX, the screen display is shown in Figure 25 on page 54.

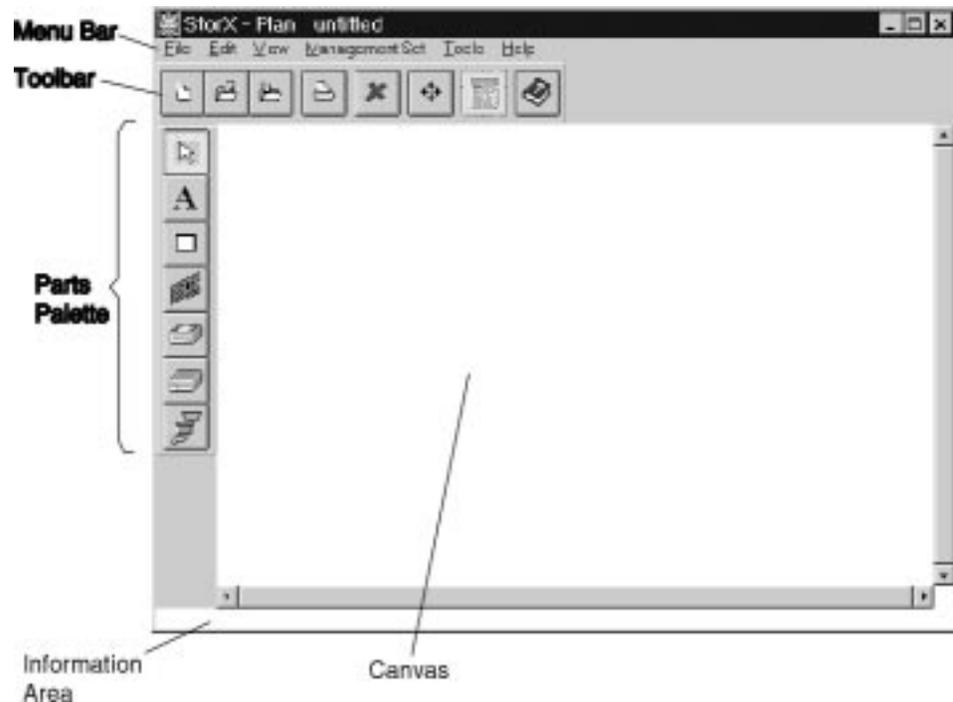


Figure 25. StorX Palette

Figure 25 contains the following items:

Menu Bar

This area contains menus so you can access several StorX functions.

Toolbar

This area contains icons that you can click on to access frequently used functions. Table 11 on page 55 describes the Toolbar icons.

Table 11. StorX Toolbar Icons

Toolbar Icon	Name	Comments
	New	Clicking this icon clears the canvas and opens a new, untitled, planner or live viewer management set. If you are creating a new management set with the live viewer, the management set hosts window is displayed. If you have an existing management set open, you are reminded to save or discard any changes before closing the management set.
	Open	Clicking this icon displays a window to select the location and the name of the management set you want to open. Only a management set that is created with the live viewer can be opened by the live viewer. The planner can open either planner or live viewer management sets. If you have an existing management set open, you are reminded to save or discard any changes before closing it. When you open a management set created by the live viewer, StorX migrates the management set to a planner management set. When the management set is migrated StorX removes the logical disk information and host-based attributes for the devices in the storage network.
	Save	Clicking this icon saves the current management set without changing the file name or directory. If the management set has not yet been saved previously, you are prompted to enter the file name and directory.
	Print	Clicking this icon prints the current management set.
	Delete	If you are using the planner, clicking this icon removes an individual object, or group of objects, from the canvas. To delete an object from the canvas, highlight the object and select Delete from the Edit menu. If you are using the live viewer, you can only delete individual objects with a status of Missing.
	Pan To	Clicking this icon causes StorX to display the entire management set. Clicking on an object in the Pan To window centers the object in the StorX window.
	Event Monitor	If you are using the live viewer, clicking this icon displays various events that are occurring in the storage network.
	Help	Clicking this icon displays the help index: this is an index of all the help topics that are available.

Parts Palette

Table 12 on page 56 contains part-palette icons that you can use to plan a storage network, or add boxes, text or enclosures to a planner or live viewer management set.

The parts palette consists of different icons, depending whether you are using the planner, or live viewer. For example, you are able to add adapters, disks and connections while using the planner.

You can display or hide the parts palette by placing a check mark next to **Show Parts Palette** on the **View** menu.

Table 12. StorX Parts Icons

Parts Icon	Palette	Name	Comments
		Select	Clicking on this icon enables you to work with the objects on the canvas.
		Add Text	Clicking on this icon places a text entry box on the canvas. When the text entry field appears on the canvas, type in the text and press Enter. To move text, click and hold mouse button 1 on the text and move the text.
		Add Box	Clicking this icon places a box on the canvas. Place the mouse pointer where you want the box, hold down mouse button 1, and draw the box you want to add. To resize a box, click and hold mouse button 1 on the border of the box, and adjust the size of the box. To move a box, click and hold mouse button 1 in the box and move the box.
		Add Adapter	If you are using the planner, clicking this icon enables you to add adapters to your planner management set.
		Add Disk	If you are using the planner, clicking on this icon enables you to add disks to your planner management set.
		Add Enclosure	When this icon is active, position the mouse pointer on the canvas and click the left button to add an enclosure
		Make Connection	If you are using the planner, clicking on this icon enables you to connect the disks and adapters that you added to your planner management set.

Canvas

StorX uses this area to display the devices, adapters, and connections that represent a planner or live viewer management set.

Information Area

This area displays status information and on-line help information for various StorX functions.

4.6 Identifying Missing Connections

StorX displays the connections between the devices in a management set as solid lines. When you are using the live viewer, StorX displays missing connections as broken (dotted) lines.

4.7 Device States

Table 13 on page 57 shows the different device states that can be shown by the live viewer display.

Table 13. StorX Device States

Device	Device State						
	Good	Changed	Degraded	Off-line	Suspect	Broken	Missing
Adapter							
Disk							
Unknown							

The meanings of the states shown in Table 13 on page 57 are:

- Good

The device is available and fully operational.

- Changed

An attribute of the device has changed. For example, the licensed internal code level may have changed. You can display the Event Monitor window and look for entries for this device to determine what has changed. You can clear the changed status of the device on the canvas by selecting **Clear Changes** on the Edit menu on the menu bar.

- Degraded

The device is not fully operational; it may operate, but may have limited capabilities.

- Off-Line

The device is not available. The device may be off-line, or the host that it is attached to may not be operational or may be powered off.

- Suspect

The status of the device was Missing or Broken, but now has been changed to Good. You can use the **Clear Changes** menu item under the Edit menu to reset the state indicators. You can display the Event Monitor window and look for entries for this device to determine what has changed.

- Broken

The device has reported a failure.

- Missing

StorX cannot locate the device. The device may have been removed from the storage network, or the host that it is attached to may not be operational or may be powered off.

When StorX detects changes in the storage network topology, the device symbol on the canvas is displayed with a cross-hatch pattern over it. This indicates that the device is unsynchronized.

A device is unsynchronized when its state or topology has not been updated to reflect its current state. The only way a management set can always be synchronized is for StorX to be constantly polling and monitoring that management set.

Selecting **Refresh Storage Networks** from the Management Set menu on the menu bar will rediscover the storage network and update, and synchronize, the device status on the display.

If You Need Help...

The StorX application contains extensive on-line help information that describes the functions of StorX and many of the tasks you will want to perform. You can access the on-line help information by clicking on the **Help** icon on the toolbar, or clicking on the **Help** icon on the menu bar.

If you need help or have any questions regarding the capabilities or use of StorX, you can contact us at the StorX Web page. You can use the Web page to download the product and sample planner management sets, contact customer support, and view additional product information. The uniform resource locator (URL) address for this Web page is:

www.ibm.com/storage/storwatch/storx

4.8 Working with the StorX Planner

The planner displays a window that you can use to plan your storage network by adding adapters, disks, connections, text annotations, and boxes.

The planner can read management sets that are created by the live viewer and the Planner, but can only create planner management sets.

A planner management set has a file extension of .SPS.

Using the StorX planner is intuitive. If you need more information, you can click on the **Help** icon, or refer to the *User's Guide* that comes with the StorX product.

4.9 Working with the StorX Live Viewer

The live viewer displays a canvas that you can use to monitor or discover an existing storage network. You can add text annotations and boxes to any live viewer management set.

A live viewer management set has a file extension of .SLS.

When StorX displays a live viewer management set, graphic symbols may be placed on the device icons on the canvas. These symbols represent the operational state of the devices on the canvas. You can right-click on the device icon and select **Properties** to view the operational state of a specific device.

Using the StorX Viewer is intuitive. If you need more information, you can click on the help icon, or refer to the *User's Guide* that comes with the StorX product.

4.10 How Do I Discover a Storage Network?

You discover a storage network when StorX queries the storage network attached to the host you have specified. StorX then displays a graphical representation of the host and the SSA devices attached to it.

To discover a storage network:

- Start the StorX live viewer.
- Click on the **New** icon in the toolbar.
- StorX opens the Management Set Hosts window.
- Type the host name, the host name and domain, or the fully qualified TCP/IP address in the **Add New Host** box. (You can add more than one host.)
- Click on the **Add** button. StorX checks to ensure that the host exists and that the SSA network agent is installed on the host. StorX then adds the host to the Hosts in the **Management Set** box.
- When you have added all of the host computers that you want to discover, click on **OK**. StorX then discovers the storage network components that are accessible from the hosts and displays them on the canvas.

If you type a name of a host that StorX cannot find, StorX issues a message that asks you if you want to add the host you specified. If you add this unknown host, StorX will try to discover the host you have specified.

4.11 Performance

The StorX product is designed to put an extremely light load on any system. One of the design goals was that there should be no (measurable) performance impact on an SSA subsystem from running StorX. For this reason, it has been designed so that:

- It would never issue more than one transaction from a host at a time, the StorX Agent network daemon (not the device driver, nor the adapter) waits patiently for one transaction to complete before issuing another.
- The transactions that are used have been carefully chosen to impose only a true negligible load on the system.
- Every transaction has carefully assessed time-outs included, so that if a device has difficulty responding, the adapter will automatically abort it. Typically, that adapter is not talked to again, unless the adapter tells StorX to.

The amount of memory StorX uses increases with the number of devices that are in the management set. For example, if you added another host, with several devices, StorX would use more system memory.

4.12 Event Monitor

When you are using the live viewer, the Event Monitor displays a list of event log entries that have occurred while StorX is running. StorX saves and restores the event log entries as part of the management set.

To view the event log entries, select **Event Monitor** from the Tools menu on the menu bar, or click on the **Event Monitor** icon on the toolbar.

Figure 26 on page 60 shows an example of the Event Monitor.

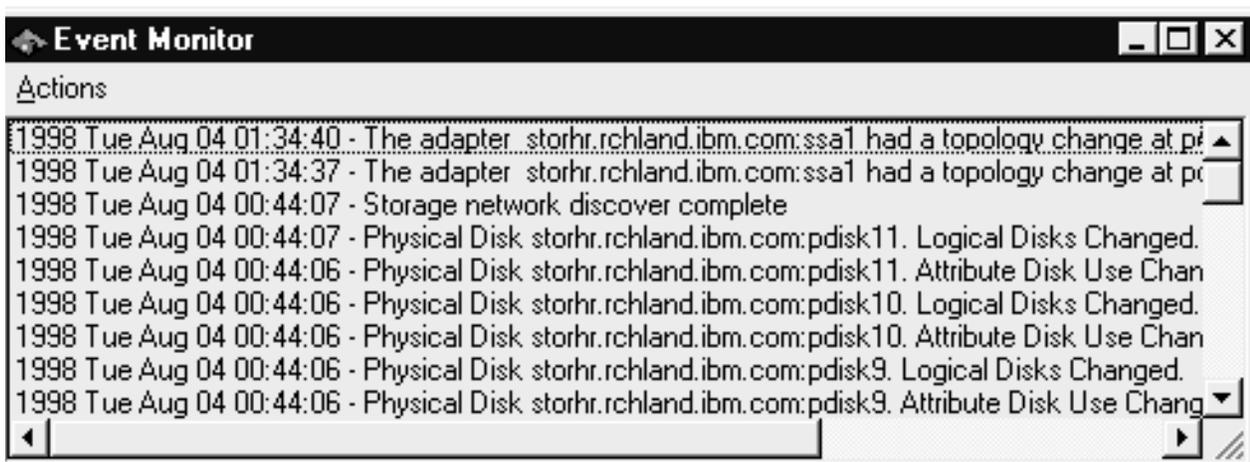


Figure 26. Event Monitor

You can also copy, export, or clear the log entries by selecting **Actions** on the event monitor menu bar. You can also use the event monitor to find a device, on the canvas, that is associated with a specific event log entry.

For more information on the event monitor, refer to *IBM StorWatch Serial Storage Expert User's Guide (Sc26-7267)*.

4.13 Installing StorX

In this section, we discuss the hardware and software requirements for using StorX, show how to install the product from the Web, and explain how to solve problems.

4.13.1 System Requirements

In this section we discuss the hardware and software you need to use StorX.

4.13.1.1 Installation Requirements

There are no special hardware requirements for the clients. They should have at least one SSA adapter.

The minimum requirements for the manager are:

- Personal computer with an Intel 486, 50 MHz processor
- 24 MB of RAM
- 25 MB of free disk space
- VGA or better display
- Windows 95 or NT 4.0

Refer to the StorX Web page (<http://www.ibm.com/storage/storwatch/storx>) for information on SSA PTFs or other software that you should install before installing the StorX.

4.13.2 Licenses

Ensure that you review the StorX license agreement.

4.13.3 Installing the IBM StorWatch Serial Storage Expert from the Web

Before you download and install StorX from the Web, exit all Windows programs.

4.13.3.1 Download and install StorX on Windows 95 or NT 4.0

To download StorX from the Web and install it on your Windows 95 or Windows NT system, follow these steps:

1. Access the IBM StorWatch Serial Storage EXpert Web page at the following URL:

`http://www.ibm.com/storage/storwatch/storx`

2. Select the option to download the IBM StorWatch Serial Storage EXpert and fill in the resulting form. When you submit the form, the displays you see depend on the browser you are using.
3. After you have downloaded the files, click on **Start**, point to **Run**, and enter the path and file name for the storx11install.exe file. For example, enter `d:\temp\storx11install.exe` and click on **OK**.
4. After reviewing the license agreement, respond to the displays that the IBM StorWatch Serial Storage EXpert presents to complete the installation process.
5. When the installation process completes, the IBM StorWatch Serial Storage EXpert prompts you to view the README file. The README file contains

last-minute changes that are important. View the contents of the README file at this time.

6. The Program Group Serial Storage EXpert contains the icons for the programs.

4.13.3.2 Using the StorX License Installation Diskette

The StorX license installation diskette enables StorX so that all of the program features remain available to you.

To use the StorX diskette:

1. Ensure you have installed StorX on your computer
2. Insert the diskette in drive a:
3. Click on **Start**, then select **Run**
4. Type

```
a:\setup.exe
```

StorX will prompt you for your license information

4.13.4 Installing the StorWatch Serial Storage Expert from the Product CD

The IBM StorWatch Serial Storage Expert CD-ROM contains the following:

- storx11install.exe: This software is the product code for the IBM StorWatch Serial Storage Expert.
- sxusergd.pdf: This is the PDF version of the IBM StorWatch Serial Storage Expert User's Guide, SC26-7267. You can view or print this file by using the Adobe Acrobat Reader 3.01 program. You can download the Adobe Acrobat Reader program from the following URL: <http://www.adobe.com/reader>

To install the files that are on the IBM StorWatch Serial Storage Expert CD-ROM:

1. Remove any previous versions of the IBM StorWatch Serial Storage Expert from your system. Use the instructions that are described in Uninstalling the IBM StorWatch Serial Storage Expert to uninstall the program.
2. Insert the IBM StorWatch Serial Storage Expert CD-ROM disc into your CD-ROM drive.
3. In Windows 95/NT, choose the **Run** command from the Start menu.
4. Type **d:\storx11install.exe** (where d: is the CD-ROM drive) and click **OK**.
5. The IBM StorWatch Serial Storage Expert installation dialog box appears. Follow the instructions on-screen to progress through the introduction, accept the license agreement, and select where you want the program installed.
6. When the installation is complete, a message appears indicating that the installation was successful.
7. Insert the StorX license installation diskette and run the program on the diskette.

4.13.5 Operating Requirements

To ensure the proper operation of the StorX Live Viewer, you should ensure the following:

- TCP/IP must be active on the system that is running StorX and the system that contains the adapters.

You can 'ping' each system to determine if TCP/IP is active on the Windows 95/NT 4.0 system or the AIX system.

- You must have one of the following operating systems installed on the host computer that contains the adapters:
 - AIX
 - Version 4.1.x
 - Version 4.2.x
 - Version 4.3.x

To determine the level of AIX on the host system, type:

oslevel

in an AIX typescript window.

- Windows NT 4.0
- You must have the latest level of SSA adapter microcode and disk microcode installed. Contact your service representative to determine if you need to update the microcode.
- Access the *AIX Fix Distribution* web site, from the StorX web site, to ensure that you have the latest versions of these device drivers. If you do not have the latest versions of these device drivers, download any of the following device drivers that need to be updated. You can search for these device drivers by name or search for the authorized program analysis report (APAR) number. The APAR numbers are, 71759 for AIX 4.1, or APAR number 71809 for AIX 4.2.
 - devices.mca.8f97.com (all AIX 4.1 and 4.2 systems, only MCA-based AIX 4.3 systems).
 - devices.mca.8f97.diag
 - devices.mca.8f97.rte
 - devices.pci.14104500.diag (PCI systems only)
 - devices.pci.14104500.rte (PCI systems only)
 - devices.ssa.IBM_raid.rte
 - devices.ssa.disk.rte
 - devices.ssa.tm.rte (for target mode applications only)
- You must have the devices.ssa.network_agent fileset on your AIX or NT systems. You can download the filesets that apply to your AIX or NT systems from the StorX web page. The AIX filesets are **installp** packages, in **tar** format.
- Install the **installp** packages by using **smit**.

Note: You must restart your system after installing these device drivers.

4.13.6 Installing the IBM SSA Network Agent for Windows NT 4.0

The IBM SSA network agent enables StorX to discover Window NT 4.0 storage networks.

To install the network agent:

1. Download the Windows NT 4.0 SSA network agent file from the StorX web site.
2. Click on Start, then select Run.
3. Type c:\agentinstall.exe (where c:\ is the drive and directory containing the SSA agent software you have downloaded).
4. Respond to the prompts that are displayed to you.

4.13.7 Starting StorX

To start StorX, click on **Start**, point to **Programs**, point to **Serial Storage EXpert** and select the appropriate icon.

You can also start the StorX by entering the following commands at an MS-DOS prompt. Ensure that you set your system to the directory that contains the StorX, (for example, cd \storx). Enter one of the following commands in an MS-DOS prompt:

- STORX: This command starts the planner.
- STORX /L: This command starts the live viewer.

4.14 Problem Analysis

The manager polls the SSA network agent to gather information about the storage network. The SSA network agent receives the manager's polling and reports the disposition of the devices in the storage network.

The manager uses remote procedure calls (RPCs) over a TCP/IP communications network to communicate with the network agent on the host system. Therefore, RPC must be enabled and active when you are using StorX.

To enable manager-to-network agent polling, the manager, SSA network agent, and the device drivers on the host systems must be at compatible software levels.

Figure 27 on page 65 shows a conceptual view of the StorX manager and the SSA network agents on the host systems.

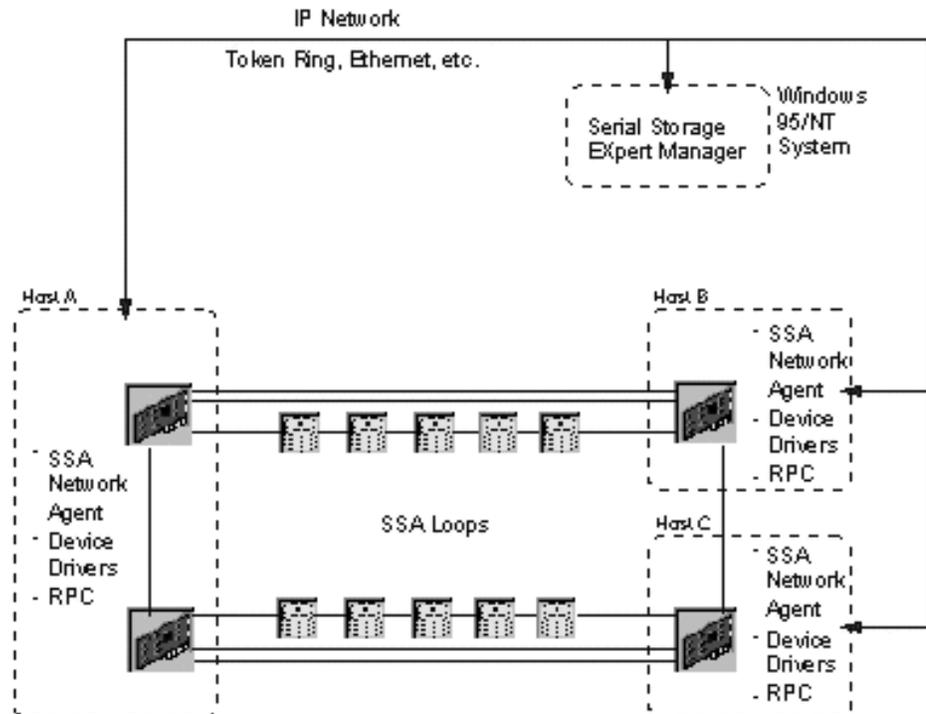


Figure 27. Structure of StorX Functioning

If you encounter a problem while installing or using the StorX, consult the Web page <http://www.ibm.com/storage/storwatch/storx>. This Web page has the latest information about the installation and use of this product.

Also ensure that:

1. Your machine meets the requirements stated in "Hardware and Software Requirements".
2. The manager that is installed on a Windows 95 or Windows NT 4.0 system. The manager contains the interface that displays the graphical information representing the storage networks.
3. TCP/IP under Windows 95 or Windows NT is installed and active, and you can communicate with the host systems through TCP/IP.
4. You have applied the appropriate SSA PTFs needed to support the system you are using.
5. You have checked, whether the daemon is available on every client by using this command:

```
rpcinfo -t <hostname> 300667
```

The result should be

```
program 300667 version 1
ready and waiting
```

6. An SSA network agent is installed on each system that is part of a storage network. To test that an SSA network agent is installed, issue this command:

```
lslpp -L devices.ssa.network_agent.*
```

7. The level of the SSA network agent matches the level of AIX. You can determine the level of AIX with:

```
oslevel
```

8. The following files are available on every client

```
/usr/sbin/rpc.ssald  
/usr/sbin/ssanetd
```

4.15 Unsupported Devices

StorX is unable to discover and display the network topology of some storage networks. When this happens, the StorX may fail to discover all the devices in a storage network.

4.16 Performance impacts of StorX

In this section we discuss the impact of StorX on the performance of the selected customer systems.

We assume that your production applications are running on these systems.

4.16.1 Impact on Disk Performance

There is no impact on disk performance. The StorX daemons receive their data only from the adapter. No input is provided by the disk drives.

4.16.2 Impact on CPU Performance

To monitor for performance impact, there are two processes on the AIX machine that need to be active: the local library `rpc.ssald` and the network daemon `ssanetd`.

If the live viewer has not been started, there are no StorX daemons running on your system.

The daemons are started by a remote procedure call from the live viewer.

If the network daemon and local library are running in transient mode, the local library will terminate 2 minutes after the last inquiry from the live viewer, and the network daemon will terminate 15 minutes after it was last accessed by the local library.

When the live viewer is running, two separate events can occur. These are:

- Discovery of storage network
- Refresh of storage network

Discovery of Storage Network

This process is used to initially discover the storage network.

Figure 28 on page 67 shows an example of a discovery process recorded with the AIX performance toolbox for Version 1.0.3 of the live viewer, for an RS/6000-G40 with two processors, 256 MB main storage, three SSA adapters, and four disks in one loop.

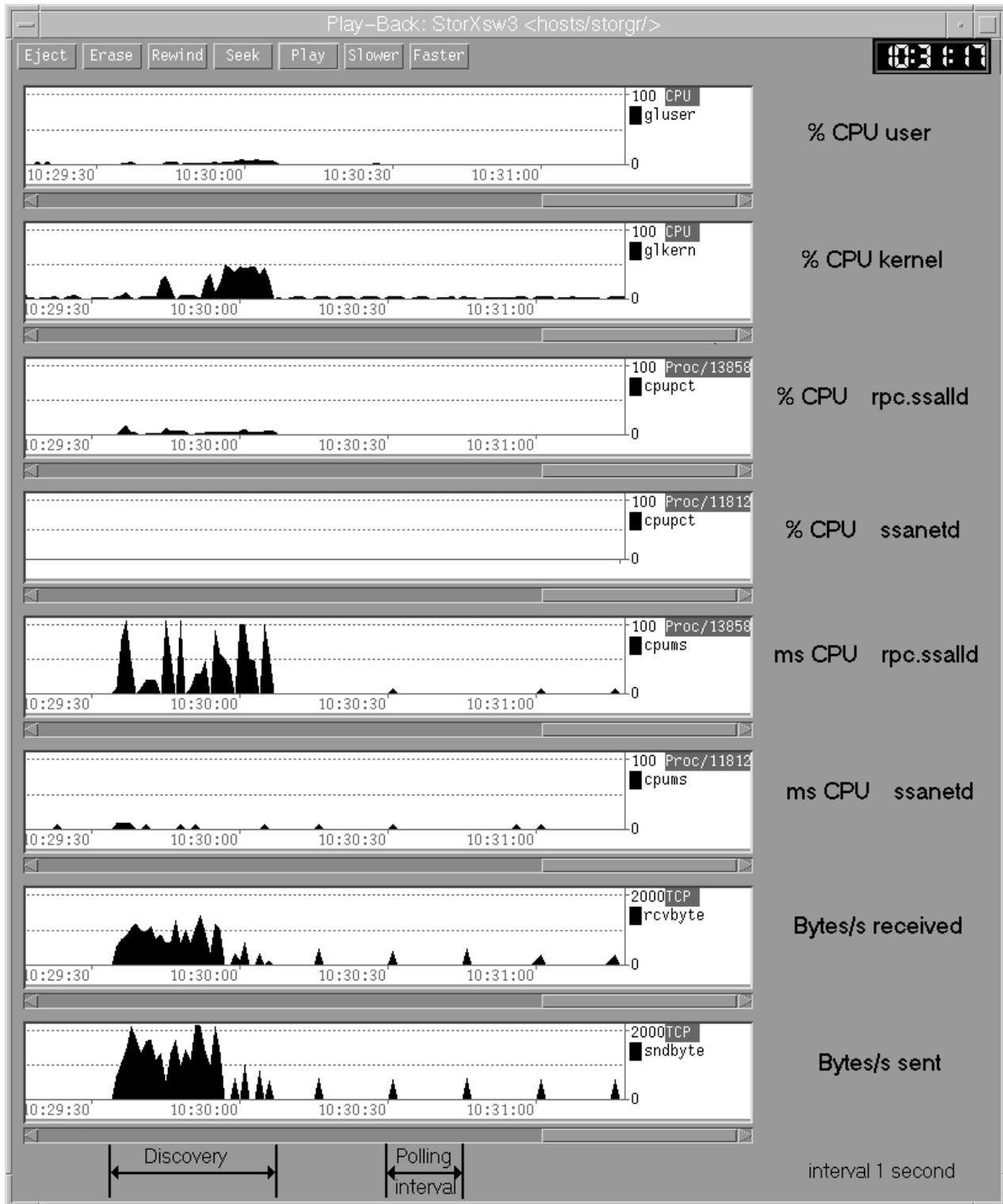


Figure 28. Discovery of a RS/6000-G40

The impact of the network daemon is so low that it is hardly measurable. The impact of the local library is lower than 5%. There are some kernel activities as the result of system calls. However, the peaks of the kernel activities are less than 50% of processor utilization.

Refresh of a Storage Network

After detecting a storage network change, the live viewer displays a message that a refresh of the storage network is required.

To get the new configuration of the storage network you must start a refresh.

Figure 29 on page 68 shows an example of a refresh process recorded with the AIX performance toolbox for Version 1.0.3 of the live viewer.

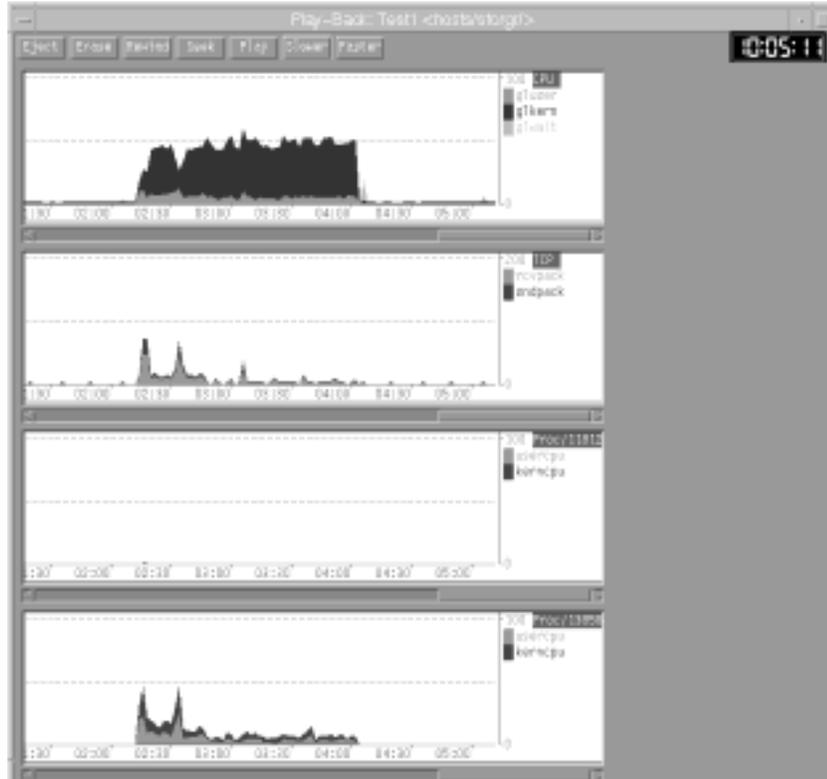


Figure 29. Refresh of an RS/6000-G40

The impact on processor performance is comparable to that of a discovery, but the refresh takes a shorter time. The entire impact on the system is less than that of a discovery.

Elapsed Time

Elapsed time is the time between discovery of a storage network and termination of the live viewer, excluding the time consumed by any refreshes.

The impact of an elapsed time of local library and network daemons is so low that it is hardly measurable.

4.16.3 Impact on Network performance

Figure 28 on page 67 shows the bytes per second received and sent for a discovery and for elapsed time. Figure on page 68 shows the packages per second sent and received for a refresh and for elapsed time.

At discovery and refresh times, the data rates are much higher than the elapsed time. In those times, the data rates for bytes sent and received are mostly less than 1 KB/s. That should not be a problem for any network, and actually facilitates the use of slow connections such as modem connections.

At elapsed time the polling time variable in the management set determines how often the live viewer sends an inquiry to the storage network. This interval can be 15 or 30 s, 1, 3, 5 or 10 min, or none. Only a few bytes of data were sent and received at an inquiry, and thus the impact on the network is hardly noticeable.

4.16.4 Summary

When you do not use the live viewer, there is no impact on performance.

Most of the time, the live viewer polls the clients at an interval specified by you. The impact on the network and on the client is hardly noticeable.

Whenever you set monitoring on, after changes to your storage network, you should discover or refresh the storage network display. The impact at this time is much higher than that of the elapsed time, but it is mostly less than 5%.

4.17 Printing

This section discusses the various tasks that are associated with printing the information in a management set.

When you select Print, StorX opens a dialog window so you can make choices concerning the print operation. You can choose:

- Printer name selection
- Select all, or a range of pages to print.
- Select collated, or non-collated print sequence.
- Select the number of copies to print.
- Select to begin printing or cancel printing.
- Select various properties that are associated with paper and the positioning.

4.17.1 Printing the Canvas

You can print the contents of the canvas, which will include all storage networks in the current management set.

To print the canvas, click on the **File** menu, click on **Print**, and select **Canvas**.

4.17.2 Printing Cable Labels

StorX aids you in documenting and labeling the connections in your storage network by providing the capability of printing labels, based on your storage network. These labels will help you:

- Connect the host adapters to the enclosures so they match your storage network plan.
- Identify the source and target of the connection if there is a problem with the storage network.

You can request to print labels by choosing one of the following:

- On the menu bar, click on **File**, then **Print**, then **Labels**. This will print labels for every connection in the current management set.

- Right click on a connection to access the canvas pop-up menu and click on Print Selected Labels. This will print labels for the individual connection.
- Select several connections by pressing and holding the Shift key while you click on the connections for which you want to print labels. Place the mouse pointer over any of the highlighted connections on the canvas and right click to access the canvas pop-up menu. Click on Print Selected Labels to print cable labels for the highlighted connections.
- On the menu bar, click on Management Set, then Storage Network Navigator, and highlight a storage network. Place the mouse pointer over any of the highlighted connections on the canvas and right click to access the canvas pop-up menu. Click on Print Selected Labels to print cable labels for all of the connections in the storage network.

Depending on the number of connections in your storage network, the labels can span multiple pages. The print window will display the number of pages necessary to print labels for all of the connections.

For each connection in the plan, StorX prints two copies (one each to attach at the source and target ends of the connection). StorX prints the same information on each side of the label, so you can fold each label in the middle. This enables you to view the cabling information on each side of the folded label. StorX prints a unique number on each label. StorX prints this same number on two labels so you can identify the pair of labels that are needed for each end of the cable.

The contents of the printed labels are:

- Enclosure, location, and port number (if the connection is from or to a device in an enclosure).
- Unique ID for the device, serial number, and port number (if the connection is from or to a device which is not in an enclosure).
- Host name, adapter name, slot number, and port number (if the connection is from or to an adapter).

When you choose to print labels, StorX checks for any disks that are not in an enclosure, for example, the disk is in a Planner management set. If StorX locates a disk that is not in an enclosure, StorX will ask you if you want to continue printing labels. If you respond Yes, StorX will print all labels, except labels for backplane connections and resiliency connections, and disks that are not in an enclosure.

You should use labels with a format of two columns and 10 rows on each sheet. For example, Avery 5161, 5261, 5961, 5661, or 8161.

4.17.3 Printing Enclosures

You can print images of the enclosures in your management set. The enclosure labels you have specified in the Enclosure View Preferences window determine the labels that are printed with the enclosures. StorX prints one enclosure on each printed page. Empty slots in an enclosure are printed as white. Slots with blank devices are printed as shaded. Slots with devices have the device icons printed in the slot with the device labels that you specified in the Disk View Preferences window.

To print the enclosures in your management set click on File, click on Print, and select Enclosures.

4.17.4 Displaying Page Boundaries

You can preview the page layout by displaying the page boundaries before printing the canvas. This enables you to make adjustments to avoid page splits that intersect the devices when the canvas is printed.

To preview a management set, select View on the menu bar, and then select Show Page Boundaries.

4.17.5 Printing a Large Management Set

If a management set is too large to fit on a single page, StorX formats the management set into page segments so it can be printed and reassembled. StorX prints information in the corners of the page segments.

When you print the management set (canvas), StorX provides information on each printed page to aid you in assembling the printed management set:

- Management set name (upper left corner).
- Page x of x (upper right corner).
- Print date, time, and management set file location (lower left corner).
- Row x and Col. x (lower right corner). The row designations and the column designations describe the relationship of each printed page to other printed pages in that management set.

You can preview the page layout by displaying the page boundaries before printing the canvas. This enables you to make adjustments to avoid page splits that intersect the devices when the management set is printed.

4.18 Updating StorX

To support new devices, a new rules file will be available on the Web.

4.19 Uninstalling the StorX

The following section describes the process for uninstalling StorX.

4.19.1 Planner and Live Viewer

Never use anything other than the Uninstall program of StorX to remove the StorX.

Follow these steps to uninstall StorX:

1. Before uninstalling the StorX, ensure that the product is not in use and that you have saved your management set files.
2. Click on the **Start** button, point to **Programs**, point to the **StorX** program, and select **Uninstall**.
3. When the Confirm File Deletion window appears, select **Yes** to remove StorX from your system.

4.19.2 Removing StorX Network Agents from an AIX Client

To remove the file sets, use **smit install_remove** and select **devices.ssa.network_agent.rte**.

Chapter 5. Monitoring and Managing SSA Environment with StorX

In this chapter we explain how to monitor and manage your various SSA configurations, perform some troubleshooting, and run diagnostics with StorX. We do not cover any other SSA tools. For a description of these other tools, please refer to Chapter 3, "Tools for Managing SSA Storage" on page 19.

5.1 Customizing a StorX Live View

The live viewer in StorX displays most of the system data that you wish to see. Each adapter is described and displayed with its host name. The disks can be fully described with serial number, logical name, size and type. A typical StorX viewer screen for SSA disks attached to a Windows NT system is shown in Figure 30. Only part of the canvas has been shown for clarity.

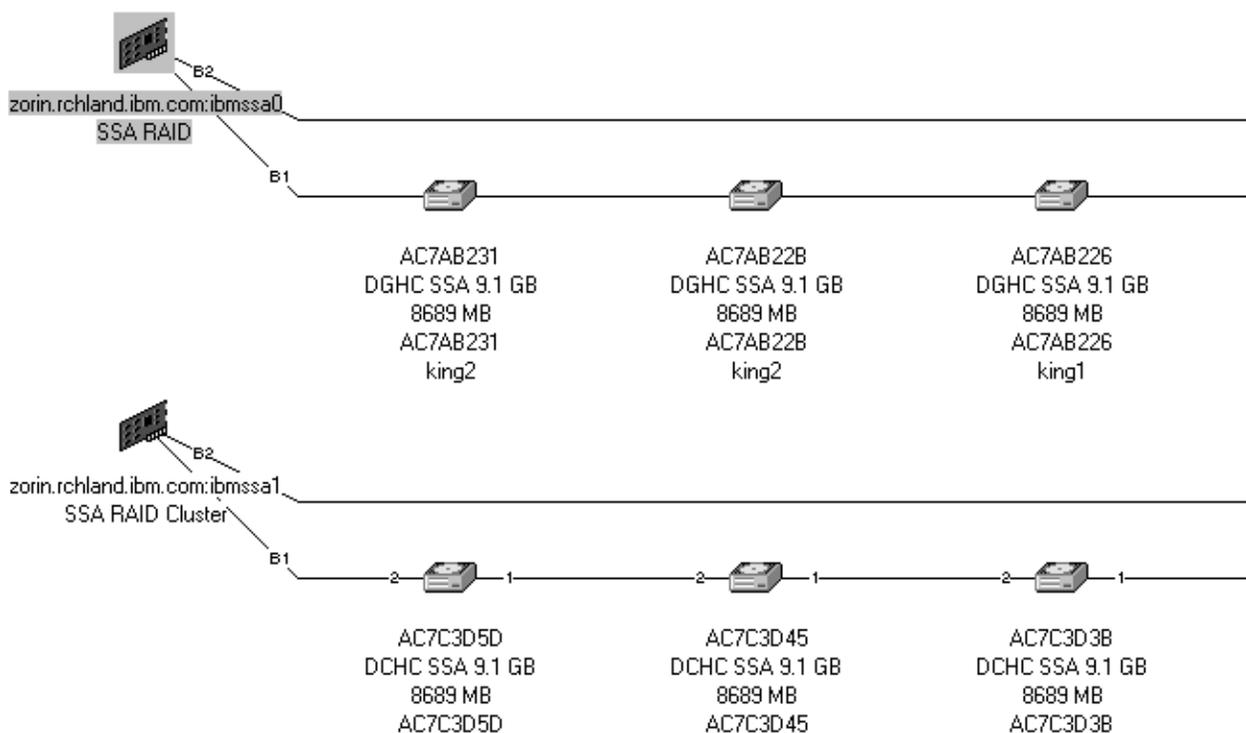


Figure 30. Uncustomized StorX Live View

The information displayed for each component in the storage network can be customized in the management set properties. The menu items are shown in Figure 31 on page 74. Any combination of the properties can be selected for display. Similar menus are available for the other components in the system.

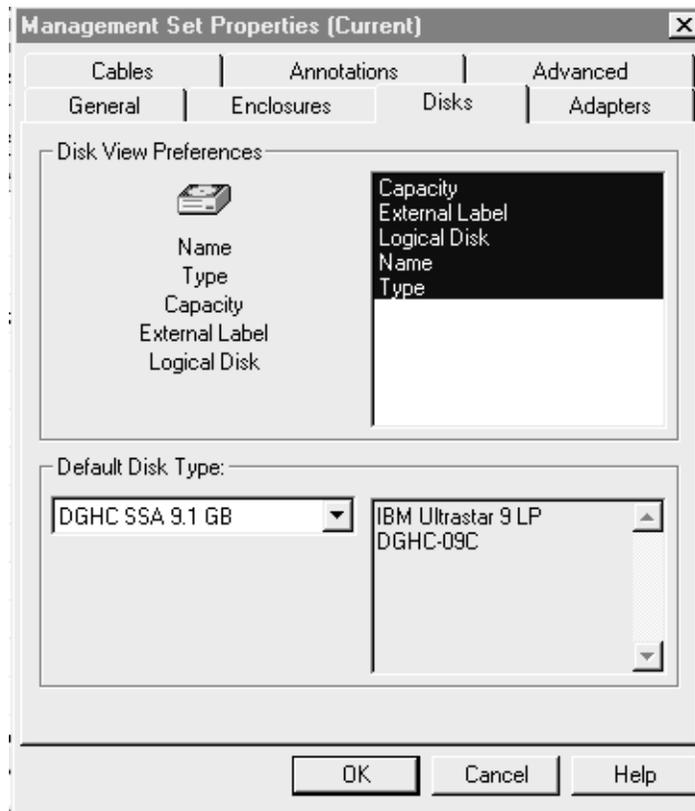


Figure 31. Management Set Properties

One of the major problems with the uncustomized canvas is that for large, complex configurations, it can be difficult to see all the devices easily. By adding the 7133 enclosure icon to the canvas, which can be obtained from the parts palette by a simple drag and drop operation, and dragging the disks into the enclosure, the size of the canvas can be reduced. Figure 32 on page 75 shows the customized StorX live view using the enclosure icon for the same configuration as the previous example.

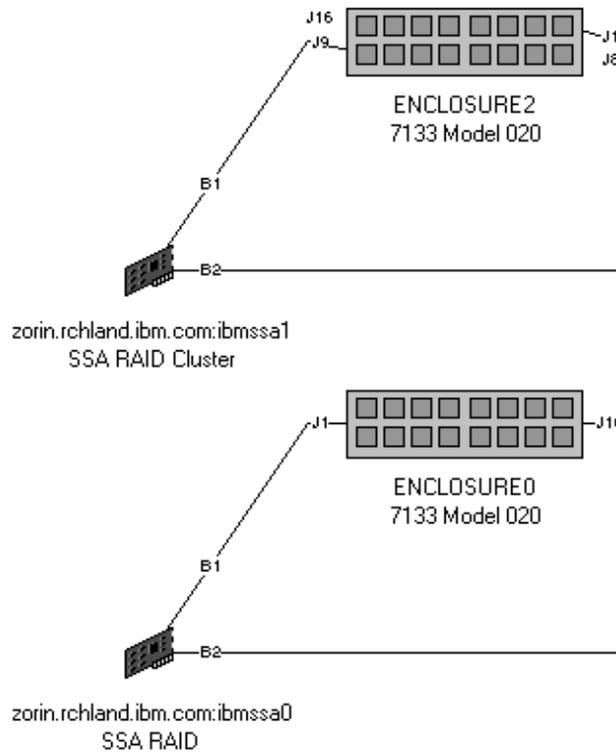


Figure 32. Customized Storx Viewer

It is possible to further customize the picture by adding descriptive text to the canvas. You may add as much text as required to suit your requirements. It is also possible to split the 7133 enclosure into its four component quadrants. This feature is useful when each quadrant is attached to different adapters. Figure 33 on page 76 shows a 7133 enclosure broken down into its quadrants

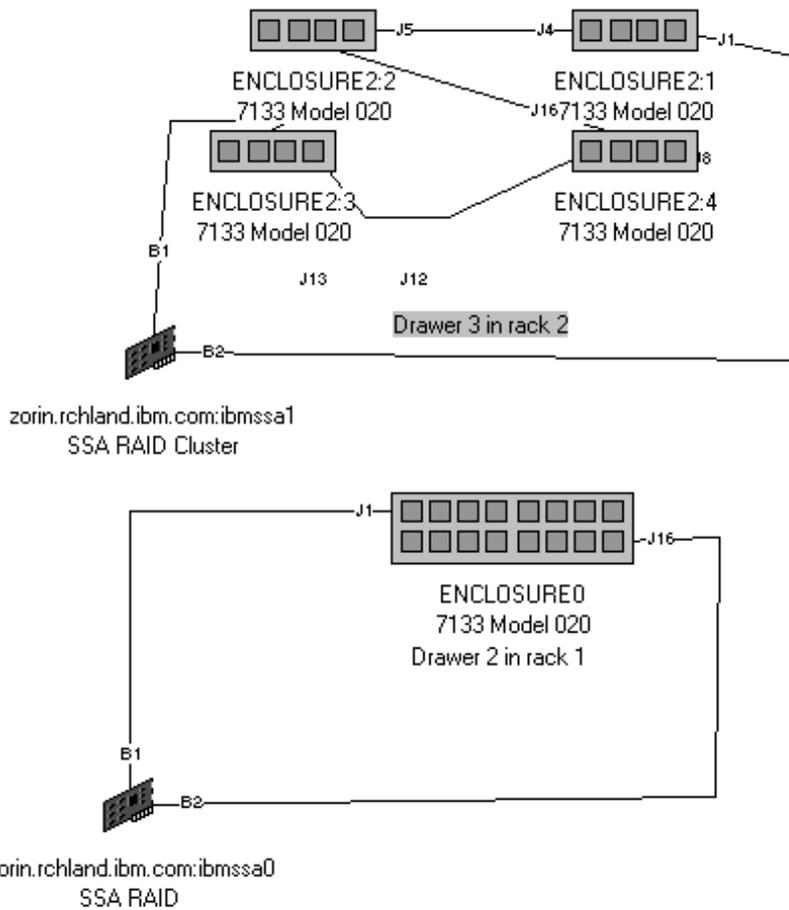


Figure 33. Customized StorX Viewer with Annotations

5.2 Event Monitor

With StorX, there is an event monitor which displays a list of every event log entry that have occurred while StorX has been running. StorX saves and restores the event log entries as part of the management set.

To view the event log entries, select **Event Monitor** from the Tools menu on the menu bar, or click on the Event Monitor icon on the toolbar.

An example of the event log monitor is shown in Figure 34 on page 77.

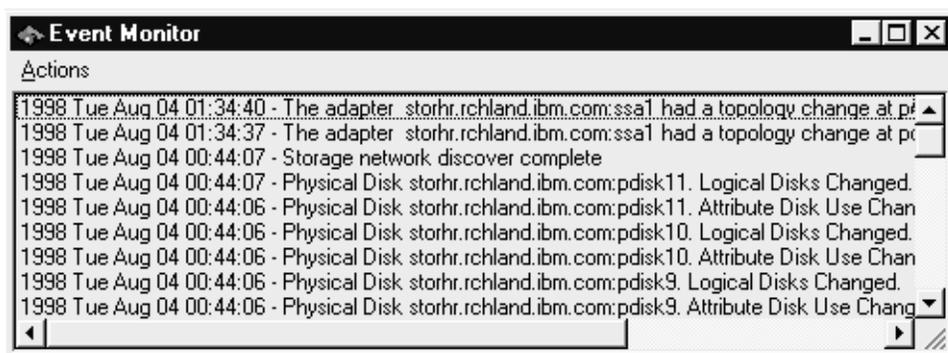


Figure 34. Event Monitor Display

It is possible to find a device on the canvas that is associated with a specific event in the event monitor window. StorX will find and highlight the associated device on the canvas and position the canvas to the location of the device. To find a device that is associated with a specific log entry, click on the event in the Event Monitor window to highlight it, and click on **Actions** on the event monitor menu bar. A drop down menu will appear, click on **Find** and the device associated with the event is highlighted. The result of this can be seen in Figure 35 on page 77.

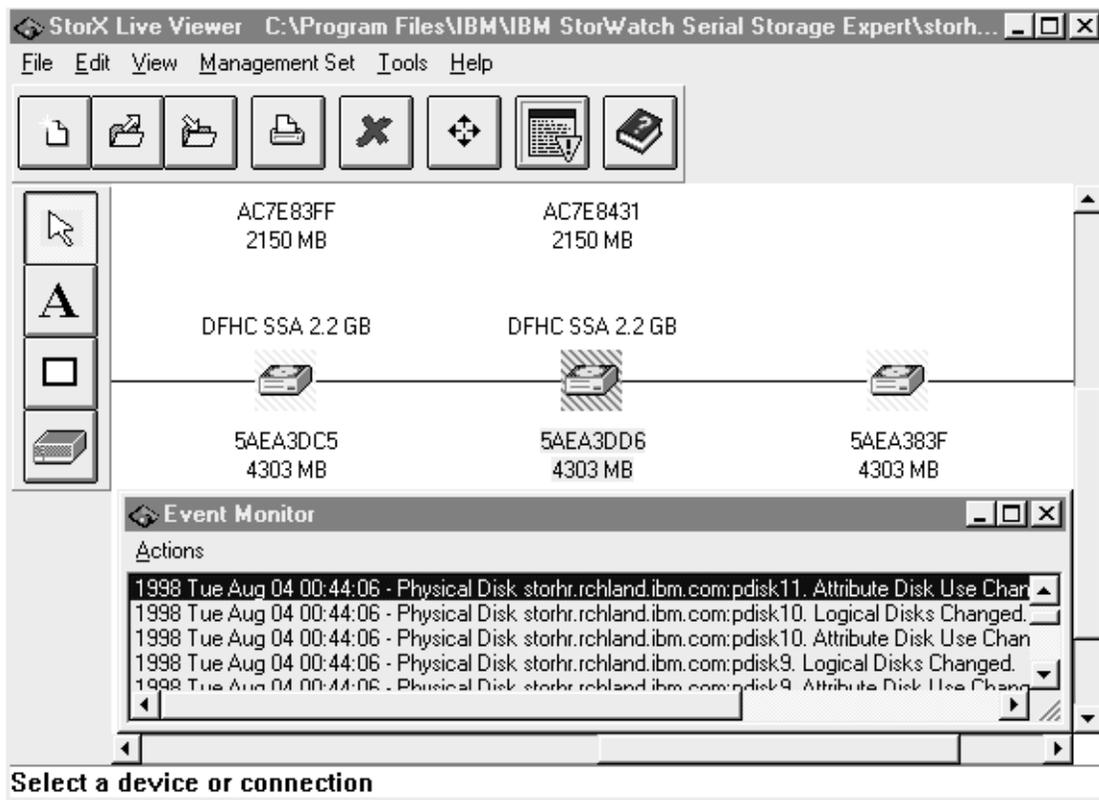


Figure 35. Event Monitor Showing the Associated Device

5.3 Monitoring and Managing a Shared Configuration

In this section, we will use a configuration with eight SSA disks in one loop connected to two host systems, RS/6000-1 and RS/6000-2. When you manage your configuration from RS/6000-1, logical disk and pdisk names differ from those used when you manage your configuration from RS/6000-2, Figure 36 on page 78 and Figure 37 on page 79, respectively, show the configuration managing from RS/6000-1 and from RS/6000-2 (logical disk option).

Note: For all subsequent examples and diagrams, for ease of use and explanation, we will show all disk drives attached to the adapters. We will not be using the 7133 enclosure icon.

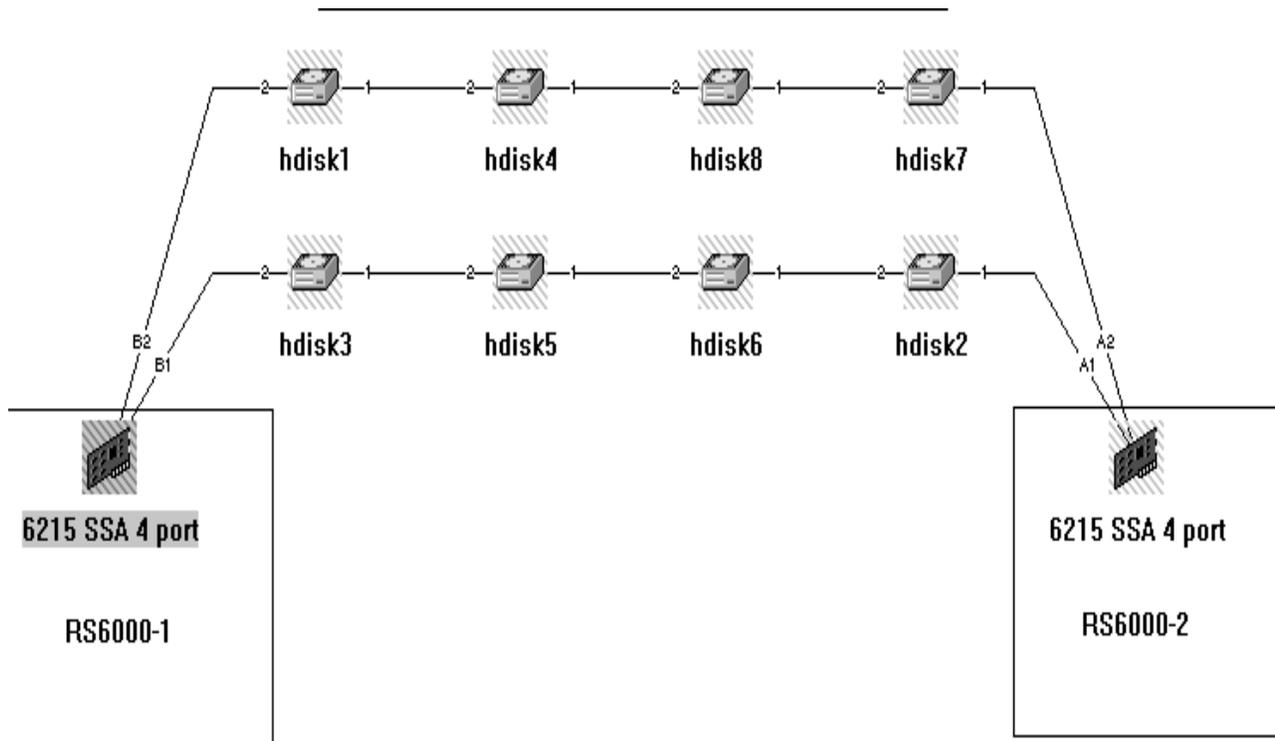


Figure 36. StorX Live View from RS/6000-1: Logical Disk Option

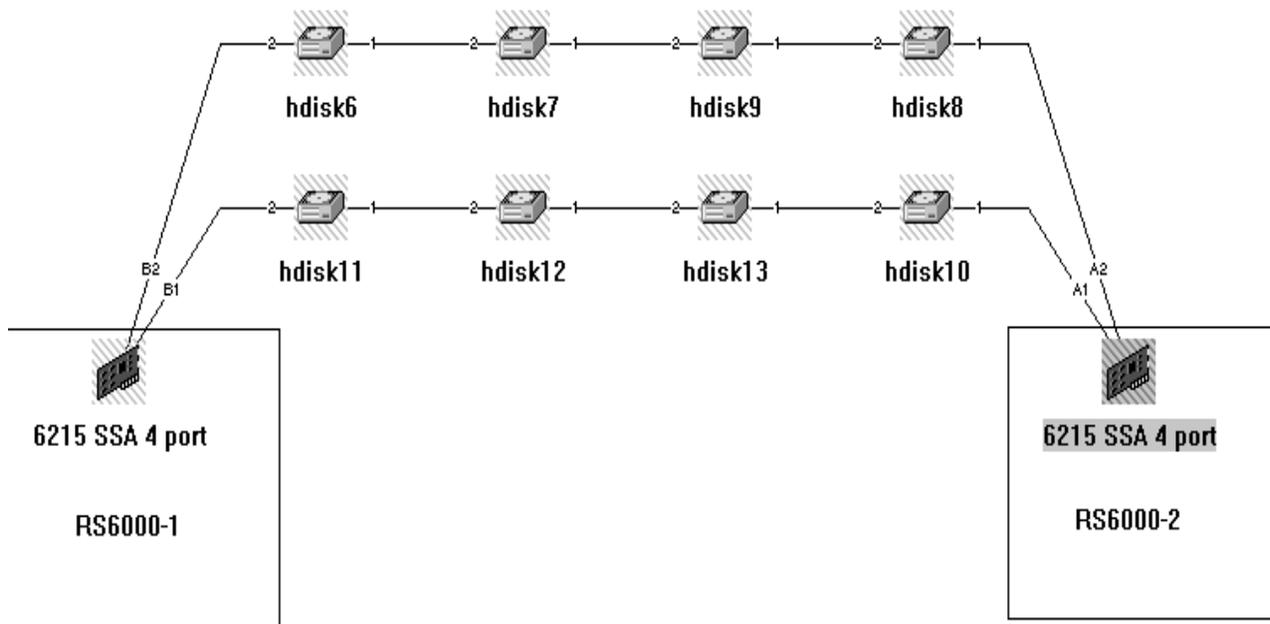


Figure 37. StorX Live View from RS/6000-2: Logical Disk Option

If you compare Figure 36 and Figure 37 you see that the names of the logical disks differ according to which station you manage your configuration from.

Figure 38 on page 79 and Figure 39 on page 80, respectively show configuration managing from RS/6000-1 and from RS/6000-2 (pdisk name option).

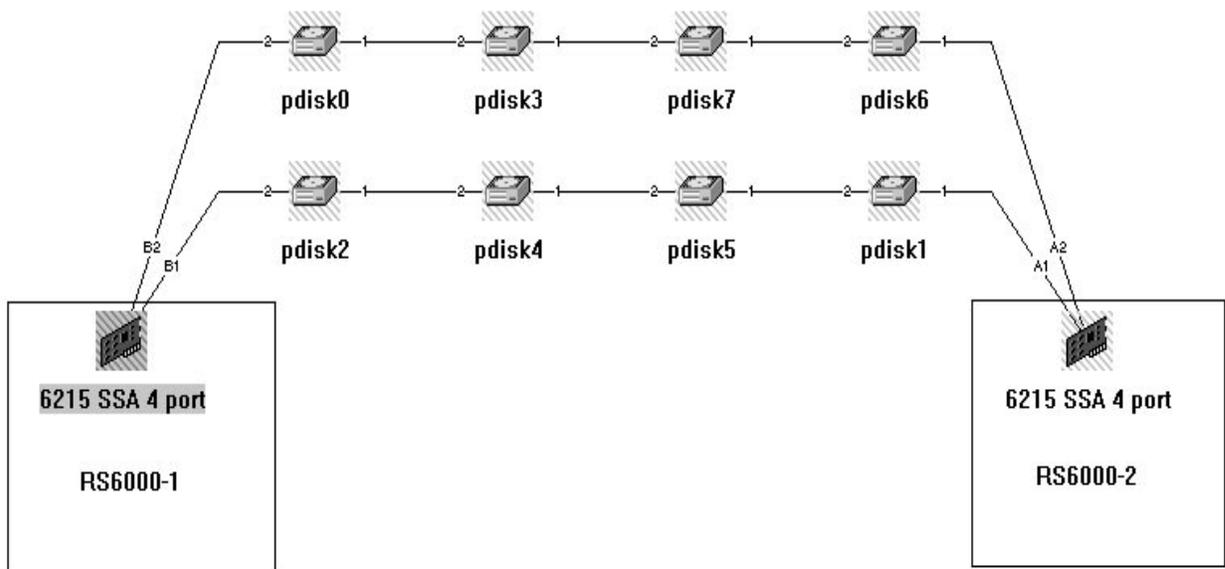


Figure 38. StorX Live View from RS/6000-1.Pdisk Name Option

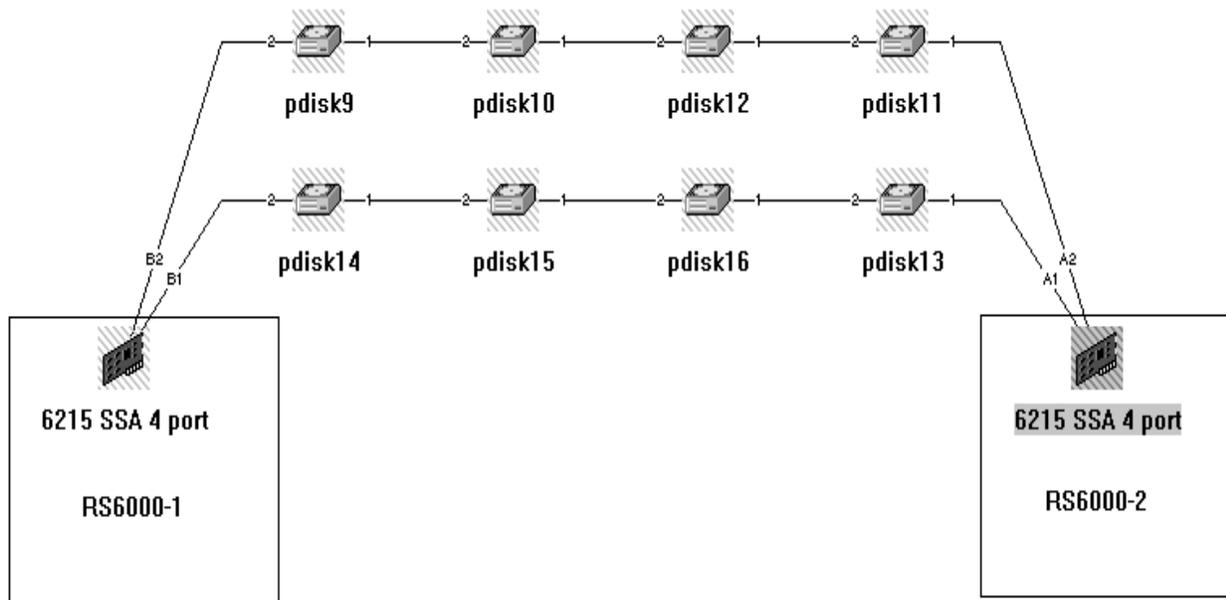


Figure 39. StorX Live View from RS/6000-2. Pdisk Name Option

Similarly, in Figure 38 and Figure 39, the pdisk names differ according to whether you manage your configuration from RS/6000-1 or RS/6000-2.

As you can see, it is very easy to manage your configuration with StorX. You can also select other SSA view preferences as we have done previously, but in this case, the external labels, capacities, and types will be the same for both systems.

5.4 Monitoring and Managing a Complex Configuration

In this section, we explain why it may be beneficial to customize your SSA configuration, especially when it is a complex configuration.

We now look at a complex configuration with 32 SSA disks from two SSA bays, on one loop and shared by two systems. Figure 40 on page 81 shows us a traditionally drawn configuration diagram.

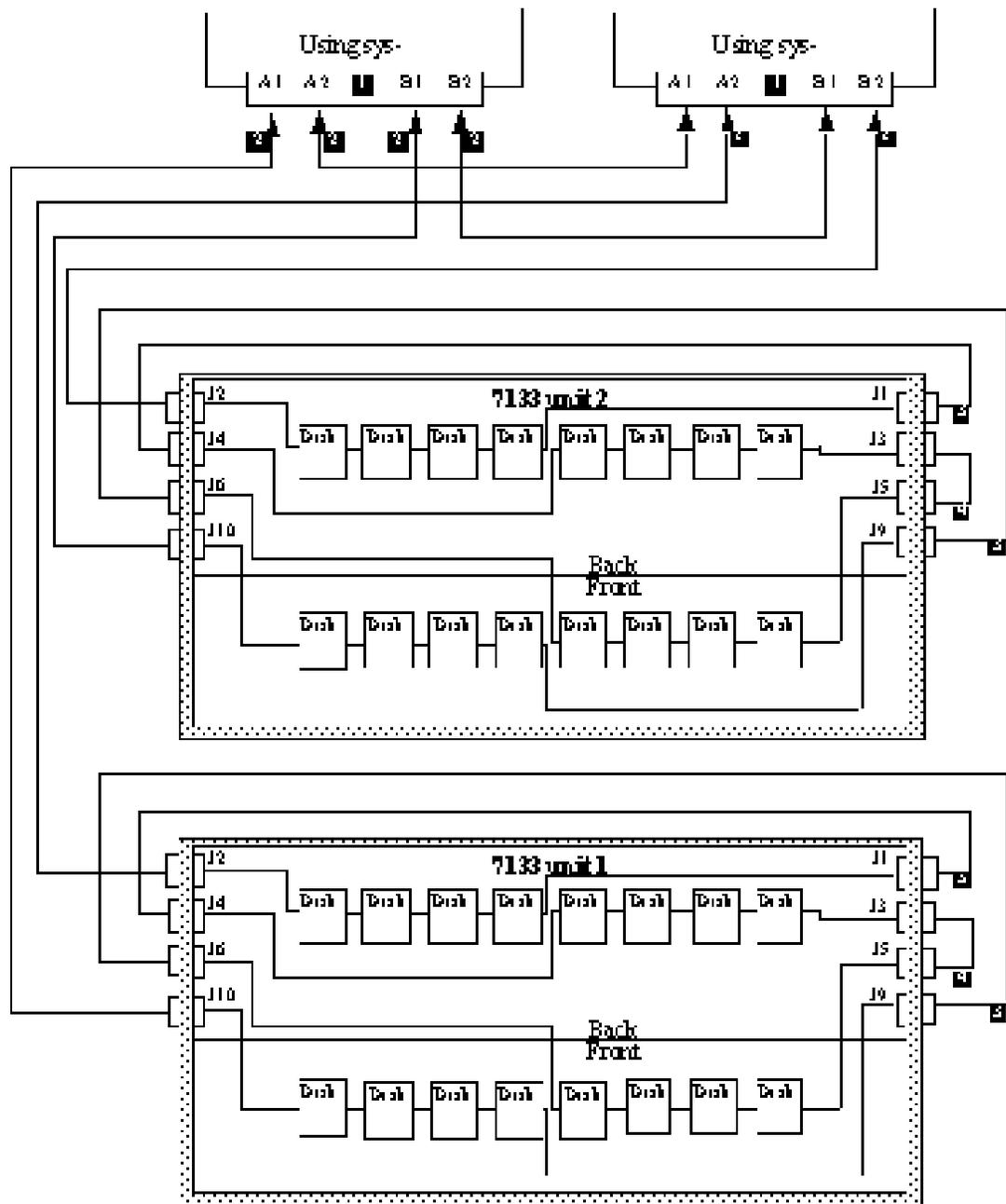


Figure 40. SSA Planning without Using StorX

Figure 41 on page 82 shows us the uncustomized StorX live view for the same configuration

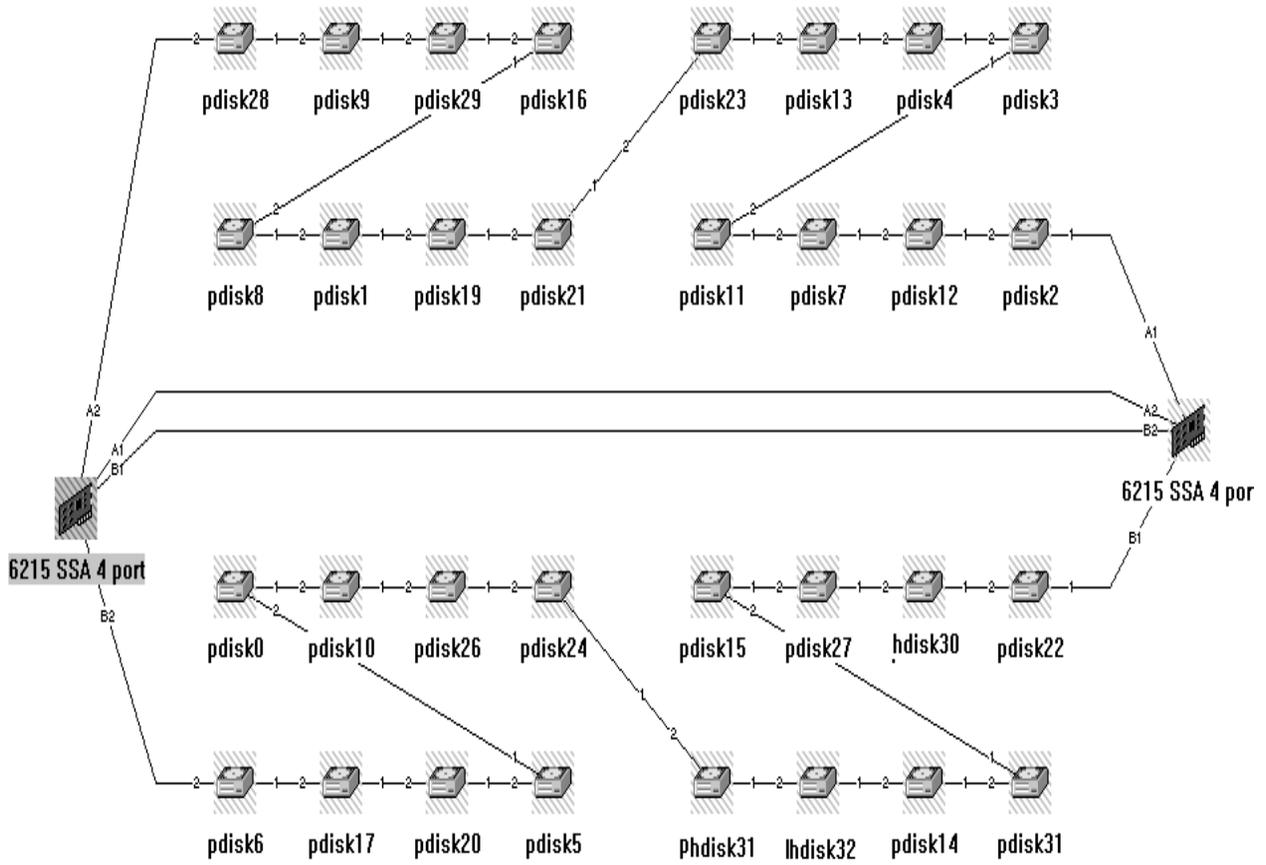


Figure 41. Live View of Uncustomised StorX

Figure 42 on page 83 shows us the same configuration, with customization.

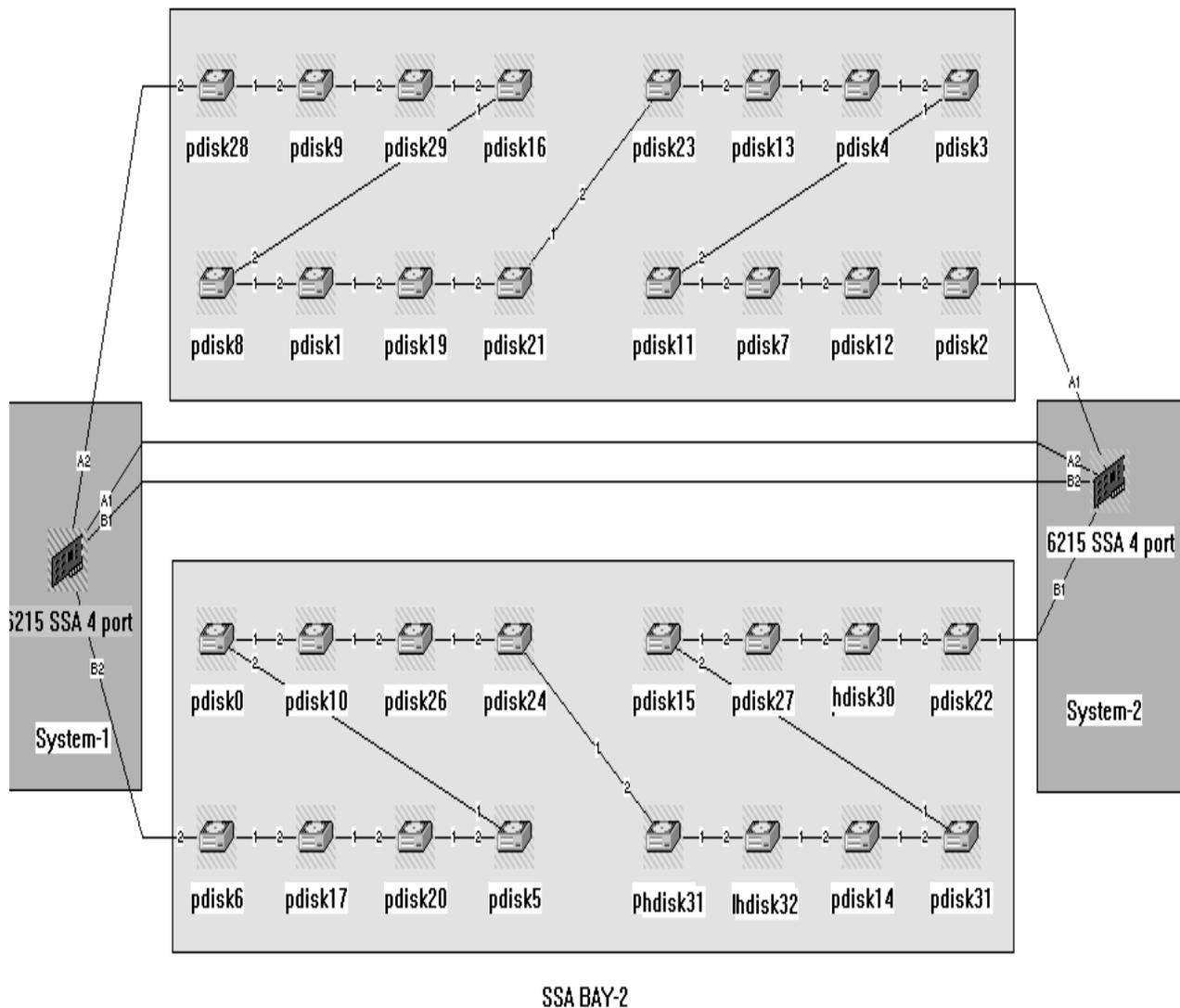


Figure 42. Live View of Customized StorX from System-1, Pdsk Name Option

There are many ways to customize a StorX live view, and it is very easy to do. With full drag and drop capabilities, it is easy to show the configuration in a manner that suits you. The important point is that if you manage many SSA configurations, you may want to customize the live view to facilitate your exploitation.

5.5 Using StorX for Subsystem Troubleshooting

In this section, we discuss SSA subsystem troubleshooting. We show what happens, and what recovery steps are required, when:

- There is a disk failure.
- There is an SSA open loop.
- There is an adapter failure.

Figure 43 on page 84 and Figure 44, respectively, show us one configuration with two loops connected to one SSA adapter, with logical disk option and with pdisk option.

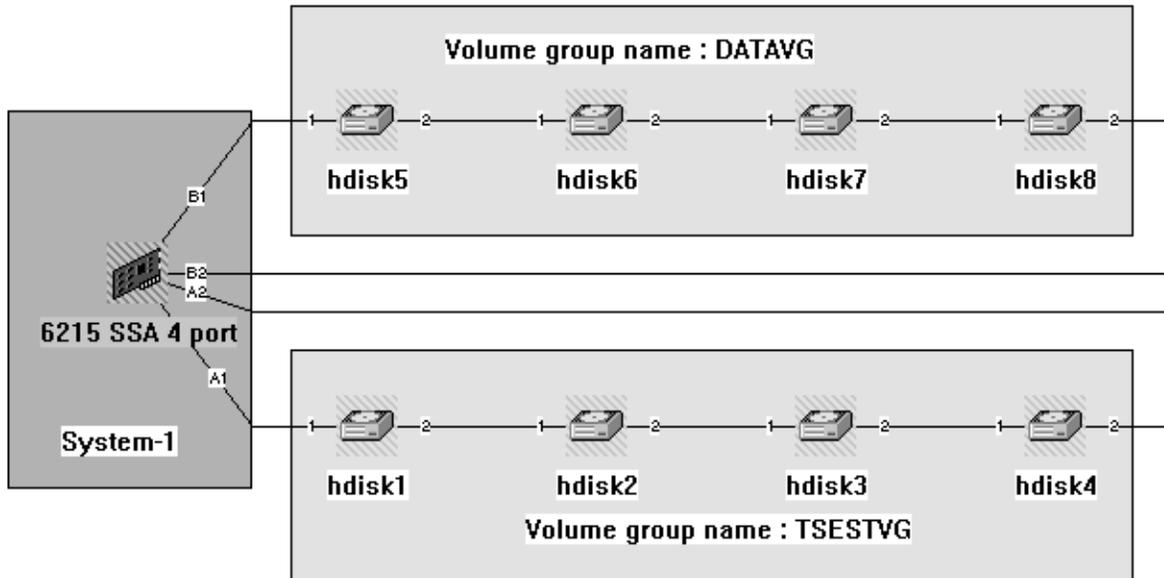


Figure 43. Initial Configuration.-Logical Disk Name Option

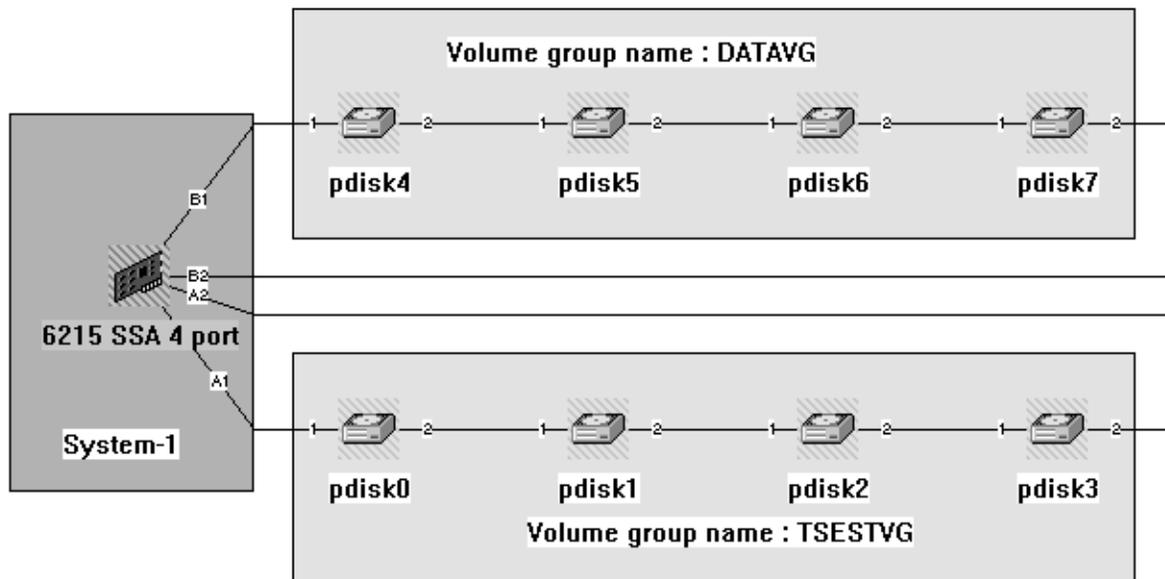


Figure 44. Initial Configuration. -Physical Disk Name Option

5.6 Monitoring and Managing an hdisk Failure

In our example, a problem occurs on hdisk 7.

The first indication is that the live viewer pops up a window indicating that a change has occurred and a refresh is required to see the event. Figure 45 on page 85 shows the result after the refresh.

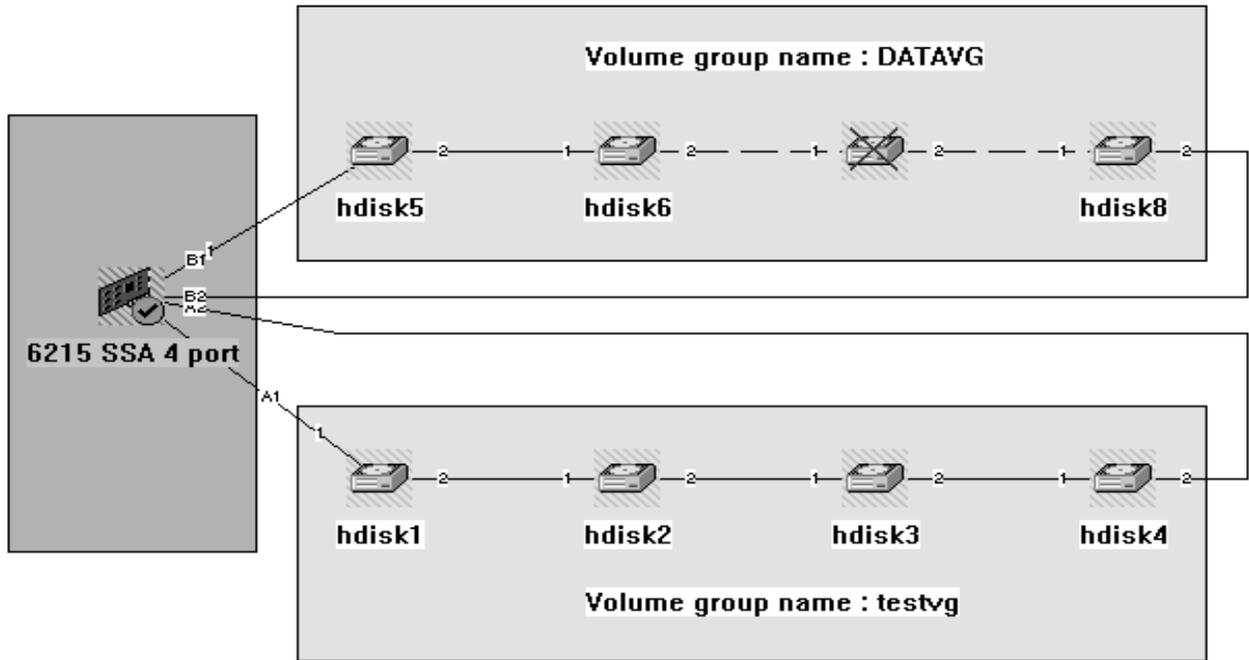


Figure 45. StorX Live View of an hdisk Problem

You can see that StorX detected the problem immediately, and indicates it by a cross on the defective hdisk (hdisk 7 and its attached pdisk 6), and also by a flag to the SSA adapter. This flag means that a change has occurred in this configuration (refer to Table 13 on page 57 for explanations of device states).

5.6.1 Recovery Procedure

In this section, we see what actions are required and the StorX view as the problem is resolved. Figure 46 on page 86 shows us the StorX view after the problem disk drive is replaced with a new one.

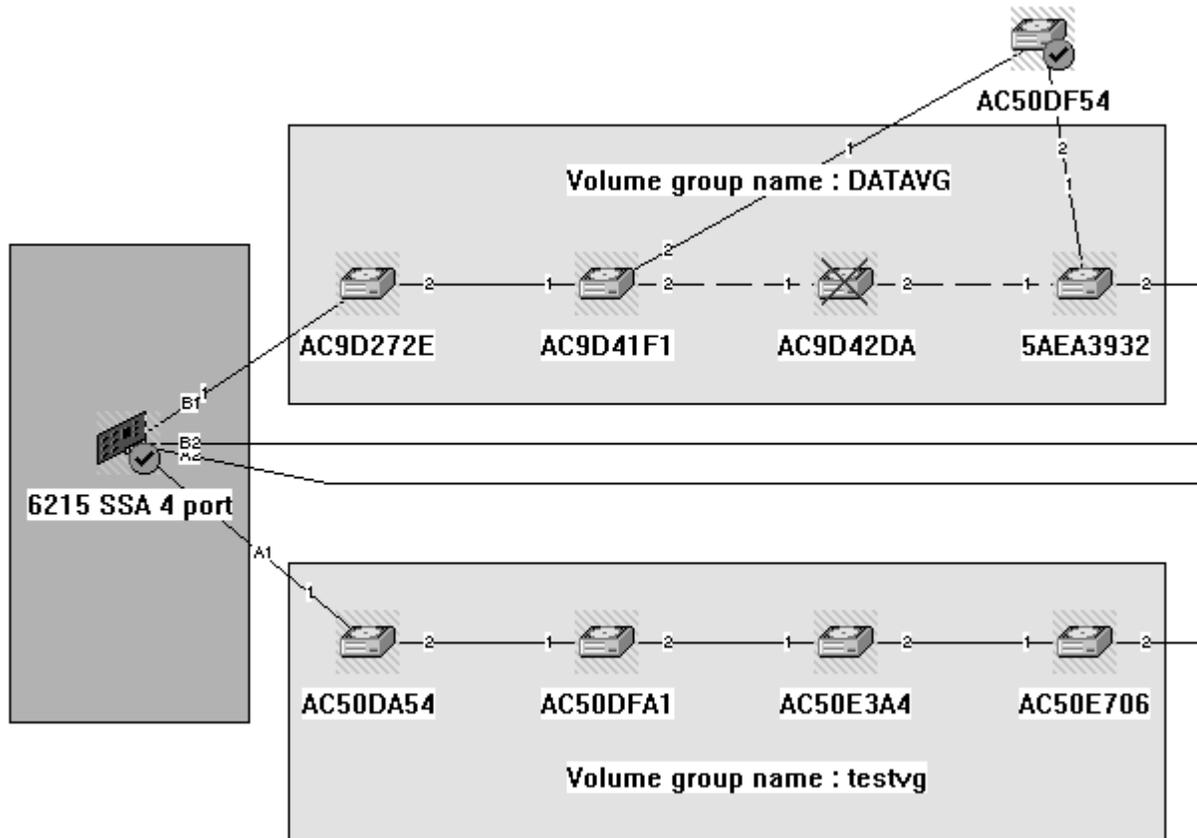


Figure 46. Hdisk Replacement StorX Live View

We see that the bad disk still shows its cross, indicating that it is out of order. There are flags on the SSA adapter and on the new SSA disk to notify us that a change has occurred in our configuration.

To complete the hardware replacement procedure, we must inform the operating system. We therefore have to type some AIX commands. The sequence is:

1. **reducevg datavg hdisk7** - to remove hdisk 7 from DATAVG
2. **rmdev -l pdisk6 -d** and **rmdev -l hdisk7 -d** - to remove pdisk 6 and hdisk 7 of the ODM.
3. **cfgmgr** - to configure the new disk
4. Select **Clear change** on the StorX management tool to redisplay the live viewer

Figure 47 on page 87 shows us the result of the **clear change** command.

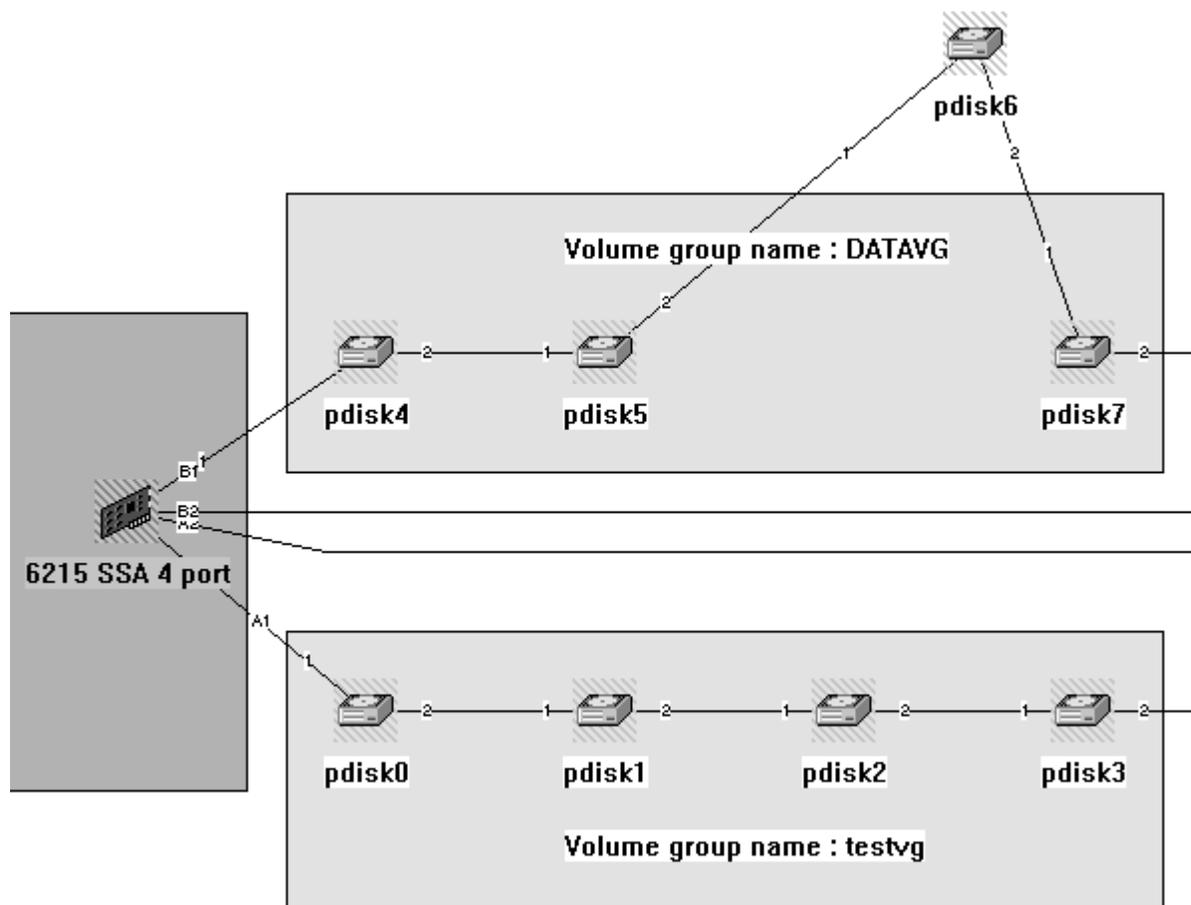


Figure 47. Configuration of the New SSA Disk.-Pdisk Name Option

The flags have been removed, indicating that the hardware procedure is complete.

Hdisk 7 and its pdisk attached (pdisk 6) has been replaced and recognized by the system. We must now complete the software procedure, by integrating hdisk 7 in the DATAVG volume group. This is done by typing the **extendvg datavg hdisk7** AIX command:

The software procedure is now complete. To clean up, we have to drag and drop external hdisk 7 in DATAVG volume group back in line with the other disk drives. Figure 48 on page 88 shows us the final display.

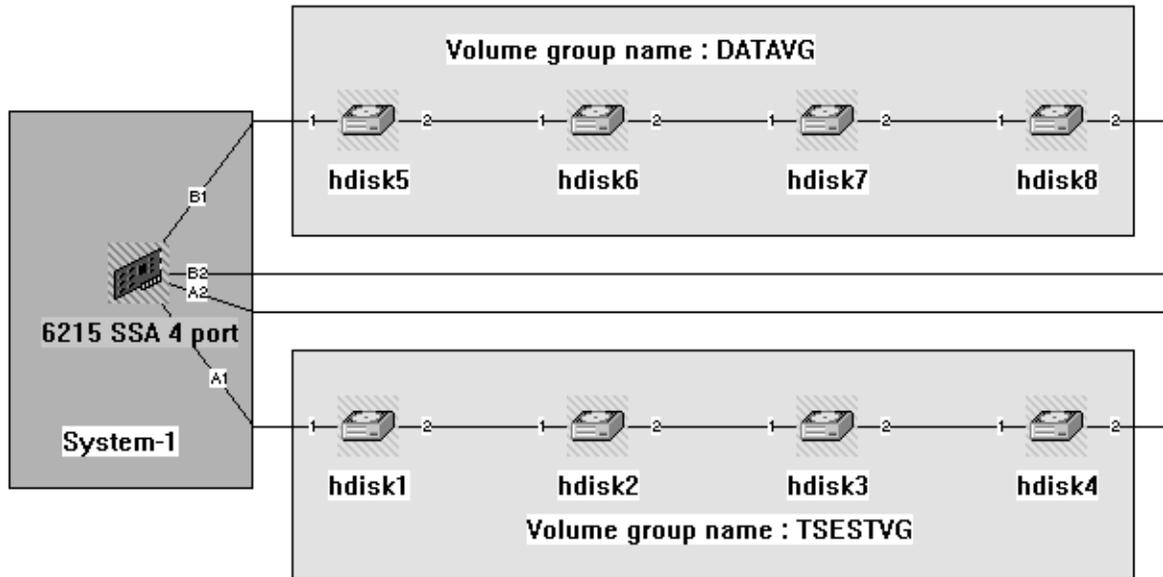


Figure 48. Recovery Procedure Completed

5.7 Monitoring and Managing an SSA Open Loop Problem

5.7.1 Initial Configuration

In this section, we show what to do in the event of an SSA open loop. Figure 49 shows the initial configuration of two SSA loops connected to one SSA adapter.

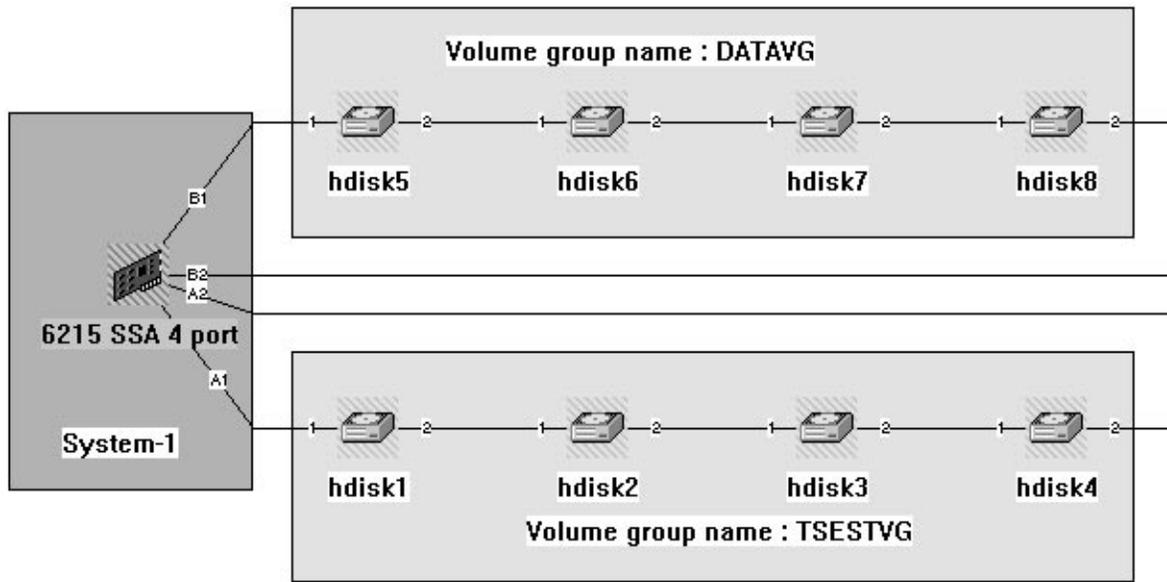


Figure 49. Initial Configuration of Two SSA Loops Connected to One SSA Adapter

Figure 50 on page 89 shows the results of an SSA open loop between Ports B1 and B2.

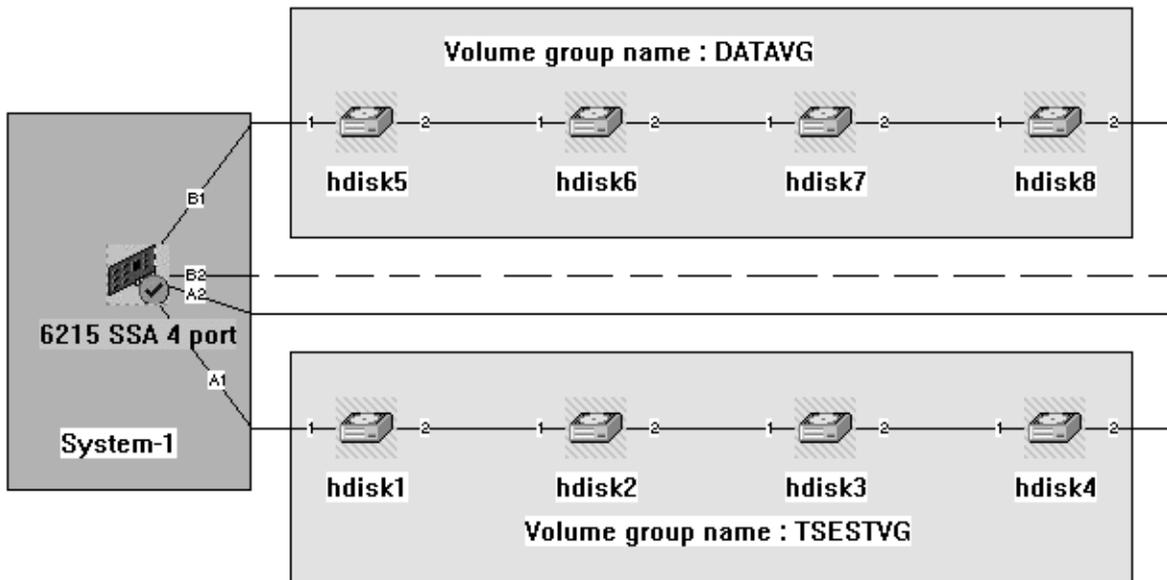


Figure 50. StorX Live View of an SSA Open Loop

We see that StorX detects the problem, and indicates the failing connection port, or cable. There is also a flag on the adapter to notify us that a change has occurred in the configuration.

5.7.2 Recovery Procedure

After replacing the defective cable and issuing the **clear changes** command on StorX, the replacement procedure is complete. We are then back to the initial configuration, which is as shown in Figure 49 on page 89.

5.8 Monitoring and Managing an SSA Adapter Failure

Because of the flexibility of SSA, and the configuration options available, an SSA adapter failure can be explained through different scenarios. In this section, we show what happens when a failure occurs for a single adapter loop and also what happens for an adapter failure when multiple adapters are on the same loop.

5.8.1 Two SSA Loops Connected to One SSA Adapter Card.

Figure 51 on page 90 shows our first configuration of one adapter with two loops.

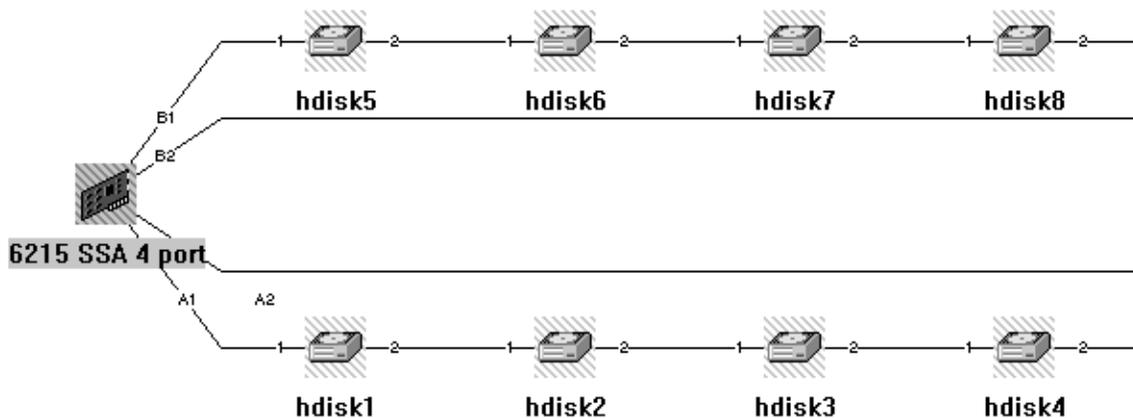


Figure 51. Initial Configuration with Two Loops - StorX Live View

An adapter failure was simulated, and Figure 52 shows the resultant StorX display.

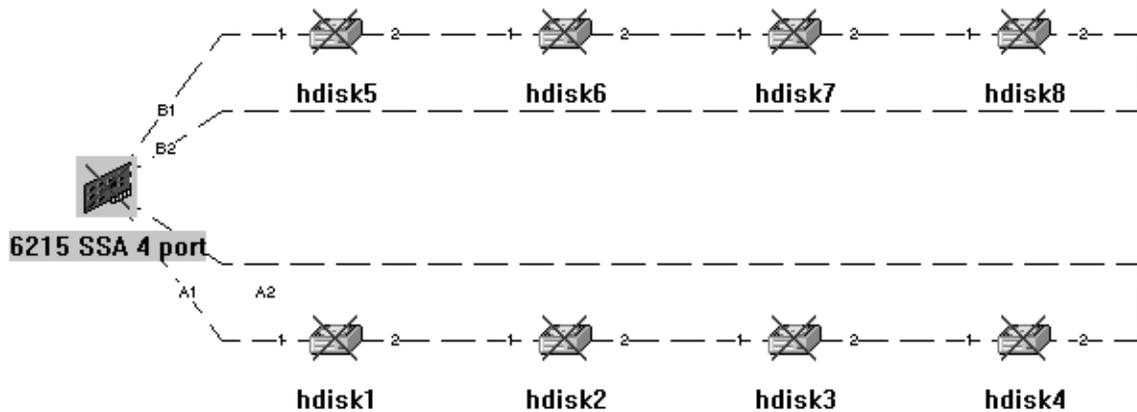


Figure 52. SSA Adapter Problem Affecting Two Loops - StorX Live View

As you can see, all the devices have a cross on them to indicate that all this configuration is completely out of order. Access to data has been lost. The resolution of this problem is to replace the broken adapter, or earplug the SSA loop into a spare SSA adapter card in the host.

5.8.2 One SSA Loop Shared by Two Systems

Figure 53 on page 91 shows our second configuration, which has one SSA loop shared by SSA adapters from different hosts.

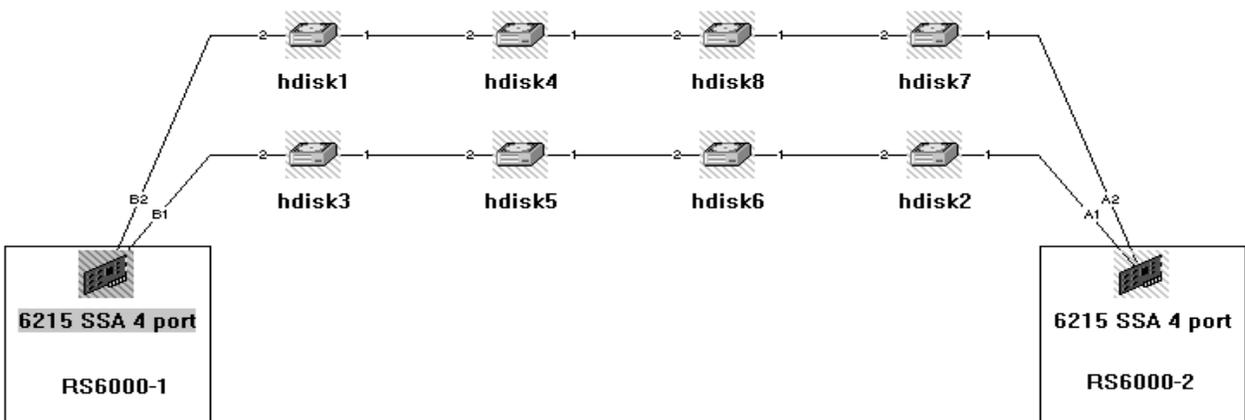


Figure 53. Initial Configuration with One Loop and Two Adapters - StorX Live View

Again, an adapter failure was simulated, and Figure 54 on page 92 shows the resultant StorX display.

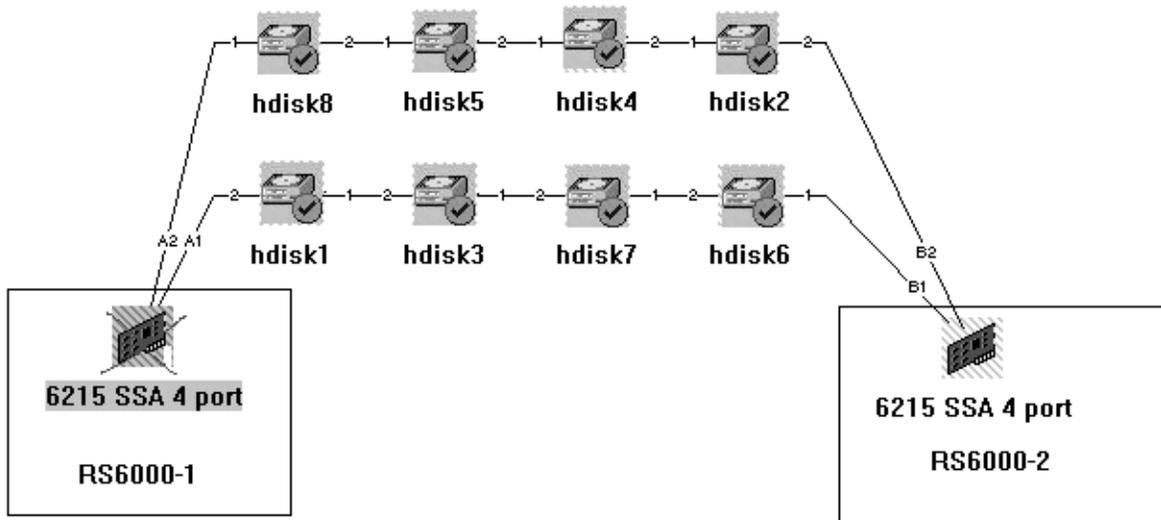


Figure 54. SSA Adapter Problem on System1 - StorX Live View

We see that StorX detects the problem and shows the broken SSA adapter card. It also indicates, by a flag on each hdisk, that changes have occurred in the configuration, but the hdisks are not broken. Data on those drives can still be accessed by RS/6000-2. To resolve this problem, we replace the bad adapter, issue the **cfgmgr** AIX command, and **clear changes** on StorX live view.

5.9 Monitoring and Managing an SSA RAID Array Creation

In this section, we explain the different steps required to create an SSA RAID array, and how StorX manages and monitors this kind of creation. We keep the two loops connected to one SSA adapter configuration shown in Figure 55 on page 93.

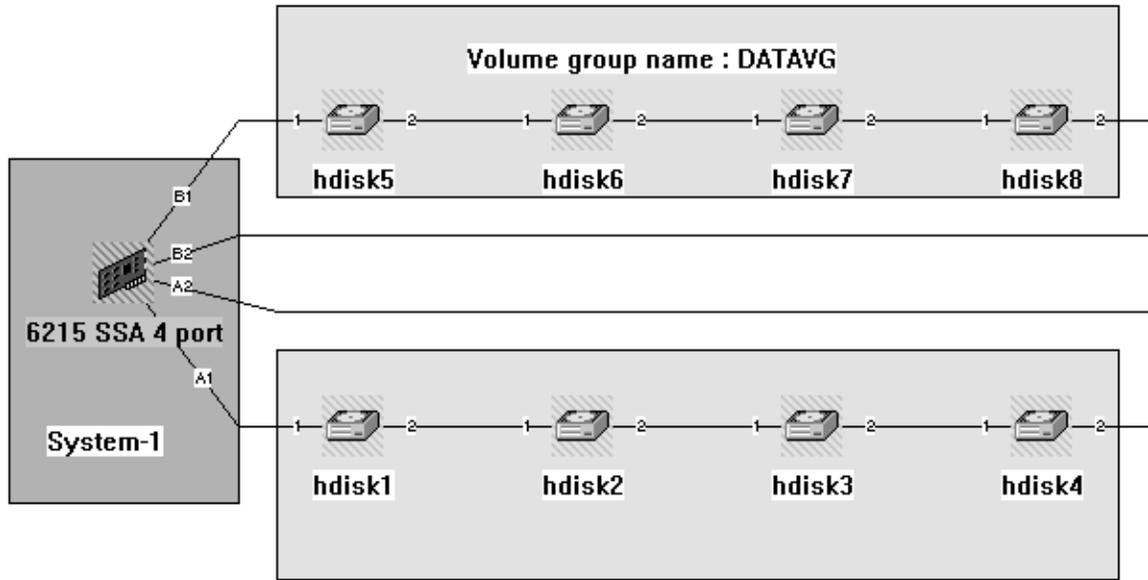


Figure 55. Initial Configuration before SSA RAID Array Creation - StorX Live View

In Figure 55 you can see that hdisk 1, hdisk 2, hdisk 3, and hdisk 4 are not included in any volume group.

For the rest of this section, we use an uncustomized StorX live view.

5.9.1 Step1: Hot Spare Disk Definition

In this step, we discuss the process to define a hot spare disk.

We define the SSA RAID array hot spare disk. The AIX command to define a hot spare disk is:

```
chgssadisk_system_command_to_exec -l'ssa2' -n 'pdisk0' '-a use=spare -u'
```

StorX detects immediately that a change has occurred, and shows it by two device state icons: one on hdisk 1 (hot spare disk), and a second one on the SSA adapter. Figure 56 on page 94 shows these changes.

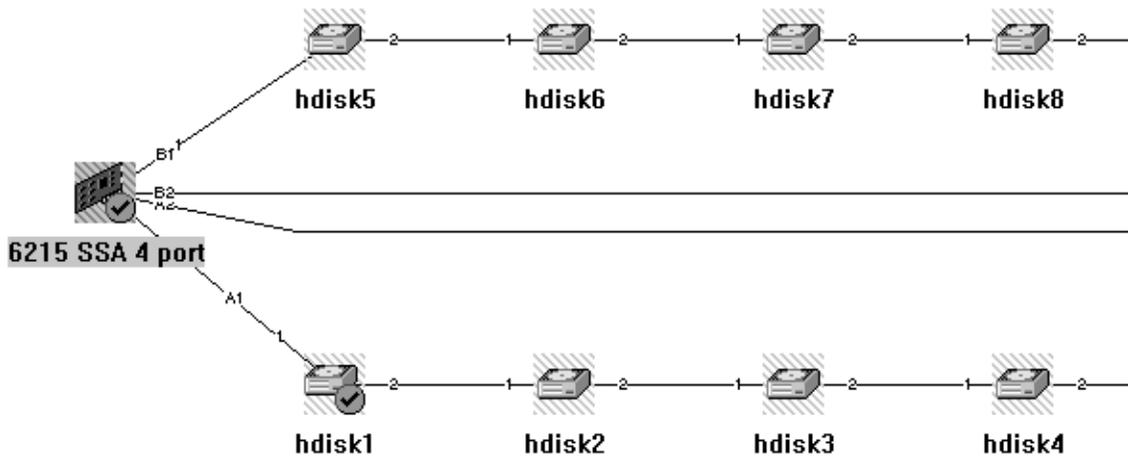


Figure 56. Hot Spare Definition StorX Live View

5.9.2 Step 2: Array Candidate Disk Definitions

The next step is to define the array candidate disks. For this example, hdisk 2 (pdisk 1), hdisk 3 (pdisk 2), and hdisk 4 (pdisk 3) are the array candidate disks.

Again, this task is performed under SMIT. The AIX command to define the array candidate disks is:

```
chgssadisk_hr_cmd_to_exec -l 'ssa2' 'pdisk1 pdisk2 pdisk3' -a use=free -u'
```

Figure 57 shows the StorX changes after the command is executed.

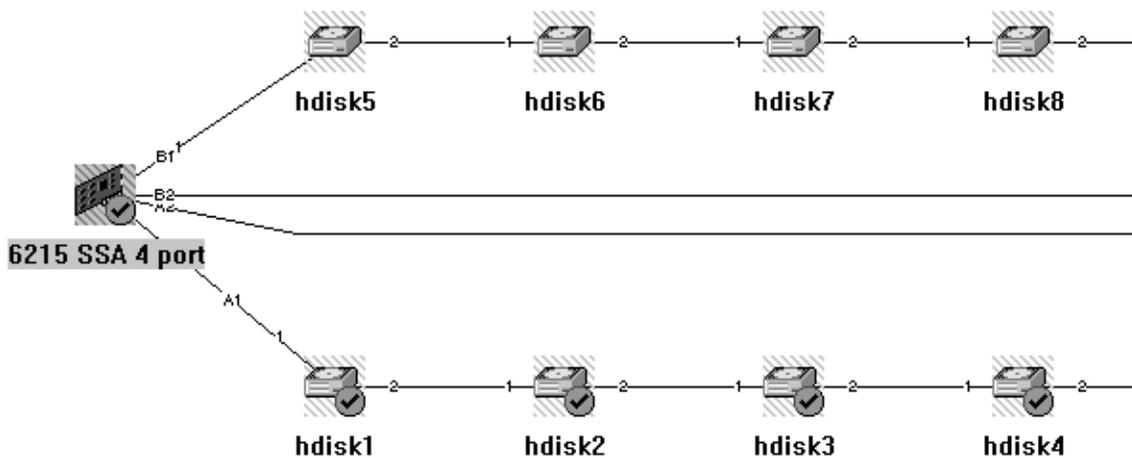


Figure 57. Hot Spare and Array Candidate Disks - StorX Live View

StorX changes the device state icons on hdisk 2, hdisk 3, and hdisk 4 which are the array candidates.

5.9.3 Step 3: SSA RAID Array Creation.

In this step we discuss how to create the SSA RAID Array. This also is done by SMIT. The command is:

```
mkssaraid_raid_5_cmd_to_exec -l 'ssa2' 'raid_5' 'pdisk1 pdisk2 pdisk3' '-a spare=true -a spare_exact=false' -a allow_page_splits=true'
```

Figure 58 shows us the new StorX live view generated by the SSA RAID Array creation.

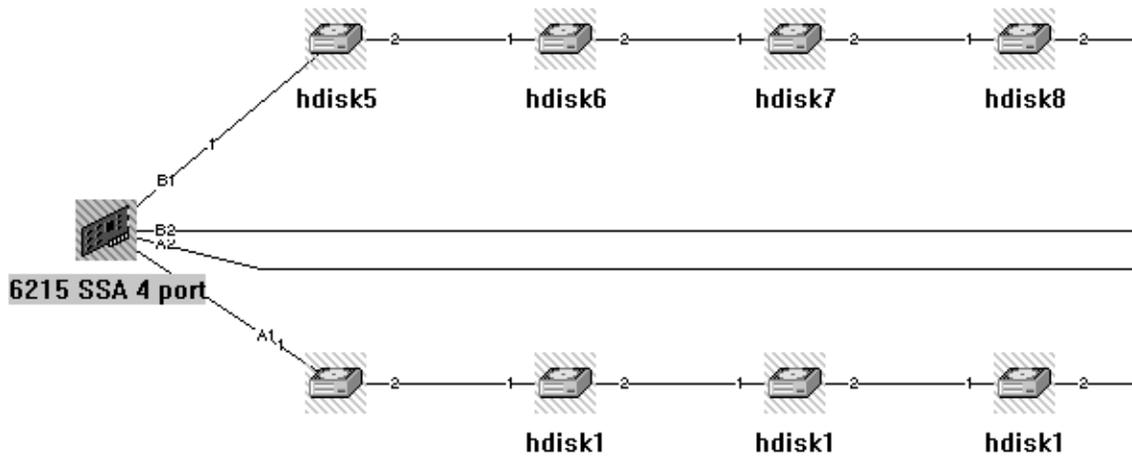


Figure 58. SSA Raid Array - StorX Live View

You can see that the SSA RAID Array contains three disks named hdisk 1, and one hot spare disk (the disk without a name). To facilitate management of the SSA RAID array, use pdisk names, as shown on Figure 59 on page 96.

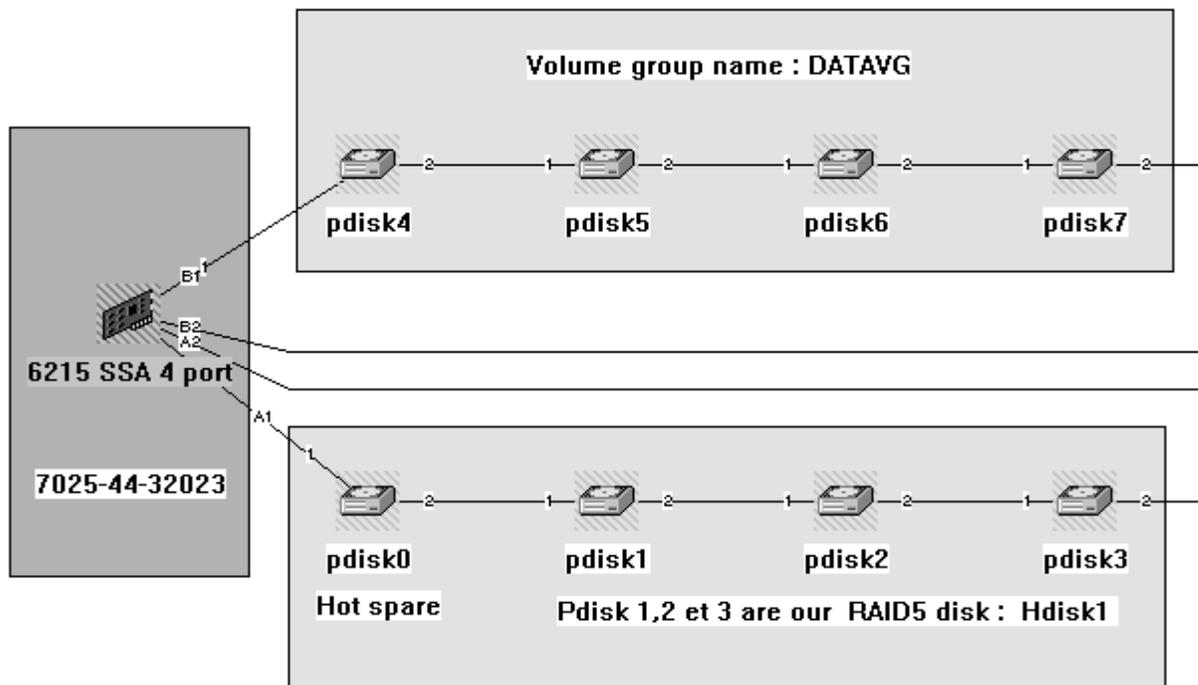


Figure 59. SSA Raid Array, Pdisk Option - StorX Live View

5.9.4 Step 4: RAID5VG Volume Group Creation

After creating an SSA Array DISK, you have to include it in a volume group. We do it by SMIT AIX command: **VGNAME=mkvg -f 'raid5vg' -s '32' hdisk1**. We have named this volume group RAID5VG, as shown in Figure 60 on page 97.

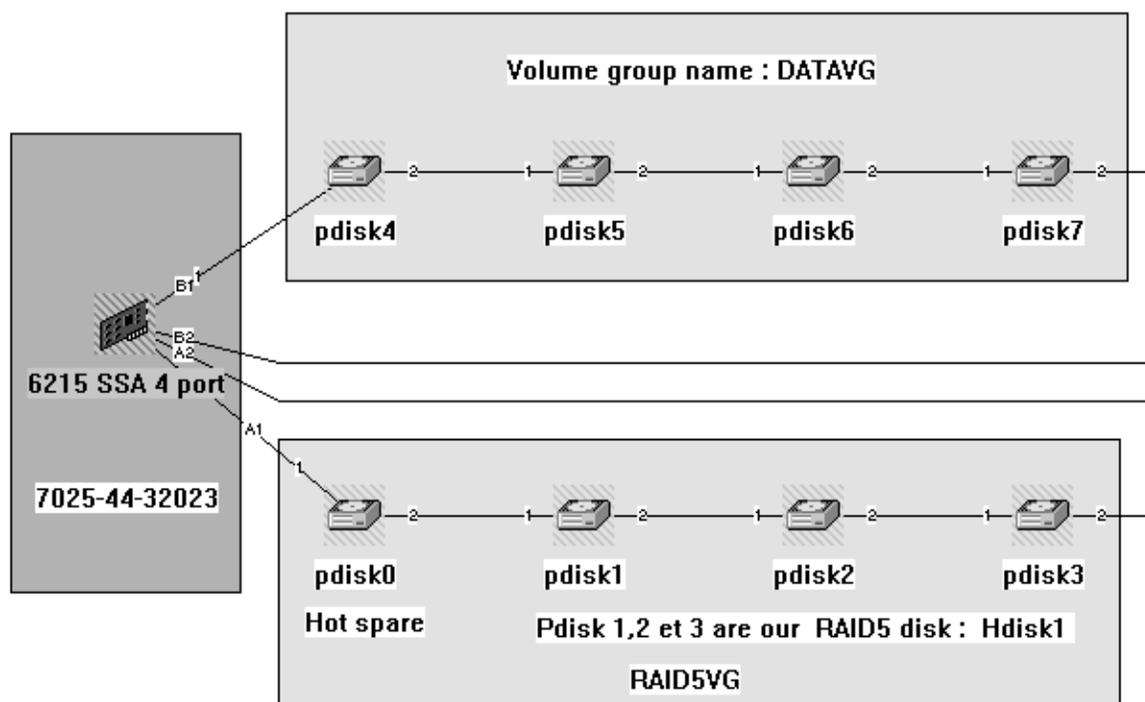


Figure 60. Volume Group RAID5VG - StorX Live View

5.10 Monitoring and Managing Disk Problems in a RAID 5 Configuration

In this section, we discuss how to handle SSA disk problems in a RAID 5 configuration. We show you the steps to follow, using StorX, for the configuration just created (Figure 60).

5.10.1 Pdisk 1 Crash

In this step, we show what happens when we lose a disk that was part of the SSA RAID array. We have selected pdisk 1 as our problem device. Figure 61 on page 98 shows the StorX live view of the broken disk.

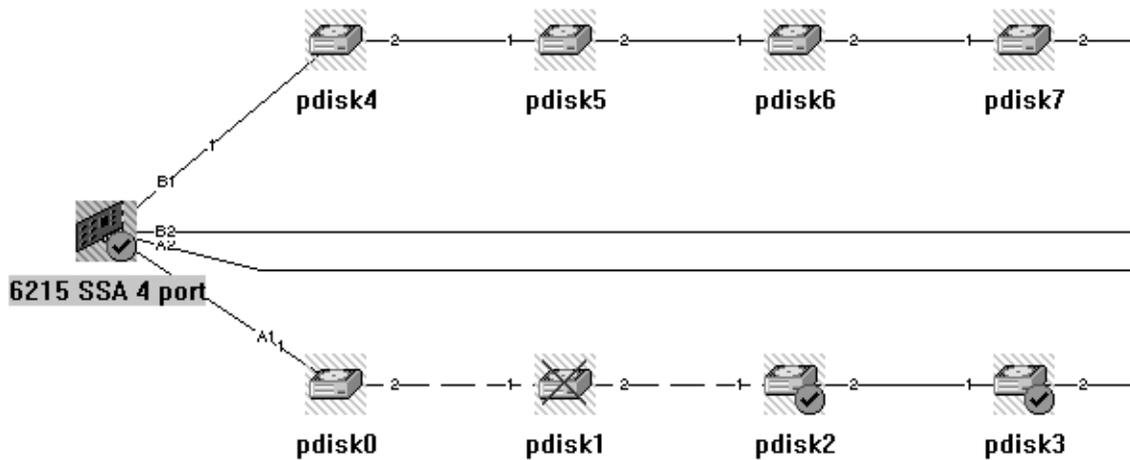


Figure 61. Crash of Pdisk 1 in a SSA RAID Array

You can see that StorX detects the pdisk crash, and shows us by a red cross on pdisk 1. The device state icons on the SSA adapter, and on pdisks 2 and 3, which are included in SSA RAID 5 hdisk 1 are also changed. These flags indicate that a storage change has occurred on this volume group, but data is still accessible.

5.10.2 Recovery Procedure

The first step in the recovery procedure is to replace pdisk 1, and then initialize the new SSA disk by the AIX command: **cfgmgr**. The result of this is shown in Figure 62 on page 99.

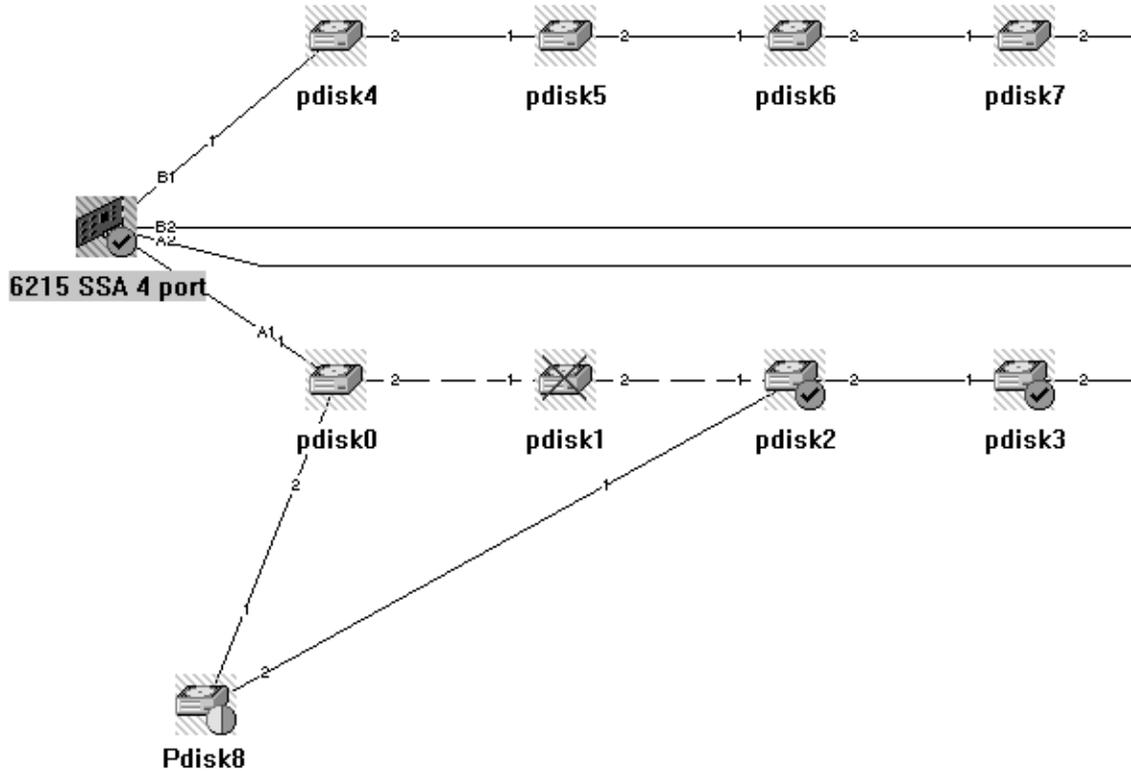


Figure 62. Replacement of Pdisk 1 - StorX Live View

The new disk, pdisk 8, is recognized by the system but not yet defined in the RAID 5 environment. At this time, pdisk 0, which was the hot spare disk, becomes a member of the SSA RAID 5 array, so currently there is no hot spare disk.

5.10.3 New Hot Spare Disk Definition

The second step is to define a new hot spare disk (pdisk 8) for the SSA RAID 5 environment. We do it by SMIT or by the AIX command:

```
chgssadisk_system_command_to_exec -l 'ssa2' -n 'pdisk8' -a use=spare -u'
```

On StorX, we do a **clear change**. The result of these two commands is shown in Figure 63 on page 100.

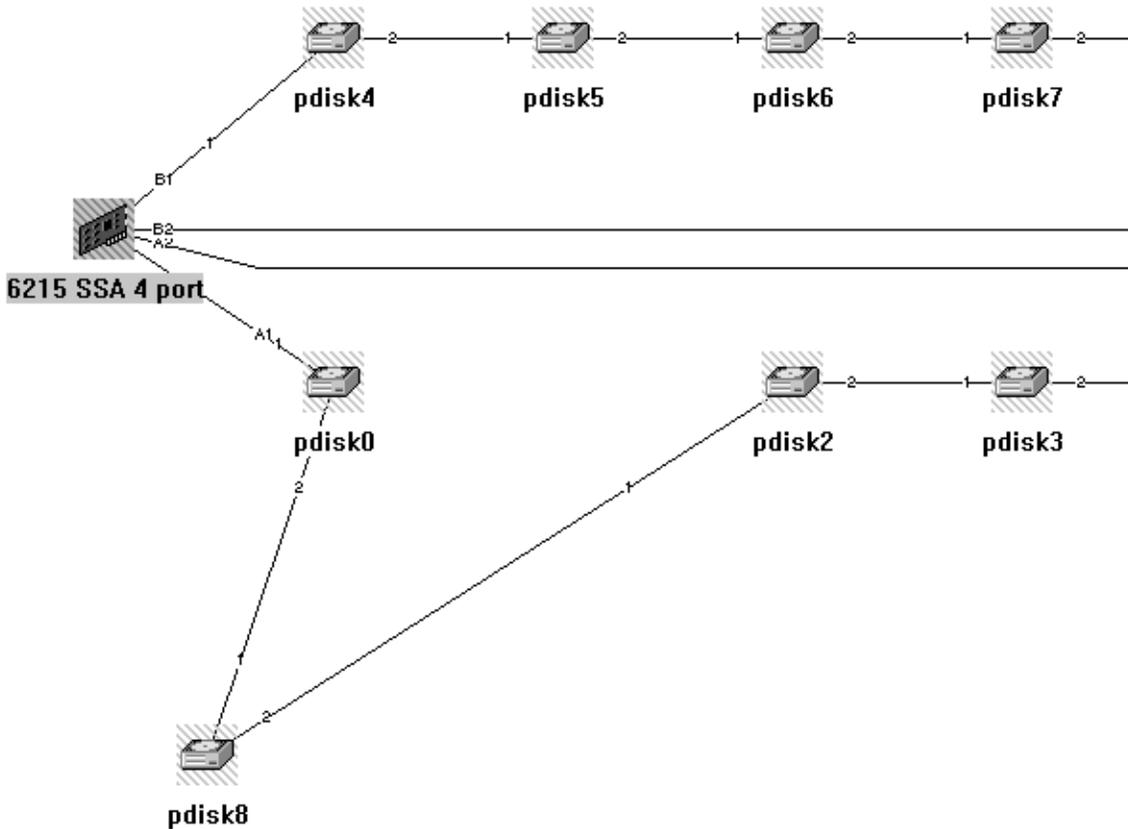


Figure 63. Pdisk 8 Definition - StorX Live View

You can see that there are no flags anywhere, and pdisk 8 has been defined as the hot spare disk.

5.10.4 Complete Recovery Procedure

The recovery procedure is now complete, so all that is left is to drag and drop pdisk 8 to its right place. Figure 64 on page 101 shows us the new SSA RAID 5 environment from the pdisk perspective.

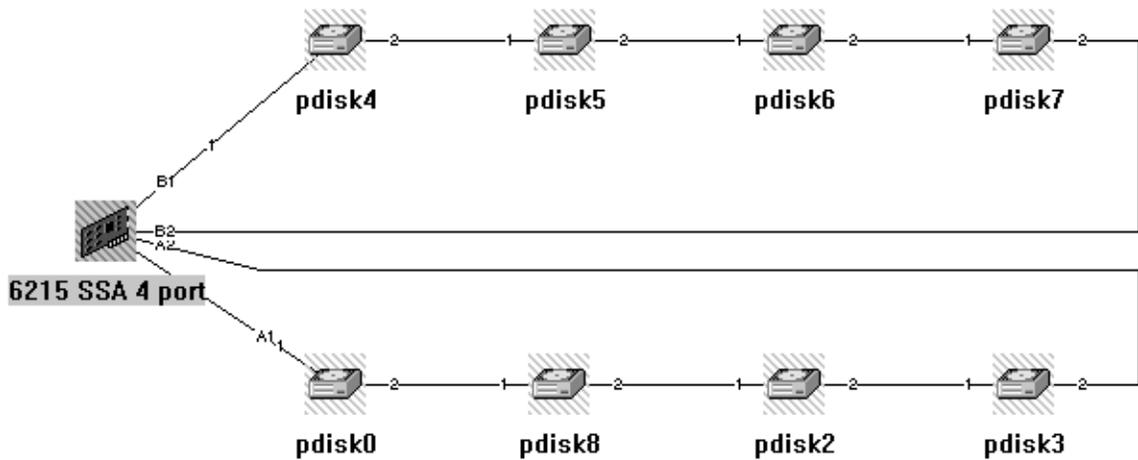


Figure 64. Figure 59. New SSA RAID 5 Environment. - Pdisk Option

You can see that pdisk 8 has replaced pdisk 1. Figure 65 on page 101 shows us the same thing from the hdisk (customized) perspective.

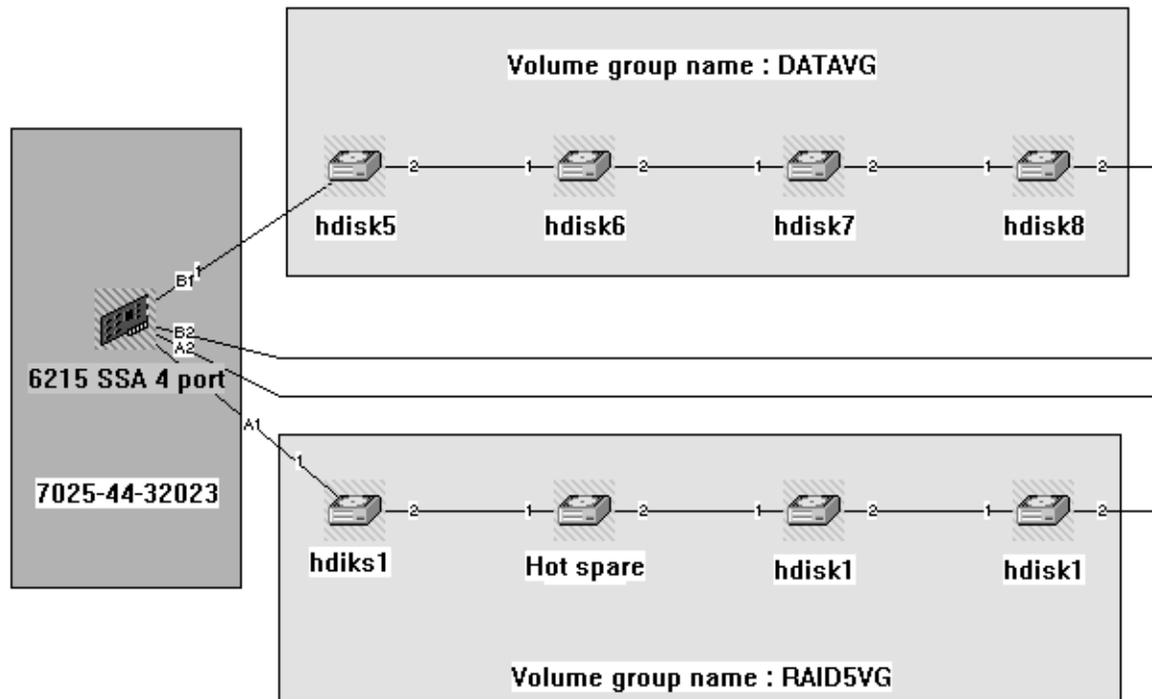


Figure 65. New SSA RAID5 Environment. - Logical Disk Option

You can see that members of RAID5VG have changed, and that a new hot spare disk has been created. You can compare it with Figure 60 on page 97, which represents this configuration before the pdisk 1 crash.

5.11 Monitoring, Managing, and Diagnosing with StorX

In this section, we show you how to use StorX to manage your SSA environment. With StorX, you can run on-line diagnostics, view vital product data (VPD), identify whether a disk is an AIX system disk or a hot spare disk, and determine whether disks are members of an SSA RAID array.

We choose the configuration of two loops connected to one SSA adapter configuration, because it is a good example of the difference in information required when you select the adapter or different disks. Figure 66 shows us the configuration.

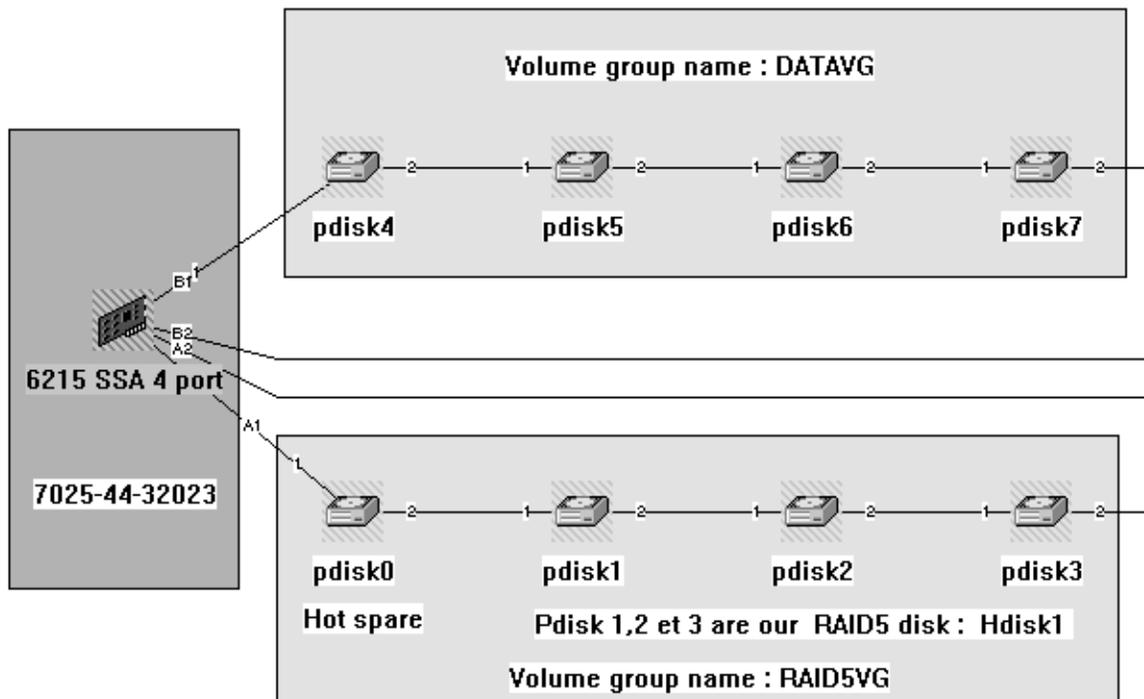


Figure 66. Two Loops Connected to an SSA Adapter with RAID5 Array as one Loop

5.11.1 Disk Properties

To get to the disk properties window (as shown in Figure 67 on page 103), double-click on any pdisk or hdisk, or click once with the right mouse button and select **Properties**. For example, we have double-clicked on pdisk 6. Figure 67 on page 103 through to Figure 71 on page 107 show us the different properties provided by the Disk Properties window.

Figure 62 shows the disk properties of pdisk 6

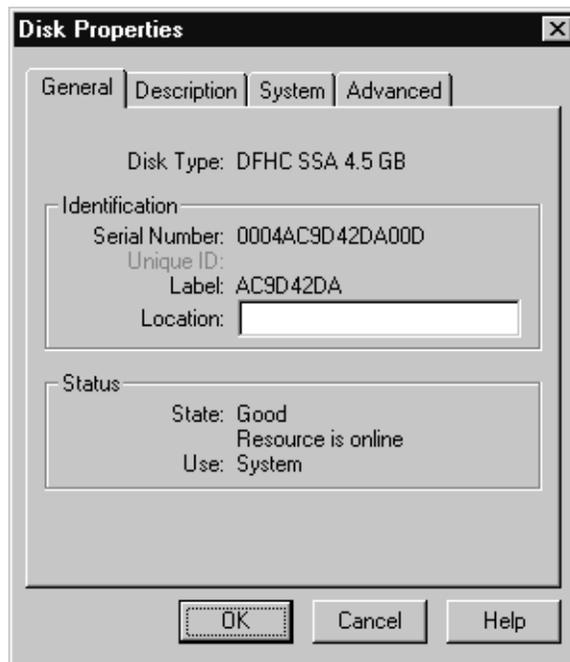


Figure 67. General Page of Disk Properties Window

.There are four different divisions (pages) in the properties window.

- General overview of the selected device
- Description of the selected device
- System information
- Advanced attributes which include identifying and diagnosing the device

We can see that the general page of the disk properties window shows:

- SSA hdisk type
- Serial number
- External label
- Location (you have to describe this location)
- State
- Current use (in our example, pdisk 6 is an AIX system disk)

Figure 68 on page 104 shows us the Description page of the Disk Properties window.



Figure 68. Description Page of Disk Properties Windows

The Description page of the Disk Properties window indicates:

- Disk type
- IBM product name
- Capacity
- Maximum speed

Figure 69 on page 105 shows us the System page of the Disk Properties window.

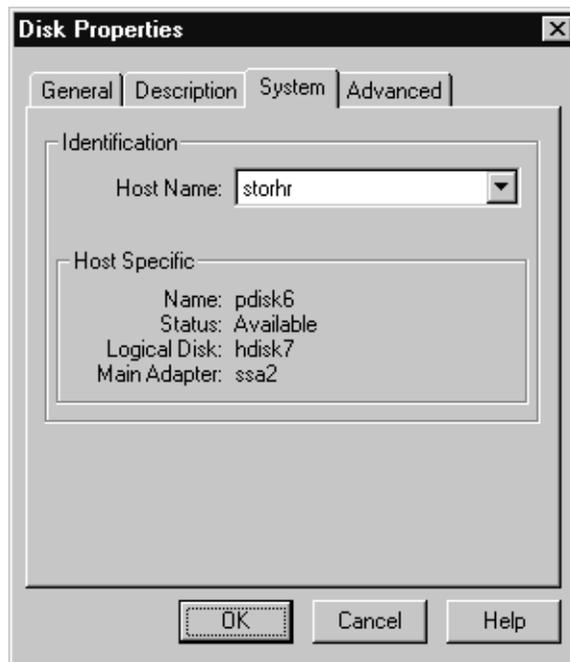


Figure 69. System Page of Disk Properties Window

The System page of the Disk Properties window indicates:

- System name
- Pdisk name
- Logical disk name
- Main SSA adapter that manage hdisk 7

Figure 70 on page 106 shows us the Advance page of the Disk Properties window.

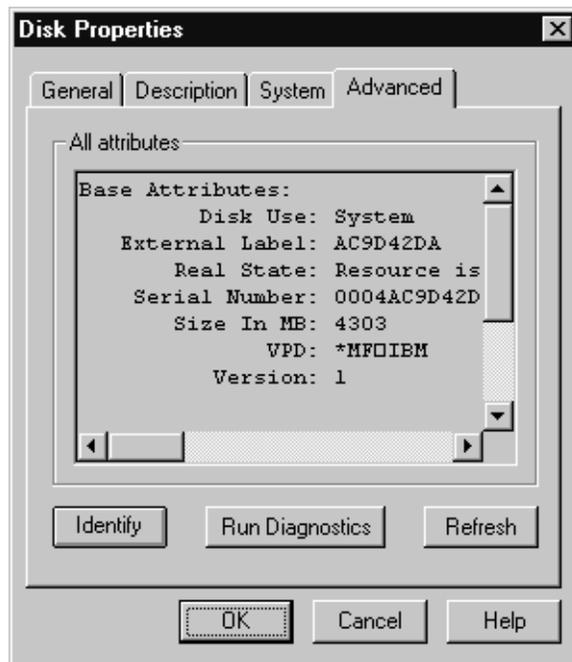


Figure 70. Advanced Page of Disk Properties Window

The Advanced page of the Disk Properties windows shows:

- Disk use
- External label
- State of the hdisk
- Serial number
- Size (of disk)
- VPD: *MFOIBM (part number)
- StorX version

We can also see that the Advanced page of the Disk Properties window permits us to make a visual identification of the SSA disk in question by blinking the orange light on the front panel of the drive.

We also can run concurrent diagnostics on SSA pdisks, and the SSA adapter.

We have seen that we can see which hdisk is used for what purpose. We now want to know if Disk Properties can detect the hot spare disk and the SSA RAID array member disks. Figure 71 on page 107 shows us the general page of the disk properties window for pdisk 0 (hot spare disk).

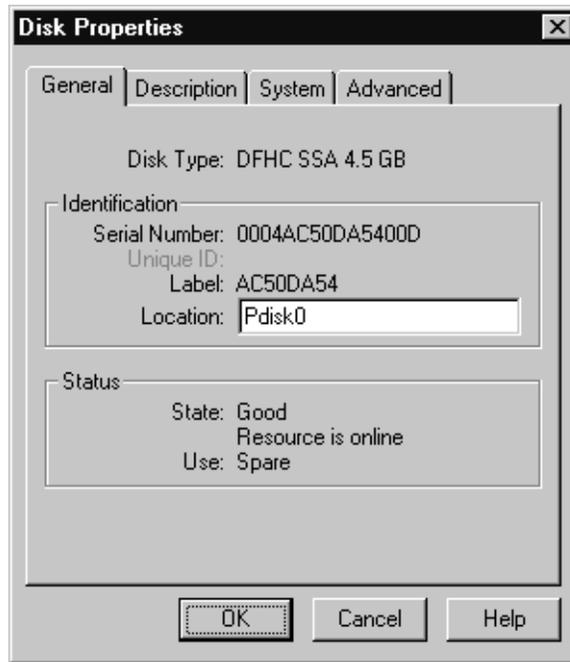


Figure 71. General Page of Disk Properties Window

We can see that the pdisk 0 usage is set to SPARE.

Figure 72 on page 107 shows the general page of the disk properties window of pdisk 3 (SSA RAID array member).

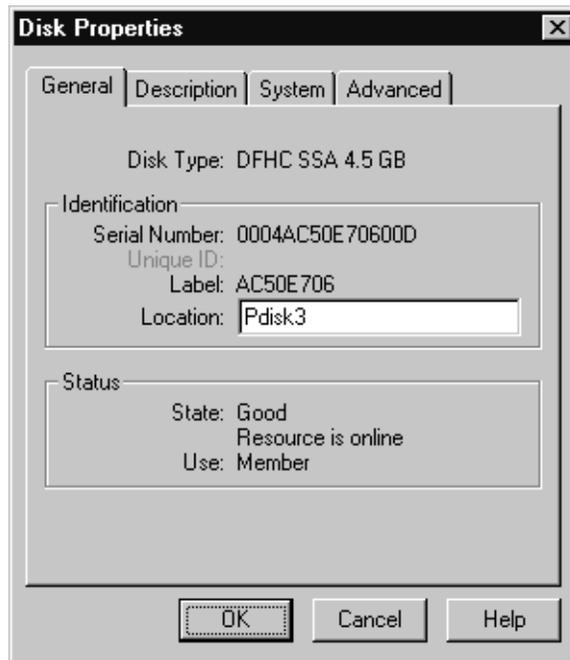


Figure 72. General Page of Disk Properties Windows

This time, we can see that Pdisk 3 is recognized to be an SSA RAID array member disk, which is correct.

5.11.2 Adapter Properties Window

The same type of window is available for the adapter properties, as shown in Figure 73 on page 108.

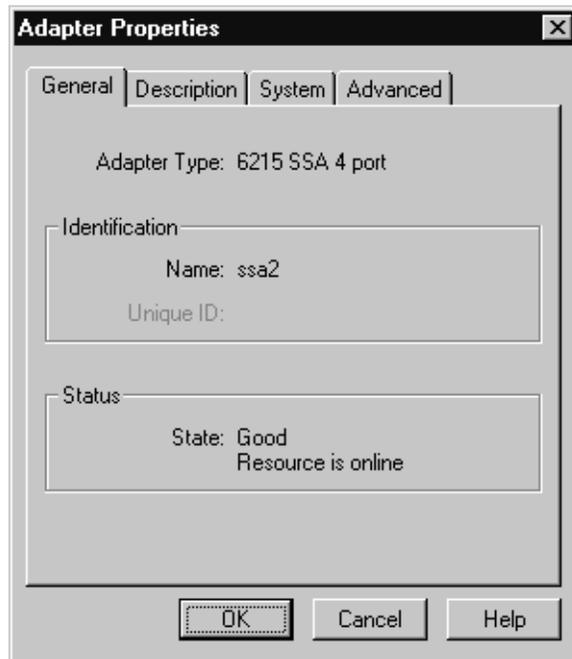


Figure 73. Adapter Properties Window

We do not detail the adapter properties in this section, because they offer the same types of information as the Disk Properties window we have just seen.

Chapter 6. Availability, Mirroring, and Clustering with SSA

In this chapter, we consider the issues of availability, mirroring and clustering in an SSA environment.

6.1 General

Neither the SBus adapter nor the 7190 supports the use of SSA subsystems to boot most operating systems. You must provide a non-SSA boot device that is supported for the specific hardware platform and operating system version you are using. This requirement provides an improved level of fault isolation between system configuration or hardware faults and hardware faults on the SSA storage subsystem. The one exception to this is systems that are running Windows NT. You can install the SSA adapter and device driver on a Windows NT system and then install the Windows NT operating system on the SSA disks.

In many respects most of the clustering and disk mirroring solutions that are available with SSA disks are the same as or similar to, the solutions that are available to normal SCSI disks. However, some basic design features of SSA and the 7133 disk subsystem that make it particularly suitable for use in these environments. These features are described in Part 6.1.2, "SSA Availability Features" on page 111. In this chapter we will consider the platforms running the following operating systems:

- AIX
- Windows NT and other PC operating systems
- Other operating systems

Before discussing operating system specifics, we must define what we mean by the above terms and how SSA technology fits into these environments.

6.1.1 Definitions

Availability

Availability is a measure of the degree to which a system can be used for its intended purposes during the times required by the business.

Service Level

Service level is the goal or target level of availability defined for a system. A service level can be negotiated between the users and managers of a system, or mandated by management for a particular purpose. Service levels are generally justified in terms of cost and resource considerations. Service levels are defined in such terms as level of availability, responsiveness, maximum permissible downtime, or maximum number of system outages over a specified period.

Outage

An outage is any loss of service, planned or unplanned. Unplanned outages are generally caused by defects in, or failure of, system components. Planned outages are generally those scheduled for systems management.

Recovery

Recovery is the process of restoring service after an unplanned outage. Recovery time is a key element of availability.

Storage

For the purposes of this document, we define storage as all persistent forms of storage. This excludes a system's real memory or RAM, but includes disks and tapes.

Backup

Backup is the process of copying selected information to some removable form of storage, so it can be retrieved later in case of a failure.

Levels of Availability

Availability can be visualized as a line where each point along it has an associated cost, and where improvements can be obtained through investment of additional resources or technology. In general, for a given system, the greater the level of availability desired, the greater the costs associated with achieving it. Many terms are used in the computer industry to define levels of availability. For the purposes of this discussion, we define availability at four major levels:

- *Base Availability*

Base availability is the level of availability achieved with a single system and basic systems management practices in place. For many people, this level of availability is sufficient.

- *Improved Availability*

Improved availability systems provide greater robustness through the application of some additional technology or resource to one or more system components. This additional technology provides greater availability at a higher cost than a similar base availability system. Techniques such as disk mirroring, the use of an uninterruptible power supply (UPS), redundant components, data journaling and check summing, hot pluggable disks, and disk sharing can each be used to help overcome certain system failures. For example, a system with mirrored disk subsystems will offer improved availability over one with nonmirrored disks, because it can overcome certain disk failures.

- *High Availability*

High availability systems attempt to provide continuous service within a particular operational window by minimizing the causes of failure and minimizing recovery time when failure occurs. Generally, this requires a high degree of redundancy in system components so that the continued operation of the entire system is protected from the failure of any one component. Providing this level of protection eliminates these single points of failure. The ultimate objective is to eliminate all single points of failure in the system. This can be accomplished by having redundant components or systems, and availability management technology that can automate the transfer of services to those redundant components or systems if a failure occurs. In this environment, it is crucial to ensure that the recovery time from any unplanned outage is minimal. These systems are still likely to require some planned outages for systems management, but these should occur outside the operational window. Recovery times in this scenario should be in the order of tens to hundreds of seconds. If applications are written appropriately, users may not actually see this loss of service as anything other than a longer than average response time.

- *Continuous Availability*

At this level of availability, the system never fails to deliver its service. These systems attempt to provide 100% availability to the end user by providing both redundancy in components and the ability to perform all error recovery and change processes online. In this scenario, planned outages may occur, but they should not be apparent to the end user. These systems are sometimes described as *fault tolerant*.

6.1.2 SSA Availability Features

There are several inherent design features of SSA that make it suitable for use in high-availability systems, regardless of the platform and operating system:

- Loop topology gives two read and two write paths to each disk.
- Reconfiguration into two strings in the event of a loop failure is automatic.
- Supports one or more adapters or SCSI-SSA converters in a loop.
- Spatial reuse is enabled.

These are shown in Figure 74.

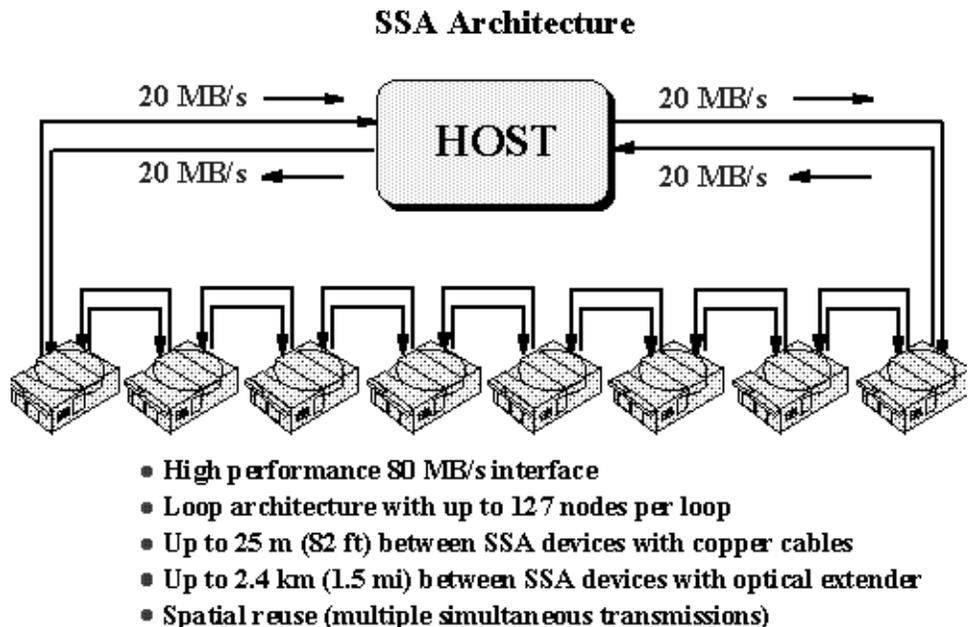


Figure 74. Basic SSA Loop Topology

The IBM 7133 disk subsystem has additional features, which enhance the availability features of SSA disk subsystems. These are:

- Disk hot plug capability
- Redundant power and cooling units within the 7133, which are also hot pluggable

These are shown in Figure 75 on page 112.

7133-020 and 7133-600 SSA Subsystem

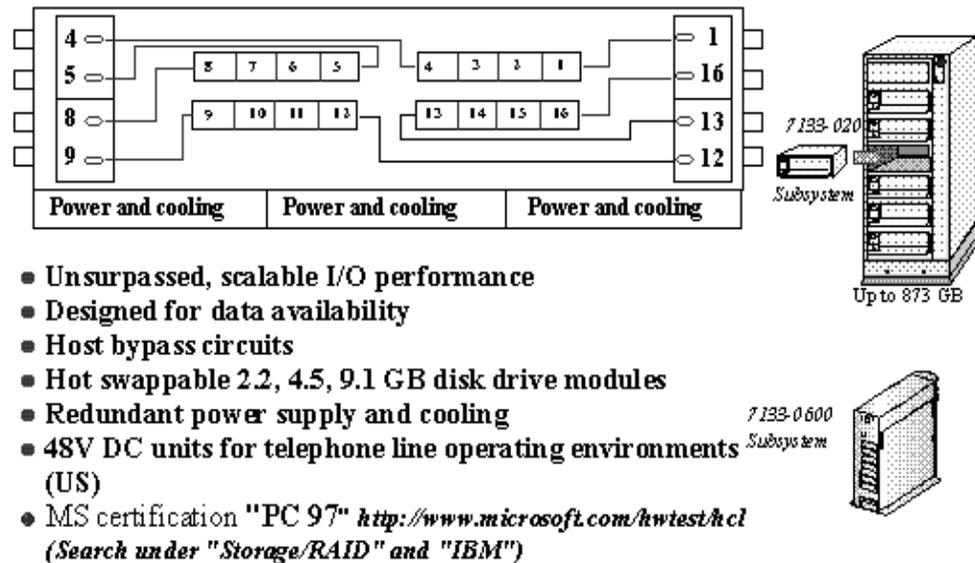


Figure 75. Features of the 7133 SSA Subsystem

6.2 AIX

Several features of AIX and the RS/6000 make them very suitable for high-availability systems.

Availability features of AIX:

- System management interface tool (SMIT)
- Logical volume manager (LVM)
- Journalized file systems (JFS)
- Dynamic AIX kernel
- System resource controller (SRC)
- Configuration manager
- AIX update facilities

Availability Features of the RISC System/6000:

- Built-in error detection and correction
- Backup power supply
- Power conditioning
- Redundant or spare disks
- Hot-pluggable disk drives
- Multitailed disks and shared volume groups

6.2.1 High Availability

6.2.1.1 High Availability for Cluster Multi-Processing for AIX.

High Availability for Cluster Multi-Processing for AIX (HACMP) 4.2.1 provides the RS/6000 platform with the industry-leading high-availability solution for mission-critical applications, according to consultants from D.H. Brown (D.H. Brown Associates report, *UNIX Competitive Assessment*, November 1995).

Other methods of achieving higher availability, such as redundant, fault-tolerant hardware implementations also provide highly available application systems support. However, during normal activity the extra hardware cannot be used to perform application work. HACMP 4.2.1 provides the capability that reconfigures the available replicated resources automatically when hardware failures or outages occur, while allowing for resources to be fully utilized when no hardware malfunctions.

In addition to providing high availability, HACMP 4.2.1 can also be configured to provide loosely coupled multiprocessing services. Such configurations, whether they are "concurrent access" or "partitioned data access" configurations, provide the ability to spread a workload across multiple RS/6000 processors, sharing the disk and CPU resource of the clustered processors. This clustered approach, together with the capability of applications failover and recovery/restart of the HACMP 4.2.1 configured machine, provides additional levels of high-availability processing. HACMP also provides a rich set of system management tools to reduce cluster configuration and administration time. These tools include: GUI-based visual system manager (VSM) icons, cluster single point of control, system management interface tool (SMIT), or command-line interfaces. Enhanced cluster system management is available through an SNMP client, which can be customized to interface with customer applications.

6.2.1.2 High Availability Geographic Clustering

High Availability Geographic Clustering (HAGEO). HAGEO 2.1 provides real-time mirroring of customer data between systems connected by local or point-to-point networks, bringing disaster recovery capability to an RS/6000 cluster placed in two widely separated geographic locations. Data mirroring with HAGEO 2.1 can be synchronous, ensuring real-time data consistency, or asynchronous for maximum performance while maintaining sequential data consistency. Combined with IBM's HACMP, HAGEO 2.1 extends HACMP's loosely coupled cluster technology, providing even greater access to mission-critical data and applications by eliminating the site itself as a single point of failure. HAGEO 2.1 automatically responds to site and communication failures and provides for automatic site takeover. Tools are available for data resynchronization after an outage, configuration management, capacity planning, performance monitoring, and problem determination.

6.2.2 RAID Arrays and Disk Mirroring

6.2.2.1 RAID Arrays

With the IBM RAID adapters, Features 6215, 6217, 6218, and 6219 you have the ability to configure RAID-5 disk arrays as well as JBOD. The adapters all support pools of hot spare disks so that in the event of a disk failing in a RAID 5 array a standby disk is brought automatically online into the array and the array rebuild process starts. The enhanced loop adapters, Features 6215 and 6219, also support a fast write cache to overcome the RAID-5 write penalty.

Adapters 6217 and 6218 are for single host attachment only, that is, there can be only one SSA adapter in the loop. Features 6215 and 6219 currently support two adapters in an SSA loop, or JBOD mode and one adapter in a loop when used with RAID-5 disk arrays in the loop. We support up to eight adapters in an SSA loop (JBOD mode) and up to two adapters in a loop when used with RAID-5 disk arrays. If the fast write cache option is used, then only one adapter in the loop will be supported.

This increases the options available to you to provide highly available disk subsystems.

6.2.2.2 Disk Mirroring

When running AIX, no IBM product allows you to have hardware disk mirroring. All disk mirroring is done in software, using the AIX operating system. AIX allows you to make single- or double-copy disk mirrors. This is all done in software and is easily accomplished using the SMIT panels. The use of double copy mirrors also known as *triple mirroring* (because it results in three copies) is increasing, as breaking off one copy of the data allows an offline backup to be performed while the operational data is still protected.

For even more function and ease of use, IBM's AIX System Backup and Recovery/6000 (Sysback) can be purchased to make creating and restoring your backup copies even easier. Offline mirror backup is an optional feature of Sysback, which allows Mirrored logical volumes or file systems to be taken offline to create a snapshot backup of data without causing any interruption to the users or system. With Sysback, after the backup completes and you want to re-establish your mirrored copy, only physical partitions changed during the backup process will be updated, making the re-creation process much quicker. Sysback allows this process to be performed for selected logical volumes or for an entire mirrored system; it can be customized for completely unattended operation.

6.2.2.3 Remote Mirroring

There is a growing requirement for remote mirroring, which is making another copy (mirror) of the data in a different location from the primary copy. This can be in a separate room, a different part of the site, or a completely separate location. Using SSA disks, various options are available:

- For local copies or copies in close proximity, standard SSA disks and cabling are sufficient. Normal SSA cabling allows you to have up to 25 m (82 ft) between SSA nodes. This is often sufficient, even if the copies are stored in an adjacent room.
- For distances up to 2.4km. (1.5 mi) the fiber optic extender can be used. Feature 5500 of the 7133-020 provides a pair of fiber optic converters. You then have to provide the fiber cable to link the converters. Two sets of Feature 5500 are required for an SSA loop.
- For distances over 2.4km (1.5 mi) and up to 45km (28 mi) there is a solution using the IBM 7190 SCSI to SSA converter and the use of SCSI extenders. The SCSI extender is attached to the host system using a standard SCSI-2 FW cable, up to a maximum of 25 m (82 ft) distance. Two extender units can be installed side by side in a standard 19 inch rack if required. Alternatively, the extender can be placed on the desktop. Connection between the extenders is via a one-fiber pair (RX/TX) for up to 10 MB/s throughput and

via two fiber pairs (RX/TX) for up to 20 MB/s throughput. The extender is completely transparent to the host system, and the extended disk storage can be managed as if the 7190 or Vicom were directly attached to the host through the SCSI cable. This means, for example, that a remote host can be attached to the SSA loop as well. When the extenders are connected together and powered on, they establish communication between themselves and will then continually monitor the status of the link. Should the link break or one of the extenders fail, the host system will report the fault as a SCSI bus error. The extenders report the error via the built-in LCD panel and can support remote dial-in for microcode update or error log analysis.

Replacing a Failed Disk

If a disk should fail, two major points have to be considered: First, what has to be done to the SSA network to enable a replacement, and second, what are the operating system considerations. Since SSA was designed with disk hot plugging capability, nothing has to be done from an SSA viewpoint other than remove and replace the failed drive. This can be done online with minimal effect on the rest of the network. A number of procedures have to be carried out within AIX to enable the new drive to be recognized and become part of the mirrored pair. However, all of these steps can be carried out with the system online and with the users active.

The basic steps to replacing a disk drive are these:

1. Deallocate all the physical partitions associated with the physical volume identifier (PVID) of that drive.
2. Remove the PVID from the volume group.
3. Remove the definition for the disk from the device configuration database.
4. Change the physical disk drive
5. Add the disk to the volume group
6. Recreate any logical volumes, paging spaces, file systems or LV mirrors
7. Ensure that the new mirror is synchronized with the primary disk.

This procedure is fully documented in Appendix A, "How to Change an AIX Mirrored Disk" on page 153.

This procedure is also shown for StorX in Chapter 5.10, "Monitoring and Managing Disk Problems in a RAID 5 Configuration" on page 97.

Automated Procedures

There is a process in development that will automate the replacement of a failed disk drive. It is based on the concept of having a pool of hot spare disks. If a disk should fail then the failure is detected and the hot spare is brought online and synchronized with the primary copy of the data. When the service engineer replaces the failed drive, he or she will run a script so that the new replacement drive is brought online and the hot spare drive is returned to the pool. This is to keep all the active drives in their original positions to ease disk management. Also the original configuration will usually be designed to obviate any single points of failure in the SSA network. If the physical position of the active drives changes then single points of failure may be introduced into the system. The Full description of the process is shown in Appendix C, "Replacing a Mirrored Disk - Documentation" on page 161.

6.2.2.4 Vicom UltraLink Series 2000

Connectivity between parallel SCSI and the high-performance SSA technology is provided by the UltraLink controller. This allows you to attach SSA disks to devices and systems that do not have native connectivity. As the UltraLink makes SSA devices appear to be parallel SCSI devices, its usage and installation is very simple. The UltraLink 2000 interfaces to the host through the Host's SCSI bus. Since the host's OS is independent of the SCSI protocol, installation of the UltraLink requires no modification to host system software or addition of host system drivers.

The UltraLink 2000 devices support multihost attachment to the SSA loop. Multihost attachment gives heterogeneous host systems concurrent access to a single SSA loop. This feature provides the serial loop interface controller (SLIC) architecture with additional benefits, such as storage sharing, host failover, and parallel servers. Storage can now be allocated to each host on an as-needed basis, regardless of the host's manufacturer.

The UltraLink 2000 can configure and manage JBOD, RAID-0, RAID-1 and RAID-0+1 disk arrays. The UltraLink supports a pool of hot spare disks; in the event of a disk failure, online disk replacement and rebuild is automatic.

6.3 Windows NT and Other PC Operating Systems

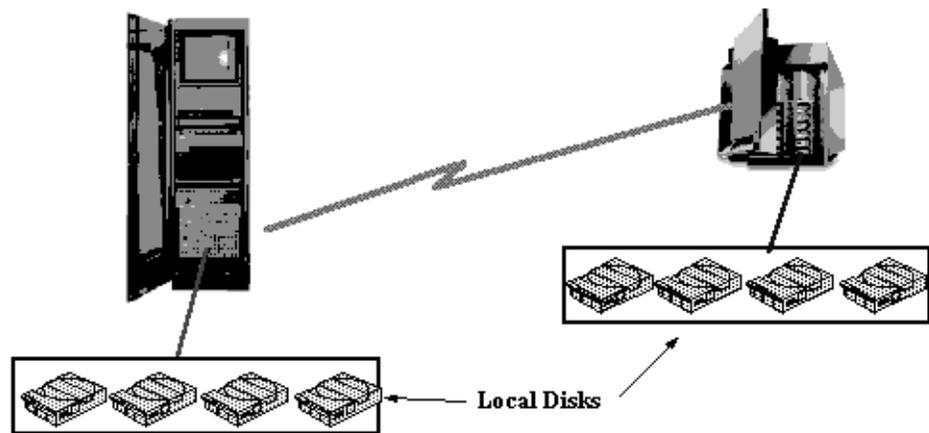
6.3.1 Standard Systems

Two IBM SSA adapters are available for use with Windows NT and other PC operating systems. There is one adapter for standard PC systems. The supported systems are

- IBM PC Server 325, 330, 520, 704, 720, Netfinity 7000
- COMPAQ Proliant 1500, 2500, 5000, 6000 servers
- HP LX PRO and LX PRO servers

This adapter supports clusters using software such as Vinca and Qualix Octopus. These clustering technologies are similar in concept in that two complete systems are mirrored. One system acts as the duty system and the clustering software replicates all the transactions so that the standby system is kept in synchronization with it. In the event of a failure within the duty system, then the standby system takes over; see Figure 76 on page 117.

Qualix/Vinca Clustering Solutions



- **Qualix Octopus** <http://www.qualix.com>
- **Vinca Standby Server** <http://www.vinca.com>
 - ▶ **IBM markets a Vinca solution**
- **PC SSA has been used with both**

Figure 76. Octopus and Vinca PC Clusters

The IBM SSA RAID adapter for PC servers, is a high performance, high connectivity, enhanced availability SSA adapter. This multitype RAID adapter can attach up to 96 SSA disk drives to PC servers that have a PCI bus. Four serial ports forming two SSA loops can be configured in up to 32 arrays; each array can be RAID-0, -1, -5, or non-RAID. Any combination of supported IBM SSA subsystems can be attached, provided that no more than 48 disk drives are configured in one loop. A RAID array can span the two loops that each adapter can support. The supported operating systems are:

- Windows NT 3.51 and 4.0; OS/2 Warp server, OS/2 SMP, OS2 Warp Server SMP and Netware 4.1 and Netware 4.1 SMP support on IBM PC server
- Windows NT 4.0 and Netware 4.1 (and Netware 4.1 SMP support for Compaq and HP)
- MS certification "PC 97"

This adapter is PC part number 32H3811 or it can be ordered as a feature of the IBM 7133 (Feature 4011), see Figure 77 on page 118.

SSA RAID Adapter For PC Servers

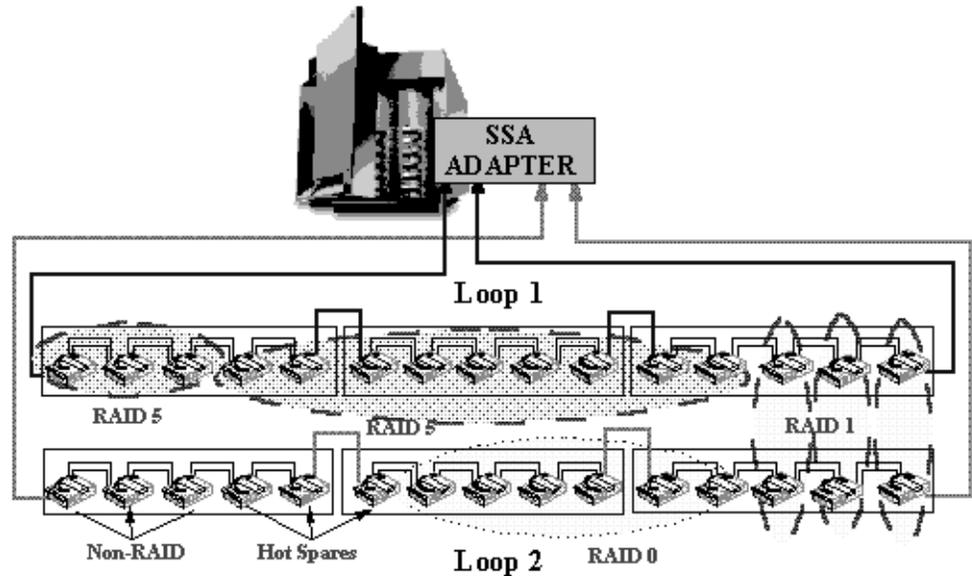


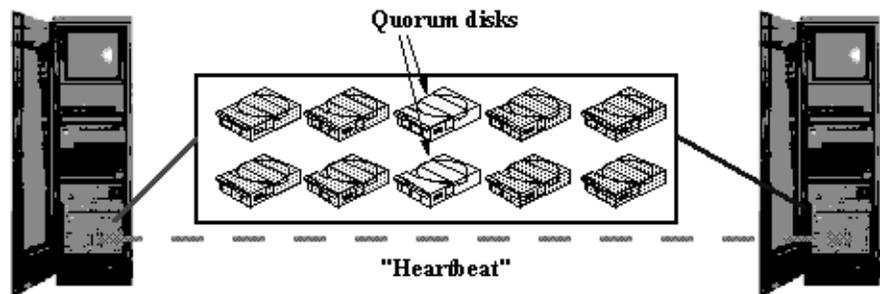
Figure 77. SSA Cluster Adapter for PC Servers

6.3.2 PC Clustering

In addition to the basic functions of SSA, a pair of SSA RAID cluster adapters provides data protection and host failover when used with Windows NT 4.0 Server, Enterprise Edition, which includes Microsoft cluster server software. The cluster adapter supports RAID-1 disk mirroring only. (It does also support the use of hot standby disks.)

The SSA RAID cluster adapter enables two-way connection of SSA disk storage between a pair of supported PC servers. When used with Windows NT 4.0 Server, Enterprise Edition, both servers are operable. If one of the servers fails, the remaining server takes on the applications of the other to ensure continued availability. When a Windows NT system is used in cluster server mode it is not possible to use the Windows NT disk concatenation feature. This means that only 22 disk drives or RAID-1 arrays can be used in this environment. This limits the capacity of the total amount of usable disk that can be attached to the cluster to 200 GB. See Figure 78 and Figure 79 on page 119.

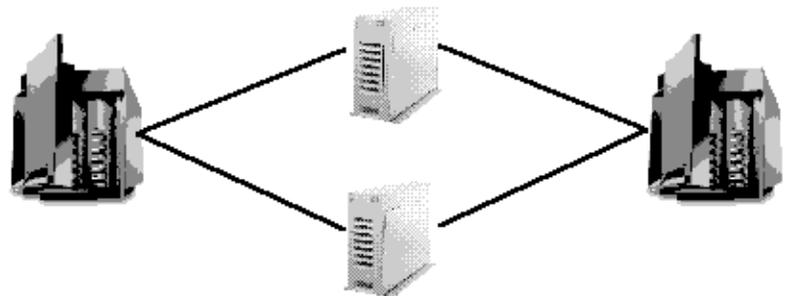
Microsoft Cluster Server Configuration



- MSCS and Windows NT can address failures (by failover) such as: server connections; server hardware (for example, CPU or memory); OS or application failure; network hub failure using redundant network connections
- MSCS is *not* designed to act as backup software and can't be used to protect your data from all sorts of problems.
- Quorum disks store information about services running on the CPUs. They are used by MSCS for failover recovery.

Figure 78. Microsoft Cluster Server Configuration: Part 1

IBM SSA RAID Cluster Adapter 96H9835



- Failover support by Windows NT 4.0 with Microsoft Cluster Server
- 22 RAID-1 arrays for each adapter pair with universal hot sparing
- Data transfer rates up to 60 MB/s
- IBM 3527 (5 disk) and IBM 7133 subsystems supported
- MS certification "Cluster" with IBM PC Server 325, 330, 704 and Netfinity 7000
 ▶ <http://www.microsoft.com/hwtest/hcl> (Search under "Cluster" and "IBM")
- See Redbook SG24-4858 "Clustering and High Availability Guide for IBM Netfinity and IBM PC Servers"

Figure 79. Microsoft Cluster Server: Part 2

6.4 Other Operating Systems

With the exception of systems using the Sun SBus to attach SSA disks to other systems, you have to use a SCSI to SSA converter. The IBM 7190 is supported on the following systems:

- **Sun Systems**
 - SPARCstation 10 and 20 running Solaris 2.4, 2.5, and 2.5.1
 - SPARCcenter 2000 running Solaris 2.4, 2.5, and 2.5.1
 - SPARCserver 1000 and 1000E running Solaris 2.4, 2.5, and 2.5.1
 - SPARCcenter 2000E running Solaris 2.4, 2.5, and 2.5.1
 - Ultra Enterprise 2, 150, 3000, 4000, 5000, and 6000 running Solaris 2.5.1
- **HP Systems**
 - HP 9000 Enterprise Servers
 - HP D Class
 - HP K Class
 - HP T Class
 - HP G Class
 - HP E Class
 - HP H Class
 - HP I Class
 - Software - HP-UX 10.01, 10.10, and 10.20.
- **DEC Platforms / OS**
 - Digital Alpha Server 300, 800, 1000, 1000A, 2000, 2100, and 2100A Series.
 - Digital AlphaServer 4000, 4100, 8200, and 8400 series.
 - Adapter KZPSA-BB Fast/Wide Differential SCSI-2
 - Software - Digital Unix 3.2B through 3.2G and 4.0 through 4.0C. Windows NT 3.5.1 through 4.0

As an alternative to the IBM 7190, there are three products from Vicom, the SLIC 1355, the UltraLink 1000, and the UltraLink 2000. These three products are supported on all host systems that implement the SCSI-2 standard. The SLIC 1355 and the UltraLink1000 support the SCSI-2 Fast and Wide Differential interface while the UltraLink2000 additionally supports the UltraSCSI interface. Table 14 details the features of the different converters.

Table 14. IBM 7190 and Vicom Product Details

	IBM 7190-100	IBM 7190-200	SLIC 1355	UltraLink1000	UltraLink2000
Platform tested	Sun, HP, DEC	Sun, HP DEC	UNIX, WindowsNT	UNIX, WindowsNT	UNIX, WindowsNT
HW RAID 1, RAID 0	No	No	No	No	Yes
Multihost support	1 -4 hosts	1 - 4 hosts	1 - 16 hosts	1 - 16 hosts	1 - 16 hosts
Fiber extender support	No	Yes	No	Yes	Yes
19 inch rack mount option	No	Yes	No	Yes	Yes
Burst data rate	20 MB/s	35 MB/s	20 MB/s	40 MB/s	40 MB/s

	IBM 7190-100	IBM 7190-200	SLIC 1355	UltraLink1000	UltraLink2000
Sustained data rate	18 MB/s	29 MB/s	18 MB/s	35 MB/s	35 MB/s
I/O throughput	1900 I/O/s	2600 I/O/s	1900 I/O/s	3000 I/O/s	3000 I/O/s
I/O throughput with 2 UltraLinks	N/A	N/A	N/A	5100 I/O/s	5100 I/O/s
Capacity per SCSI channel	1 -48 drives	1 -48 drives	1 -64 drives	1 -64 drives	1 -64 drives
Capacity per SCSI channel with multiple UltraLinks	N/A	N/A	N/A	1 -120 drives	1- 512 drives

6.4.1 RAID and Disk Mirroring

The UltraLink 2000 supports hardware RAID levels 0, 1, and 0+1. The other SCSI-to-SSA converters and the Sun SBUS adapter do not support the hardware implementation of RAID. They do support software RAID implementation. This is usually accomplished by using facilities within the host operating system or by additional software such as Veritas Logical Volume Manager or DiskSuite.

6.4.2 Clustering and High Availability Support

No specific device support is provided with either the Sun Sbus adapter, the IBM 7190, or the Vicom product set to aid clustering. Experience from customers indicates that commonly used clustering software technologies such as Solstice High Availability and Veritas Storage Replicator for File Systems will work with Vicom products. The general rule is that if it works with SCSI disks, then it will work with Vicom as the host sees the SSA disks as generic SCSI disks attached by means of a SCSI adapter.

Chapter 7. Managing Backup and Recovery

It is critically important to provide recovery information in the event of a disaster or loss of access to data. This data could be the operating system or application data. Different forms of data may require different backup requirements and this should be allowed for.

In this chapter we explain how to back up application and system data using different methods to provide backup and restore services, and data management for your SSA environment.

This can be done by using mirroring, RAID-5, HACMP, a remote backup center, or software products such as ADSTAR Distributed Storage Manager (ADSM), or you can delegate this responsibility to another company, such as IBM Global Services (IGS). We discuss mirroring, RAID-5, and HACMP in Chapter 6, "Availability, Mirroring, and Clustering with SSA" on page 109, so they are not covered in this chapter.

It should be noted here that the backing up of SSA data and its recovery are no different from those for data stored on other forms of disk, such as SCSI. However, the ability to boot from SSA depends on the hardware and software installed. Please refer to Chapter 7.2.1, "Backing Up the AIX Operating System" on page 124 for the necessary requirements for this function

For more information on ADSM, please refer to the redbooks, *ADSM Concepts* (SG24-4877) and *Using ADSM Hierarchical Storage Management* (SG24-4631), and the ADSM Web site, www.storage.ibm.com/adsm.

7.1 Generalities

The backup facilities provided by the operating system enable all information (including both user and operating system data) to be copied, generally to removable media such as tape. In the event of a major disaster, the information can be easily restored.

7.1.1 Basic Backup Strategy

To define your backup strategy, you have to determine which data you want to back up, when and how often you want to back up the data, and the support you want to use for the backups.

7.1.1.1 What Data Must Be Backed Up

The minimum prerequisite for a backup strategy is to save ROOTVG. Then depending on your configuration, you need to define which other data you have to back up. If you have only one volume group, ROOTVG, you need not do a complete system backup everyday. Inside ROOTVG, however, you can back up one or more customer logical volumes more frequently. If you have a larger configuration with more volume groups, you have to back up those other volume groups also, but you may not need to back up the complete volume groups every day.

7.1.1.2 Decide When and How Often the Backups Should Occur

For ROOTVG, there is no strict timing. You need to do a backup every time you change system definitions. This could be when you add some users, add another disk, or modify some file system parameters.

For customer processes and data, there are several options, depending on the nature of the information produced by the business.

For example, sites that have a great deal of static reference information with only a small percentage of day-to-day changes would benefit from an incremental backup process that only backs up data that has changed. Sites processing large amounts of information daily may choose to back up more frequently. The choice of backup frequency should reflect the business information cycle and be strongly tied to the criticality of the information. The decision is simple: how long can the business survive without key information? If the answer is one day, the backups must be scheduled to run at least once a day.

7.1.1.3 Determine the Backup Media to Use for the Backups

Generally the choice of backup medium is disk, tape drive, or optical drive.

Backing up to another disk on your system generally is similar to mirroring on RAID-5 which is described in Chapter 6, "Availability, Mirroring, and Clustering with SSA" on page 109.

Backing up to a tape drive is inexpensive and offers the most capacity. Tape is the most common backup medium today.

Backing up and restoring using an optical drive is generally slower than using tape, and does not have the capacity of tape, but it does have the advantage of being defined as online, so that access to the backup data is instantaneous.

You can also choose to use a remote center and duplicate your backup, or you can leave this responsibility to another company, such as IBM Global Services.

7.2 System Backup and Restore Using Standard Utilities

In this section, we explain how to manage your backup and restore process using standard AIX Version 4. With AIX V4 you can back up the AIX system volume group (ROOTVG) and all the other volume groups such as DATAVG and TESTVG using SMIT. On AIX Version3, you can back up only your ROOTVG, with the SMIT startup command.

7.2.1 Backing Up the AIX Operating System

The ability to boot from SSA is provided by many combinations of RS/6000, software, and installed adapters. We provide here the information necessary to determine if your system offers SSA boot capabilities.

7.2.1.1 Hardware Prerequisites for SSA Boot Support

One of the following is required if you want to boot from an SSA disk drive:

- 6214 SSA four-port Adapter (4-D)
- 6216 SSA Enhanced four-Port Adapter (4-G)
- 6217 SSA RAID four-port Adapter (4-I)

- 6219 Micro-Channel SSA Multi-Initiator/RAID EL Adapter (4-M)
- 6221 SSA Enhanced four-port Adapter (4-G)
- The system cannot be booted from a disk attached to an SSA RAID adapter, if the disk is part of a RAID logical unit number (LUN). Booting from the 6215 or 6218 adapters is not supported.
- An SSA loop that has a disk with the boot logical volume on it can only be attached through a single adapter on a single machine.
- Only certain models of the RS/6000 can boot from SSA disks (see Table 15).

Table 15. RS/6000 Models that Support Booting from SSA disks

Machine Type	Model	6214	6216	6217	6219
7012	G30	Yes	Yes	No	No
	G40	Yes	Yes	No	Yes
7013	591	Yes	Yes	Yes	Yes
	595	Yes	Yes	Yes	Yes
	J30	Yes	Yes	No	No
	J40	Yes	Yes	Yes	Yes
	J50	Yes	Yes	Yes	Yes
7015	R21	Yes	Yes	Yes	Yes
	R30	Yes	Yes	No	No
	R40	Yes	Yes	Yes	Yes
	R50	Yes	Yes	Yes	Yes

7.2.1.2 Software Prerequisites for SSA Boot Support

AIX Version 4.1.4 or higher is required.

If you are mirroring ROOTVG with SSA disks, we recommend that each copy of the boot logical volume be on a separate SSA loop.

7.2.1.3 SMP Firmware Requirements for SSA Boot Support

Two versions of the Micro-Channel SMP firmware provide boot support for SSA. The first version (9.23) is the absolute minimum required. The second version (A923) is recommended, as it addresses some problems encountered with the first version.

They function as follows:

- Version 9.23 (0923) enables you to boot off SSA disks. However, only one SSA adapter from the system can be attached to the loop. If two or more adapters from the same system should be attached to the loop, Version 9.23 does not support the boot capability.
- Version A9.23 (A923) fixes the multiple SSA adapter problem of Version 9.23.

7.2.1.4 Back up and Restoration of AIX System Volume Groups

Your backup system must be bootable, and the best way to make it bootable is through SMIT (for IBM systems).

In this section we go through the different SMIT menus one at a time. The sequence is:

1. Type **SMIT** on your command line and press Enter. This brings up the main SMIT menu
2. Select **System Storage Management** and press Enter.
3. Select **System Backup Manager** and press Enter. This brings up the backup screen
4. Fill in the **Back Up the System** screen as shown in Figure 80 on page 126.
5. Press Enter to start the system backup.

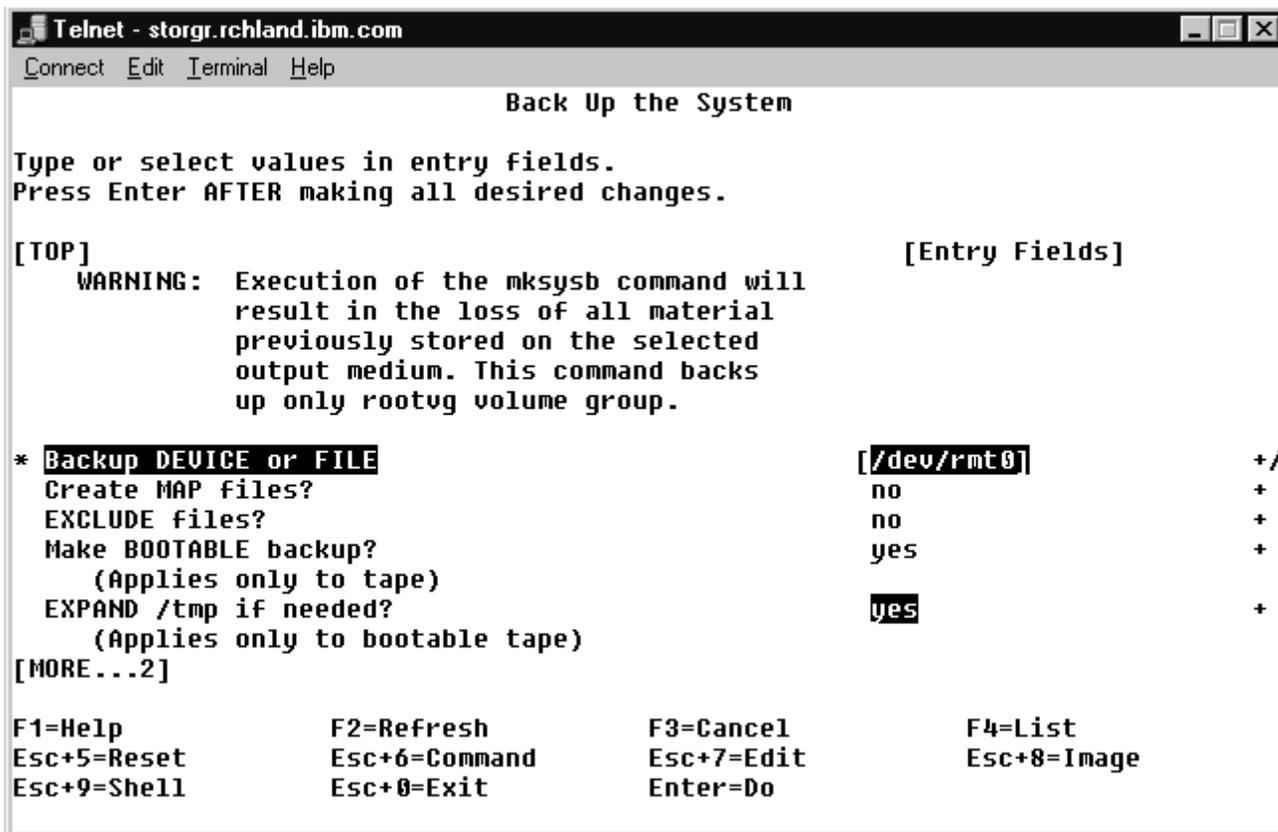


Figure 80. Back Up the System SMIT Screen

7.2.2 Backing Up a Customer Volume Group: DATAVG

The following steps show how to back up a volume group:

1. Type **SMIT** and press Enter. This brings up the main SMIT menu.
2. Select **System Storage Management** and press Enter.
3. Select **Logical Volume Manager** and press Enter
4. Select **Volume Groups** and press Enter.

5. Select **Backup a Volume Group** and press Enter. This brings up the **Back Up a Volume Group** screen.
6. Fill in the Back Up a Volume Group screen as shown on Figure 81.
7. Press Enter to start the back up function.

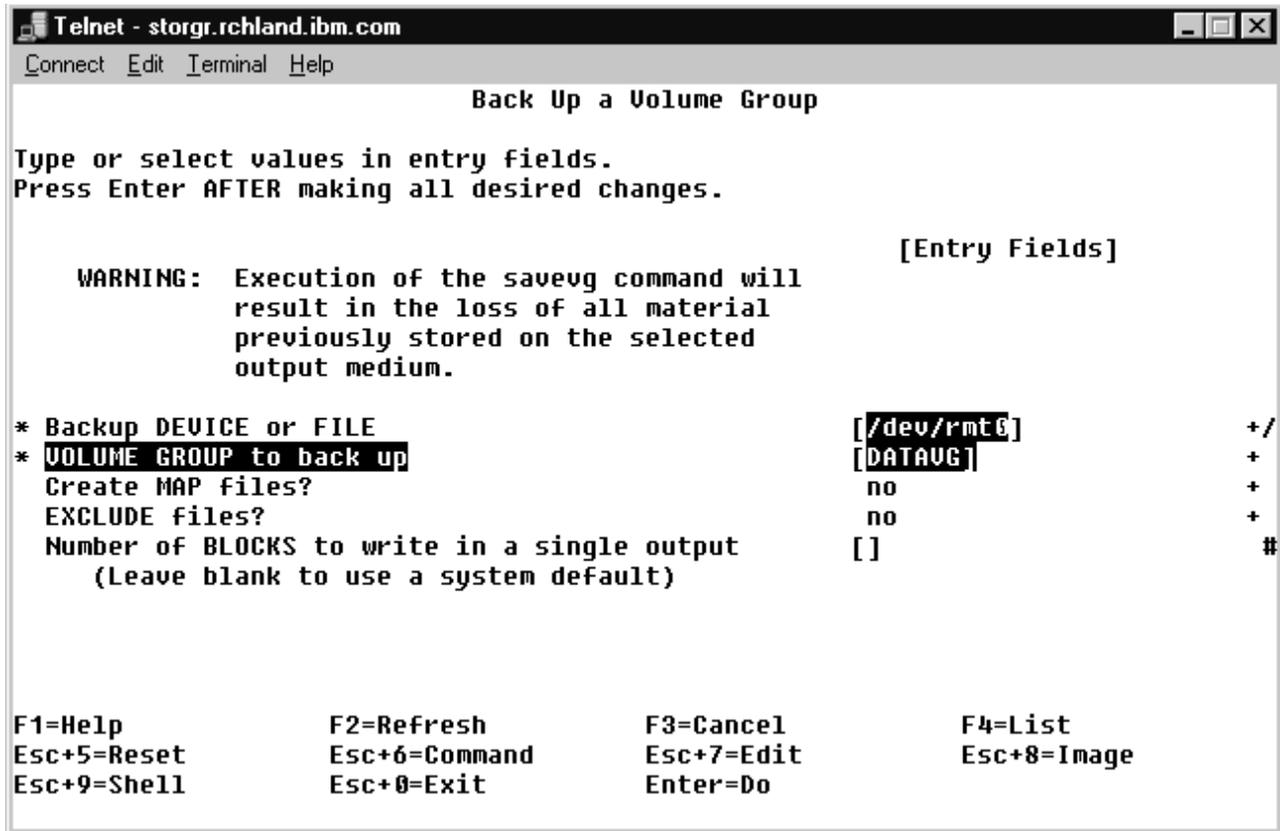


Figure 81. Back Up a Volume Group SMIT Screen

7.2.3 Restoring the AIX Operating System: ROOTVG

To restore ROOTVG, see Appendix G, “Backing Up and Restoring the Operating System in AIX 4.2” on page 175 or connect to this Web page:
<http://w3-3.austin.ibm.com/afs/austin/depts/aixserv/faxes/mksysb.basics.42.bak>

7.2.4 Restoring a Customer Volume Group: DATAVG

To restore a volume group, follow these steps:

1. Type **SMIT** and press Enter. This brings up the main SMIT menu.
2. Select **System Storage Management** and press Enter.
3. Select **Logical Volume Manager**.
4. Select **Restore Files in a Volume Group Backup**.
5. Fill in the Restore Files in a Volume Group Backup SMIT screen shown in Figure 82 on page 128.
6. Press Enter to start the restore function.

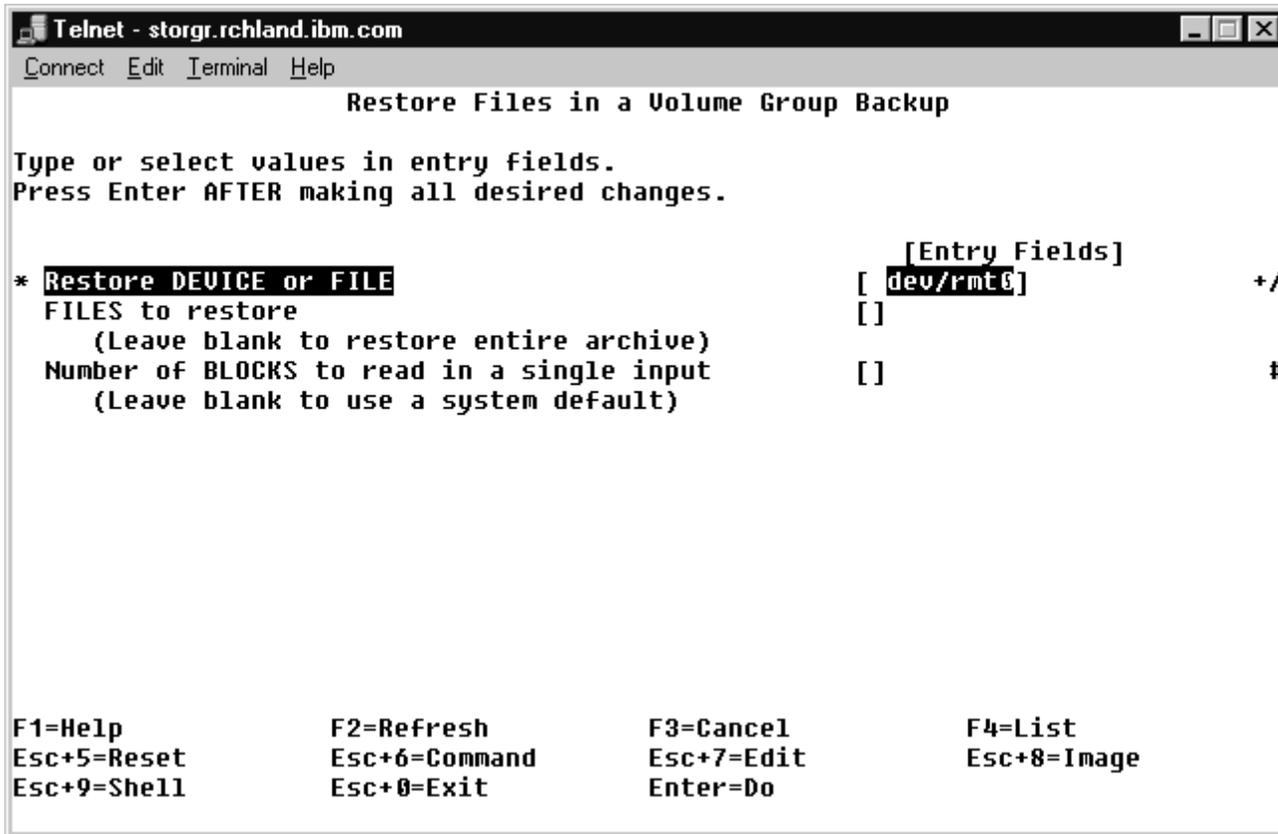


Figure 82. Restore Files in a Volume Group SMIT Screen

7.3 Managing Backup and Recovery with ADSM

ADSTAR Distributed Storage Manager, an IBM client server solution for distributed data management, supports a wide variety of IBM and non-IBM platforms for small, medium, and large systems. ADSM facilitates the management of your storage environment (both SSA and non-SSA disks), provides facilities such as backup and restore, archive and retrieve, and a Hierarchical Storage Manager (HSM) tool.

ADSM provides a centrally administered service based on a set of policies specifying how data will be treated during a backup or archive operation. Many services can be automated and scheduled to facilitate management of your SSA (and non-SSA) environment. When you install and customize ADSM, you define which devices and media ADSM is to manage.

Figure 83 on page 129 shows the multiplatform support that ADSM provides.

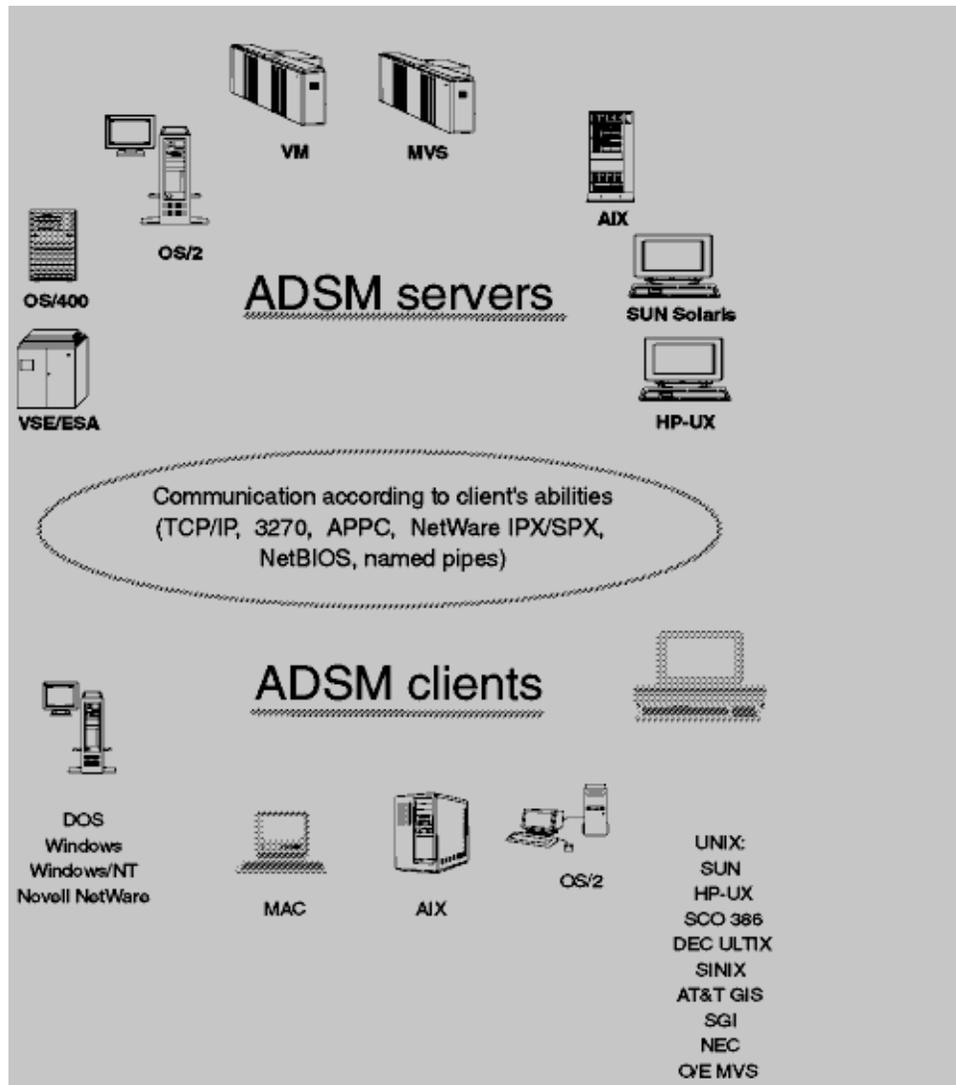


Figure 83. Platforms Supported by ADSM

The principal components of ADSM are:

- **Server** - Provides ADSM services to client workstations. The server maintains a database and recovery log for ADSM resources, users, and user data.
- **Server storage** - Consists of storage pools of random and sequential access media. The storage pools contain files that are backed up, archived, and migrated from client workstations.
- **Server utilities** - Provides an interface to help configure the ADSM server.
- **Administrative client** - Provides an interface to control the ADSM server.
- **Backup/archive client** - Provides backup and restore, archive and retrieve services for workstations and file servers.
- **HSM client** - Provides space management services for workstations on some platforms.
- **Server storage migration**- One goal of system-managed storage is to ensure the most efficient use of your storage resources. You can structure server

storage into a hierarchy. You can then define criteria by which data is migrated from faster devices (disk, for example) to slower devices (tape, for example). You can also use virtual volumes to store data on another server.

- **ADSM policy** - Governs storage management, including:
 - **Backup**- Copy files from client workstations to server storage to ensure against lost of data. Copies of multiple versions of a same file can be stored.
 - **Archiving**- Copy files from client workstations to server storage for long-term storage.
 - **Space management** - Free client storage space by copying a file from client workstations to server storage. The original file is replaced with a much smaller stub file that points to the location of the original server storage. This is also called *Migration* or *HSM*.

7.3.1 How ADSM Store Client Data?

When clients back up, archive, or migrate files, ADSM carries out three steps (see Figure 84 on page 130).

1. Determine where to store the file.

ADSM checks the management class bound to the file to determine the destination of the file. A destination is a group of disks or tape volumes. These groups of volumes are called *storage pools*. Copy groups, which are in management classes, specify the destination for backed up and archive files. The management class specifies destinations for space-managed files.

2. Store information about the file in the ADSM database.

ADSM saves information in the database about each file in server storage.

3. Stores the files in ADSM server storage.

4. ADSM stores the client files in disk or tape storage pools according to the management class specification.

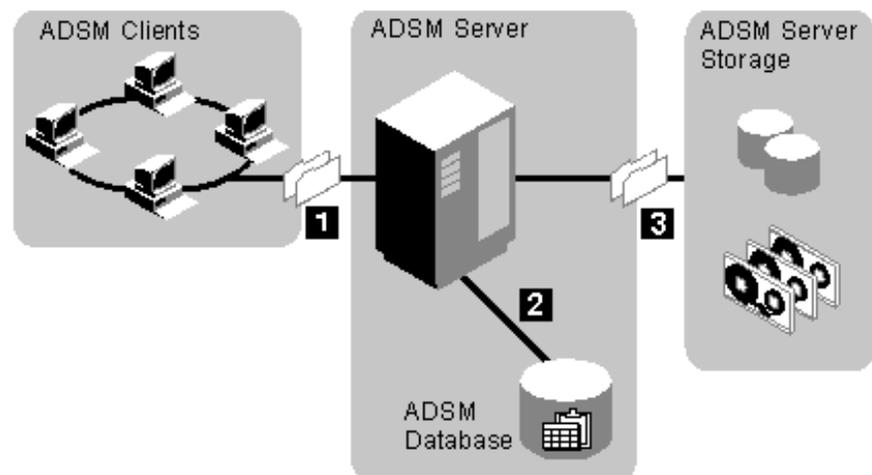


Figure 84. How ADSM Stores Client Data

7.3.2 How ADSM Controls Backup, Archive, and Space Management

ADSM carries out these four steps to control backup, archiving, and space management (see Figure 85 on page 131)

1. When an ADSM client backs up, archives, or migrates a file, the file is bound to either the default management class or a management class specified in the client's include-exclude list.
2. If, according to the management class, the file is eligible for backup, archive, or space management, the client sends the file and file information to the server.
3. The server checks the management class or copy group to determine where in server storage to store the file. If there is not enough space in the initial storage pool, the server examines the next pool in the hierarchy and places the file there, if space is available.
4. The server stores the file in the appropriate storage pool and stores information about the file in the database.

When files in server storage are migrated from one pool to another, the server updates the file information in the database.

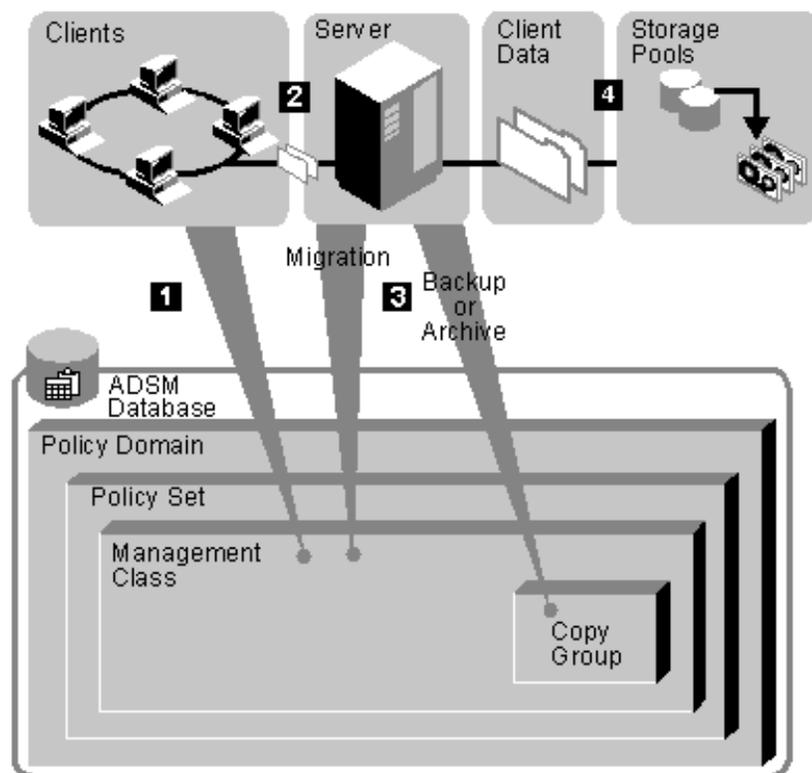


Figure 85. How ADSM Controls Backup, Archive, and Space Management

7.3.3 ADSM Database

The ADSM database is the heart of the server. It is highly tuned for storing and retrieving data from storage pools. It is a true database except that it does not

have a general query language and does not allow users to create tables. The features are:

- **Database** - The database keeps track of where all backed up and archived data is located in the storage pools. All policy and schedule information is stored in the database. Because the location and attributes of a file are stored in the database, you can easily view the attributes and versions of a file when you have to restore it, without reading and searching through all of your tapes to find the correct one.
- **Recovery log** - The recovery log is used to help maintain the integrity of the database. It keeps track of all changes made to the database, so that if a system outage were to occur, a record of the changes would be available in the log. When a change occurs to the database, the recovery log is updated with some transaction information before the database is updated. Thus uncommitted transactions can be rolled back during recovery so that the database integrity remains.
- **Mirrored for high availability** - Mirroring is used to configure ADSM to maintain as many as three copies of the recovery log, database, or both on independent storage volumes. The mirrored copies are treated equally; there is no concept of primary copy and alternate copies. Therefore the server reads from the database with the best response time.
- **Automatic Alternate Volume Switching** - If a physical volume on which the database or recovery log resides fails, ADSM takes the failing volume offline and continues to perform operations using the remaining intact copy or copies. This is done concurrently so that ADSM services remain online. After the failed volume is repaired, ADSM allows the volume to be brought on-line and synchronized with the intact volumes, (as in concurrent mode).
- **Dynamically expand or contract** - Space can be dynamically added or deleted from the database or recovery log, as needed. For instance, you can start with one volume and, as the system grows, you can allocate more space to the database or recovery log. This can be done dynamically, without any need to bring down ADSM.
- **Full or incremental backup while server is active** - A full backup takes longer to run than an incremental backup because it copies the entire database. However, recovery time is faster with a full backup because only one set of volumes has to be loaded to restore the entire database. In contrast an incremental backup takes less time because it copies only those database pages that have changed since the last time the database was backed up. However, incremental backups increase the time it takes to recover a database because a full backup must be loaded first, followed by all incremental backups in the same database backup series. The norm is to perform incremental database backups daily and a full backup weekly.

7.3.4 Optional ADSM Tools

ADSM provides optional tools to improve the management of your SSA (and non-SSA) environment such as Hierarchical Storage management (HSM), Disaster Recovery Management (DRM), Andrew File System(AFS), Distributed File System Support (DFSS), and Network Storage Management (NSM).

In this section we discuss HSM, and DRM. These two tools can be used to directly manage your SSA environment.

7.3.4.1 Hierarchical Storage Management

HSM is an optional ADSM tool that enables you to defer the cost of logical disk upgrades as well as the interruption of the server when large disks are required (see Figure 86 on page 133). HSM is also called *space management*.

With HSM, new files and the files you use most frequently remain on your local file systems, while those you use less often are automatically migrated to distributed storage devices through an ADSM server. You can also specify file priorities for migration according to file size, the number of days since they were accessed, or both.

By migrating eligible files to distributed storage devices, HSM frees space for new data on your local file system and takes advantage of lower cost storage resources available in your SSA (and non-SSA) environment.

Files migrated to ADSM storage are easily accessible. When you ask for a migrated file, HSM automatically recalls it to your local file system. If you choose, HSM can recall a migrated file temporarily. You can also choose to have HSM read a migrated file from ADSM without storing it on your local file system.

HSM is integrated with backup. You can specify not to migrate a file until it has been backed up. If you migrate a file and then take a backup the next day, ADSM is smart enough to copy the file in ADSM storage and not require a recall to back it up.

HSM is also controlled by policy to ensure the integrity of your data.

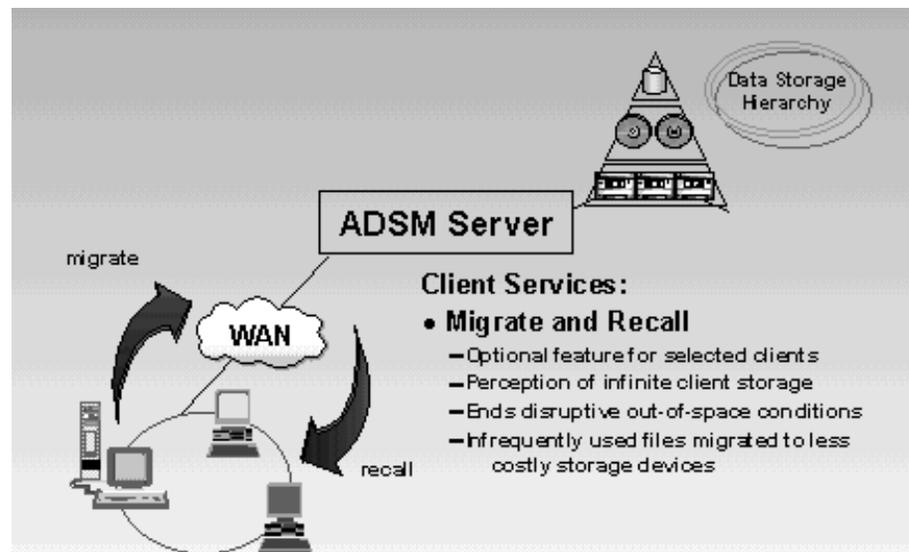


Figure 86. Hierarchical Storage Management

HSM is a technology in client/server environment. When designing your environment, keep in mind that migrated data is active data that must be treated differently from backup data. ADSM server availability becomes more critical when it is storing active data. It is also important to ensure that the ADSM server

database and storage pools are backed up on a regular basis, with a copy stored on external devices or offsite for disaster recovery purposes.

When HSM tuning is optimized, its activities (migration, transparent recall) have minimal impact on your users.

7.3.4.2 Disaster Recovery Manager (DRM)

It is important to manage your SSA environment and its backup and restore functions, especially when a catastrophic interruption of business occurs and destroys your ADSM server, ADSM client, or both. We recommend that you store backup data offsite on external devices such as tapes, libraries or optical devices to protect it from damage. DRM helps you plan, prepare, and execute a disaster recovery plan (see Figure 87 on page 135 for a schematic view of DRM).

Before planning a disaster recovery procedure, you must regularly:

1. Back up client data to the ADSM server.
2. Save the client recovery information in the ADSM database.
3. Back up the primary storage pools on external and removable media.
4. Back up the server database on external and removable media.

The external and removable media can be stored offsite for further security.

Once you have satisfied the above prerequisites, you can work with DRM to reduce the time to develop and maintain an ADSM disaster recovery plan.

The steps to manage your disaster recovery are as follows:

Disaster Recovery Plan File - Run a query to find all of the required information from the ADSM server and generate a recovery plan file that is based on a predefined recovery strategy for the server. The required information includes:

- Site-specific server recovery instructions, defined by the administrator, such as contact names, phone numbers, and internal company procedures
- Sequential procedure to recover an ADSM server
- List of ADSM database backup and copy storage pool volumes required to perform the recovery procedure
- Devices required to read the database backup and copy storage pool volumes
- Space requirements for the ADSM database and recovery log
- A copy of ADSM server options file, device configuration file, and volume history file
- Shell scripts and ADSM macros for performing server database recovery and primary storage pool recovery.

Offsite Recovery Media Management - In this example, the primary storage pool backup is offsite. Offsite media management is used during routine operation and defines a process for moving ADSM data user backup and copy storage pool volumes offsite for disaster recovery protection. Backup volume location is included in the disaster recovery plan file. If the ADSM server is destroyed, the disaster recovery plan file can be used to provide a list of offsite volumes required at the recovery site.

Recovering the Server - To do a complete ADSM server recovery you need the disaster recovery plan file, the backup volumes, and an operational AIX/UNIX system. If these three conditions are met, the disaster recovery plan file automatically performs these steps:

1. Restores the ADSM server and the administrative client software.
2. Restores the ADSM server options, volume history, and device configuration files.
3. Creates the ADSM server volumes for the database and recovery log.
4. Restores the ADSM database.
5. Starts the server.
6. Creates the ADSM server volumes for new primary storage pools.
7. Defines the primary volumes to the ADSM server.
8. Restores the primary storage pools from the backup volumes.

Recovering the Client - The disaster recovery plan file performs the followings steps:

1. Restores the operating system.
2. Restores the communication protocol.
3. Restores the ADSM client code.
4. Restores the ADSM client option files.
5. Restores the client file systems from the primary storage pools.

We highly recommend that you test your recovery procedures periodically to test recovery of your ADSM server and ADSM client.

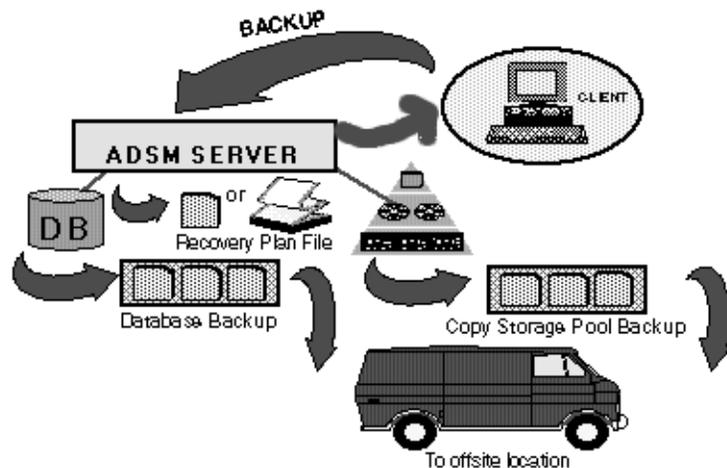


Figure 87. Disaster Recovery Manager Overview

7.3.5 ADSM Device Support

ADSM provides device support for many different types of disk, optical, and tape drives. For the most up-to-date list contact your local IBM representative, or go to this Web site: www.storage.ibm.com/adsm

Chapter 8. Performance and Tuning

In this chapter, we discuss how to improve performance and provide some basic rules. However, data integrity is most important and this should not be compromised in order to improve performance.

8.1 Introduction

I/O performance depends on many different factors, including the hardware, the configuration of the hardware, the operating system with its settings, the application itself, and how the application uses the system resources. Any of these values can produce a bottleneck for system performance. It is important to remember that removing one bottleneck almost inevitably leads to one elsewhere. A balanced system design is the ultimate goal of all performance tuning.

8.1.1 Performance at Disk Level

Usually, more than one disk drive is installed in any given system. If all your data is on one disk, the performance cannot be better than the Media Data Rate in Table 16 on page 137. For random operations with small block sizes, the seek and latency times are very important.

Table 16. SSA Disk Parameter

Formatted Capacity (GB)	1.1	2.2	4.5	4.5	9.1	9.1	18	9.1
Media data rate (Banded) in MB/s	9.59-12.58	9.6-12.58	9.59-12.58	10.2-15.42	10.2-15.42	11.5-22.4	11.5-22.4	16.16-25.6
Sustained data rate (MB/s)				9.2	9.2			
Average read (ms)	6.9	7.5	8.0	7.5	8.5	6.5	7.5	6.3
Average write (ms)			9.5					
Track-to-track read (ms)			0.5	0.5	0.5	0.7	0.7	0.7
Rotational Speed (rpm)	7200	7200	7200	7200	7200	7200	7200	10020
Latency (Average)	4.17	4.17	4.17	4.17	4.17	4.17	4.17	2.99
Buffer size (KB)			512	512	512	1024	1024	1024
Interface Transfer Rate (max) MB/s	20	20	20	80	80	160	160	160
Name	Starfire 1100	Starfire 2200	Starfire 4320	Scorpion 4500	Scorpion 9100	Sailfin 9100	Marlin 18000	Thresher 9100

The Starfire disks are older and can no longer be ordered.

Further information about the disks can be found
<http://www.storage.ibm.com/hardsoft/diskdrdl.htm>

For sequential applications with large block sizes, the media data rate is the most important factor. However, while for online transaction types of workload, which typically use small block sizes with lots of small seeks, the seek and latency times

of the disks are the most important factors. If you can stripe the data over more than one disk, you will usually obtain better performance than by having all the data on one disk. If you use AIX disk mirroring you get better read performance than by using nonmirrored disks.

8.1.1.1 Tuning

All disks have a memory buffer on the disk. Normally this cache is used as read cache. Some operating systems, but not AIX, allow you to use this memory as write cache. This way you get better write performance, but you must use this option carefully as the cache is not protected and in the event of power failure or certain error conditions on the SSA loop you can lose data.

Disk performance is also affected by the area of the disk (inner, middle, outer) in which the data resides. The sustained data rate from a disk is greatest on the outside diameter of the disk (the lowest logical block addresses on the disk). Thus data that needs to be read or written at a high data rate is best situated at low logical block addresses (LBA). The access time of a disk is best midway between the outer and inner edges of the disk. Data that is accessed frequently should be placed in this region. There are several tools on the different platforms to see the location of the data or file systems on the disk.

On AIX you can use **xlvm**. To change the area expectations of a logical volume on AIX you can use **smit chlv1**. After this it is necessary to reorganize the volume group of the logical volume by **smit reorgvg**. The parameter **POSITION on physical volume** is an expectation that not all data can be in one area.

8.1.1.2 Performance Implications of Disk Mirroring

If mirroring is being used with non-RAID disks and Mirror Write Consistency is turned on (as it is by default), you may want to locate the copies in the outer region of the disk, since the Mirror Write Consistency information is always written in Cylinder 0. This region corresponds to the lower logical block addresses near 0. From a performance standpoint, mirroring write operations is costly; mirroring with Write-Verify is costlier still (extra disk rotation per write), and mirroring with both Write Verify and Mirror Write Consistency costs most of all (disk rotation plus a seek to Cylinder 0). To avoid confusion, we should point out that although an **lslv** command will usually show Mirror Write Consistency to be on for nonmirrored logical volumes, no actual processing is incurred unless the **COPIES** value is greater than one. Write Verify, on the other hand, defaults to off, since it does have meaning (and cost) for nonmirrored logical volumes.

It should be remembered that for read operations, mirroring can have significant performance advantages since the LVM has a choice of two disks from which it can read the desired data and it can therefore choose whichever disk has the smaller queue or shortest seek time.

These comments are not specific to SSA, as SCSI disks see the same effects.

The block size that you use depends upon your application. Large blocks are more effective than smaller blocks for transferring large amounts of data. But if your application mainly reads and writes small amounts of data, it makes no sense to use large block sizes. Small block sizes are more effective if high I/O rates are required.

queue_depth

The main parameter that can be changed is the *queue_depth* for the hdisks. However, we do not recommend changing the queue depth parameter as the system default is probably optimum for most environments.

On AIX You can see the parameter value by the following command:

lsattr -El hdiskn -a queue_depth

To change this value the file systems on this disk must be unmounted and the hdisk must have the status defined. If it is deemed necessary to change this value then:

1. **umount /filesystem_name** (unmount the file systems.)
2. **rmdev -l hdiskn** (set the hdisk from available to defined.)
3. **chdev -l hdiskn -a "queue_depth=x"** (x is the new queue depth.)
4. **cfgmgr** (run configuration manager to make disk available.)
5. **mount /filesystem_name** (mount the file systems unmounted in step 1.)

8.2 Performance at Adapter Level

In this section, we discuss how you can improve performance by configuring your loops and selecting the right type of adapter.

8.2.1 Number of Disks in the Subsystem

In any disk subsystem, the larger the number of independent actuators per megabyte, the smaller is the impact of seek time. This effect is more pronounced as the number of start I/Os per second increases, the transfer length of each operation decreases, and the size of the seek increases. When approaching limiting cases it will be better to use a larger number of small capacity drives (for example 4N x 4.5GB drives) rather than a few high capacity drives (N x 18GB). However, be sure to split the access (see Table 16 on page 137).

If your I/O mix consists of many long sequential transfers, then 6 to 12 disks per SSA loop is a good choice. In the limit, this number of files can sustain sufficient data throughput to saturate the bandwidth available to the adapter at micro channel (at approximately 35 MB/s). The maximum sustained data rate of about 35 MB/s can be achieved with as few as six drives, provided the transfers are large sequential blocks of data. The six drive load was measured with 196 KB block sizes. This requires using raw I/O to non-JFS logical volumes, or raw I/O to hdisks.

The more typical situation is using the JFS. The default block size transferred with JFS is 4 KB. Figure 88 on page 140 shows the operations per second on numbers of disks.

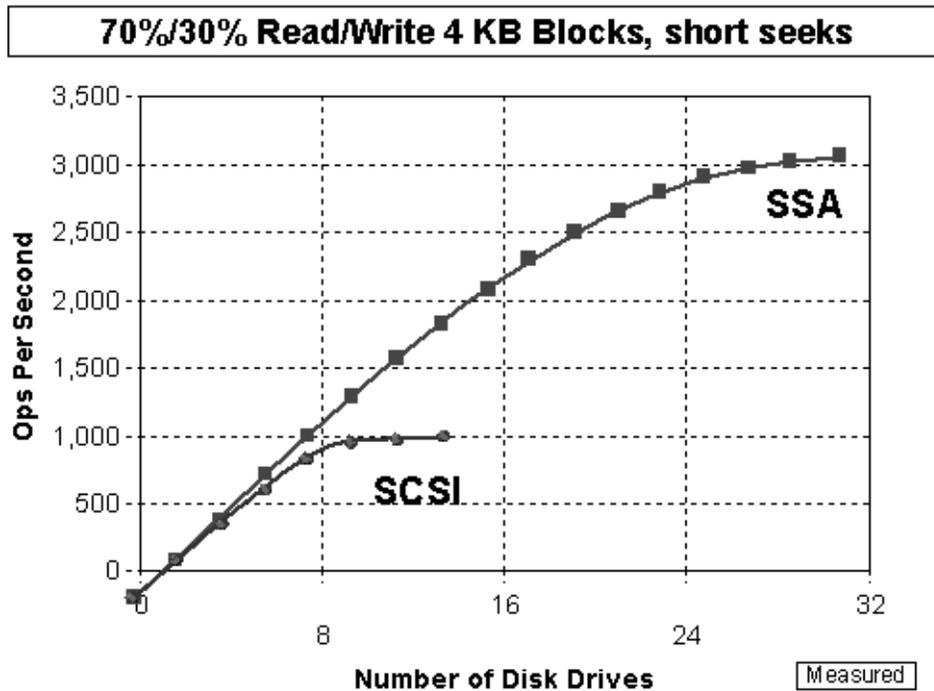


Figure 88. Performance Chart 1 - SSA versus SCSI

If your I/O mix consists of many short transfers distributed with random seeks across the file, then rather more files can be comfortably handled by a single adapter. Two loops of 16 or 24 disks would be appropriate (making a total of 32 or 48 disks per adapter).

Frame Multiplexing

The SSA adapter card is limited at any one instant in time to transferring data over four duplex links (two per port). There is a physical limit of eight data transfer operations at any one instant. However, because the adapter data bandwidth greatly exceeds the bandwidth available from a single drive, this is not really relevant when considering the number of drives that an adapter or host can support.

Transfers from each disk are broken up into 128 byte frames, individually addressed. It is possible to interleave frames for different transfers, so that the frames from many different data transfers can be intermixed and sent down the same link, one after the other. This is called *frame multiplexing*. The SSA hardware automatically manages this, deciding according to well defined rules when to insert a new frame, or when to allow existing traffic to flow through. Similarly, the hardware automatically interprets the frame address, and demultiplexes the data stream, separating it out into its individual logical transfers.

The number of data streams that SSA can support on a port is very large in theory. The SSA adapters under consideration here adapters support ten data transfers, plus two control message frame transfers, per port. Hence 40 data transfers could be streaming between an adapter card and the SSA network at one time.

This flexibility in the SSA architecture means it can be quite difficult to state categorically the exact number of disks that an adapter can keep busy, since this is often highly dependent on the particular workload being executed at any particular time.

8.2.2 Basic Configuration Rules

Every link in the loop has a 20 MB/s read and a 20 MB/s write wire. As each adapter has two links, an adapter can perform 40 MB/s reads and 40 MB/s writes per loop.

Fairness Algorithm

To ensure that no one device dominates the loop, SSA implements a fairness algorithm. Simply stated, the algorithm involves the circulation of tokens (called SAT tokens, for satisfaction). The tokens circulate in both directions around a loop, or from end to end on a string. The principle of the fairness algorithm is that each node will, over time, accumulate a queue of I/O, either data or commands and responses, which it wants to transmit. Typically, if a node is not passing messages from the adjacent node on one side to the adjacent node on the other side, it sends its I/O on the appropriate link, which would be quiet. If the node is busy passing traffic from one side to the other and still wants to transmit I/O, it queues the I/O. In theory, a node could be stuck passing traffic for other nodes and be unable to transmit its own. Enter the SAT token. One token circulates in each direction around the loop (or along the string). Simply stated, if the node receives a SAT token and has a queue of I/O waiting to transmit, it is allowed to transmit the I/O at that time. This transmission is allowed up to a level at which the node is considered "satisfied". At this point, the node must again wait until either until the link is quiet or it receives another SAT token.

Thus, in theory, you want to place the disks with the lowest I/O nearest the adapter, and the disks with the highest I/O farthest from the adapter. The increase in latency, for placing disks further from the adapter, is only about five ms per hop, so even if a disk is 20 hops from the adapter, the increased latency is only 0.1 ms. However, the SSA loop must be heavily loaded for disk placement within the loop to matter.

In a loop with a single initiator, it is more important to balance the I/O load across both halves of the loop than to worry about the position of the disk within the loop.

In a loop with multi-initiators, place the disks adjacent to the adapter that controls them.

In Figure 89 on page 142, Disks 0-7 are using Port A1 and Disks 8-15 are using port A2. If the loop is broken, disks automatically take the only possible way to the adapter. The performance impact can be 0% if the link between disk 7 and disk 8 is broken and it can be up to 50% if one of the links to the adapter is broken.

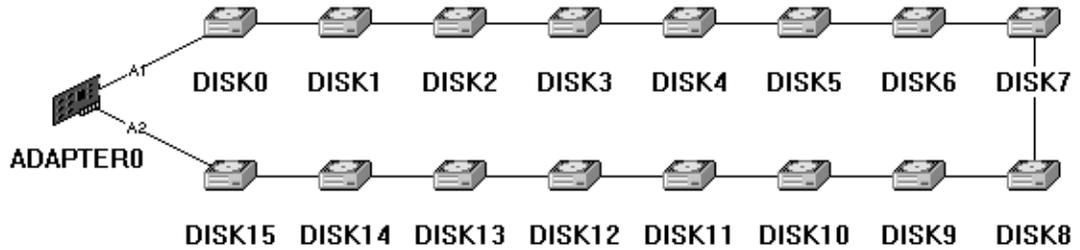


Figure 89. Performance Sample 1 - Single Loop

To get a better performance you can change the configuration on Figure 89 on page 142 to Figure 90 on page 142.

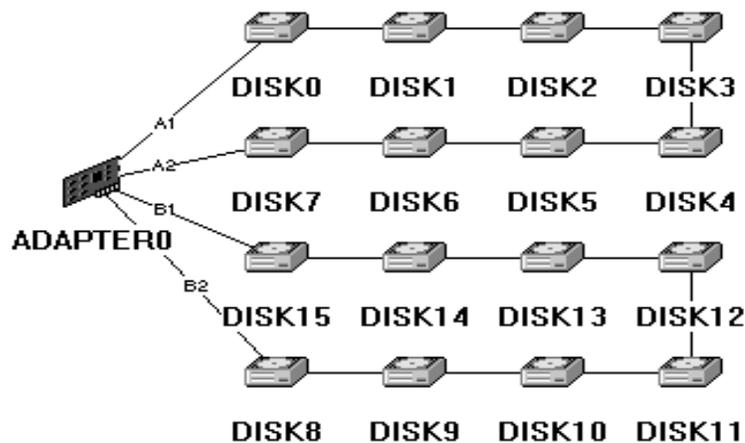


Figure 90. Performance Sample 2 - Two Loop

Now you use four links therefore the adapter can 80 MB/s read and 80 MB/s write from the disks. So you get a theoretical bandwidth from 5 MB/s for each disk at the same time. In Figure 89 on page 142 it was 2.5 MB/s. You have two separate loops. If one loop breaks you have between 50 and 100% performance on the defective loop and full performance on the other loop.

8.2.3 Number of Adapters

The next bottleneck can be the adapter. The throughput from a SSA Micro channel adapter to a system is not higher than 35 MB/s and from a SSA PCI adapter it can be 60 MB/s. For higher data rates, you need more adapters.

How many adapters you can use in your system depends on the slots in your system and on other system specifications. In Table 17 on page 143, you can see how many adapters of which type you can use.

Table 17. Comparison of Adapters

Feature Code	6214	6215	6216	6217	6218	6219	4012	4011	4003	7190-100	7190-200
Operating System	AIX	AIX	AIX	AIX	AIX	AIX	Windows NT	Windows NT, OS/2, NetWare, DOS	Solaris 2.4, Solaris 2.5.1	SUN, HP, AIX, NT, other	SUN, HP, AIX, NT, other
Adapter Description	Classic	Enhanced	Enhanced	RAID-5	RAID-5	Enhanced			Classic	external Box	external Box
Bus	MCA	PCI	MCA	MCA	PCI	MCA	PCI	PCI	SBus	SCSI-FW	SCSI_UW
HW RAID Types	n/a	5	5	5	5	5	0,1,5	0,1,5	n/a	n/a	n/a
Adapter/system	4	4	4	4	4	4	3	3	2	SCSI	SCSI
Loops	2	2	2	2	2	2	2	2	1	1	1
Disks per Loop	48	48	48	48	48	48	22	48	48	48	48
Maximum JBOD Multi-Attach	2	2	8	1	1	2*	2	1	4*	4	16
Maximum RAID Multi-Attach	n/a	1 RAID-5	0	1 RAID-5	1 RAID-5	1 RAID-5	2 RAID-1	1	n/a	n/a	n/a
Read Cache	n/a	32 MB	n/a	8 MB	8 MB	32 MB		8 MB			
Fast Write Cache (optional)	n/a	4MB	n/a	n/a	n/a	4MB			n/a		
FW Feature Code	n/a	6222	n/a	n/a	n/a	6222			n/a		
Intermix with FC	6216	6219	6214	n/a	n/a	6215			n/a	7190	7190
HACMP Qualified	Yes	Yes	Yes	No	No	Yes	n/a	n/a	n/a	Yes	Yes
Target Mode Support	No	No	No	No	No	Yes			n/a		
I/O Operations/s Non-RAID	3000	3000	3000	3000	3000	3000		3000	3000	1900	3000
I/O Operations/s RAID-5	n/a	3000	n/a	3000	3000	3000	2000 RAID-1		n/a	n/a	n/a
I/O Operations/s RAID-5	n/a	1000	n/a	1000	1000	1000		1000	n/a	n/a	n/a
MB/s (r/w) Non-RAID)	35/35	35/35	35/35	35/35	35/35	35/35		60/40	35/35	18	35
MB/s (r/w) (RAID-5)	n/a	29/13	n/a	29/7	29/29	29/13	60/RAID-1	35/7	n/a	n/a	n/a
Microcode Level	2401	1801	2402	2904	3700	1801					

Feature Code	6214	6215	6216	6217	6218	6219	4012	4011	4003	7190-1 00	7190-2 00
Adapter Type	4-D	4-N	4-G	4-I	4-J	4-M					
Diagnosis & Maintenance	diag SMIT	SSA RSM	SSA RSM	SSAU SSAC F							
Management Layouts	StorX maymap	SSA RSM	SSA RSM								

A good solution for performance and availability is to use two adapters in the same loop for the same system, but not all adapter types allow this. If you do this, it is better for performance to put disks between those adapters as shown in Figure 91 on page 144.

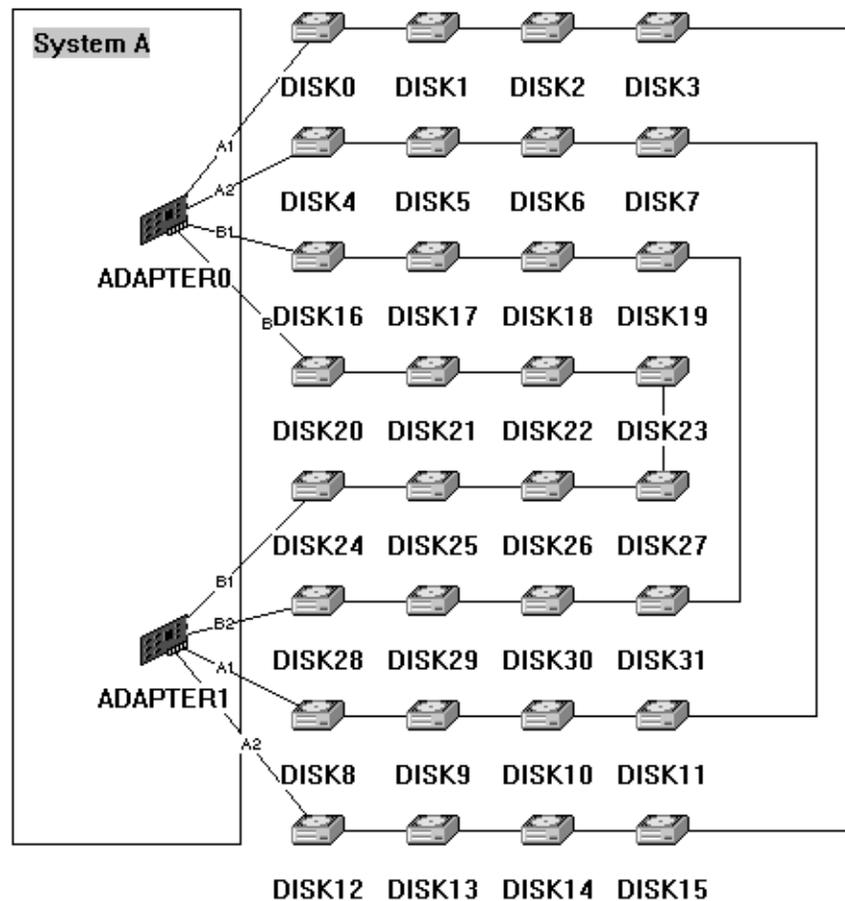


Figure 91. Performance Sample 3 - Four Loops and Two Adapters

In this configuration, four disks are using one port of the adapter. There are a theoretical bandwidth of 5 MB/s for read and 5 MB/s for write for each disk at the same time. If one adapter fails, the disks are still available with up to 50% performance. In this configuration, the throughput from disk to system can be 70 MB/s on Micro channel.

Figure 92 on page 145 shows connection to the adapter directly.

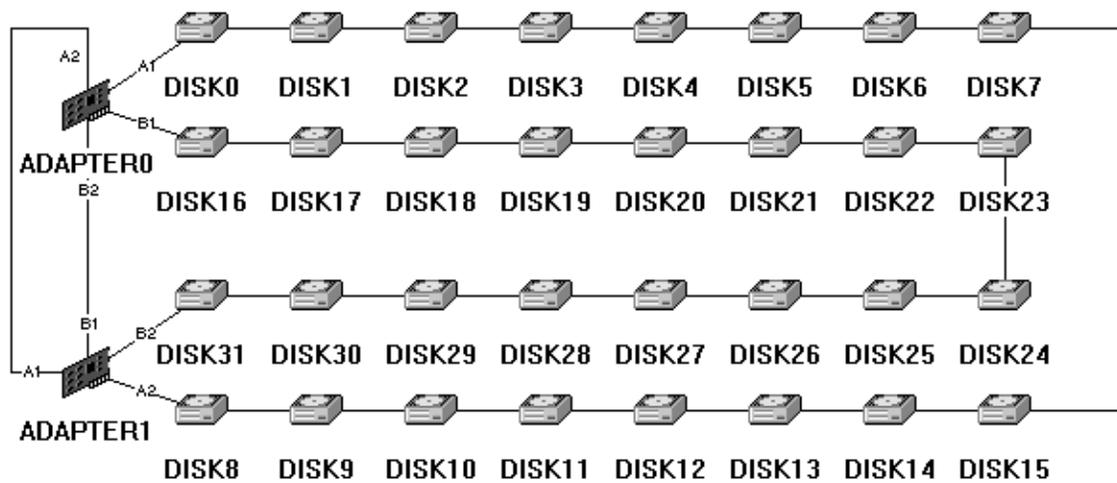


Figure 92. Performance Sample 4 - Two Loops on Two Adapters

This configuration is not as good. The availability and the throughput from the adapter to the system is like Figure 91 on page 144, however, the disks use only one link per loop per adapter. So you get only a theoretical bandwidth of 2.5 MB/s for read and 2.5 MB/s for write for each disk at the same time.

Using three adapters in one system, as shown in Figure 93 on page 146, is a good solution.

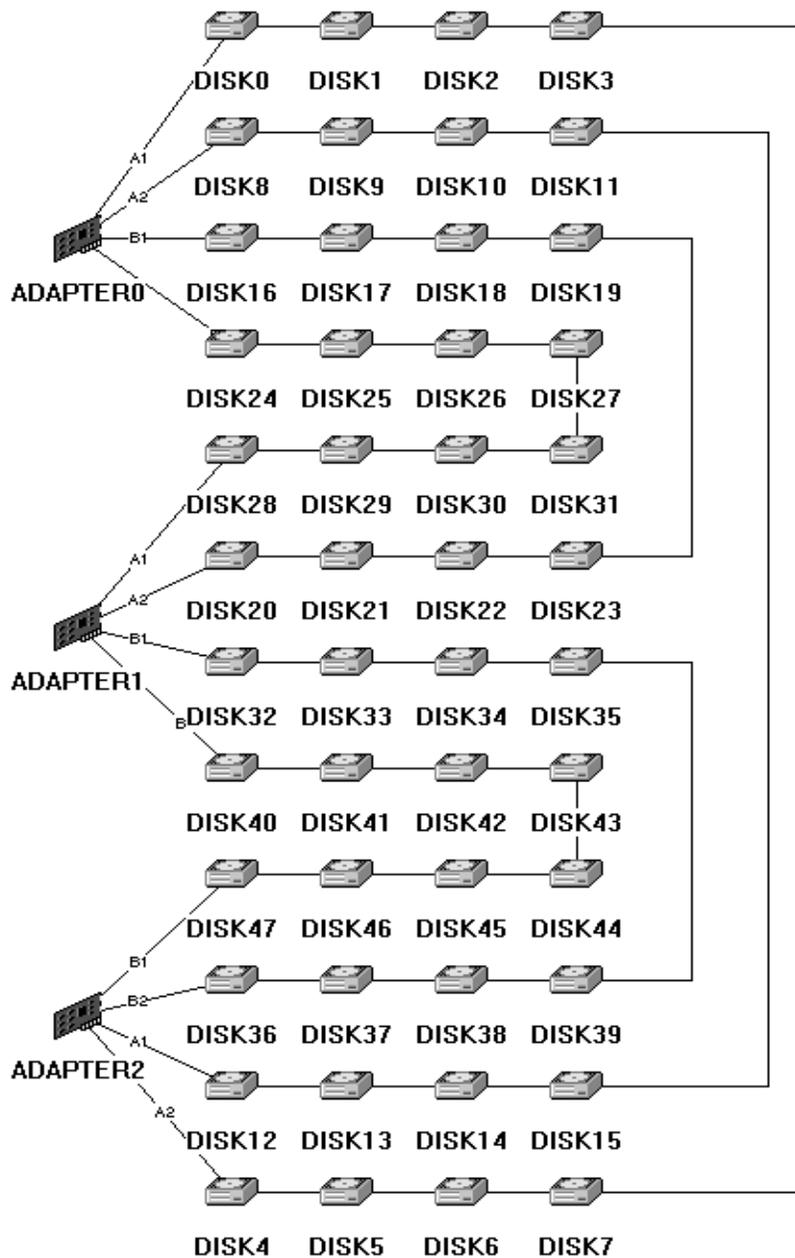


Figure 93. Performance Sample 5 - Four Loops on Three Adapters

If you use multisystem attached SAA configurations, the same rules apply.

8.2.4 Adapter Types

In this subsection, we discuss the use of adapters with RAID hardware. The RAID array is configured on the adapter, not by the operating system.

An overview of the differences between the RAID levels can be found in Table 18 on page 147.

Table 18. Overview RAID Hardware

RAID Level	Common Name	Description	Disks required	Data Availability	Data transfer capacity	I/O request rate
0	Striping	Data distributed across the disks in the array. No redundant information provided.	N	Lower than single disk	Very high	Very high for both read and write
1	Mirroring	All data replicated on N separate disks. N is most commonly 2.	2N, 3N	Higher than RAID level 3,5 and single disk	Higher than single disk for read, similar to single disk for write	Up to N times that of a single disk for read, less than a single disk for write
5	Raid-5	Data sectors are distributed as with disk striping, redundant information is interspersed with user data.	N+1	Much higher than a single disk; comparable to RAID-3	Similar to disk striping for read; lower than single disk for write	Similar to disk striping for read; generally lower than single disk for write

Non-RAID

Non-RAID configurations are also called JBOD (Just a Bunch of Disks) and are supported by all types of adapters.

RAID-0

Adapters that support Raid-0 are available only for Intel-based systems.

Select RAID-0 for applications that would benefit from the increased performance capabilities of this RAID level but, because no data redundancy is provided, do not use RAID-0 for mission-critical applications that require high availability.

RAID-1

Adapters that support Raid-1 are available only for Intel-based systems.

Select RAID-1 for applications where data availability is a key concern, and which have high levels of write operations such as transaction files, and where cost is not a major concern.

RAID-5

Raid-5 is supported by several AIX adapters (see Table 17 on page 143) and also by adapters for Intel-based systems.

Select RAID-5 for applications that manipulate small amounts of data, such as transaction processing applications. Since all the disk heads move independently (to satisfying multiple requests) it is appropriate for multiuser applications. Also, the selection of RAID-5 is a good compromise between performance, data protection, and costs.

Number of components in a RAID-5 array

In a RAID-5 array, when read requests map to separate disks within the array, they can be processed in parallel. This striping effect means that arrays with a large number of components can offer better read performance than those with fewer components. To offset this, the larger the number of components in a given array, the longer the rebuild time required when a component fails, and the poorer the performance of the array while the component is missing. On the other hand, the larger the number of components in a single array, the lower the cost of providing parity data. Most users will prefer to use relatively small arrays to conserve performance. We recommend three to eight disks.

We also recommend using spare disks.

It is possible to define several arrays in a loop or on one adapter.

It is also possible and better for performance to span RAID arrays over the loops on one adapter.

Use of fast write cache

Fast write cache can significantly improve the response time for write operations. However, care must be taken not to flood the cache with write requests faster than the rate at which the cache can destage its data.

Fast write cache can adversely affect the maximum I/O rate however, since additional processing is required in the adapter card to determine if the data that is being transferred is in the cache or not.

Fast write cache typically provides significant advantages in specialized work loads - for example, copying a database onto a new set of disks. If the fast write cache is spread over multiple adapters this can multiply the benefit.

Fiber extender

The use of fiber extenders reduce the throughput per loop over distance. At 2.4 km (1.5 mi) distance, the performance can be reduced to 12 MB/s.

Below are some recommendations for the use of SSA RAID adapters.

Non-RAID operation

- Distribute the disks evenly over the available SSA loops.
- If possible, distribute read and write data evenly throughout the SSA loops.
- For applications needing a high rate of operations per second, use logical volumes made up of small disks. The adapter can sustain a 70/30 mix of read/write operations, with 4 KB transfer lengths, at up to 3000 operations per second. This I/O rate can be produced by relatively few disks, say 32. Further disks can be attached to the adapter to provide additional connectivity and storage; however, this will not increase the overall operations per second rate of the subsystem. A second adapter should be installed if higher I/O rates are required and the load is balanced between the two adapters.
- When using logical volumes with mirroring, arrange to have the mirror copies on different disks on different loops.

RAID-5 operation

- For transaction processing applications, where a large number of small, unrelated I/O requests are made:
 - Use smaller (1 or 2 GB) rather than larger disks (4 GB)
 - If a database log file is being used, place it in a separate non-RAID disk, and mirror it if required.
 - Increase the queue_depth attribute on each hdisk from the default of 3 to 2N or 3N for a N+1 array. (This maximizes the chance of reading data from each component of the array in parallel.)
- For data intensive applications, where a small number of large, possibly related I/O requests are made:
 - It is tempting to use a large number of components in each array (to maximize the effect of striping)
 - When the array is operating with one component missing, however, it must read every other component to reconstruct the missing data. This means that when reconstructing data, performance reduces as the number of components in the array increases. It is better to limit the number of components in the array to a relatively small number, to make, say, a (4+P) array.

Having selected a particular SSA RAID configuration it is important not to neglect the conventional LVM tuning techniques (for example spreading a logical volume across multiple disks, using striped logical volumes or multiple JFS logs). As ever, it is also important to avoid file system fragmentation.

8.3 Performance on OS Level

Some RAID types are not implemented on the adapter, but you can use the tools of the operating system - for example, (AIX: Logical Volume Manager, Solaris: DiskSuite) or additional software (for example, Veritas Volume Manager) to realize RAID configurations.

The use of hardware is always better for system performance, than software.

The tools for tuning the disk operations differ on the operating systems, but there are some common rules:

- Distribute the throughput evenly to adapters.
- Distribute the heavily used data in a balanced way over the access ways and the disks.
- If you have heavily used small files - for example, database logfiles - use separate disk.
- Avoid fragmentation. For positioning of data on the disks, see "Performance at Disk Level" on page 137

Journalled File Systems (JFS)

The case of doing sequential I/O using JFS is kind of a cross between the sequential and random cases above, in that a sequential request will translate into a series of 4 KB block requests that will often be sequential to one another on the disk, depending upon file system fragmentation. LVM and JFS tuning will be as important as ever.

A few performance tips for LVM:

- Logical volumes that are spread among disks perform better than if on one disk
- Place the jfslog on a different disk than the one containing the file system
- Consider using multiple jfslogs.
- Watch out for fragmented file systems, and consider using striped file systems (AIX 4.1 only)
- Compressed file systems (AIX 4.1 only) improve disk performance but increase CPU use to do the compression/decompression.

Tuning

A simple tool available on all UNIX platforms and a good starting point for analysis is IOSTAT.

iostat *interval count*

```
tty:   tin      tout  avg-cpu:%user %sys %idle %iowait
        0.0    220.7      0.0  3.7  89.9  6.5
```

```
Disks: % tm_act  Kbp/s  tps  Kb_read  Kb_wrtn
hdisk0   4.6   11.1   1.8    24     0
hdisk1   0.0    0.0   0.0     0     0
hdisk2   1.8    1.8   0.5     4     0
```

IOSTAT indicates the most used disks and what kind of use they receive, read or write.

From the results, you can start to analyze the heavily used disks to see if:

- There are several heavily used files (such as log files) you can distribute on separate disks
- There is one file that is mostly read, so that you can mirror this file.
- There is one file that is mostly write, which can be striped, over RAID disks for necessary data protection.
- Also, look at other recommendations in this chapter that may apply for your situation.

Another very good tool is the AIX Performance Toolbox with agents also for SUN and HP systems. With this tool you can create your own graphical views for monitoring performance parameters. You can also record these views over time, to analyze disk activities or other parameters at critical situations as they occur, whether you are there or not, rather than at a fixed time.

Performance tuning is not only tuning of I/O performance. There are correlations to available main memory, CPU and memory performance.

There are many other tools for the different operating systems.

Literature for Performance Tuning

AIX

Redbooks:

- *AIX 64-bit Performance in Focus*, SG24-5103
- *RS/6000 Performance Tools in Focus* SG24-4989
- *Understanding IBM RS/6000 Performance and Sizing*, SG24-4810
- *A Practical Guide to SSA for AIX*, SG24-4599
- *AIX Storage Management*, GG24-4484

on the Web

- <http://service.boulder.ibm.com/rs6000/>
- <http://www.redbooks.ibm.com>
- http://www.rs6000.ibm.com/resource/aix_resource/Pubs/
- <http://sysperf.austin.ibm.com/> AIX Support Line -- Systems & Performance Group
- http://www.austin.ibm.com/doc_link/en_US/a_doc_lib/aixgen/

Windows NT

- *Windows NT Integration Guide*, SG24-4763 (IBM Redbook)
- *Implementing PC Server RAID SCSI and SSA RAID Disk Subsystems*, SG24-2098
- Several Microsoft publications and other literature

Novell

- *Novell IntranetWare and BorderManager for IBM Netfinity*, SG24-2145

SUN

- *Exploiting SSA in Sun Solaris Platform Environments*, SG24-5083 (IBM Redbook)
- *Sun Performance and Tuning: SPARC and Solaris*, ISBN 0-13-149642-5

HP Web site

- <http://docs.hp.com/hpux/os/#specific>

8.3.1 Guidelines for improving I/O Performance

Examine these guidelines in terms of your situation:

- Limit the number of disks per adapter so that the adapter is not flooded. With high throughputs using large block sizes, five to six disks can flood the adapter.
- Mirror across different adapters.
- The performance will be affected by the location of the logical volume with the disk. A contiguous, unfragmented partition in a logical volume will improve performance.
- You can turn off mirror write consistency cache for logical volume mirroring, but doing so removes the guarantee of consistency of data in case of a crash. In this case, you would have to recopy all logical volumes to make sure they

are consistent. The removal does provide a 20% + improvement in performance, however.

- For mirrored environments, make sure you are using the parallel scheduling policy.
- If any of the logical volume exists on more than one disk, stagger the partitions over the disks. This is automatically accomplished when the logical volume is created with a "maximum" INTER-POLICY.
- Balance the I/Os across the loop; you do not want to place all the workload on one half of the loop.
- In a multi-initiator environment, place the disks adjacent to the adapter that is using them.

Appendix A. How to Change an AIX Mirrored Disk

This is a full description of the changes that have to be made to the logical volume manager when a failed SSA disk drive that is part of a mirrored pair is changed.

8.3.2 Fixed Disk - Removing and Replacing a Fixed Disk

The following procedure will remove and replace a fixed disk on all levels of AIX on a RISC System/6000. Do not use this procedure if the ROOTVG disks are not mirrored and if any of the following logical volumes are on the drive:

- hd2
- hd3
- hd4
- hd5
- hd6
- hd8
- hd9var (at AIX 3.2 or higher).

If you are not using mirroring, and if any of these LVs are on this drive, then the user can use the *migratepv* command to move the LVs to another drive in the same volume group (this assumes the disk is still accessible). If this is not possible, the rootvg disks will need to be restored from a mksysb or Sysback backup file, or reinstalled from installation media. Information on migrating the 'rootvg' to another disk can be obtained by requesting document 2488 "Migrating rootvg to Another Disk Drive" from 800-IBM-4FAX.

Please make sure that the "/" and "/tmp" file systems have enough space prior to starting this procedure. If AIX 3.1.7 or earlier is being used, contact the AIX Support Center to order APAR IX20478 before following this procedure, or upgrade to AIX 3.2 or higher. Please do not use this document if the system is a "/usr client," "diskless client," or "dataless client."

Before running any commands listed here, we recommend that the following fixes be installed, if the system is earlier than AIX 4.2.1, to avoid a situation where a physical volume identifier (PVID) of 0000000000000000 is placed in the volume group descriptor area (VGDA):

<u>AIX Level</u>	<u>APAR</u>	<u>PTF/Fileset Level</u>
3.2.5	IX66215	U447739
4.1.5	IX66105	U447978 bos.rte.lvm 4.1.5.7
		or
		U448077 prpq.clvm 4.1.5.2
4.2.0	IX66226	U447812 bos.rte.lvm 4.2.0.12

Back up the system before any changes are made, so the original status can be restored if anything should go wrong.

8.3.2.1 Basic Steps

The basic steps to replacing a disk drive are as follows:

1. Deallocate all the physical partitions associated with the PVID of the drive concerned.
2. Remove the PVID from the volume group.
3. Remove the definition for the disk from the device configuration database.

If there is just one disk in the volume group, proceed to the next discussion, entitled "Steps to Take if There Is Only One Disk in the Volume Group" on page 154. Otherwise, proceed to the section titled "Determining Which Logical Volumes Will Be Affected" on page 154.

Steps to Take if There Is Only One Disk in the Volume Group

If the drive that is about to be replaced, or has already been replaced, is the only drive in the volume group, then simply run the following command, replacing "VGname" with the name of the volume group:

```
exportvg VGname
```

At this point, the user should be able to remove the disk definition, if it hasn't already been done. Instructions to do this step are included in the subsection "Removing the Disk Definition from the System" on page 156.

Determining Which Logical Volumes Will Be Affected

Every physical partition (PP) on the disk allocated to any logical volume (including file systems or paging devices) must be deallocated, either by moving the contents of those PPs to another disk or by removing them.

To determine what logical volumes have PPs allocated to that disk, run the following command, where 'hdisk#' is the 'hdisk' name:

```
lspv -l hdisk#
```

In some cases, the hdisk name no longer exists, and the disk is identifiable only by its 16-digit PVID. If this is the case, substitute the PVID for the hdisk name:

Example:

```
lspv -l 0123456789abcdef
```

If an error message is returned stating that the device ID is not in the device configuration database, then perform the following steps to resolve this problem.

1. Run the following command:

```
odmget -q "name=VGname and attribute=pv" CuAt > /tmp/filename
```

where "VGname" refers to the volume group.

2. Manually edit the file, and delete all the stanzas, except one. Each stanza will look similar to the following:

```
CuAt:
```

```
name = "rootvg"
```

```
attribute = "pv"
```

```
value = "0123456789abcdef0000000000000000"
```

```
type = "R"
```

```
generic = ""
rep = "sl"
nls_index = 0
```

3. Change the 32-digit value to show the PVID of the drive to remove, leaving the 16 trailing zeros intact. Also make sure the name line shows the volume group.
4. Run the following command:

```
odmadd /tmp/filename
```

The user should now be able to run **lspv -l PVID**, where PVID is the 16-digit identifier.

Deallocating Physical Partitions from the Disk

If there is another disk in the volume group with enough space to contain the partitions on the disk that is to be replaced, and the disk being replaced has not failed, then the user can use the **migratepv** command to move the data from the hdisk to be replaced to the other hdisk. It can be determined how much space is available on the other hdisk by running the following:

```
lspv hdisk# | grep FREE
```

To migrate individual logical volumes, run the following command:

```
migratepv -l LVname sourcehdisk# targethdisk#
```

There is more information on **migratepv** in the manual pages. If the original disk is still accessible, then logical volumes can only be migrated to hdisks in the same volume groups, .

If the user is unable to migrate the data, the partitions allocated to the drive to be replaced or PVID must be removed. If the LV is mirrored onto this drive, remove the LV mirror. To determine which LVs are mirrored, run the following command on each logical volume, substituting the name of each LV for LVname:

```
lslv LVname | grep COPIES
```

If the value for COPIES is 2 or higher, remove only the copy of the logical volume that is mapped to the old hdisks PVID. These copies can be removed by running the following command, substituting either 1 or 2 for # (where # is equal to COPIES-1) and substituting the disks or PVIDs from which to remove the copy for hdisk#:

```
rmlvcopy LVname # hdisk#
```

If the copy intended for removal resides on multiple drives, all those drives should be specified in the **rmlvcopy** command, separated by spaces. It can be determined if that copy contains more than one drive by running

```
lslv -m LVname
```

and looking under either the PV1, PV2, or PV3 columns (whichever is appropriate).

If running AIX 4.1, you should install IX71533, which will prevent **rmlvcopy** from removing the last accessible copy of a PP, should you fail to specify all the disks that make up one mirror. At AIX 4.2, this fix is IX71749. The file set versions that

contain this fix are bos.rte.devices 4.1.5.3, prpq.clvm 4.1.5.9, and bos.rte.lvm 4.2.1.8. The file set level can be determined with the following command:

```
lspp -l <fileset>
```

Note: prpq.clvm will be installed only in configurations using concurrent access HACMP. If the LV is not mirrored, the entire logical volume must be removed.

At this point, run the following:

```
lsvg -l VGname
```

Note the LV type and whether there is a mount point.

If the LV is associated with a file system mount point, unmount the file system and remove it by running the following commands, substituting the directory mount point for **mntpt**:

```
umount /mntpt
```

```
rmfs /mntpt
```

If the LV is of type 'paging,' run **lsps -a** to determine if the paging space is active. If so, it must first be deactivated. Run the following command:

```
chps -a n LVname
```

```
shutdown -Fr
```

The system will then reboot.

After rebooting, remove the paging space with the following command:

```
rmpps LVname
```

Remove any other logical volume with the following command:

```
rmlv LVname
```

Note: If the user tries to remove a logical volume that is serving as a dump device, the dump device must first be reassigned. The same is true if the copy of a LV that serves as a dump device is being removed (like a mirrored paging space). Do this with the following command:

```
sysdumpdev -Pp /dev/sysdumpnull (for the primary device)
```

```
sysdumpdev -Ps /dev/sysdumpnull (for the secondary device)
```

Removing the Identifier from the Volume Group

Using either the PVID or the hdisk name in place of hdisk# (depending on which of the two is used to obtain a list of LVs), run the following command:

```
reducevg VGname hdisk#
```

When using the PVID value, if the command complains that the PVID is not in the device configuration database, rerun **lsvg -p VGname** to determine if the PVID was actually removed. If it has not been removed, then proceed to the next step, "Removing the Disk Definition from the System."

Removing the Disk Definition from the System

If the disk is an SSA disk, determine which physical disk corresponds to the logical disk. One way is with the following commands:

Isdev -Cc disk -F name' 'connwhere

Isdev -Cc pdisk -F name' 'connwhere

See which SSA disk serial number coincides with the 'hdisk' to remove. If the 'hdisk' does not appear, or the user has been working with a PVID value up to this point, the 'pdisk' whose serial number does not coincide with any of the 'hdisks' is likely the disk to remove. Before physically removing the SSA disk, perform the following from within Diagnostics (run **diag**):

1. Select **Service Aids**
2. Select **SSA Service Aids**
3. Select **Set Service Mode**. This ensures that transactions in the loop will not be lost when the disk is removed.
4. Select the 'pdisk' determined above. The yellow check light on the drive will be turned on.
5. Remove the SSA disk. It is not necessary to power off any devices. If more than one SSA disk is being replaced, do not physically remove more than one disk at a time, because other disks in the loop may become inaccessible if you do.
6. Exit the diagnostic service aids.

If the user has been working with the hdisk# instead of a PVID value, now delete the hdisk name from the system configuration with the following command:

rmdev -dl hdisk#

If the user has a SSA disk, delete the pdisk from the system configuration by running the following:

rmdev -dl pdisk#

If the user has been working with the PVID value, rather than an hdisk name, make sure that it is removed from the device configuration database, replacing VALUE below with the 16-digit PVID followed by 16 zeros. Run the following command:

odmdelete -q value=VALUE -o CuAt

If it is a non-hotswappable SCSI disk, shut down the machine. An IBM hardware representative should remove and replace the drive. When completed, reboot the machine or, if replacing an SSA drive, run **cfgmgr**.

Adding the New Drive Back to an Existing Volume Group

1. Run **lspv** to ensure that the hdisk that was added is listed. Note whether there are holes in the list of hdisks (for example, hdisk0, hdisk1, hdisk3, and hdisk5'). AIX will name the replaced disk with the first available name (hdisk2 in this case). The same is true with pdisks, or any other device. If the hdisk is still not listed, the problem may be hardware. Run diagnostics.

2. Add the disk to the volume group with the following command:

extendvg VGname hdisk#

3. Recreate any logical volumes, paging spaces, file systems, or LV mirrors. Recreate the file system using one of the following:

mklv [-OPTIONS...] VGname #PPs hdisk# ...

crfs -v jfs -d LVname -m /<mountpoint> -A yes

or

crfs -v jfs -g VolumeGroup -a size=<#blocks> -A yes

4. Recreate a paging space with the following:

mkps -ans #PPs VGname hdisk#

5. Recreate a logical volume with the following:

mklv [-OPTIONS...] VGname #PPs hdisk# ...

6. Recreate an LV mirror with the following:

mklvcopy LVname newcopy# hdisk# ...

Use the **syncvg** command to resynchronize newly created logical volume copies. This can be done with one of the following commands:

syncvg -l LVname

syncvg -v VGname

syncvg -p hdisk#

If jfslog device has been recreated, format it for use by running the following command:

logform /dev/logLV (say "Yes" when prompted to destroy it)

See InfoExplorer or the manual pages for the above commands for correct syntax or more options. SMIT may also be used to perform the above tasks by running **smit <command>** to bring up the appropriate menu.

Appendix B. Booting from SSA Disks

The following conditions apply if you want to boot from an SSA disk drive:

- One of the following adapters is required:
 - 6214 SSA 4-port Adapter (4-D)
 - 6216 SSA Enhanced 4-port Adapter (4-G)
 - 6217 SSA RAID 4-port Adapter (4-I)
 - 6219 Micro-Channel SSA Multi-Initiator/RAID EL Adapter (4-M)
 - 6221 SSA Enhanced 4-port Adapter (4-G)
- The system cannot be booted off of a disk attached to an SSA RAID adapter, if the disk is part of a RAID logical unit (LUN). Booting from either the 6215 or the 6218 adapters is also not supported.
- Only certain models of the RS/6000 family can boot from SSA disks. Please refer to “RS/6000 Models that Support Booting from SSA” on page 159 for a list of machines and adapters that are supported.
- Models of the RS/6000 SMP family with Micro Channel interface must be at a firmware version of A9.23 or higher. (more details)
- AIX Version 4.1.4 or higher is required.
- If mirroring rootvg with SSA disks, we recommend that each copy of the boot logical volume be on a separate SSA loop.
- We recommend that an SSA loop that has a disk with the boot logical volume on it be attached through a single adapter on only a single machine.

RS/6000 Models that Support Booting from SSA

Table 19 shows which models of the RS/6000 support booting from the different SSA adapters:

Table 19. RS6000 Models that Support Booting from SSA Disks

Machine Type	Model	6214	6216	6217	6219
7012	G30	YES	YES	NO	NO
	G40	YES	YES	NO	YES
7013	591	YES	YES	YES	YES
	595	YES	YES	YES	YES
	J30	YES	YES	NO	NO
	J40	YES	YES	YES	YES
	J50	YES	YES	YES	YES
7015	R21	YES	YES	YES	YES
	R30	YES	YES	NO	NO
	R40	YES	YES	YES	YES
	R50	YES	YES	YES	YES

SMP Firmware Requirements

There are two versions of the Micro Channel SMP firmware that provide boot support for SSA; the first version is the absolute minimum required. We recommend the second version, as it addresses some problems encountered with the first version. The versions are:

Version 9.23 (0923)

Version 9.23 was the first version of the Micro Channel SMP firmware which allowed you to boot from SSA. However, the restriction was that only one SSA adapter from the system could be attached to the loop. If two or more adapters from the same system were attached to the loop, it would not work.

Version A9.23 (A923)

This is the second version of the firmware. This version fixed the multiple SSA adapter problem discovered with Version 9.23.

Appendix C. Replacing a Mirrored Disk - Documentation

This is a copy of the documentation related to the replacement of mirrored disks. This information is provided here to help you understand the process involved. The scripts mentioned below will be available on a CD-ROM and will also be available on a few Web sites (for example, the AIX fix distribution Web site). For complete documentation, please refer to these Web sites. At the time of publication, the Web sites were not available. Please contact your IBM representative for the locations of the script files source.

C.1 SSA spare tool Scripts

This edition applies to Version 1 Release 1 Modification 0; of the SSA spare tool and to all subsequent releases and modifications until otherwise indicated in new editions.

C.1.1 Working with the SSA spare tool Scripts

The SSA spare tool is a tool for managing serial storage architecture (SSA) storage networks. This tool functions with the family of SSA adapters, SSA disk units, and SSA enclosures that are developed by the IBM Storage System Division.

This document explains how to install and enable the scripts for the SSA spare tool in a mirrored environment.

This document is intended for storage network administrators who are familiar with SSA, the Logical Volume Manager (LVM), and general storage network concepts.

The SSA spare tool works with the LVM to automatically identify stale partitions or missing physical volumes in LVM mirrored volume groups. If the SSA spare tool finds any stale partitions or missing physical volumes, it will:

- Automatically resynchronize the stale partition, if this can be done without requiring a hardware replacement.
- Issue an e-mail message, indicating what has been done, or if a disk must be replaced.
- Logically replace the failed disk in the volume group with a preassigned spare disk.
- Resynchronize the new disk with the remaining disks in the volume group.
- Notify the user when it has logically replaced a failed disk and resynchronized the replacement disks with the other disks in the volume group.
- When the failed disk is physically replaced and the data is resynchronized, other scripts are provided to:
 - Move the data from the temporary spare disk to the new replacement disk.
 - Prepare the spare disk for future use as a spare if another disk failure occurs.

The purpose of these scripts is to maintain a protected, software mirrored, environment and automatically move data from a failed disk to a spare disk when

required. The SSA spare tool supports all disks that are contained in a 7133 Disk Subsystem. To enable the SSA spare tool you must install AIX scripts on all machines that share the SSA disks.

C.1.2 Operating Requirements

To ensure the proper operation of the SSA spare tool, you should ensure the following:

- All the disks in the storage networks that are monitored have the logical volumes 100% mirrored and are on separate physical volumes.
- All disks in the storage network are part of a volume group.
- The volume group is not varied on in concurrent mode.
- Nodes must have rsh access between them.
- Quorum should be turned off.
- These scripts do not support boot or dump logical volume types.
- No more than one physical volume can be stale for each volume group.
- Appropriate spare disks are available when needed.

Note: The SSA spare tool does not support RAID protected volume groups.

C.1.3 SSA spare tool Scripts

notify_server

This script is invoked by the AIX error notify function when LVM_SA_PVMISS, LVM_SA_STALEPP, LVM_SA_MWCWFAIL, or LVM_SA_WRTERR events occur.

invoke_replace_disk

This script is run, on the server, by the daemon, when the server has been notified by the client of an LVM error. The script checks to determine if a physical volume is missing. If a physical volume is missing, this script attempts to recover the physical volume by varying on the volume group. If the script is unable to initiate a recovery, it chooses an appropriate spare disk and invokes the *replace_disk* script.

replace_disk

This script creates the spare disk to replace the failed disk. It removes the failed disk from the active volume group and adds the spare disk to the volume group. The script then synchronizes the volume group so the volume group is returned to the level it was at, before the disk failure occurred.

ce_replace_disk

This script is run by the service representative. This script ensures that a spare is available and then replaces a failed disk with a new disk. This script then makes the spare available for the next failure.

cust_notify

This script is used to produce different types of customer notification, such as, e-mail and pager notification.

load_diskdb

This script must be run on the server to add the SSA loops and SSA disks to the Object Data Manager (ODM). Also, anytime that changes have been made to the system configuration, this script should be run to update the ODM.

For example, you would run **load_diskdb -h host1 host2 ...** to create the SSA disks and SSA loops for the ODM classes.

choose_spare

This script is used by the `invoke_replace_disk` script to choose a spare disk. This script attempts to get find a disk of the same size, and in the same SSA loop. If it cannot find a disk of the same size, the script tries to locate a disk with a larger size, than the failed disk, in the same SSA loop.

If the script cannot find a spare on the same loop, the script will try to locate a spare on one of the other SSA loops.

check_spare

This script checks to ensure that the spare disks are in a usable condition. This script creates a logical volume on a spare disk and then removes the logical volume. This tests and verifies the usable condition of a spare disk.

C.1.4 Installation Requirements

The SSA spare tool contains two different installp packages, one for servers, and one for clients. To use the SSA spare tool, you must install each of these installp packages.

The SSA spare tool requires certain software to be available on the host system. Ensure that you have the latest versions of the following device drivers. If you do not have the latest versions of these device drivers, download any of the following device drivers that need to be updated. You can search for these device drivers by name or search for the authorized program analysis report (APAR) number 71759 for AIX 4.1.

- devices.mca.8f97.diag
- devices.mca.8f97.com
- devices.mca.8f97.rte
- devices.ssa.disk.rte

Licenses

Ensure that you review and comply with the SSA spare tool license agreement.

C.1.5 Installing the SSA spare tool

To install the files that are on the SSA spare tool diskette:

1. Remove any previous versions of the SSA spare tool from your system.
2. Install the installp packages by using SMIT.
3. You must restart your system after installing these device drivers
4. To uninstall this product, use the `installp` command.

C.1.6 Setting Up the Daemons on the Client

To install the SSA spare tool on the server, perform the following steps.

1. Run the **odmadd PVMISS.add** command.
2. Run the **odmadd STALEPP.add** command.
3. Run the **odmadd MNCWFAIL.add** command.
4. Run the **odmadd WRTErr.add** command.
5. Make an entry in `/etc/services` to point to a port for the daemon. For example:
`ddaemon 37001/tcp #Disk`
6. Change the entry in the `notify_server` script to let the client know what the server's name is: `SERVER=<SERVERNAME>`
7. You may also, optionally, want to change the `/etc/syslog.conf` file. This will enable the server to log messages that are sent by using the socket into the `/tmp/syslog.out` file. An example of this change would be:

```
*.debug /tmp/syslog out *.daemon /tmp/syslog.out *.user /tmp/syslog.out
```

C.1.7 Setting Up the Daemons on the Server

1. To install the SSA spare tool on a server, perform the following steps.
2. Run the **odmadd PVMISS.add** command if the server is also a node.
3. Ensure that the `syslogd` daemon is running by entering the command: **# `Issrc -s syslogd`**
4. Make an entry in `/etc/services` to point to a port for the daemon. For example:
`ddaemon 37001/tcp #Disk`
5. Start the server daemon by entering the following command: **# `dd_daemon`**
This command should be an entry in the `/etc/inittab` file as follows:
`ssasapre:2respawn:/usr/lpp/hotsparring/dd_daemon` When you complete this command, the daemon will appear in the process table as `dd_serv`.
6. You may also, optionally, want to run `setup_services` script to perform the previous steps.
7. You may also want to change the `/etc/syslog.conf` file. This will enable the server to log messages that are sent by using the socket into the `/tmp/syslog.out` file. An example of this change would be: **.debug /tmp/syslog out *.daemon /tmp/syslog.out *.user /tmp/syslog.out*
8. Setup the e-mail notification by entering: **mail_id=user mail_domain=(your company domain)**
9. Run **load_diskdb -h host1 host2 ...** to create the SSA disks and SSA loops for the following Object Data Manager classes.
SSAdisks: `diskSN = "AC7E3ED2" loopNum = "1" function = "spare"`
SSALoops: `loopNum = "1" hosts = "lvm1 lvm2"`
This will create the following output files:
 - `/tmp/rdssa/invoke_replace.log`: This contains a running log of every time `invoke_replace_disk` was run.
 - `/tmp/rdssa/rdsk_invoke.out.$$`: This contains the output from the `replace_disk` command.

C.1.8 Handling Problems

- You should also ensure that your system meets the following conditions:

- Your machine meets the requirements that are stated in Installation Requirements and Operating Requirements.
- TCP/IP is installed and active, and you can communicate with the host systems by using TCP/IP.
- You have applied the appropriate SSA PTFs that are needed to support the system you are using.
- Before you report a problem, you should gather the following information:
 - The version of AIX that you are using.
 - The error logs in /tmp/rdssa.

Appendix D. Tuning RAID 5 Arrays and Using Fastwrite Cache

Two little-known parameters exist for SSA RAID 5 arrays which should be modified for performance: `queue_depth` and `max_coalesce`.

- `queue_depth`: Specifies the maximum number of commands that the SSA disk device driver dispatches for a single disk drive for an `hdisk`.
- `max_coalesce`: The maximum number of bytes that the SSA disk device driver attempts to transfer to or from an SSA logical disk in one operation.

You can change these attributes with `chdev` or via `smitty chgssardsk`. Set `queue_depth` from the default of 3 to $2N$ or $3N$ for a $N+P$ array. This maximizes the chance of reading data from each component of the array in parallel. Set `max_coalesce` to $64 \text{ KB} \times N$. The default is `0x20000` which corresponds to two 64 KB strips (and applies to a $2+P$ array). So for an $7+P$ array of eight disks, set the value to `0x80000`. Currently this value shouldn't exceed `0xFF000` (1 MB) even though the SSA device driver allows values up to 2 MB.

If you are using the RAID EL adapters with Fast Write cache, be sure to set the fast write attribute to on (the default is off) for the `hdisks` you want to use the cache.

This use of this cache is split between all `hdisks` with fast write on, so don't indiscriminately turn it on for all `hdisks` on the adapter. The fastwrite attribute is set via `smitty ssaraid` or `smitty chgssardsk`.

Finally, the `dbmw` attribute of the SSA adapter can be changed, although I have little information on any performance benefits this might provide. The manual states:

"Holds the size of the DMA area the SSA adapter device driver for this adapter will use. You can use the `chdev` command to change the value of this attribute. The default value provides a DMA area that is large enough to allow the adapter to perform efficiently, yet allows other adapters to be configured. The default value is practical for normal use. If, however, a particular SSA device attached to the using system needs large quantities of outstanding I/O to get best performance, a larger DMA area might improve the performance of the adapter."

Appendix E. Disk Striping and Sizing

E.1 Stripe Size

With AIX we can create striped logical volumes (LVs) with a stripe size from 4 KB to 128 KB, but we can't mirror them. We can spread a LV across many disks, and then mirror the LV. In this case the stripe size is the same as the physical partition size (in the case of 4.5 GB drives the stripe size is 8 MB). To distinguish between the two cases, we will call the first one *LV striping* and the second one *LV spreading*.

The best stripe size is difficult to determine as trade-offs exist. There are two appropriate ways to evaluate this: one uses queueing theory, such as a Best1 analysis; the other uses testing of your application with various stripe sizes. Both options take considerable time, money, and effort. Here we try to provide some insight into the issues.

AIX has a read-ahead feature such that when a program reads sequential pages of a file, the virtual memory manager schedules additional sequential reads of the file. This feature, along with a small stripe size, causes (for the case of a single user) reads to occur simultaneously over several disks; thus increasing the throughput to several times that of a single disk. With many users, however, smaller stripe sizes can cause performance to be worse as we will explain later. We also make some assumptions about disk performance that vary depending upon disk type, but are good enough for the purposes of examining stripe size.

Disks have several limiting factors. Seek (the time to move the head to the appropriate track) and latency (the time to rotate the disk to the sector we want to read) are commonly referenced and measured in milliseconds (ms). Let's assume a seek+latency time of 12ms. The media or data rate (how fast we can read/write data given that we don't have to worry about seek and latency) is another limiting factor, which is measured in megabytes per second (MB/s) - approximately the same as kilobytes per millisecond (KB/ms). We'll assume a data rate of 8 MB/s or 8 KB/ms. A less commonly known limit is the maximum number of I/O/second, that is, given a particular disk and assuming a specific mix of random/sequential and specific I/O size, a disk can perform only so many I/O/s. If all I/O to a disk is random, then a disk can only perform only 80 I/O/s (assuming seek+latency of 12 ms and reads of 4 KB). Assuming all sequential I/O, with I/O sizes of 4 KB, then a disk can perform 2000 I/O/s. Obviously we want to avoid moving the head; that is, we want to avoid "start I/O".

Now, how could smaller stripe sizes and many users cause performance to get worse? Consider a simple two user and two disk scenario with each user performing a 64 KB I/O when using a 32 KB stripe size (assume sequential I/O and assume that the I/O occurs on two stripes only). In this case, each user will have to perform two start I/Os, one to each disk. This will cause two seeks per disk for four start I/O total. For two users using 8 MB stripe sizes, we will either get two seeks on one disk or one seek per disk on two disks, for a total of two start I/O. In either case with 8 MB stripe sizes, we have we have 1/2 the start I/O. Once a disk's start I/O capacity is exceeded, performance suffers. To summarize this concept:

Start I/Os increase proportionally with the number of I/Os overlapping stripe boundaries. Smaller stripes increase the probability that an I/O overlaps stripe boundaries.

The reality is many users with many disks and a mix of I/O sizes and types. Thus queueing theory, statistical modeling, or actual testing are the only ways to determine the optimum stripe size.

The key point for efficient use of the disk subsystem is the even distribution of the I/O across as many spindles as possible. Assuming your I/O patterns are random across your LV structures (or data files in Oracle terminology), any stripe size that spreads the LVs across the disks will evenly distribute the I/O among the disks. This indicates that a wide range of stripe sizes may be equivalent and optimal for a given environment - that is, stripes greater than the average I/O size but small enough to spread the data across the disks.

Oracle provides further guidance in setting up logical volumes. For example, Oracle recommends putting data tables and indices on different disks. This assures that the index lookup and the data read occur on different disks, resulting in better response time. This spreads data files across one set of disks and indices across another. Since the I/O to the indices typically uses more bytes than the data, we must often put the indices on more disks than physically necessary. For example, if all the indices take 4 GB, they might be put on three 4.5 GB disks as the I/O rates to the indices are higher, thus eliminating a bottleneck and helping balance the I/O between the two groups of disks.

IBM could have invested in providing LV striping and mirroring function in the LVM; however, given that "even distribution of the I/O across as many spindles as possible" is the goal, and that LV spreading accomplishes this, IBM has not needed to invest in this functionality.

E.2 Disk Sizing

Disk sizing consists of more than buying enough disks to hold the data. One must also assure that performance and availability goals are met. Trade-offs exist between performance, availability and cost. For example, RAID-5 implementations provide relatively lower cost, but reduced performance: a single write from the application generates four I/Os on a RAID array. Mirroring, or RAID-1, provides availability and high performance but requires buying twice as many disks.

Other factors are important. Additional memory (and clever programming) can reduce I/O. Indeed, some say the best I/O is no I/O. Oracle uses the SGA to reduce I/O, while AIX uses extra memory to cache file system I/O. The AIX device drivers and disk also use techniques to improve disk performance.

Keep in mind that a bottleneck is not necessarily bad. While you will always have bottlenecks, and removing one creates another, removing a bottleneck improves performance. Other factors provide additional cushion: increased memory will reduce the I/O load. Also, you may be experiencing SCSI bottlenecks that won't exist with SSA, and new disks probably have better speeds and feeds.

Appendix F. Laying Out an Oracle Database on AIX Disks

Designing a data layout strategy for Oracle databases requires working with the Oracle database administrator, as the AIX system administrator's volume group (VG) and logical volume (LV) layout strategy affects the DBA's strategy. For example, Oracle can stripe data between data files (generally not recommended for AIX), but you can also spread LVs across disks using LV striping, LVs with a maximum physical volume (PV) allocation policy, or RAID arrays. Striping twice would add an unnecessary layer of overhead in translating from a file to its physical location on disk.

Oracle recommends splitting the drives into five to eight groups, for AIX and Oracle binaries, tables, indexes, redo logs, rollback segments and archive logs (and possibly temporary space for Oracles). Doing so helps reduce head movement and fragmentation, AIX is usually in rootvg. The Oracle binaries can exist in rootvg or another VG. The database can be in one or more volume groups, independent of the five disk groups. Usually, one VG simplifies administration, but AIX limits volume groups to 32 disks. When working with the DBA, the ability to translate between AIX and Oracle terms is helpful.

An Oracle database usually comprises one or more table spaces: a system table space and usually one or more user table spaces. Each table space consists of tables and indexes (neither of which can span table spaces). Table spaces are placed into data files. From an AIX viewpoint, a data file is either a raw LV or a JFS file. The choice between raw LV and JFS is a decision to be made early in the data layout process. JFS provides availability features, although with a performance cost. It journals the metadata, so that AIX can rebuild the file system to a consistent state in the event of a system crash. Because it journals the metadata in the jfslog, it can be an I/O bottleneck in itself. Journaling also places a load on the CPU and the JFS metadata structures take up disk space. However, JFS does offer the advantage of allowing the use of AIX tools to back up the data. Customers who use raw LVs generally use the Oracle back up utilities to back up the data, although Sysback or dd can be used to back up raw LVs.

In choosing the number and size of drives for each group, it is necessary to consider not only the size of the structures (tables, indexes, and the like), but also the I/O rate and the access patterns (random, sequential or, more probably somewhere in between). A typical drive can perform about 80 random 4 KB I/O per second, or about 1800 sequential 4 KB I/O/per second. A file monitor (filemon) analysis of an existing workload can be instrumental in determining the I/O load for LVs and PVs, and the access patterns. Oracle has tools that are also useful in determining the I/O load. Without an existing workload analysis, one must make an educated guess. Indexes generally have higher I/O loads than the other groups. As a result, it's often better to have more smaller disks for indexes (often with lots of unused space on the drives) to provide the I/O bandwidth necessary for balanced performance. The goal is to spread the workload equally among all the disks. The benefit of putting the drives into the five groups is twofold: to reduce head movement, and to allow multiple disk actuators to satisfy a transaction's I/O requests at the same time.

Once the number of drives for each group is determined, there is then the issue of physical attachment, including the number of disk adapters, the number of SSA

loops, and the number of drives per loop. Here again, the goal is to balance the I/O among the adapters and loops (and provide any availability characteristics desired.) RAID arrays further complicate the issue, as one hdisk comprises several physical disks (pdisks in the case of SSA). For RAID arrays other than SSA, the physical disks may be hidden from AIX; nevertheless, the physical to logical mapping is generally known. RAID-5 arrays provide availability and also usually spread the workload evenly among the disks in the array. Generally, RAID-5 is appropriate for data that is seldom changed; mirroring costs more but offers better performance. Once the physical attachment is completed, we assign disks to volume groups and LVs to disks in the volume groups. For SSA, we put the disks with highest read rates at the end of the adapter port domains.

For logical volume layout, one generally uses the maximum physical volume allocation policy to spread the LVs across all the disks in the group (for example, all the index LVs are spread across all the disks in the index group) and then mirror them for availability. Spreading the LVs across the disks in the group balances the load among these disks. If you do mirror logical volumes, those LVs with mirror-write consistency (MWC) turned on should be placed on the outer edge of the disk, as writes to LVs with MWC on also write to the mirror write consistency cache, on the outer edge of the disk. (The outer edge of the disk usually has a faster data transfer rate than the center or inner edge.)

In general, one should turn on MWC only for redo logs and archive logs. The purpose of mirror write consistency is to assure that mirrored data is consistent in case a system crash occurs. If there is a system crash, Oracle uses the redo logs and possibly the archive logs to bring the database back to a consistent state (so we don't have to worry about the tables and indexes).

Oracle recommends striping and mirroring. AIX can create striped LVs (with the strip/stripe size equal to 4 KB to 128 KB but currently doesn't support mirroring these LVs. AIX can create striped LVs, with the strip/stripe size equal to the physical partition size, and mirror them. (We call these *spread LVs* to distinguish them from the striped LVs with the smaller stripe size.) Oracle recommends striping for one purpose: to balance the I/O load among the disks. Thus, an appropriate strip/stripe size is small enough that we can spread the LV across all the disks we want.

(The phrase "strip/stripe" is used because the documentation for different layout mechanisms differs in usage, as for LVM, SSA RAID, 7135 RAID, and other disk subsystems, including OEM.)

We don't, however, want the strip/stripe size to be smaller than our average I/O size, because one I/O at the application level could generate more than one I/O at the disk level. This is potentially bad because, with a multiuser workload, it can drive the rate of I/O per second beyond the capacity of the disk.

With RAID-5 arrays, one array per group is generally adequate, resulting in one hdisk. Assuming that the data strips are assigned round robin from each physical disk (this is the case for SSA, and the data strips are 64 KB in size), then the LVs with the highest expected I/O load would be placed in the center of the hdisk and the other LVs would be placed around the center to minimize seek times.

Oracle documentation recommends a tuning strategy that addresses the following in the order shown:

1. Business rules
2. Data design
3. Application design
4. Logical structure of the database
5. Structured query language (SQL)
6. Access paths
7. Memory allocation
8. I/O and physical structure
9. Resource contention
10. The underlying platform

Note that Steps 1 through 6 concern the architecture, which in turn affects how one tunes in Steps 7 through 10. Also, experienced tuners say the most benefit comes from tuning the SQL and the next from other changes in Steps 1 through 6.

These points are very important for two reasons: If performance is not adequate, it is best to begin by examining the SQL. Attempting to tune the disks first could waste time if you later decide to tune the SQL, modify the database structure, add or delete indexes, or take other actions that result in different I/O patterns. Second, to improve performance further, it is best to look at Steps 1 through 8 before examining the disk layout.

Memory, or disk cache tuning is slightly related to disk layout. Oracle, believing that it can perform disk cache management better than operating systems, keeps its own cache in the SGA. If you are using JFS rather than raw LVs, this can conflict with the JFS cache in the AIX memory. If so, and Oracle doesn't find the data in its cache, AIX then looks in the JFS cache. The data probably won't be in the JFS cache if it isn't in the SGA, but that's not important. What is important to ensure that the SGA remains in memory and is not paged out to page space. To ensure that the page replacement algorithm steals JFS cache pages, rather than SGA pages, use the **vm tune** command to lower maxperm toward numperm. The difference is the amount of working storage (think SGA and programs) AIX can steal, rather than persistent storage (think JFS disk).

Finally, if you are using AIX 4.2 and JFS rather than raw LVs, make sure that IX67978 is installed, as this fixes a problem with synchronous writes that will otherwise affect performance.

Appendix G. Backing Up and Restoring the Operating System in AIX 4.2

This document details the commands for creating, verifying, and restoring a system backup in AIX 4.2.

The IBM AIX version of UNIX is different from other UNIXs in two main ways: the object database manager (ODM) and the logical volume manager (LVM). It is because of the ODM and the LVM, as well as the ability to have multiple volume groups, that a complete system archive made with **cpio** or **tar** will not restore properly. Attempting to restore such an archive on a running system can potentially crash the machine.

G.1 MKSYSB

G.1.1 MKSYSB Tape Images

Creating a mksysb on a tape drive will result in a bootable tape. There will be four images on the tape, and the fourth image will contain only rootvg JFS mounted file systems. The target tape drive must be local to produce a bootable tape.

The following is a description of mksysb's four images.

(The bosboot image, the Mkinsttape image and the dummy TOC image must have a block size 512 KB. Rootvg data has its block size defined by the device.)

Image 1: The bosboot image contains a copy of the system's kernel and specific device drivers allowing the user to boot from this tape.

blocksize: 512 KB
format: raw image
files: kernel device drivers

Image 2: The mkinsttape image contains files to be loaded into the RAM file system when booting in maintenance mode:

blocksize: 512 KB
format: backbyname
files: ./image.data, ./tapeblksz, ./bosinst.data and commands

Image 3: The dummy image contains a single file containing the words "dummy toc". This image is used to make the mksysb tape contain the same number of images as a BOS Install tape.

Image 4: The rootvg image contains data from the rootvg volume group (mounted JFS file systems only):

blocksize: determined by tape drive configuration on create
format: backbyname
files: rootvg, mounted jfs filesystems

WARNING: If the device block size is set to 0, mksysb will use a hard-coded value of 512 KB for the fourth image. This can cause the create and restore to take 5 to 10 times longer than expected.

Creating a 'mksysb' to a file will create a non-bootable single image tar archive containing ONLY rootvg jfs mounted, file systems.

G.1.2 Writing to a Tape Drive

A mksysb tape is bootable only when it has been written to a tape drive.

The procedure is as follows:

1. Using SMIT

Fill in the correct device name to be used. Press Enter to start the backup. If more than one tape is required, SMIT will prompt the user to change the tape.

```
smit mksysb
Backup DEVICE or FILE           [/dev/rmt#]
Create MAP files? ..... no
EXCLUDE files? ..... no
List files as they are backed up? ..... no
Generate new /image.data file? ..... yes
EXPAND /tmp if needed? ..... yes
Disable software packing of backup? .....no
Number of BLOCKS to write in a single output [ ]
                (Leave blank to use a system default)
```

2. Enter the following from the command line:

```
mksysb -i /dev/rmt# 2>/tmp/mksysb.err
```

G.1.3 Creating a MKSYSB

The procedure is as follows:

1. The file system /tmp must have at least 12 MB free prior to creating the mksysb.
2. Unmount all nonessential file systems to reduce the size of the mksysb backup. Try to limit the mksysb to only system file systems if possible.
3. Note how many volume groups the system has, what disks they are located on, and the location of each disk.

Hdisk numbers are not retained.

```
lsvg
```

```
lsvg -p <vgname>
```

```
lsdev -Cc disk
```

G.1.4 Writing to a File

The procedure is as follows:

1. From SMIT

A mksysb image file is created in the same manner as the bootable tape described above. The only difference is:

```
smit mksysb
```

```
Backup DEVICE or FILE                [/path/file]
```

```
Number of BLOCKS to write in a single output [ ]
```

(Leave blank to use a system default)

2. Enter the following from the command line:

```
mksysb -i /path/file 2>/tmp/mksysb.err
```

G.1.5 Verifying

The only method to verify that a system backup will restore correctly with no problems is to actually restore the following according to each company's disaster recovery plan.

To minimize problems due to tape media damage, the following tests may be performed:

WARNING: These tests verify only that the tape media can be read and do not guarantee that a mksysb will be restored successfully.

Data verification

1. Using SMIT

```
smit lsmksysb
```

```
DEVICE or FILE                        [/dev/rmt#]
```

```
Number of BLOCKS to read in a single input [ ]
```

(Leave blank to use a system default)

Type or select values and press Enter only after making the desired changes.

2. Enter the following from the command line:

```
tctl -f /dev/rmt# rewind
```

```
restore -s4 -Tvqf /dev/rmt#.1 > /tmp/mksysb.log
```

G.1.6 Boot Verification

The only way to verify that the mksysb tape will boot successfully is to bring the machine down and boot from the tape. No data needs to be restored.

NOTE: To boot a keyless system or a system with multiprocessors in service mode, consult the model's documentation or call 1-800-call-aix for assistance.

WARNING: Having the PROMPT field in the bosint.data file set to "no" causes the system to begin the mksysb restore automatically, using preset values with no user intervention.

If the state of prompt is unknown, this can be set during the boot process. After answering the prompt to select a console during the boot up, a rotating character will be seen in the lower left of the screen. As soon as this character appears, do the following:

1. Check PROMPT

To check a mksysb tape to see how the prompt is set, run the following while in normal mode:

```
chdev -l rmt# -a block_size=512
tctl -f /dev/rmt# rewind
cd /tmp
restore -s2 -xvqf /dev/rmt#.1 ./bosinst.data
```

Check the prompt field in the control_flow stanza.

2. Boot in service mode

1. If the system has a key, turn it to service. If the system is keyless, consult the model's documentation or contact 1-800-call-AIX for assistance.
2. Insert the mksysb tape into the tape drive.
3. Reboot the system (**shutdown -Fr**).

3. The system should now boot from the mksysb tape.

Note: Multiprocessor systems display a maintenance menu called a Bump Menu. This is a hardware menu. For more information, consult the model's documentation or contact 1-800-call-AIX for assistance.

4. A message should appear that says: "press F1 or 1 for Console." Press the F1 key for a graphics display and the 1 key for an ASCII display, and press enter.
5. Press 1 for English, if asked.
6. The Installation and Maintenance menu should display.
The system has booted successfully.
7. Turn the key to normal (or follow the nonkey instructions for rebooting into normal mode), and reboot the system.

G.1.7 Restoring a MKSYSB

Be sure to select all physical volumes required for the root volume group. This is especially important if there is mirroring. Mirrored disks must be selected at the time of installation or an error message such as the following will appear:

"not enough physical volumes."

1. Boot in service mode

1. If the system has a key, turn it to service. If the system is keyless, consult the model's documentation or contact 1-800-call-AIX for assistance.
2. Insert the mksysb tape into the tape drive.
3. Reboot the system by running the command **shutdown -Fr** or by pressing the **Reset** button twice.

2. The system should now boot from the mksysb tape.

If the system will not boot, remove the mksysb tape and boot from install media such as a CD-ROM or tape.

Note: Multiprocessor systems display a maintenance menu called a Bump Menu. This is a hardware menu. For more information, consult the model's documentation or contact 1-800-call-AIX for assistance.

3. Next, at the message stating "press F1 or 1 for Console," press the F1 key for a graphics display (or 1 for an ASCII display), and press Enter.
4. Press 1 for English if asked.
5. The Installation and Maintenance menu should appear. If the system was booted from media other than the mksysb tape, the tape can now be inserted into the tape drive. Follow the steps listed Section, "Restore Menus" on page 180.
6. After the restoration is complete, turn the key to normal. If the system is keyless, this is automatic.
7. The system will reboot once. The system displays events from inittab, and a login prompt should appear.

Note: If the system has volume groups other than rootvg, there may be error messages prior to a login prompt. Ignore these messages unless they prevent you from obtaining the prompt.

Total restore time varies from system to system. A good rule of thumb, provided the device block size is not set to 0, is twice the amount of time it took to create the mksysb.

8. If the block size of the source tape drive was 0, the mksysb would have been created with a blocksize of 512 KB. Restore time is 5 to 10 times longer than can be normally expected.
9. If there are other disks which contain other volume groups, they will need to be imported into the newly created ODM. No references to the other volume groups exist prior to this step.
 1. Match the newly labeled hdisk numbers to the appropriate SCSI location IDs. Do not assume that the hdisks will be at the same location. The disks are renumbered from lowest to highest location number and labeled accordingly.

Enter:

lsdev -Cc disk

2. Import each volume group into the new ODM. The following must be run for each non-rootvg volume group, although, only one disk per volume group need be selected:

importvg -y vname hdisk#

For example, if there is another vg named "data" and it resides on "hdisk3" and "hdisk4," enter:

importvg -y data hdisk3

3. Activate each non-root volume group by entering:

varyonvg vname

4. Restart the system to resynchronize the ODM and mount all file systems.

Enter:

shutdown -Fr

Note: If the system is not to be reboot again, use "mount -a" in place of "shutdown". This will mount all the new file systems. However, a system reboot should be scheduled as soon as possible to be sure that everything is synchronized.

Appendix H. Special Notices

This publication is intended to help UNIX storage administrators to manage and monitor their installed SSA disk subsystems. The information in this publication is not intended as the specification of any programming interfaces that are provided by SSA. See the PUBLICATIONS section of the IBM Programming Announcement for SSA 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, 500 Columbus Avenue, Thornwood, NY 10594 USA.

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 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 performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

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

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

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

IBM ®
RS/6000
AIX

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc.

Java and HotJava are trademarks of Sun Microsystems, Incorporated.

Microsoft, Windows, Windows NT, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

Pentium, MMX, ProShare, LANDesk, and ActionMedia are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

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

Appendix I. Related Publications

The publications listed are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

I.1 International Technical Support Organization Publications

For information on ordering these ITSO publications see "How to Get ITSO Redbooks" on page 187.

- *Clustering and High Availability Guide for IBM Netfinity and IBM PC Servers*, SG24-4858 (*Introducing IBM and Wolfpack*, SG24-4858)
- *Exploiting SSA in Sun Solaris Platform Environments*, SG24-5083 (available at a later date)
- *Implementing PC Serveraid SCSI and SSA RAID Disk Subsystems*, SG24-2098
- *OS/2 Warp Server Integration Guide for IBM Netfinity and IBM PC*, SG24-2125
- *Novell Intranetware and Bordermanager For IBM Netfinity and IBM Servers*, SG24-2145
- *A Practical Guide to Serial Storage Architecture for AIX*, SG24-4599

I.2 Redbooks on CD-ROMs

Redbooks are also available on CD-ROMs. **Order a subscription** and receive updates 2-4 times a year at significant savings.

CD-ROM Title	Subscription Number	Collection Kit Number
System/390 Redbooks Collection	SBOF-7201	SK2T-2177
Networking and Systems Management Redbooks Collection	SBOF-7370	SK2T-6022
Transaction Processing and Data Management Redbook	SBOF-7240	SK2T-8038
Lotus Redbooks Collection	SBOF-6899	SK2T-8039
Tivoli Redbooks Collection	SBOF-6898	SK2T-8044
AS/400 Redbooks Collection	SBOF-7270	SK2T-2849
RS/6000 Redbooks Collection (HTML, BkMgr)	SBOF-7230	SK2T-8040
RS/6000 Redbooks Collection (PostScript)	SBOF-7205	SK2T-8041
RS/6000 Redbooks Collection (PDF Format)	SBOF-8700	SK2T-8043
Application Development Redbooks Collection	SBOF-7290	SK2T-8037

I.3 Other Publications

These publications are also relevant as further information sources:

- *7133 SSA Subsystem: Operator Guide*, GA33 3259
- *7133 SSA Subsystem: Hardware Technical Information*, SA33 3261
- *7133 Models 010 and 020 SSA Subsystems: Installation Guide*, GA33 3260
- *7133 Models 500 and 600 SSA Subsystems: Installation Guide*, GA33 3263
- *7133 SSA Subsystem: Service Guide*, SY33 0185

- *SSA 4 Port Adapter and Enhanced SSA 4 Port Adapter: Technical Reference*, S31H 8612
- *SSA RAID Subsystems: Planning SSA RAID Subsystems*, GA33-3271
- *SSA Adapters: User's Guide and Maintenance Information*, SA33-3272
- *MicroChannel SSA RAID Adapters Technical Reference*, SA33-3270
- *PCI SSA RAID Adapters Technical Reference*, SA33-3225
- *AIX Versions 3.2 and 4 Performance Tuning Guide*, SC23-2365

How to Get ITSO Redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, CD-ROMs, workshops, and residencies. A form for ordering books and CD-ROMs is also provided.

This information was current at the time of publication, but is continually subject to change. The latest information may be found at <http://www.redbooks.ibm.com/>.

How IBM Employees Can Get ITSO Redbooks

Employees may request ITSO deliverables (redbooks, BookManager BOOKs, and CD-ROMs) and information about redbooks, workshops, and residencies in the following ways:

- **Redbooks Web Site on the World Wide Web**

<http://w3.itso.ibm.com/>

- **PUBORDER** – to order hardcopies in the United States

- **Tools Disks**

To get LIST3820s of redbooks, type one of the following commands:

```
TOOLCAT REDPRINT
TOOLS SENDTO EHONE4 TOOLS2 REDPRINT GET SG24xxxx PACKAGE
TOOLS SENDTO CANVM2 TOOLS REDPRINT GET SG24xxxx PACKAGE (Canadian users only)
```

To get BookManager BOOKs of redbooks, type the following command:

```
TOOLCAT REDBOOKS
```

To get lists of redbooks, type the following command:

```
TOOLS SENDTO USDIST MKTTOOLS MKTTOOLS GET ITSOCAT TXT
```

To register for information on workshops, residencies, and redbooks, type the following command:

```
TOOLS SENDTO WTSCPOK TOOLS ZDISK GET ITSOREGI 1998
```

- **REDBOOKS Category on INEWS**

- **Online** – send orders to: USIB6FPL at IBMMAIL or DKIBMBSH at IBMMAIL

Redpieces

For information so current it is still in the process of being written, look at "Redpieces" on the Redbooks Web Site (<http://www.redbooks.ibm.com/redpieces.html>). 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.

How Customers Can Get ITSO Redbooks

Customers may request ITSO deliverables (redbooks, BookManager BOOKs, and CD-ROMs) and information about redbooks, workshops, and residencies in the following ways:

- **Online Orders** – send orders to:

In United States
In Canada
Outside North America

IBMMAIL

usib6fpl at ibmmail
caibmbkz at ibmmail
dkibmbsh at ibmmail

Internet

usib6fpl@ibmmail.com
lmannix@vnet.ibm.com
bookshop@dk.ibm.com

- **Telephone Orders**

United States (toll free)
Canada (toll free)

1-800-879-2755
1-800-IBM-4YOU

Outside North America
(+45) 4810-1320 - Danish
(+45) 4810-1420 - Dutch
(+45) 4810-1540 - English
(+45) 4810-1670 - Finnish
(+45) 4810-1220 - French

(long distance charges apply)
(+45) 4810-1020 - German
(+45) 4810-1620 - Italian
(+45) 4810-1270 - Norwegian
(+45) 4810-1120 - Spanish
(+45) 4810-1170 - Swedish

- **Mail Orders** – send orders to:

IBM Publications
Publications Customer Support
P.O. Box 29570
Raleigh, NC 27626-0570
USA

IBM Publications
144-4th Avenue, S.W.
Calgary, Alberta T2P 3N5
Canada

IBM Direct Services
Sortemosevej 21
DK-3450 Allerød
Denmark

- **Fax** – send orders to:

United States (toll free)
Canada
Outside North America

1-800-445-9269
1-800-267-4455
(+45) 48 14 2207 (long distance charge)

- **1-800-IBM-4FAX (United States) or (+1) 408 256 5422 (Outside USA)** – ask for:

Index # 4421 Abstracts of new redbooks
Index # 4422 IBM redbooks
Index # 4420 Redbooks for last six months

- **On the World Wide Web**

Redbooks Web Site <http://www.redbooks.ibm.com>
IBM Direct Publications Catalog <http://www.elink.ibm.com/pbl/pbl>

Redpieces

For information so current it is still in the process of being written, look at "Redpieces" on the Redbooks Web Site (<http://www.redbooks.ibm.com/redpieces.html>). 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.

LIST OF ABBREVIATIONS

ADSM	ADSTAR Distributed Storage Manager	VG	volume group
CPU	central processing unit	VPD	vital product data
DASD	direct access storage device		
FW	fast write		
HACMP	high availability for cluster multiprocessing		
HAGEO	high availability geographic cluster		
HSM	hierarchical file system		
HW	hardware		
ID	identification, identity		
I/O	input or output		
ioctl	I/O control		
ITSO	International Technical Support Organization		
JBOD	just a bunch of disks		
JFS	journaled file system		
LED	light-emitting diode		
LUN	logical unit, logical unit number		
LV	logical volume		
LVM	logical volume manager		
MSCS	Microsoft cluster server configuration		
MWC	mirror write consistency		
OS	operating system		
PCI	personal computer interconnect		
PP	physical partition		
PTF	Program Temporary Fix		
PV	physical volume		
RAID	redundant array of independent disks		
RPC	remote procedure call		
RSM	remote systems management		
SCSI	small computer system interface		
SLIC	serial link interface controller		
SMP	symmetrical multiprocessor		
SMIT	system management interface tool		
SSA	serial storage architecture		
StorX	IBM Serial Storage Expert		
Sysback	AIX System Backup and Recovery 6000		
TCP/IP	transmission control protocol/ Internet protocol		
UID	unique identifier		

Index

A

- abbreviations 191
- acronyms 191
- Adapter Types 146
- ADSM 128
- advantages of SSA 3
- AIX SMIT Interface 23
- ANSI 1
- Availability 109
 - AIX 112
 - Disk Mirroring 114
 - HACMP 113
 - HAGEO 113
 - Remote Mirroring 114

B

- Backup 110
 - AIX System Volume Groups 126
 - Customer Volume Group 126
- Backup and Recovery 123
- Backup Media 124
- Backup Strategy 123
- Basic SSA Configuration 1
- Booting from SSA Disks 159

C

- Clustering and High Availability Support 121
- Configuration Rules 141
- Configuring Drives 22
- Customized StorX Viewer 75

D

- Device Drivers 21
- Disaster Recovery Manager 134
- Disk Fencing 22
- Disk Mirroring 138
- disk mirroring 114
- Disk Properties 102
- Disk Striping and Sizing 169
- Disks 7

E

- Enhanced Loop Adapters 22
- Event Monitor 60, 76

F

- Fairness Algorithm 141
- fast write cache 148
- Fiber extender 148
- filemon 41
- Frame Multiplexing 140

G

- Guidelines for improving I/O Performance 151

H

- HACMP 113
- HAGEO 113
- Hot Spare Disk Definition 93
- How many adapters? 143

I

- I/O performance 137
- iostat 38

L

- Levels of Availability 110

M

- Management Set Properties 74
- Managing SSA Disks 19
 - AIX Environment 19
 - HP-UX Environment 44
 - StorX 51
 - Sun Solaris Environment 43
 - Windows NT Environment 45
- Managing SSA Disks in Other Environments 49
- Maymap 27
- MKSYB 175
- Monitoring and Managing
 - Complex Configuration 80
 - Diagnosing with StorX 102
 - Disk Problems in a RAID 5 Configuration 97
 - hdisk Failure 84
 - Shared Configuration 78
 - SSA Adapter Failure 90
 - SSA Open Loop Problem 88
 - SSA RAID Array Creation 92

N

- Non-RAID 147

O

- Octopus 116
- Online Service Monitoring 13
- Oracle Database 171
- Outage 109

P

- Pathlight Adapters 49
- PC Clustering 118
- Performance Toolbox 41

R

- RAID 113, 167

- RAID Arrays 113
- RAID Volume Group Creation 96
- RAID-0 147
- RAID-1 147
- RAID-5 147
- Recovery 109
- remote mirroring 114
- Removing and Replacing a Fixed Disk 153
- Replacing a Failed Disk 115
- Replacing a Mirrored Disk 161
- Replacing a Mirrored Disk - Script 161
- Restoring a Customer Volume 127
- Restoring the AIX Operating System 127
- Rules
 - adapters 19
 - IBM 7190 4
 - SBus adapter 19
 - SSA 3
- Rules for SSA Loops 3

S

- SBus 19
- SCSI-2 mapping, 3
- Service Level 109
- Spatial Reuse 2
- Special Tools 27
- SSA 1, 7
 - advantages 3
 - performance 137
 - Subsystems 8
- SSA Adapter Features 20
- SSA Adapter Features Table 20
- SSA AIX diagnostics 26
- SSA Boot Support 124
- SSA Cable 32
- SSA Configuration 1
- SSA Hardware Components 7
- SSA RAID Array Creation 95
- SSATools 29
- Storage 110
- Storage Network 53
- StorWatch Serial Storage Expert 51
- StorX 51, 73
 - Customizing Live View 73
 - Device States 57
 - Display 54
 - event monitor 76
 - Installing 61
 - Overview 52
 - Parts Palette 56
 - Performance 60
 - Printing 69
 - Starting 64
 - Toolbar 55
- StorX Live Viewer 59
- StorX Planner 58
- Subsystems 8
 - 3527 8
 - 7131 9
 - 7133 10

- 7190-100 10
- 7190-200 14

T

- Tools
 - Disksuite 44
 - filemon 41
 - Other 38
 - Remote Systems Management 45
 - SSA Cable 32
 - Veritas 44
 - vmstat 39
 - WebSSA 45
- Tools for Managing SSA Storage 19
- Tuning 138
- Tuning RAID 5 Arrays 167

V

- Vicom and IBM Product Comparison 50
- Vicom UltraLink Series 2000 116
- Vinca 116
- vmstat 39

W

- Windows NT 116

ITSO Redbook Evaluation

Monitoring and Managing IBM SSA Disk Subsystems
SG24-5251-00

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at <http://www.redbooks.ibm.com>
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?

Customer **Business Partner** **Solution Developer** **IBM employee**
 None of the above

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction _____

Please answer the following questions:

Was this redbook published in time for your needs? Yes___ No___

If no, please explain:

What other redbooks would you like to see published?

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

SG24-5251-00
Printed in the U.S.A.

