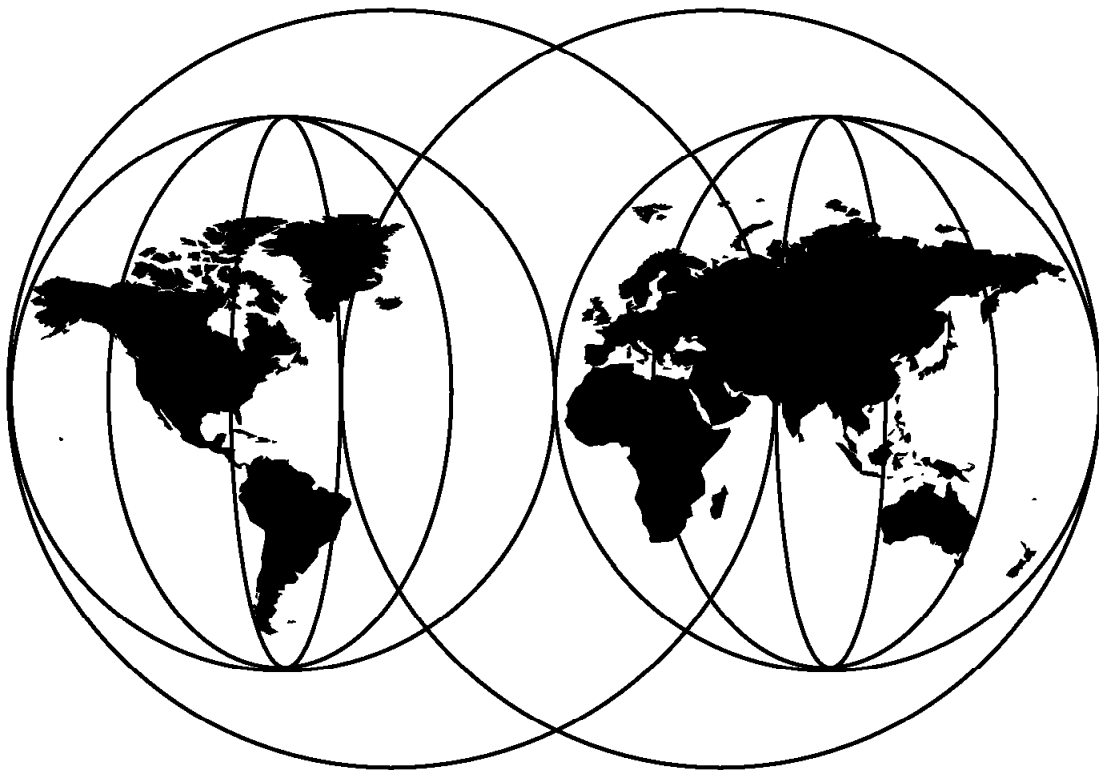




IBM Versatile Storage Server

*Dave McAuley, Peter McFarlane, Yuki Shigeiwa, Pat Blaney,
Barry Mellish, Alexandra Arnaiz, Mark Blunden*



International Technical Support Organization

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



International Technical Support Organization

SG24-2221-00

IBM Versatile Storage Server

August 1998

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix A, "Special Notices" on page 361.

First Edition (August 1998)

This edition applies to the IBM Versatile Storage Server storage subsystem. See the PUBLICATIONS section of the IBM Programming Announcement for IBM Versatile Storage Server for more information about what publications are considered to be product documentation.

Note

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

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.

© **Copyright International Business Machines Corporation 1998. All rights reserved.**

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

Contents

Preface	xvii
The Team That Wrote This Redbook	xvii
Comments Welcome	xviii
Chapter 1. Introduction	1
Acknowledgments	2
Agenda	4
Current Open Storage Disk Products	5
IBM 7204 External Disk Drive	5
IBM 7131 105, SCSI Mixed Media	5
IBM 7027 High Capacity Storage Drawer	5
IBM 7131-405 SSA High Performance	6
IBM 7133, Model 020 and Model 600	6
IBM 7135 RAIDiant Array	6
IBM 7137 RAID	6
Customer Requirements	8
Centralized storage	8
Storage partitioning with heterogeneous environments	9
Data sharing	9
Easy user access	9
High Availability	9
Investment Protection	9
Data and File Systems	10
Classes of Sharing	11
Storage	11
Copy	11
Data	12
Storage Sharing	13
Data Sharing	14
Storage Infrastructure Requirements	15
Growth	15
Access	15
Management	16
Movement	16
Security	16
Versatile Storage Server Overview	17
Chapter 2. Versatile Storage Server Architecture	21
Versatile Storage Server Architecture	21
3466 Network Storage Manager	23
RS/6000 processor	23
SSA disks	23
Magstar tape library	23
ADSM	24
WebShell client	24
3466 Web Cache Manager	24
RS/6000 processor	24
SSA disks	24
Magstar tape library	24
ADSM/HSM	24
IBM Virtual Tape Server	25

ESCON host attachments	26
Tape volume cache	26
Stacked volumes	26
3490E emulation	26
Versatile Storage Server Architecture	27
Host adapter bus	27
Storage processor bus	27
Disk adapter bus	28
Disk racks	28
Definition	28
Disk partitioning	29
Host	29
VSS	30
Common Parts	31
RISC planars	31
SSA adapters	31
7133	31
Rack and power supplies	32
Host interface cards	33
Ultra-SCSI adapter	33
Cross-Cluster Interconnection	34
Route storage server operations	34
Heartbeat monitor	34
Host Adapter PCI Bridge Link	36
Disk Adapter PCI Bridge Link	37
Dual Clusters	38
Four-way high-performance SMP standard	38
Read/Write cache	38
RS-232 ports	38
Ethernet port	39
CD-ROM, diskette drive, and internal hard disk	39
SSA Disk Adapters	40
Two to eight adapters	40
Fast-write cache	40
RAID-5 data protection	41
Disk Emulation	42
AS/400	42
UNIX	42
Chapter 3. Versatile Storage Server Technology	43
Overview	44
Versatile Storage Server Racks	45
2105-B09	45
2105-100	46
Subsystem rack configuration	46
The 2105-B09 Rack	47
The 2105-100 Rack	49
Storage Server (Front)	50
Storage Server (Rear)	51
The Versatile Storage Server	52
Symmetric multiprocessing (SMP) processors	52
332 MHz RISC CPUs	53
CPU level 1 cache	53
CPU level 2 cache	53
Snooping	54

Packaging	54
The Versatile Storage Server ...	55
Internal SCSI adapter	55
SCSI disk drive	55
CD-ROM	55
Diskette drive	56
Ethernet	56
The Versatile Storage Server ...	57
Serial ports	57
LCD display	57
Service processor	58
Storage Server Cache	59
Error correcting code	59
Synchronous DRAM (SDRAM)	59
Up to 6 GB of read/write cache (3 GB per storage server)	60
Performance Enhancing Caching Algorithms	61
Least recently used (LRU) destage	61
Staging	62
Adaptive cache algorithm	62
Sequential prediction on read	62
Fast Write Cache bypass on greater than full-stripe writes	63
PCI Local Bus	64
Adapters	65
Host Adapter	65
Disk Adapter	66
SSA Relationships	67
Disk Adapter Memory	68
Volatile storage	68
Nonvolatile storage	69
Exclusive-Or Function	70
Parity calculation	70
The Disk Drawer	72
7133 SSA disk drawer	72
7133-010	73
7133-020	73
7133 Power and Cooling	74
Three power supply modules	74
7133 Model Differences	76
Automatic host bypass circuits	76
Third power supply module	76
Redesigned front cover and on-off switch	76
Different cable connector numbering	77
7133-020 Host Bypass Circuits	78
Drive Modules	79
Disk drive module	79
Dummy drive module	80
Array Definition	81
6+P+S	81
7+P	82
Sparing	83
Disk Drives	85
Features	85
Areal density	86
Capacity	86
Performance	86

Reliability	86
Thin Film Disks	87
Substrate	87
Recording layer	87
Protection layer	88
Landing zone	88
MR Head Technology	89
Separate read and write elements	89
Magnetic recording and reading process	90
PRML Read Channel	91
Peak detection	91
Viterbi detection	92
Zoned Bit Recording	93
Zones	93
ID Sector Format	95
MR Head Geometry	97
Rotary actuator	97
ID sector format	98
MR Head Effect on ID Sector Format	99
No-ID Sector Format	100
No-ID Sector Format ...	101
RAM-based, servo-generated sector IDs	101
Other advantages	102
Predictive Failure Analysis	103
Overview	103
Processes	104
Predictive Failure Analysis ...	105
Error logs	105
Channel calibration	105
Disk sweep	105
Ultrastar 2XP in SSA Configuration	107
Three-way router	107
Versatile Storage Server 524-Byte Sector Format	108
Chapter 4. Versatile Storage Server Data Flow	109
Versatile Storage Server Data Flow Overview	110
RAID-5 Terminology	110
Hardware	110
Concepts	110
Predictive cache algorithms	110
Input/output operations (I/Os)	111
RAID-5 Terminology	112
RAID Advisory Board terms	112
RAID array	112
Strip	112
Stripe	113
Redundancy group stripe	113
Virtual disk	113
Versatile Storage Server Hardware Overview	114
Storage server cache	114
SSA adapter nonvolatile storage (Fast Write Cache)	114
SSA adapter volatile cache	115
SSA disk buffer	115
Versatile Storage Server Data Flow Architecture	116
Storage server cache	117

512 MB to 6 GB volatile memory	117
Managed with LRU algorithm	118
Read and write cache	118
Data only	118
SSA Adapter Nonvolatile Storage (Fast Write Cache)	119
Protection of fast write data	119
4 MB of SRAM	120
SSA Adapter Volatile Cache	121
Improved performance for RAID-5 update writes	121
32 MB of DRAM	121
SSA Disk Buffer	123
Improved sequential read throughput	123
Improved sequential write throughput	123
512 KB buffer in each SSA disk	124
Independent data transfer across SSA bus	124
Interleaved data transfer on SSA bus	124
SSA Frame Buffer	125
Used for SSA frame management	125
Versatile Storage Server Concepts	126
524 byte disk sector	126
Cache concepts	126
Cache transparency	126
Front end and back end I/O	126
524 Byte Disk Sector	127
SSA disks formatted with 524 byte sectors	127
System additions	127
Cache Concepts	128
Cache segment	128
Disk track	128
Cache Transparency	130
VSS cache managed by the subsystem	130
No external control of caching	130
Front End and Back End I/O	131
SCSI front end	131
SSA back end	131
Predictive Cache Algorithms	133
Used for management of storage server cache	133
Sequential predict and sequential prestage	133
Adaptive cache	134
Sequential Predict and Sequential Prestage	135
Used for management of storage server cache	135
Sophisticated sequential access pattern detection	135
Anticipatory staging of data	135
Storage server prestages a sequential staging unit	136
SSA back end reads are of 32 KB tracks	136
Parallel access increases sequential bandwidth	136
Sequential Prestage Cache Management	137
Sequential prestage synchronized with host access	137
Sequential data LRU-destaged	137
Adaptive Cache Concepts	138
Management of storage server cache	138
Dynamic selection of best caching algorithm	138
Partial and full track staging	138
Adaptive caching managed in track bands	139
Adaptive Cache Algorithm	140

Storage server generates cache statistics	140
Algorithm adapts to changing data access patterns	141
Algorithm not biased by sequential access	141
Input/Output Operations	142
Random reads	142
Sequential reads	142
Writes	143
Random Read Cache Hit	144
Cache directory searched	144
Transfer from Storage Server cache to host	144
Random Read Cache Miss	145
Cache directory searched	145
Read from SSA disk	145
Sequential Read	147
Cache hit possible if data recently written	147
Data prestaging begins when sequential access is detected	147
Storage server overlaps prestage with host access	148
Data read sequentially preferentially destaged from storage server cache	148
Fast Write	149
Fast Write bypass when appropriate	149
Optimum data availability – three copies of data	149
Data destaged asynchronously from Fast Write Cache	150
RAID-5 Write Penalty	151
Update writes subject to RAID-5 write penalty	151
Fast write masks the write penalty on update writes	152
Data Integrity for RAID-5 Parity	153
RAID-5 update write integrity challenge	153
Fast Write Cache assists parity integrity	153
Parity regenerated if SSA adapter cache malfunctions	154
Fast Write Cache Management	155
Asynchronous destage to disk	155
Fast Write Cache managed by threshold	155
Write preempts	156
Stripe Writes	157
RAID-5 write penalty avoidance	157
Sequential write throughput	158
Write Preempts	159
Update write to a particular block	159
Fast Write Hits	159
Multiple updates processed by single destage	159
Write Blocking	160
Adjacent blocks destaged together	160
Fast Write Cache Destage	161
Fast Write Cache destage triggered by threshold	161
Other destage triggers	161
Destage from SSA adapter cache	162
Writes from Fast Write Cache to Disk	163
Destage uses hierarchy of VSS storage	163
Data removed from Fast Write Cache following destage	163
SSA adapter cache copy retained to improve hit rates	164
Transfers from Storage Server Cache to SSA Adapter	165
VSS data transfer	165
Fast Writes	166
Stripe Write to SSA Adapter	167
Bypass Fast Write Cache	167

Data is written from storage server cache to adapter cache	167
Data and parity written to disk buffers	168
Task complete when data written to physical disk	168
Versatile Storage Server Data Flow Summary	169
Storage server cache	169
Adaptive cache	169
Sequential prediction and sequential prestage	170
Fast writes	170
SSA adapter nonvolatile storage	170
Chapter 5. Drawer Design	171
Drawer Design	171
Component	172
Drawer	173
Internal SSA path	173
Redundant power supply and cooling fan	174
Host Bypass Circuit	175
Disk Drive	176
Ultrastar 2XP disk drive	176
Buffer	176
4.5 GB disk or 9.1 GB disk	177
Routing function	177
SMART/PFA	178
PFA	178
How the PFA works	179
Monitoring	179
SMART	179
Disk Drive Reliability	180
PFA and SMART	180
Channel calibration	181
Log recording	181
Disk sweep	181
Configuration	182
Automatic SSA loop configuration	182
An example of the loop reconfiguration	183
New disk installation	183
Sparing	184
Sparing Procedure	186
Sparing with hot spare disk	186
Sparing without hot spare disk	186
Component Interaction	188
SSA connection	188
Power supply and cooling fan	188
Interface between the SSA RAID adapter and the disk drive	189
Chapter 6. Versatile Storage Server Configuration Options	191
Versatile Storage Server Configuration Options	191
Host connectivity	191
Storage server	191
Disks	192
Racks	192
Power supplies	192
Configuration	192
Host Connectivity	193
Host attachment cables	193

Connectivity Considerations	194
Number of hosts	194
Volume of data	194
Data rate	195
I/O rate	195
Availability	195
Backup requirements	195
Summary	195
Host Support	196
Advanced software	196
SMP Cluster Options	197
Processors	197
Read cache size	197
Standard features	198
Disk Adapters	199
Number of adapters	199
Guidelines	199
RAID Arrays	201
Maximum Availability	202
Disk Sizing	203
State-of-the-art technology	203
High performance	204
Logical Volume Allocation	205
Server	205
VSS	205
Multiple access	206
2105-B09 Storage Server Rack	207
Disk storage	207
2105-100 expansion rack	209
Power supply	209
Disk drawers	210
SSA cables	210
Investment Protection	211
Versatile Storage Server Maximum Configuration	212
Power Options	213
Redundant power	213
Disk drawer power cords	213
Optional battery	213
Access	214
Service indicators	214
VSS Enclosure	215
Configuration management	216
Intranet access	216
Web interface	217
Configuration Considerations	218
Number of hosts	218
Volume of data	218
Data rate	218
I/O rate.)	219
Availability	219
Chapter 7. Migrating Data to the Versatile Storage Server	221
Overview	222
Issues	223
Connection	223

Replacement for existing storage	224
Reformat of existing drives	224
Method of transferring data	224
Migration of existing 7133s and drives	225
One or two spare drives per loop	225
Supported Servers and Prerequisites	226
IBM RS/6000	226
IBM AS/400	227
Supported Servers and Prerequisites ...	228
Sun Microsystems	228
Hewlett-Packard	228
Compaq	228
Data General	229
Host View of the Versatile Storage Server	230
UNIX systems	230
AS/400 systems	231
Transferring data—UNIX Systems	232
Volume management software	232
Direct copy	233
Backup and restore	233
AIX Logical Volume Manager	235
Terms	235
Complete Logical Volume Copy	237
Migrate Physical Volume	239
Mirror Migration	241
Mirror Migration ...	243
Impact on Availability	244
Volume management methods	244
Direct copy method	245
Backup and restore methods	245
Use of Existing 7133 Drawers with Versatile Storage Server	246
7133s can be migrated to new racks	246
Existing 7133s can be migrated to new racks	246
Existing 7133s can remain in 7105 or 7202 racks	247
Use of Existing Disk Drives in VSS	248
Strict drive requirements	248
All drives in an array and drawer must be the same size	248
Drives must be reformatted	249
524-byte Sector Format	250
Chapter 8. Versatile Storage Server Performance	251
Versatile Storage Server Performance Overview	252
VSS Performance Highlights	252
Performance Planning Information	252
Total System Performance	252
VSS I/O Operations	253
Guidelines for VSS Configuration	253
Workload Characterization	253
Other Performance Considerations	253
Summary	253
Performance Highlights	254
Storage server cache	254
Adaptive cache	254
Sequential prediction and sequential prestage	255
Fast write	255

SSA adapter nonvolatile storage	255
Total System Performance	256
Disk subsystem performance as part of overall system performance	256
System performance considerations	256
I/O Operations	258
Front end I/O	258
Back end I/O	258
Guidelines for Configuration—Overview	260
Storage server cache size	260
Number of SCSI ports per host	260
Choice of 4.5 or 9.1 GB disk drives	261
Number of RAID-5 arrays per SSA loop	261
Storage Server Cache Size	262
Storage server cache size determined by two factors	262
Host Caching	263
Effective Use of Host Cache	264
Effectiveness of host caching depends on several factors	264
Host Caching Environments	266
UNIX file systems	266
Database management systems	267
DB2 parallel edition	267
Oracle Parallel Server	267
Subsystem Effect of Host Caching	268
Where host caching is highly effective	268
Where host caching is not highly effective	269
Storage Server Cache Size Guidelines—Part I	270
Host caching	270
Storage Server Cache Size Guidelines—Part II	272
Where host caching is effective	272
Storage Server Cache Size Guidelines—Part III	274
Where host caching is effective	274
Four-Way SMP	276
Where both storage servers are four-way SMPs	276
Rules of thumb	276
Consider Storage server capability during failover	277
Versatile Storage Server SCSI Ports	278
Emulates multiple LUNs	278
Supporting multiple SCSI initiators	279
UltraSCSI adapters	279
Throughput considerations	279
Number of SCSI Ports per Host	280
VSS SCSI adapter throughput	280
Consider high-availability SCSI connection	280
Multiple SCSI attachment options	281
Consider virtual disk partitioning	281
Disk Capacity Selection	282
Disk specifications	282
Access density	282
Disk Specifications	283
Disk can perform 50 I/Os per second	283
Access Density	284
I/Os per second per gigabyte	284
Selection of capacity depends on access density	284
Rule of thumb	285
Select 4.5 G disk drives where appropriate	285

Number of RAID Arrays per SSA Loop	286
SSA adapter supporting one or two loops	286
Rules of thumb	287
Workload Characterization Overview	288
Read to write ratio	288
Synchronous write I/O content	288
Sequential I/O content	288
Caching characteristics of data	289
Read to Write Ratio	290
Significant in RAID disk subsystems	290
Random writes	290
Sequential writes	291
Synchronous Write I/O Content	292
Used by database management systems	292
Response time sensitive	292
Benefit significantly from fast write	293
Fast Write Cache usage	293
Sequential I/O Content	294
Sequential I/O	294
Sequential detect	295
SCSI host attachment utilization	295
Caching Characteristics of Data	296
VSS adaptive caching	296
VSS caching of data written	297
Dependent on the access pattern	297
Other Performance Considerations	298
Race conditions	298
Migrating data from RAID-1 file systems to RAID-5	298
Parallel query processing	298
Race Conditions	299
Heterogeneous environments	299
Shared SCSI bus environments	299
Migrating Data from RAID-1 File Systems	301
RAID-1 (software mirroring or in hardware)	301
Consider read bandwidth	301
Consider 4.5 GB disks	302
Use of Parallel Query Processing	303
Database management parallel query processing	303
A VSS virtual disk defined as a physical disk to the host system	303
A VSS virtual disk as a partition of a RAID-5 array	303
A VSS RAID array equivalent to physical disk	304
Plan mapping of virtual disks to RAID arrays	304
Performance Review	305
Storage server cache	305
Adaptive cache	305
Sequential prediction and sequential prestage	306
Fast write	306
SSA adapter Fast Write Cache	306
Summary	307
Overall system performance	307
Disk subsystem cache	307
Improved performance through Fast Write	307
Configuration options	308
Configuration flexibility	308

Chapter 9. Versatile Storage Server Maintenance	309
Overview	310
Philosophy	310
Repair Actions	311
Interfaces	311
Overview ...	312
Reporting	312
Sparing	313
Upgrades	313
Code EC management	313
VS Specialist	314
HTML browser	314
Status screen	315
Onsite Maintenance Interface	316
ASCII terminal	316
Character based	317
Remote Support Interface	318
Call home	318
Support Center	319
Reporting – Error Log	320
Error log	320
Error log analysis	321
Problem record generation	321
Reporting – SNMP	322
Simple network management protocol	322
Two management information bases	322
Reporting – E-mail	324
Reporting – Call Home	325
Remote service facility	325
Information provided.	325
Repair Actions – Customer	327
Console specification	327
Logical configuration	327
Limited physical configuration	328
Limited repair	328
Subsystem code EC management	328
Repair Actions – CE and PE	329
Service procedures	329
Code EC Management	331
Supported interfaces	331
Code EC process	332
Release process and media types	332
Chapter 10. Subsystem Recovery	333
Subsystem Recovery	333
Types of Failure	334
Data integrity	334
Data availability	334
Concurrent maintenance	334
Remote Services	335
Access through phone line and modem	335
Call-home support	335
Remote diagnostics and support	336
Host Connection Failure	337
Disk Mirroring	339

High Availability with Multihost Dual Access	340
HACMP with Mirroring	341
Storage Server Operation	342
Storage Server Failure	344
Storage server cache	344
Storage server processors	345
Disk Subsystem Failure	346
Disk adapter failure	346
Disk drive failure	347
SSA cable or connection failure	347
SSA disk drawer	347
Power System Failure	348
Redundant power	348
DC power control unit	349
Battery backup	349
7133-020 disk drawer	349
Disaster Recovery	350
Data Recovery	350
Configuration data	351
Chapter 11. Software Support	353
Software Support	353
SSA RAID management	353
Device Driver Support	353
SSA Software Support	354
Configuration methods software	354
Device Drivers	354
Adapter	355
Current Support	356
Planned Support	357
Device Driver Function (UNIX)	358
Device Driver Function (OS/400)	359
Appendix A. Special Notices	361
Appendix B. Related Publications	363
Redbooks on CD-ROMs	363
Other Publications	363
How to Get ITSO Redbooks	365
How IBM Employees Can Get ITSO Redbooks	365
How Customers Can Get ITSO Redbooks	366
IBM Redbook Order Form	367
Index	369
ITSO Redbook Evaluation	371

Preface

This redbook gives a broad understanding of the new architecture of the Versatile Storage Server (VSS). This book provides an introduction, and describes in detail the architecture, technology, data flow, configuration, migration and recovery aspects for the VSS.

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 San Jose Center.

Dave McAuley is the project leader for open systems disk storage products at the International Technical Support Organization, San Jose Center. He has written eight ITSO redbooks and several international conference papers, among them a paper on image processing and pattern recognition in IBM's *Journal of Research and Development*. Dave teaches IBM classes worldwide on all areas of IBM disk storage products. Before joining the ITSO in 1994, he worked in the product support center in Scotland as a large systems storage specialist. Dave has worked at IBM for 19 years, with international career assignments in manufacturing, design and development, and marketing. His background in computer science spans a varied portfolio including supercomputing application development, process control and image analysis systems design and microcode development, automation and robotics systems project management, and large systems marketing and technical support.

Peter McFarlane is currently Customer Services Manager and a Senior Technical Specialist for OSIX Pty. Ltd. in Sydney, Australia. He is a shareholder and member of the board of directors of OSIX. OSIX is an IBM Business Partner, focussed on providing AIX solutions and services. Peter has 13 years of experience in the Unix and open systems field. He has worked at OSIX for two years in his current role. Prior to joining OSIX, Peter worked at Sequent Computer Systems for one year and Unisys Corporation for 13.5 years as a hardware and software support specialist. His areas of expertise include AIX and Unix, storage products, networking and education. He attained the IBM certification of AIX Support Specialist (ASP) in December of 1996. He has written Shell Programming and TCP/IP courses for OSIX.

Yuki Shigeiwa is a Information Technology Engineer in Japan. He has two years of experience in ADSM and open systems storage. He has worked at IBM for two years. His areas of expertise include marketing, design, and implementation support for ADSM and planning support for Open attach disk and tape subsystems.

Pat Blaney is a Technical Support Rep for Storage Systems with Advanced Technical Support in San Jose, California, USA. He joined IBM in 1977 as a Systems Engineer in New York and moved to San Jose in 1982, working in internal I/S for the San Jose plant. In 1987, he became the product planner for the 3390 and later worked on the introduction of the 3990-3 Extended Platform and the 3990-6. He joined what is now Advanced Technical Support in 1993 when the support mission for storage systems moved from Gaithersburg to San Jose. He has worked with all members of the RAMAC Array Family, most recently with the introduction of the RAMAC Virtual Array and the RAMAC

Scalable Array to the RAMAC Array Family. He represents Advanced Technical Support at Share and GUIDE in the US, and is frequently a speaker at symposiums presented by IBM Education and Training. He is now adding Open Systems disk products to his areas of expertise.

Barry Mellish is a Senior Storage Specialist in the UK. He has worked for IBM for the last 14 years. Barry joined IBM as a Property Services Engineer responsible for IBM premises in Central London. Barry moved into system engineering 10 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 two 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 technical support group specializing in Open System Storage Solutions.

Alexandra Arnaiz is a Marketing Specialist in the Philippines She has 6 years of experience in Computer software. She has 2 years of experience in Storage. Her areas of expertise include mid-range storage devices and programming software.

Mark Blunden is the project leader for Open Systems at the International Technical Support Organization, San Jose 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.

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

John Aschoff, Storage Systems Division, San Jose
Brent Beardsley, Storage Systems Division, Tucson
Maggie Cutler, ITSO Technical Editor
Mike Hartung, Storage Systems Division, Tucson
Paul Hooton, Storage Systems Division, Havant
Dick Johnson, Storage Systems Division, San Jose
Marc Roscow, Storage Systems Division, San Jose
Dan Smyth, Storage Systems Division, San Jose
Michelle Tidwell, Storage Systems Division, San Jose
Viet Tran, IBM Global Services Planning
Dean Underwood, Storage Systems Division, Tucson
Steve Van Gundy, Storage Systems Division, San Jose

Thanks are also due to the many other collaborators and reviewers who contributed to the production of this document.

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 371 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. Introduction

Introduction



IBM Versatile Storage Server (VSS)

International Technical Support Organization
San Jose

June 1998



© IBM Corporation 1998

Acknowledgments



- Project leader
 - Dave McAuley - ITSO-San Jose Center
- Residents
 - Peter McFarlane - Business Partner, OSIX Pty Ltd Australia
 - Yuki Shigeiwa - IBM Japan
 - Pat Blaney - IBM US
 - Barry Mellish - IBM UK
 - Alexandra Arnaiz - IBM Philippines



© IBM Corporation 1998

Acknowledgments

This book is the result of a residency project run at the International Technical Support Organization-San Jose Center in San Jose California. This project was designed and managed by:

Dave McAuley, International Technical Support Organization-San Jose Center

The book was written by:

- Dave McAuley, IBM ITSO-San Jose
- Peter McFarlane, Business Partner - OSIX Pty. Ltd., Australia
- Yuki Shigeiwa, IBM Japan
- Pat Blaney, IBM AT-San Jose
- Barry Mellish, IBM United Kingdom
- Alexandra Arnaiz, IBM Philippines

This book was last updated by Mark Blunden, ITSO-San Jose

We are grateful for the contributions of many people, but particularly acknowledge:

- Dean Underwood, Storage Systems Division, Tucson

Audience



- This presentation is written for:
 - Customer personnel
 - IBM and business partner technical and marketing specialists
 - IBM client representatives



© IBM Corporation 1998

Trademarks



The following products and features referenced in this presentation are trademarks of corporations in the United States and/or other countries:

IBM

- ES/390
- ESCON
- IBM
- System/390
- MVS
- AS/400
- RS/6000
- AIX
- NetView
- PowerPC
- Ultrastar
- PFA
- VS 2100
- VSS

OTHER

- UNIX
- HP
- SUN Sparc
- Windows
- Windows NT
- Intel



© IBM Corporation 1998

Versatile Storage Server - Agenda



- Introduction
 - Architecture
 - Technology
 - Data Flow
 - Drawer Design
 - Configuration Options
 - Migration
 - Performance
 - Maintenance
 - Subsystem Recovery
 - Software Support
 - Summary
-

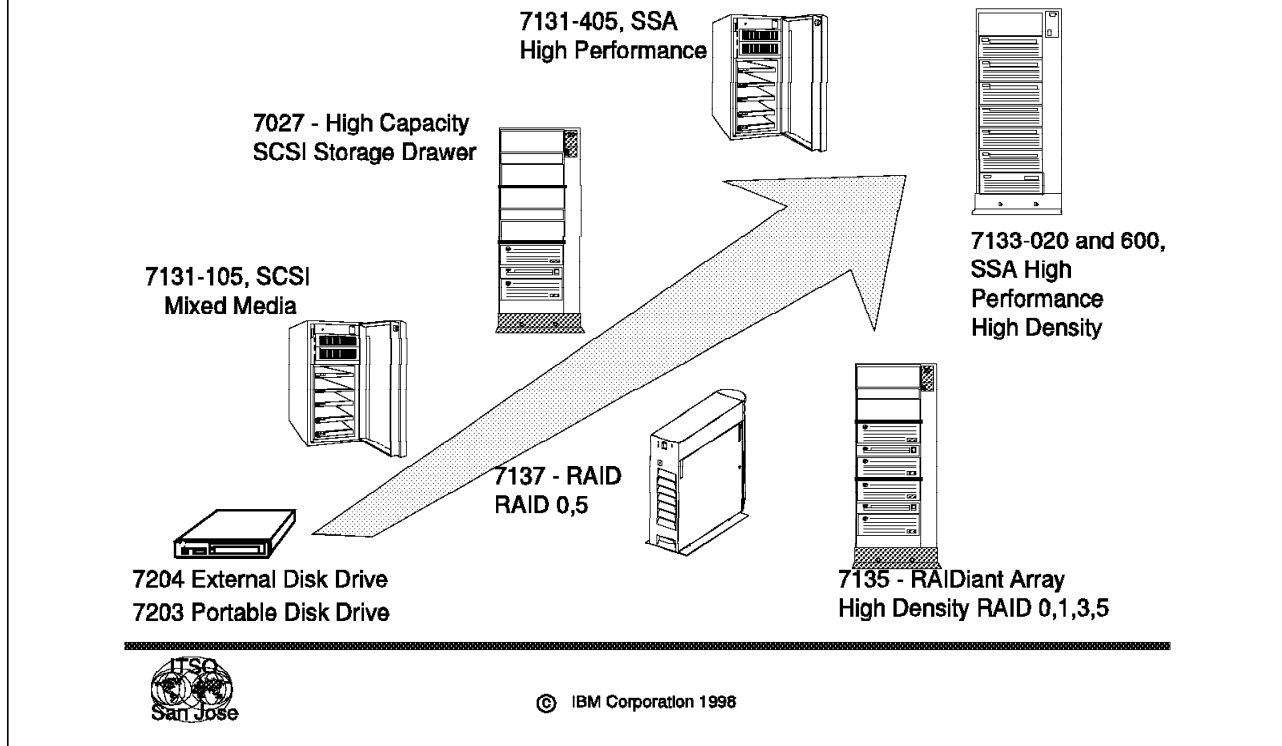


© IBM Corporation 1998

Agenda

This foil lists the topics covered in the presentation guide. The topics cover the Seascapes architecture, how it is implemented in the VSS, data flow, configuration options, installation, migration, maintenance, and subsystem recovery.

Current Open Storage Disk Products



Current Open Storage Disk Products

The foil shows the open storage disk products available as of April 15, 1998.

IBM 7204 External Disk Drive

The IBM 7204 is a small external disk enclosure. Depending on the model, it can house either a 2.2 GB, 4.5 GB or 9.1 GB single disk drive. It attaches to a SCSI-2 Fast/Wide differential adapter in the RS/6000.

IBM 7131 105, SCSI Mixed Media

The IBM 7131 105 is a low cost, stand-alone SCSI storage tower that provides expansion storage for hard disk drives as well as tape or CD-ROM drives. In total, the tower provides up to 45.5 GB of storage in five hot-swappable disk bays with the capability of housing up to two tape drives or four CD-ROM drives. The media bays can be used to house an additional three non-hot-swappable disk drives, giving a total storage capacity of 63.7 GB.

IBM 7027 High Capacity Storage Drawer

The IBM 7027 fits into a standard 19 inch rack, 7015-R00, or the 7014-S00. It can house up to 67.3 GB in a drawer which is seven EIA units high. In addition, there are three tape or CD-ROM drive bays; some tape units occupy two of these bays. Attachment to the RS/6000 is with single-ended or differential SCSI-2 F/W connections.

IBM 7131-405 SSA High Performance

This provides up to 45.5 GB of disk storage using IBM's SSA hard disk drives. The disk drives are housed in five hot-swappable bays and can be a mixture of 2.2 GB, 4.5 GB and 9.1 GB drives.

IBM 7133, Model 020 and Model 600

The IBM 7133 has two models, the 020 and 600, that are functionally equivalent and can be populated with up to 16 customer-replaceable disk drives. The supported disk drives come in three capacities: 2.2 GB, 4.5 GB, and 9.1 GB. Four 2.2 GB disk drives modules are standard; however, select features allow these modules to be changed at the time of order to 4.5 GB or 9.1 GB modules. Features are provided to add any combination of supported SSA disk drive modules providing up to 145 GB of disk storage capacity. The 7133 can be directly attached to the RS/6000 using an SSA adapter or attached to a variety of other UNIX hosts using either an external Vicom SLIC or IBM 7190 SCSI to SSA converter attached to a SCSI-2 differential fast and wide (SCSI-2D F/W) adapter in the host, or internally using the Sun SSA adapter for relevant Sun machines. A variety of Windows NT servers are supported through either native PCI bus adapters or the SCSI to SSA converter. In addition, RAID capability can be given to the 7133 either by appropriate software or in the case of the RS/6000, by a native SSA RAID adapter.

IBM 7135 RAIDiant Array

The IBM 7135 can be configured as a fully fault-tolerant RAID array with dual active controllers. The dual active controller configuration provides enhanced performance as each controller can handle I/O requests. It also provides automatic controller switching (called *failover*) if a controller error is detected. Each controller can be connected to up to two RS/6000s, which makes the 7135 useful in clustering and high availability scenarios. The controllers can be installed with 16 MB of read or write cache. The write cache is mirrored between controllers to ensure data integrity in the event of controller failure. The cache is protected from loss of power by an on-board battery. Read-ahead capability is supported in the controllers. With applications with substantial sequential processing this cache can improve performance by up to 70% compared with a similar system without controller cache. The IBM 7135 is equipped with either five 2.2 GB or 4.5 GB disk drives and can be expanded to thirty drives. With 30x4.5 GB drives, this gives a total storage capacity of 135 GB, with 108 GB usable when operating in RAID-5 mode. The 7135 attaches to IBM RS/6000 processors only.

IBM 7137 RAID

The IBM 7137 is a RAID storage device for open systems platforms and is functionally equivalent to the 9337 that attaches to the AS/400. The 7137 can contain a maximum of eight disk drives and is offered in three deskside models (413, 414 and 415) and three rack-mountable models (513, 514 and 515). The rack-mountable models fit into a standard 19-inch rack and require four units of EIA space. All of the disk drives are mounted on hot-swappable carriers. The 2.1 GB, 4.3 GB, and 8.8 GB disk drives are supported, but they cannot be mixed in the same unit. The arrays can be configured as RAID-5 or RAID 0. A maximum of two arrays can be attached to a single SCSI bus. One of the drives can be designated as a hot spare and will not be included in the RAID group unless one of the other drives fails. As standard, each model in the range comes with three drives in a two drive plus parity RAID-5 configuration. The

7137 is available with up to 4 MB of battery backed up, nonvolatile (Fast Write Cache) read or write cache. The 7137 has redundant power supplies so that operation can continue if one supply fails. The controller card is not redundant; in the event of a controller card failure, the Fast Write Cache is removed from the failed card. The Fast Write Cache is then placed into the new card, which is inserted into the 7137. The 7137 resumes processing using the data stored in the Fast Write Cache. Thus data integrity is ensured. In addition to attaching to the RS/6000, the 7137 is supported on most models of Sun, HP, and NCR machines.

Customer Requirements



- Centralized storage
- Storage partitioning with heterogeneous environments
- Data sharing
- Easy user access
- High Availability
- Investment Protection



© IBM Corporation 1998

Customer Requirements

Customer requirements have evolved in response to the explosive growth in storage requirements brought in by:

- Network computing
- Online storage of voice, image, and video data
- Data mining and data warehousing
- Data collection and point-of-sale terminals

Centralized storage

During the past 10 years, one of the main goals within the information technology organization was the decentralization of data. The idea was to place the data as close to the user as possible. Decentralization was a result of the introduction of client/server computing and advancements in technology that reduced the size and cost of computers and peripheral devices while increasing performance. With this decentralization of processing power have come added costs in managing installations and networks. Now the direction for many customers is to recentralize, while the processing power remains separate, whether in the same room or remotely.

Storage partitioning with heterogeneous environments

The majority of customers have multivendor environments. Many have created a data processing environment that has several servers located in a computer room. Often there is a requirement for servers to share storage so that as storage needs change, the storage can be reassigned.

Data sharing

Many customers need to share data among several servers. There is confusion as to exactly what is meant by data sharing and there are problems at a technical level such as differing file formats and data structures. These are discussed in the next four foils.

Easy user access

Customers need and expect configuration methods to be easy and user friendly.

High Availability

Data is a key business resource. As business becomes more reliant on computer systems for all its activities, it is becoming essential that business data be available at all times. With the greater advent of globalization, more and more companies are trading 24 hours a day. The window for data repair and maintenance is becoming smaller and thus the emphasis is on high availability or highly resilient and available systems with built-in redundancy. There are three main components:

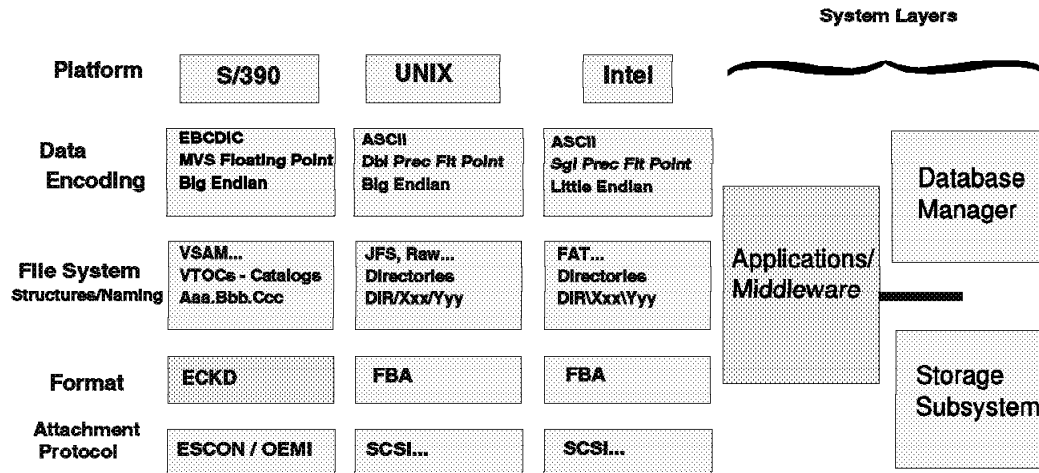
- Protection against data loss
- Protection against loss of data integrity
- Protection against loss of data availability

Any failures in these areas will cause downtime on customer computer systems with possible adverse business effects.

Investment Protection

With a growing investment in storage, customers want to be able to retain existing storage systems and incorporate them into the overall strategy.

Data and File Systems



© IBM Corporation 1998

Data and File Systems

This foil shows the types of data and file systems that are associated with the main classes of host servers. For sharing a true single copy of data, the data will have to be stored in such a way that it can be presented to each different host in the manner it expects. The storage system will have to present different views of the same piece of data to each host.

- S/390 and AS/400 use EBCDIC data format while most other platforms use ASCII.
- Data is stored in completely different structures. For example, MVS uses data sets and catalogs and UNIX uses file systems with directories. Intel-based machines also use file systems and directories, but they use the file allocation table (FAT) file system and have a different file system naming convention that uses “\” instead of the UNIX “/.”
- The methods of storing data on disks are different. For example, MVS uses extended count key data (ECKD) and UNIX, AS/400, and Intel use fixed block architecture (FBA).
- Attachment protocols from the host to the disk subsystem are different. For example MVS uses ESCON, and UNIX, AS/400 and Intel uses SCSI.

The Data Explosion



Storage Sharing



- Partitioned hardware
- No common access

Copy Sharing

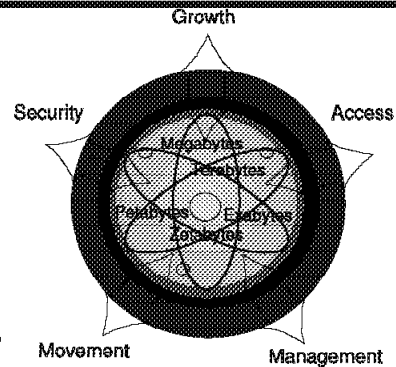


- Replicated data, fast transfer

Data Sharing



- Simultaneous multi-platform read/write access
- Lock/unlock support for integrity required



© IBM Corporation 1998

Classes of Sharing

Three kinds of sharing can take place within the storage environment. The terms are often misused, and misleading claims are made.

Storage

With storage sharing, there is a common pool of storage. This is divided into parts dedicated to individual servers. No part of the data is common. The disk controller is shared. The effect is that each host has its own "disks," although it is possible to reassign disks from one host to another.

Copy

Copy sharing is the most common type of sharing used today. One host creates or modifies the data, which is then broadcast to all other hosts that need to access that data. The mechanism for transferring the data in a UNIX environment is usually either file transfer protocol (FTP) or a database replication service. Mainframe storage systems have a remote copy facility that enables distant systems to have a read copy of data.

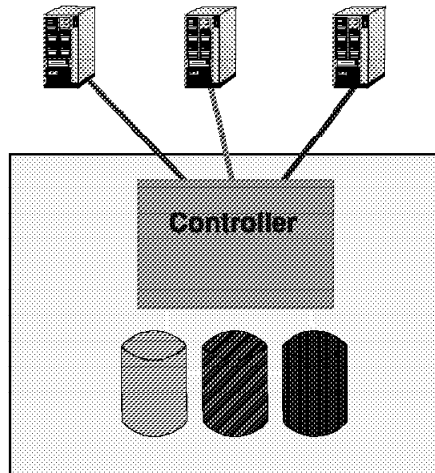
Data

With data sharing, two or more hosts have direct access to the data and they all have the ability to create and modify data. Mechanisms have to be put in place to ensure the integrity of the data.

Storage Sharing



Storage Sharing



Disk farm is shared among multiple servers, each server owns a portion

Logical drives assigned to LUNs.

- Size can vary, based on the needs of the server
- Logical drives cannot span RAID-5 arrays
- Logical drives can be changed (add, delete) dynamically



© IBM Corporation 1998

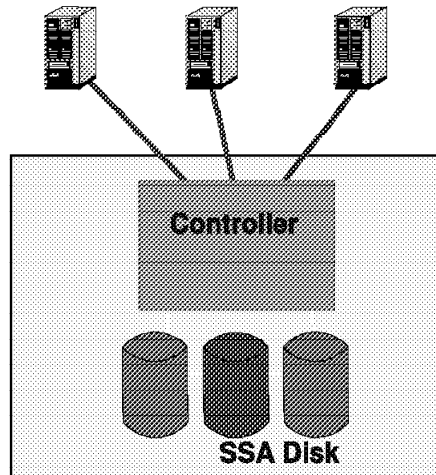
Storage Sharing

This foil shows how three servers can share a pool of disks. The disk farm, in this case, is protected storage in RAID-5 arrays. Each array can be split into logical partitions that can be assigned to each server. The size of each partition can be changed and these changes can be carried out dynamically, without shutting the disk subsystem down. Changes should not be made to a server's disk subsystem while the server is trying to access it. Therefore any applications that access the subsystem should be quiesced while the changes are made.

Data Sharing



Data Sharing



LUN can be assigned to multiple servers.

- Storage Subsystem provides the common access to the data
- The applications must provide the locking mechanism to ensure data integrity.



© IBM Corporation 1998

Data Sharing

In true data sharing, the physical or logical disk is assigned to more than one host. Procedures must be in place to prevent a second host from reading or writing data that is already being updated by another host. This locking is generally provided by the application such as Oracle Parallel Server.

Customer Requirements



- Growth
- Access
- Management
- Movement
- Security



© IBM Corporation 1998

Storage Infrastructure Requirements

When developing a storage infrastructure strategy that can handle both present and future requirements, you must consider issues of growth, access, management, movement, and security.

Growth

As applications evolve, transaction workloads, and storage capacity requirements can grow explosively and unpredictably. To enhance value, many applications are designed to be platform and vendor independent. Flexible storage systems must address this need for independence while offering granular capacity growth and the ability to move to newer, faster technologies as they become available for host adapters and storage controllers.

Access

The exponential growth of the Internet and client/server applications has led many organizations to rapidly adopt the network computing model. As mobile employees, business partners, and customers demand global access to data, storage systems must provide heterogeneous attachment for multiple platforms and address the requirements to share data across those platforms. Storage systems must also be flexible enough to easily incorporate new types of host attachments.

Management

The cost of managing many widely distributed servers can be very high. To reduce costs, the storage system must provide the option of remote management and automated operations in most computing environments, giving the flexibility to locate staff at the most cost effective location.

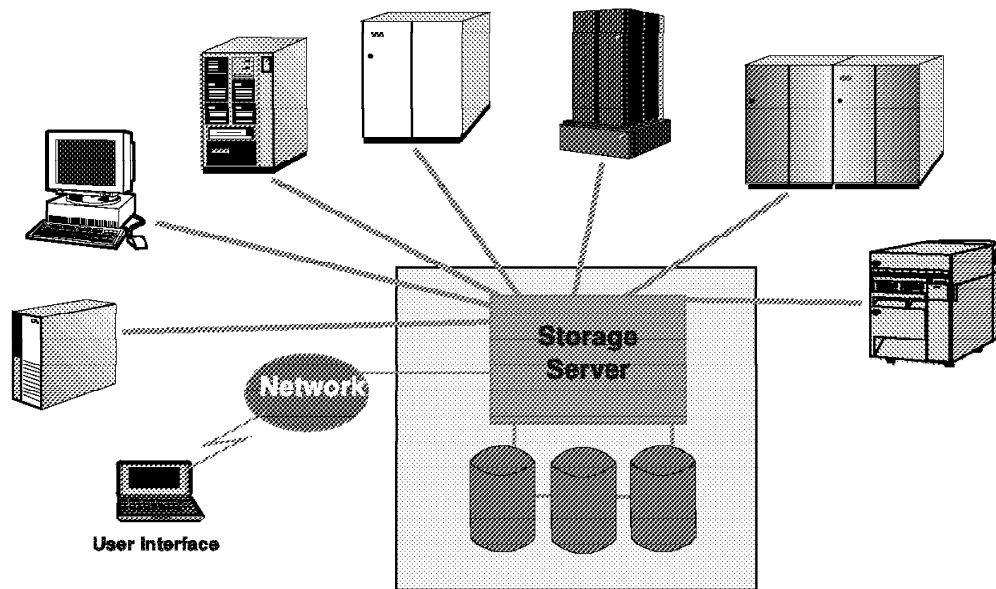
Movement

Data is being moved in large quantities over networks for many purposes, such as backup and archive, disaster recovery, and data warehousing. With the rapid growth in the quantity and types of data that must be moved, data movement can quickly become a bottleneck in the network environment. To prevent data transfer from inhibiting growth, the storage system should provide the capability for automated data movement and management that is transparent to the platform and common to all heterogeneous environments.

Security

Enterprise data is a vital corporate asset. Storage systems must address data availability and security concerns. The importance of today's data demands continued innovation in self-diagnosis and error correction, remote monitoring, performance monitoring, automatic reconfiguration, and tamper-proof access. All this has to be provided in an affordable, auditable solution.

Requirements



© IBM Corporation 1998

Versatile Storage Server Overview

VSS is designed to offer several new features and benefits to satisfy customers who require a central pool of storage. It is the first disk product to use the Seascapes architecture, (see Chapter 2, "Versatile Storage Server Architecture" on page 21). The philosophy behind the architecture is that of common parts; that is, to take parts and subsystems from other divisions within IBM and build new products with them. The advantage to the customer is that the components within the product are tried and tested and have a known high reliability. Using existing components and subassemblies also helps to keep the cost down, as there are reduced development costs and the benefits of large-scale manufacture.

VSS is designed to meet both customer and storage infrastructure requirements; see "Customer Requirements" on page 8 and "Storage Infrastructure Requirements" on page 15.

VSS offers the customer:

- A central pool of storage
- Offers the ability to distribute the storage pool among up to 64 heterogeneous hosts and reassign the storage distribution as requirements change. All activity can be carried out with the storage system online.
- Allows assigning a storage partition to multiple hosts so that true data sharing can take place. File locking and record-level locking are the responsibility of the application program.

- Provides an easy-to-use web-based browser interface, so that configuring and maintaining the system is as simple as possible.
- Offers built-in protection against data loss, and loss of data integrity.
- Can maintain access to the data under almost all circumstances.
- Through its “plug and play” I/O architecture, components can easily be changed as technology changes without disruption. The ability to use existing 7133 disk drawers and racks demonstrates IBM’s commitment to protecting customer investment in storage.

VSS uses existing RISC processor boards and chips, SSA disks and adapters, racks and power supplies from ES/390 servers, coupled with the resource and expertise of more than 40 years of making disk subsystems.

VSS consists of a high-performance storage server complex with 512 MB cache, expandable to 6 GB, redundant power supplies, and one or more drawers of high-performance SSA disks. The controller complex has built-in redundancy to protect against possible failure. Data is always protected from disk failure by the use of RAID-5. The capacity is highly scalable. The minimum configuration is two drawers (32*9.1 GB disks) with two RAID arrays configured. The usable storage capacity is 230 GB. The maximum number of disk drawers is 18. The supported disk sizes are 4.5 GB and 9.1 GB. The total disk storage, using 9.1 GB drives, is 2.62 TB with a usable capacity of 2.0 TB. (The usable capacity calculation is 18 drawers * 13 drives * 8.77 GB per drive = 2.05 TB.)

VSS is designed for use in multihost, heterogeneous environments. Up to 64 hosts can be attached concurrently. Supported hosts are the RS/6000, RS/6000 SP, AS/400, Sun Ultra series, the HP 9000 800 series, some Data General models and also PC support on various IBM and Compaq servers. Support for additional hosts is planned to be announced in the coming months. Host attachment is by the use of UltraSCSI adapters; this is backward compatible with SCSI-2D F/W. The VSS presents a known disk image to the attached host. The AS/400 is presented with the image of a 9337 disk subsystem while UNIX machines “see” a generic SCSI disk. The disk image can be divided into LUNs.

VSS has two interface types, a web-based browser interface, known as the IBM StorWatch Versatile Storage Specialist (VS Specialist for short), for user configuration and control of the subsystem, and a serial attach interface for use by service personnel. Additionally a modem connection for remote diagnostics and download of fixes can be made.

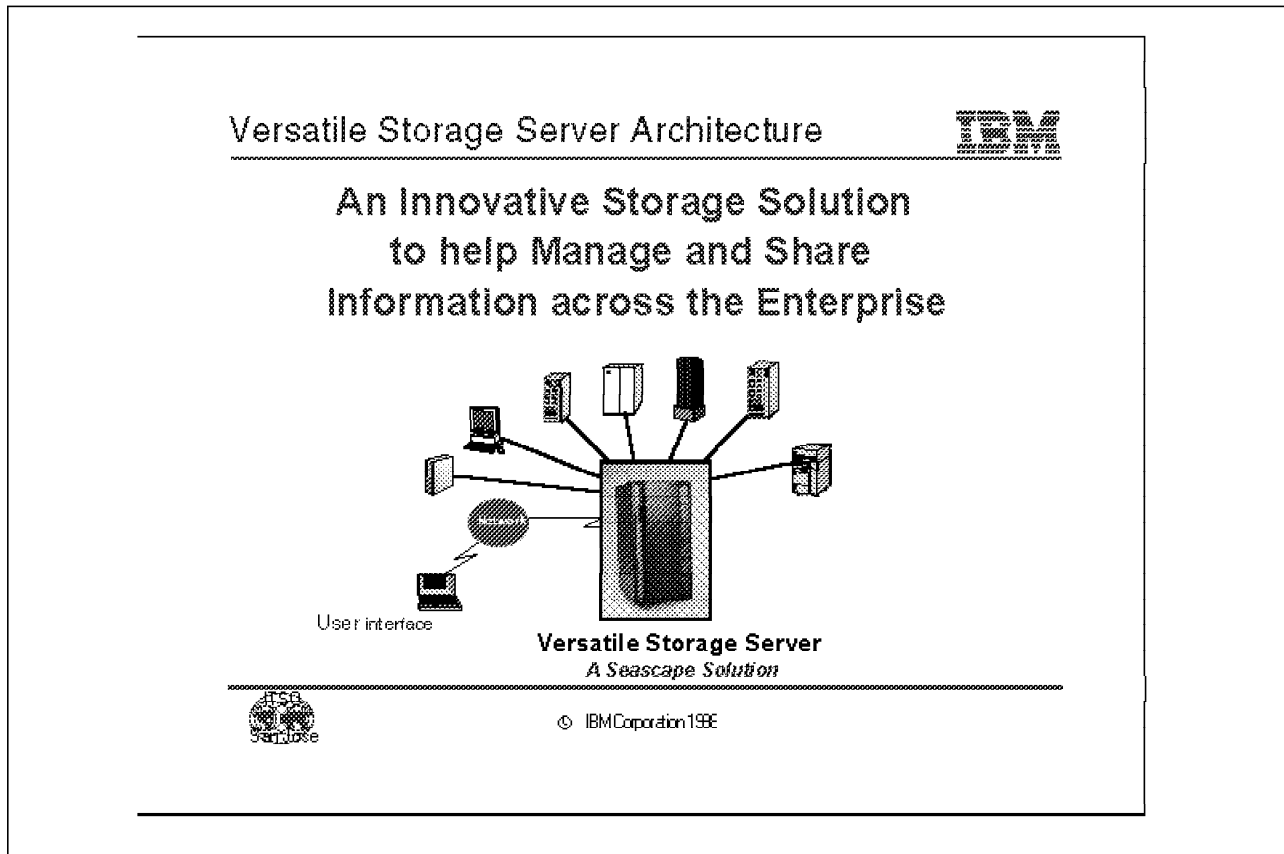
A full description of the architecture and implementation is in Chapter 2, “Versatile Storage Server Architecture” on page 21 and Chapter 4, “Versatile Storage Server Data Flow” on page 109. VSS is positioned to complement rather than replace existing IBM products. The typical VSS customer will have:

- 450+GB of storage attached to several different servers
- Common location of servers
- Common management and control or the desire to create them

For a business requiring storage of large amounts of data using a single host or all RS/6000 hosts, other storage solutions should be considered. For the homogeneous RS/6000 environment, native attached 7133 SSA disks should be considered, especially for stand-alone servers, or small clusters. For UNIX environments where a common storage pool is not required, 7133 SSA disks

attached using the Vicom SLIC, IBM 7190, or IBM SSA adapter for Sun S-Bus should be considered.

Chapter 2. Versatile Storage Server Architecture



Versatile Storage Server Architecture

In this chapter we detail the VSS architecture and the common parts philosophy. We review existing products that use the common parts philosophy and discuss how VSS meets customers storage requirements, (see "Storage Infrastructure Requirements" on page 15).

The VSS is the one of the Seascape family of products which are integrated storage servers used for the attachment of storage devices to various host computer systems. IBM's new Seascape architecture helps organizations implement a simplified, yet flexible storage infrastructure that helps them derive the most value from their data assets. Seascape includes integrated storage solutions based on a set of common building blocks, interchangeable components that can be easily matched to changing storage requirements. With excellent flexibility, Seascape helps provide a reliable storage infrastructure that can deliver performance, scalability, affordability and ease of management today and in the future.

The Seascape enterprise architecture highlights are:

- Industry leading technology integration, base product function and advanced software features

- Leveraging of interchangeable technology components that can be quickly combined into comprehensive solutions
- Highly scalable products that can attach to and share data between multiple heterogeneous computing platforms and across a variety of networking technologies
- Quantification and reduction of data management costs through advanced storage system storage server software and new operations management tools

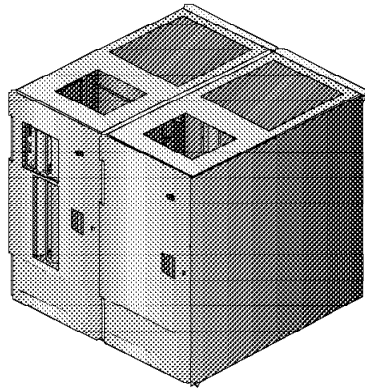
Seascape is based on the exploitation of RISC hardware components, existing SSA disks and other hardware and IBM's proven storage subsystem expertise. VSS attaches to multiple heterogeneous hosts: IBM RS/6000, IBM AS/400, Sun Ultra series, the HP family, Data General, Compaq and IBM Netfinity. Additional UNIX and NT servers will be supported upon request. VSS has been designed to meet the needs of customers who have a multihost environment and need a centrally managed common storage pool.

Management of the storage pool is carried using a web-browser based interface configuration manager. The use of web technology enables the system administrator to use the company intranet or LAN so that management can take place from the usual place of work. Security mechanisms are in place to prevent unauthorized use.

3466 Network Storage Manager and 3466 Web Cache Manager

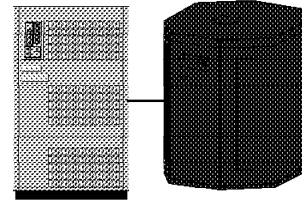


Integrated Solutions



Network Storage manager

- RS/6000 processor
– AIX preloaded
- SSA Disk
- Magstar tape library
- ADSM
- WebShell client



Web Cache Manager



© IBM Corporation 1998

3466 Network Storage Manager

The 3466 Network Storage Manager, Netstore, is based on the common parts philosophy. It provides a packaged ADSM solution in a “black box.” The customer does not have to configure and integrate each individual part of the system as this is done prior to shipping.

RS/6000 processor

A rack mounted RS/6000 Model R20 forms the central control unit of Netstore. It comes with the AIX operating systems preinstalled and preconfigured.

SSA disks

The minimum configuration includes 16 x 4.5 GB SSA disks to give 72 GB of storage. This can be increased to 144 GB.

Magstar tape library

Netstore uses the 3494 tape library with high speed, high capacity 3590 Magstar tape drives.

ADSM

ADSM provides automated backup and recovery capability for all major distributed platforms, including Sun, HP, DEC, IBM, Novell, Microsoft, Apple, SGI, and all other managed clients. The Disaster Recovery Management (DRM) feature provides system management tools to perform enterprise-wide disaster recovery management for IBM Netstore and its client platforms.

WebShell client

The ADSM WebShell client and server code support functions such as backup, client code distribution, and a “help desk” function to be run over a company’s intranet.

The IBM 3466 Network Storage Manager brings to the customer a complete packaged solution for managing and protecting data in a network.

3466 Web Cache Manager

The 3466 Web Cache Manager is also based on the common parts philosophy.

The function of the Web Cache Manager is to store web objects locally, so that multiple requests from the user community do not consume bandwidth needlessly. In performing this function, the Web Cache Manager reduces the bandwidth needs for connecting those users to the Internet backbone or an upstream access provider. It provides the capability to cache objects that flow under the HTTP and the FTP protocols. Because this product has integrated the IBM Web Traffic Express software, it can also proxy, or screen, Internet requests of end users.

RS/6000 processor

A rack mounted RS/6000 forms the central control unit. It comes with the AIX operating system preinstalled and preconfigured.

SSA disks

The 7133 disk subsystem uses high-speed serial storage architecture (SSA) for disk I/O.

Magstar tape library

Netstore uses the award-winning Magstar MP tape drives to store data migrated from the disk drives.

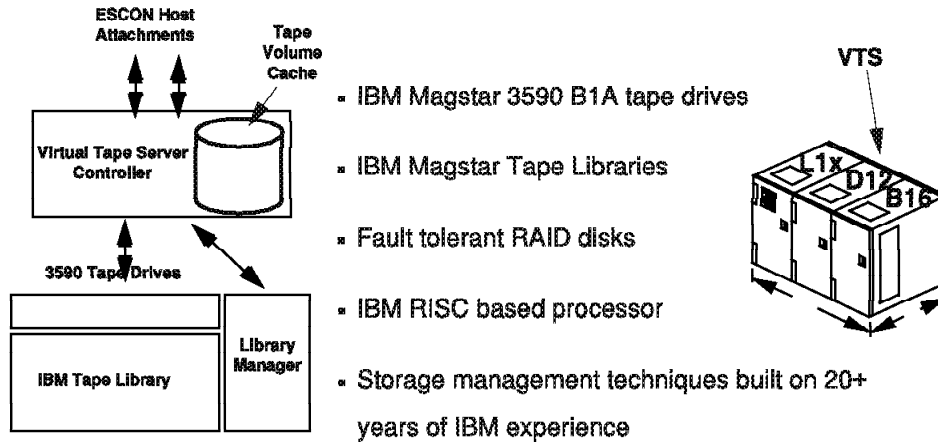
ADSM/HSM

ADSM/HSM transparently moves the older and larger web objects from disk to the tape library. This migration of objects frees up the disk space for smaller, more popular objects.

IBM Virtual Tape Server



Integration of Proven IBM Technologies



© IBM Corporation 1998

IBM Virtual Tape Server

The Magstar Virtual Tape Server is designed to stack multiple host-created tape volumes on a Magstar 3590 cartridge. It provides ES/3090, ES/9000, and S/390 enterprise servers with a revolutionary enhancement to tape processing by integrating the advanced technologies of:

- IBM 3590 Magstar tape drives
- IBM attached RAID disk storage
- IBM Magstar tape libraries
- IBM Magstar Virtual Tape Server storage server
- Robust, proven storage management software

The result is an integrated hierarchical storage management system that addresses a variety of storage requirements. The Magstar Virtual Tape Server, when integrated with Magstar tape libraries, offers a new class of device compared to traditional tape storage products. With the dramatic increase in storage capacity of the Magstar 3590 tape technology, the performance and random access capabilities of integrated fault-tolerant RAID disks, and intelligent outboard hierarchical storage management, the Magstar Virtual Tape Server significantly reduces costs associated with tape processing.

ESCON host attachments

The Magstar Virtual Tape Server provides two ESCON channel adapters for host attachment, with each adapter supporting 64 logical channel paths and a maximum channel distance of 43 km. It appears to the attached hosts as two 3490E tape subsystems, each with 16 devices.

Tape volume cache

The Magstar Virtual Tape Server's intelligent storage management software manages fault-tolerant RAID disks as tape volume cache. Access to all data is through the tape volume cache, which extends many of the performance benefits of cache disk to tape. For example, creating new tape volumes (nonspecific mounts) is done solely to the tape volume cache.

The cache can hold hundreds of virtual volumes. The content of the cache is managed to retain the most recently accessed virtual volumes so that numerous subsequent mount requests can be satisfied very quickly from the cache, similar to a cache request on DASD. If a requested volume is not present in the cache, the required Magstar 3590 cartridge is mounted, and the logical volume is moved back into the cache from a stacked volume.

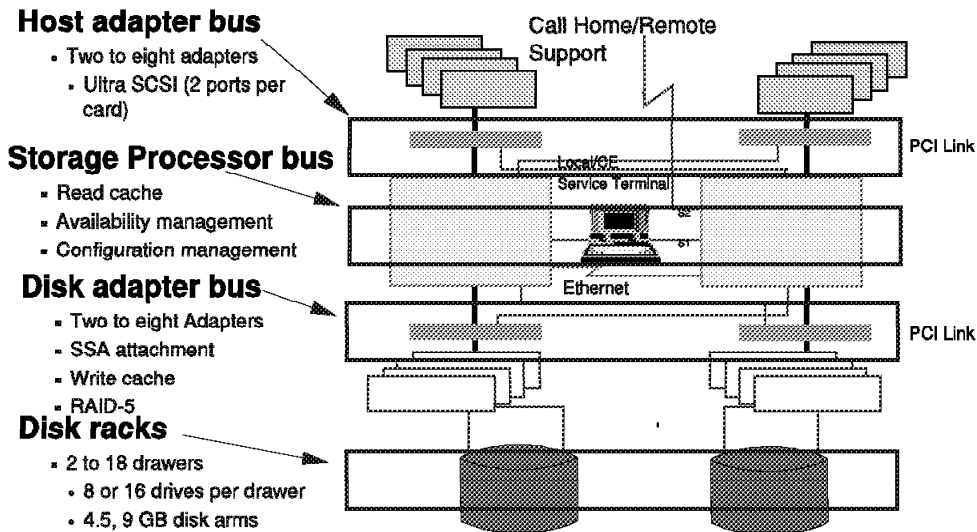
Stacked volumes

Stacked volumes are Magstar 3590 volumes that contain several logical volumes. After a virtual volume is unloaded, it is copied onto a stacked volume. The virtual volume then remains in the cache until its space is required for another virtual volume. The content of stacked volumes is managed by the Magstar Virtual Tape Server such that partially full stacked volumes are consolidated to free up space.

3490E emulation

From a host perspective, data is processed as if it resides on actual devices and cartridges. This representation of tape devices enables transparent use of Magstar tape technology. It also enables use of the Magstar Virtual Tape Server without requiring installation of new software releases. Within the Magstar Virtual Tape Server, data is stored on disk as images of either virtual Cartridge System Tape (CTS) or enhanced Capacity Cartridge System Tape (ECCST). All 3490E-type commands are supported. Tape motion commands are translated into disk commands, resulting in response times much faster than in conventional tape drives.

Versatile Storage Server Architecture



© IBM Corporation 1998

Versatile Storage Server Architecture

The basic architectural building blocks and concepts of the VSS are simple. In essence, there are three interconnected buses and racks of SSA disk drawers.

Host adapter bus

The host adapter bus is a PCI bridge, into which host interface cards can be plugged. As new and faster technologies emerge, new host interface cards can be plugged in.

Storage processor bus

The storage processor bus is a backplane into which the SMP cluster cards are plugged. In VSS the storage server is based on a RISC planar with four-way RISC symmetrical multiprocessor (SMP) engines. As RISC technology changes so the planar cards can either be upgraded or changed and new cards inserted into the backplane. The storage server contains the system read cache, (minimum size 512 MB, maximum size 6 GB). In addition to controlling the disk I/Os, the storage server carries out availability and configuration management.

Disk adapter bus

As SSA technology changes, the SSA cards can be changed. The adapter cards used are similar to the PCI SSA RAID adapters used in the RS/6000 with the exception that these have a fast write cache backed up by battery-backed Fast Write Cache.

Disk racks

The disk racks contain from 2 to 18 drawers of 7133 disks. There can be either 8 or 16 disks in a drawer. There are two types of rack. The first rack must be the storage server rack 2105-BO9, which contains the storage server, power management system and up to four disk drawers. The second and third racks are disk expansion racks 2105-100, and contain up to seven drawers of SSA disks each. The maximum supported configuration is 18 disk drawers. The disks are configured into RAID-5 arrays of eight disks. Each array is in the form of six disks plus parity plus hot spare (6+P+S) or seven disks plus parity (7+P). No other disk configurations are supported. The first RAID array in a drawer must always be configured as a 6+P+S array. The disk drives that are supported are the 4.5 GB and 9.1 GB drives. The SSA technology architecture allows new and old devices of the same capacity to coexist on the same SSA loop.

A web-based interface allows the administrator to control the VSS and to assign storage to individual hosts as required. By using web technology for the configuration tool the administrator can sit anywhere on the intranet site.

Storage can be shared (partitioned) between the attached servers or can be assigned to multiple servers. With the storage partition "shared" among several UNIX based servers, applications can access the same physical copy of data, thus achieving real data sharing. Access management and locking are provided by the application or database manager, not by the VSS.

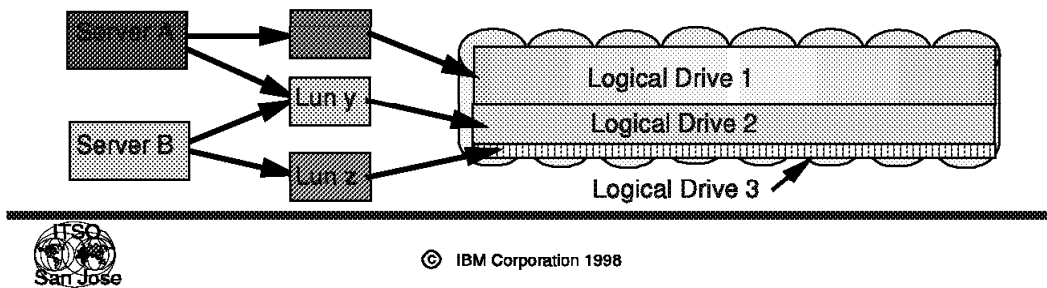
Definition

There is a single storage server in the VSS, consisting of two processing clusters with RISC based four-way SMP configurations. The storage server consists of the host adapter, PCI Link technology, SMP cluster, SSA adapter and communication ports.

Disk Partitioning



- Host
 - VSS presents logical view of storage to the host server.
 - ▶ *AS/400 - seen as 9337 580 or 590*
 - ▶ *UNIX - seen as generic SCSI disk*
 - ▶ *Multiple LUN sizes are supported*
- VSS
 - Storageserver manages the logical to physical relationship
 - Data is striped across the RAID-5 array



© IBM Corporation 1998

Disk partitioning

VSS manages the relationship between the RAID arrays and the logical view of storage presented to the host servers. RAID arrays can be divided into logical units of varying sizes, from 0.5 GB to 32 GB. This foil shows how a RAID-5 disk array on VSS can be subdivided into three logical disks. The logical disks are shown attached to two host servers. In the picture, each host server has one logical disk of its own, defined as LUN x and LUN z. LUN y is shared between both servers. File and record level locking is the responsibility of the application running on the servers such as Oracle Parallel Server. Alternatively LUN y could be a shared disk for use in a high availability environment.

Note A LUN cannot span more than one disk array.

Host

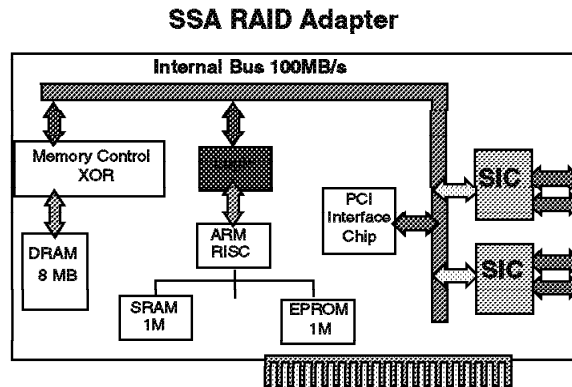
The host sees the logical disk partition as a familiar entity. In the case of the AS/400 the storage server emulates a 9337, so the AS/400 sees a 9337 580 or 590 disk array. The UNIX host server is presented with a generic SCSI disk

VSS

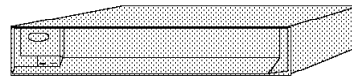
The storage server manages the relationship between the logical disk and the physical RAID arrays. Logical disk sizes and host attachments are defined during system configuration. Definitions can be changed without taking the system down so host storage pools can be easily reassigned to match changing business needs. VSS does not know whether data is stored within it. Before the system administrator reassigns storage pools, host applications accessing the pool should be quiesced and appropriate host operating system action taken so that the host is aware of what has happened.

Common Parts

- PowerPC planars
- SSA adapters
- 7133
- Rack and power supplies



7133-010/020



© IBM Corporation 1998

Common Parts

One of the main features of the VSS is the use of common parts. Using parts that are used elsewhere in IBM shortens development times, ensures the reliability of the parts, and minimizes costs.

RISC planars

The planar boards that are used in VSS are RISC planars with RISC chips. They are based on the boards that are used in the RS/6000.

SSA adapters

The adapters used in VSS are based on the PCI bus RAID-5 adapter used in the RS/6000.

7133

The 7133 SSA disk drawer and 4.5 GB and 9.1 GB disks are used in VSS. The drawer is not modified in any way except that new power cords are supplied to connect to the two 350 V DC power buses. The disk drives for VSS are specially formatted with 524-byte sectors for AS/400 compatibility, compared with 512-byte sectors that are used in the native version of the 7133. If existing disk drives are to be used in VSS, they have to be reformatted with the new disk sector sizes. Both existing 7133-020 and 7133-010 can be used in the VSS. **Note**The drawer configuration may need to be modified to conform to the VSS rules. Jumper

cables and a third power supply may be needed to make the 7133-010 compatible. There is also a need to ensure that the EC and PTF levels within the 7133 are up to date.

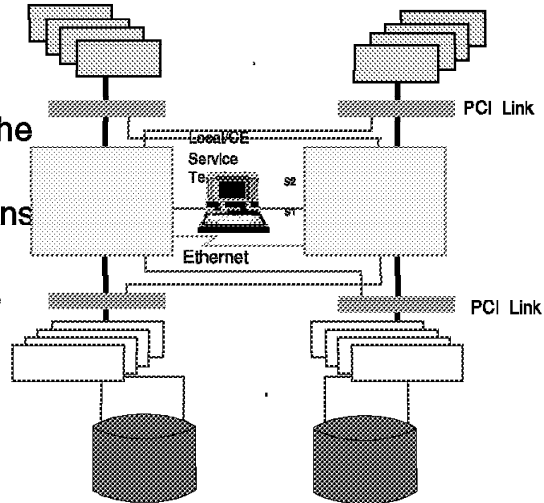
Rack and power supplies

The rack and power supplies use technology similar to that in existing ES/390 products. The internals of the rack have been altered with a DC power rail for each side. The storage server rack 2105-BO9 houses the storage server (dual processing clusters, read/write cache, PCI Link technology, input and output adapter bays and adapters), power supplies, and two drawers of SSA disk drives (32 9.1 GB drives). There is space for two additional 7133 drawers (either Model 10 or 20). The expansion rack 2105-100 houses two DC power rails and up to seven 7133 disk drawers.

Host Interface Card



- Ultra-SCSI card
 - Up to 8 adapters (2 min)
 - Supports Ultra-SCSI or SCSI-2D F/W adapters in the host
 - Two independent connections per adapter
 - Both ports can read or write data concurrently



© IBM Corporation 1998

Host interface cards

Host interface cards are housed in one of four I/O adapter bays on the PCI bridge. Each I/O adapter bay can house up to two adapters.

Ultra-SCSI adapter

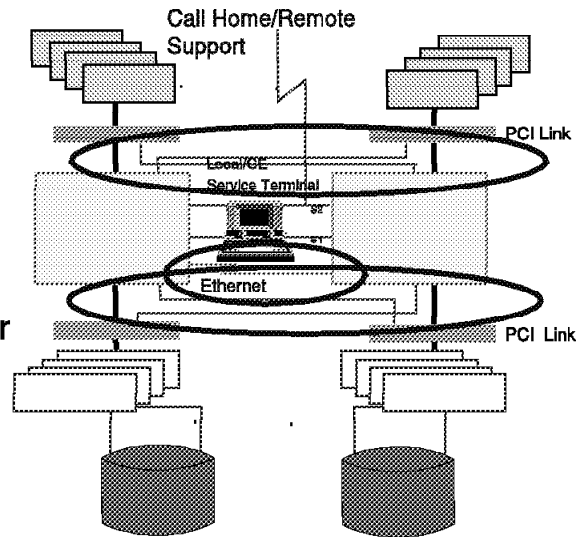
The host interface adapters are the connection point between VSS and the hosts.

The host interface adapters that are used are Ultra-SCSI adapters. They conform to the SCSI-3 standard of which SCSI-2 F/W is backward compatible. Thus SCSI-2 differential F/W adapters in the host can connect to VSS albeit at a connection speed of 20 MB/s compared with 40 MB/s for Ultra-SCSI. There are four bays, each of which can hold two host and two SSA adapters, giving a total of eight adapters that can be used. Each adapter has two independent ports that can both read or write data concurrently. Any port on any adapter can be configured to communicate to either processing cluster. Once the connection is defined, it talks to only that cluster unless a failover occurs. If a host needs to connect to disks that are attached to both clusters, then there must be two host-to-VSS connections. One connection is to be defined to one cluster, the other connection is to be defined to the other cluster. Up to 64 hosts can be connected at once.

Cross-Cluster Interconnection



- Route operations
 - Send operations to one side or the other
 - Based on setup during installation
- Heartbeat monitor
- Failover support
 - Route operations to other cluster if primary cluster fails



© IBM Corporation 1998

Cross-Cluster Interconnection

The cross-cluster interconnection enables much of the functionality and resilience of VSS.

The interconnections occur at three levels:

- A link between the two host adapter PCI bridges
- A “heartbeat” link between the two storage servers
- A link between the two disk adapter PCI bridges

Route storage server operations

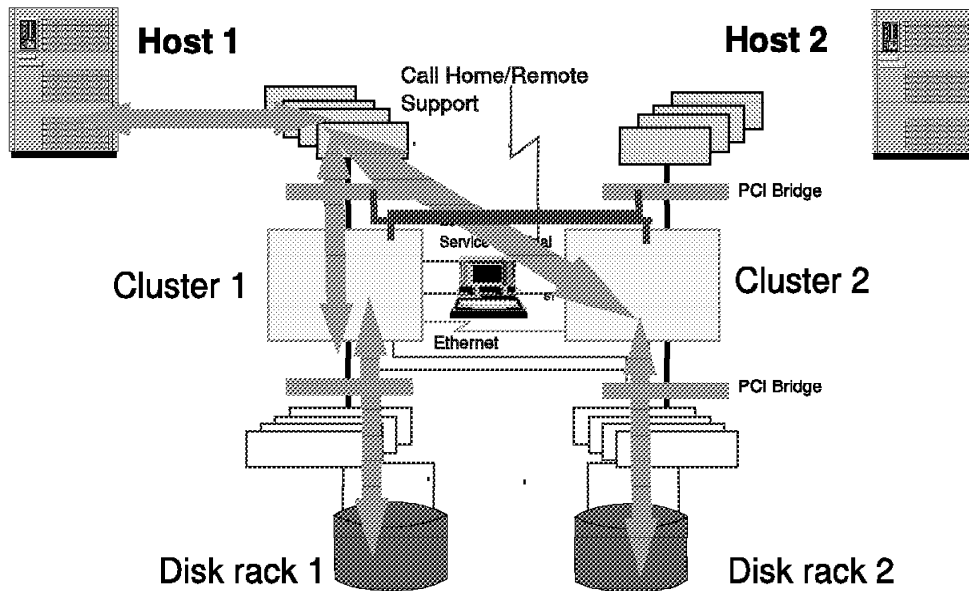
The link between the two PCI bridges enables host adapter ports to send I/O requests to either side of the storage cluster. The cluster to which the host adapter port sends its I/O requests is set up during the initial installation.

Heartbeat monitor

The heartbeat monitor enables each cluster to monitor the other and failover if a problem is detected, (see Chapter 10, “Subsystem Recovery” on page 333). Failover support is provided by the link between the two disk adapter PCI bridges and the host bridge. This link enables one cluster to route operations to both sets of disks. This link also enables online cluster microcode updates to take place. One of the storage servers is deliberately shut down and all transactions are failed over to the other. The microcode update is applied and

the cluster is then brought back online and resumes its transaction processing, (see Chapter 10, “Subsystem Recovery” on page 333).

Host Adapter PCI Bridge Link

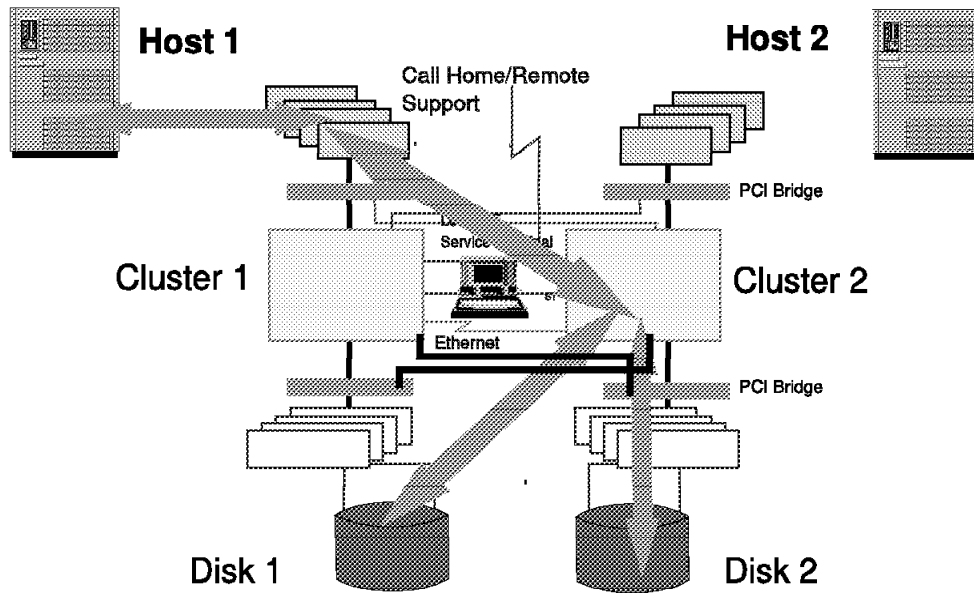


© IBM Corporation 1998

Host Adapter PCI Bridge Link

This foil shows how the host adapter PCI link operates. The host adapter PCI-bridge link enables hosts which are attached to "side 1" of VSS to see disks that are attached to "side 2." This crosslink enables disk arrays to be configured so that any host can access any disk array. VSS presents each host with its own storage pool. Each RAID array can be subdivided into partitions or virtual disks of variable size. The host operating system assigns SCSI target or LUN IDs to the partitions or virtual disks. The host sees each virtual disk as an independent "physical" disk and it can manipulate each of them as it would a physical disk. Host 1 can route I/O requests to both clusters and hence to disk racks 1 and 2 even though its host adapter is in the IOA bay attached to cluster 1.

Disk Adapter PCI Bridge Link



© IBM Corporation 1998

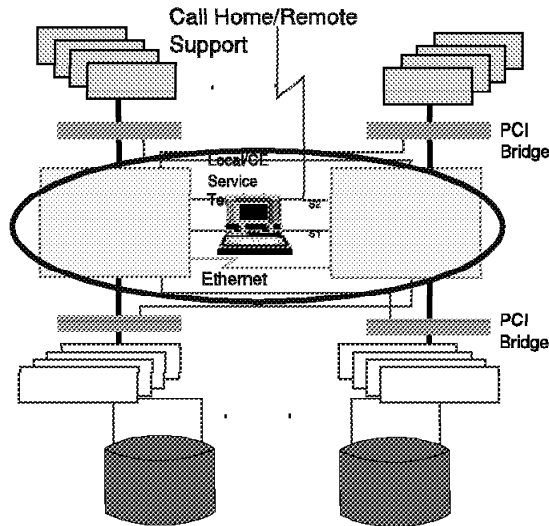
Disk Adapter PCI Bridge Link

The disk adapter PCI bridge link enables the VSS to be resilient and makes online microcode updates possible. In normal operations, Disk 1 can be accessed only by Cluster 1 and Disk 2 by Cluster 2. If the heartbeat monitor detects a problem in one of the clusters, then that cluster is shut down and its disks are taken over by the other cluster. When the problem has been rectified, the cluster is brought back online and normal operations resume. Online microcode updates can take place by failing over one cluster. The microcode to this cluster can then be updated while the others carry out all the disk I/O processing. Once the update or change has been made, the cluster is brought back online and resumes normal processing. The microcode is designed so that clusters can run at different code levels, although we do not recommend doing so for long periods.

Dual Clusters



- Four-way high-performance SMP available (four-way on each side)
 - PCI bus structure
- Read/write cache (512 MB to 6 GB)
- RS-232 ports
- Ethernet port
- CD-ROM, diskette drive, and internal hard disk



© IBM Corporation 1998

Dual Clusters

The clusters are sometimes known as *logical control units* (LCUs).

They control the I/O operations of the system and are the vehicle for availability and configuration management.

Four-way high-performance SMP standard

Each cluster, as standard, contains RISC processors in a four-way high-performance SMP configuration. The clusters use a PCI bus structure.

Read/Write cache

Each cluster contains a read/write cache to improve system response time, (see Chapter 8, "Versatile Storage Server Performance" on page 251). The minimum cache size is 512 MB and the maximum is 6 GB, split evenly between the two clusters.

RS-232 ports

The RS-232 port is used to attach a modem to the system so that in the event of a system malfunction the service team can be notified directly. The routing and severity of the alerts are set up during installation. A service engineer can use the second port to attach a local monitor to carry out diagnostic and remedial work.

Ethernet port

The ethernet port is used to connect VSS to an intranet or local LAN. This connection enables the system administrator to control the system from his or her usual place of work. System administration and maintenance tasks are carried out with a web-browser-based configuration tool.

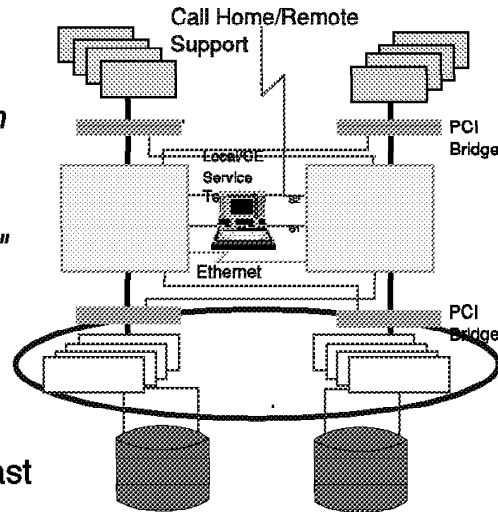
CD-ROM, diskette drive, and internal hard disk

The CD-ROM and diskette drive are used in installing microcode and applying updates. The internal hard disk is where the UNIX kernel and microcode are stored. If the hard disk should malfunction, its cluster is failed, and processing is routed to the other cluster.

SSA Disk Adapters



- Two to eight adapters
 - Two loops per adapter
- Fast write cache
 - 8 MB per adapter
 - ▶ 4 MB used as write through cache
 - ▶ Volatile memory
 - ▶ 4 MB used as "permastore"
 - ▶ Battery backup
 - ▶ Removable
 - Proven technology
- RAID-5 data protection
 - RAID-5 penalty masked by fast write cache



© IBM Corporation 1998

SSA Disk Adapters

The SSA disk adapters are based on the adapters used in the RS/6000, with the addition of a fast-write cache and new microcode.

Two to eight adapters

Each adapter can support two SSA loops. The maximum configuration has eight adapters which means that 16 SSA loops are supported. The maximum number of disk drawers that can be attached is 18. This means that a full system there will have 14 loops of 16 disks, (a single SSA drawer) and two loops of 32 disks (two SSA drawers).

Fast-write cache

The total fast-write cache in each SSA RAID adapter is 8 MB. Half of this, 4 MB, is used as a write-through cache and is in volatile memory. The other half is a mirror copy of this stored in a battery backed up "permastore" or Fast Write Cache, (see Chapter 3, "Versatile Storage Server Technology" on page 43). The "permastore" is proven technology used in the 7137 RAID array.

RAID-5 data protection

All data is protected against disk failure by the use of RAID-5 arrays. No other configurations are allowed. The RAID-5 write penalty is masked by the fast write cache.

Disk Emulation



- **AS/400**
 - 9337 emulation
 - Logical 4 GB drives
 - Logical 9 GB drives
- **UNIX**
 - Generic SCSI device drivers
 - 15 SCSI targets
 - 64 LUNs
 - LUN capacity 0.5 GB to 32 GB



© IBM Corporation 1998

Disk Emulation

To support the variety of attached hosts, VSS emulates existing devices for both UNIX and AS/400 systems.

AS/400

VSS emulates the 9337 when attached to AS/400 systems. The following emulation is provided:

- Logical 4 GB drives emulate the 9337-580. A minimum of four 4 GB drives must be configured. The actual capacity of each logical drive is 4.194 GB.
- Logical 9 GB drives emulate the 9337-590. A minimum of four 9 GB drives must be configured. The actual capacity of each logical drive is 8.59 GB.

OS/400 expects to see a separate device address for each disk drive, logical or physical. VSS will report unique addresses for each arm that is defined to the AS/400. OS/400 behaves as if 9337s are attached to the system.

UNIX

For UNIX-based systems VSS emulates the generic SCSI device drives that are supported by drivers found in most systems. The VSS enables each LUN can have a capacity ranging from 0.5 GB to 32 GB (valid LUN sizes are: 0.5, 1, 2, 4, 8, 12, 16, 20, 24, 28, and 32 GB). There does not have to be a direct correlation between physical disk drives and the logical LUNs seen by the UNIX-based host.

Chapter 3. Versatile Storage Server Technology

Technology

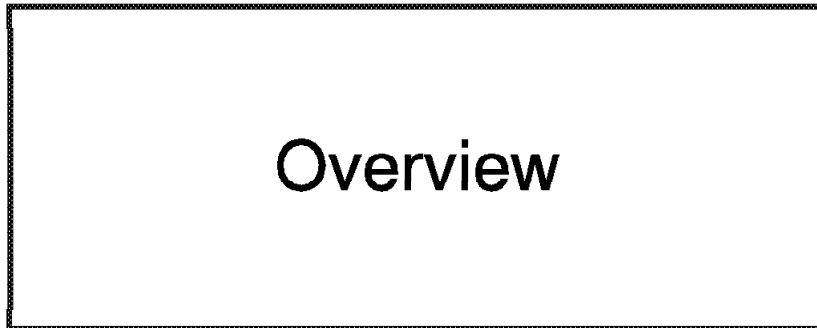


Versatile Storage Server Technology



© IBM Corporation 1998

Overview



© IBM Corporation 1996

Overview

In this section we cover the technology used in the VSS subsystem. We discuss the VSS racks, the VSS storage server and its functional components, the 7133 disk drawers, and the Ultrastar 2XP disk drives.

The racks are more than metal chassis; they contain power control and sequencing logic to ensure high availability.

The storage server uses Power Performance Chip (Power PC), Serial Storage Architecture (SSA) and Ultra SCSI technologies to provide high performance, reliable, sharable access to customer data.

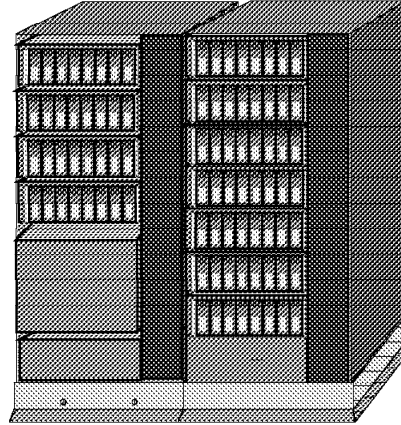
The disk drawers are state-of-the-art SSA drawers, providing disk insertion and removal features that allow configuration and maintenance to be performed without affecting host system up time.

The Ultrastar disk drives are IBM's latest drives, using such innovative features as thin-film disk, magnetoresistive heads, zoned recording, No-ID sector formats, and predictive failure analysis.

Versatile Storage Server Racks



- 2105-B09
 - Contains storage servers
 - Dual SMP clusters
 - Read Cache
 - I/O adapters and bays
 - 32 9.1GB drives
 - Redundant Power
 - Space for two 7133s
- 2105-100
 - Up to seven 7133 disk drawers
 - Fully redundant power supply
- Subsystem rack configuration
 - One 2105-B09 and up to two 2105-100s
 - Can connect existing 7015-R00 and 7202 racks, 7014-S00



© IBM Corporation 1996

Versatile Storage Server Racks

The VSS racks are an integral part of the complete VSS design. In other words, they are designed to house a high-availability storage subsystem such as the VSS. This foil presents an overview of the racks and how they make up the subsystem.

Two types of racks are used in the VSS subsystem: the 2105-B09 and 2105-100.

2105-B09

The 2105-B09 rack contains a storage server, dual SMP clusters, read cache, I/O adapters and bays, 32 9.1 GB drives, redundant power supplies and space for two 7133s.

Power requirements and specifications of the 2105-B09 rack are fully discussed under "The 2105-B09 Rack" on page 47. The VSS storage server is discussed in greater detail under "Storage Server (Front)" on page 50. The VSS storage server and disk drawers are discussed in greater detail under "The Versatile Storage Server" on page 52 and "The Disk Drawer" on page 72.

2105-100

The 2105-100 rack is similar in configuration to the 2105-B09, containing the same fully redundant power supply and space for seven disk drawers. It does not contain a storage server. The 2105-100 is fully explained in detail under “The 2105-100 Rack” on page 49.

Subsystem rack configuration

A subsystem always consists of one 2105-B09, and optionally up to two 2105-100 racks. Existing 7014-S00, 7015-R00 and 7202-900 racks can also be used as expansion racks to a 2105-B09 rack, but they do not support the power sequencing features and redundancy of the 2105-B09.

When positioning the VSS racks, any 2105-100 racks are physically placed next to the 2105-B09 rack, to allow connection of the power sequencing cables (the frames are bolted together). If existing 7015-R00 or 7202-900 racks must be used in the subsystem, they should be placed next to any 2105-100 racks.

Where there are only 2105-100 expansion racks in the subsystem, they can be placed up to 20 m (65 ft) away from the 2105-B09 if the appropriate RPQ is specified (this orders the cable that enables the 20 m distance). The standard configuration will have the 2105 racks bolted together.

The 2105-B09 Rack

0000 0000 0000 0000
00 00 00 00
00 00 00 00
00 00 00 00
00 00 00 00
00 00 00 00
00 00 00 00
00 00 00 00

Supports 19 in. drawers

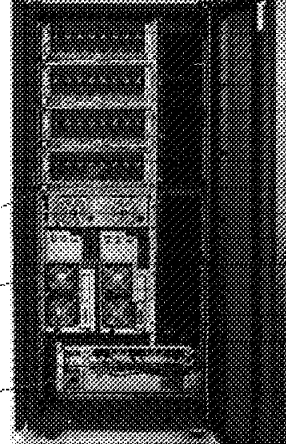
Single- or three-phase power
with dual line cords

Drawer signal and power cables
included

Up to four disk drawers

Electronics drawer

Fully redundant power supply



© IBM Corporation 1996

The 2105-B09 Rack

The 2105-B09 rack is 0.8m (33 in.) wide but supports 0.5 m (19 in.) drawers. It houses up to four disk drawers, a fully redundant power control subsystem, and a storage server containing two SMP clusters and four adapter bays. The four disk drawers provide a usable storage size of 456 GB.

Power is supplied to the 2105-B09 rack through two power cords carrying either 50 ampere single-phase or 50/60 ampere three-phase alternating current (AC) power, with each cord providing 100% of the needed power.

There are two 350 V DC bulk power supplies. One of these is enough to supply power to the subsystem, which provides for uninterruptible service if one of the DC power supplies fails. A 48 V DC transformer on each 350 V rail drops the voltage to the correct level for the individual components such as the SMP clusters and the adapter bays. The drive drawers are supplied with 350 V DC provided by the bulk power supplies. Because the drawers must have at least two of their three power supplies available to power up, each drawer has a single connection to each 350 V rail, plus a Y connection that connects to both rails. In the event that one of the 350 V bulk power supplies fails, the drive drawers can still power up.

In case of emergency, an emergency power off (EPO) switch is located on the rack, as well as power and service required indicators.

To guard against AC voltage sags and short-term outages, an optional battery can be fitted to the cabinet, providing 5 to 15 minutes of power in the event of AC power loss, depending on the configuration. The battery is not meant to provide uninterruptible service indefinitely. The storage servers will not initiate data destage or shutdown procedures in the event of an AC power loss for longer than the battery can provide backup. Typically, the host will initiate shutdown procedures, which will in turn ensure destaging of any modified data.

Power cables for the four disk drawers are included in the base rack, regardless of how many drawers are initially configured in the rack.

The 2105-100 Rack



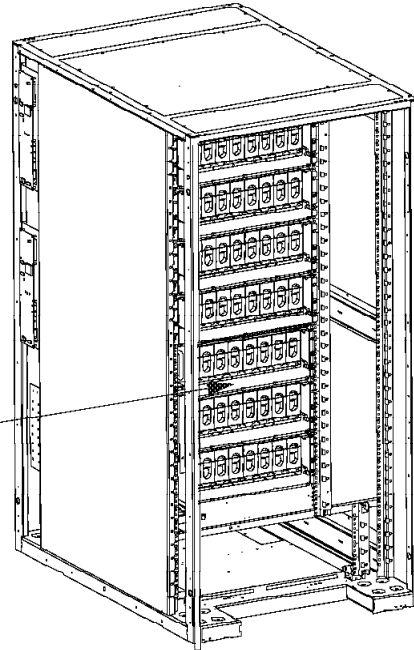
Supports 19 in. drawers

Single or three phase power
with dual line cords

Power sequenced from 2105-B09

Drawer signal and power cables
included

Up to seven disk drawers



© IBM Corporation 1998

The 2105-100 Rack

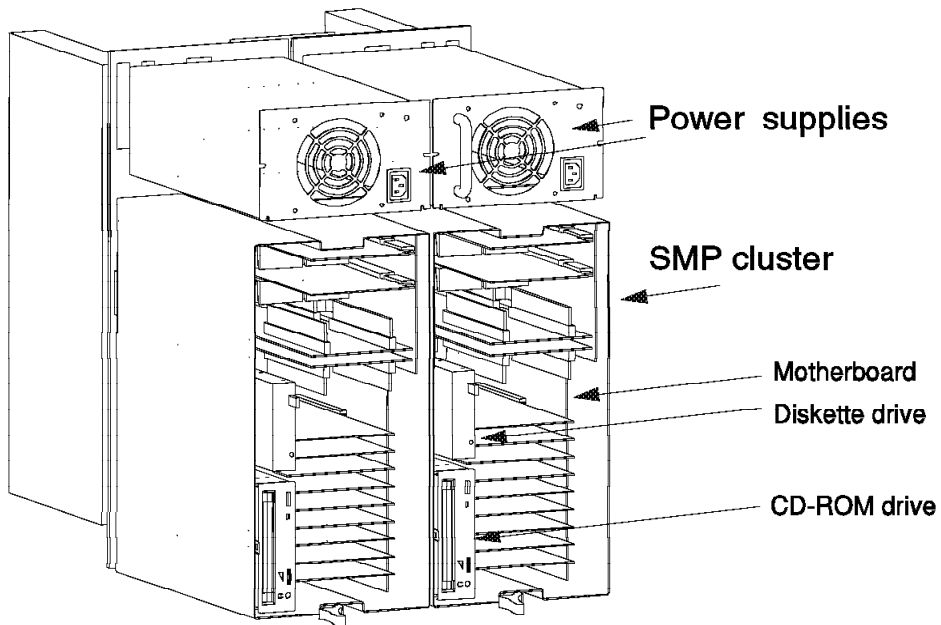
The 2105-100 rack is similar in configuration to the 2105-B09, except that it does not have the storage server that houses the SMP clusters and adapter bays. Instead, three more disk drawers can be installed in the space vacated by the storage server. A total of seven disk drawers can be installed in the 2105-100 rack, giving a maximum usable storage size of 800 GB.

The 2105-100 rack contains the same fully redundant power subsystem as the 2105-B09 rack. Power is supplied to the 2105-100 rack through two power cords carrying either 50 ampere single-phase or 50/60 ampere three-phase AC power, with each cord providing 100 % of the needed power.

The 2105-100 rack has no operator power-on, power-off switch. Power is sequenced from the 2105-B09 rack. For emergencies, an EPO switch is located on the rack, as well as power and service-required indicators.

The 2105-100 rack also has an optional battery that provides the same function it provides for the 2105-B09 rack: several minutes of backup to facilitate host shutdown.

Storage Server (Front)



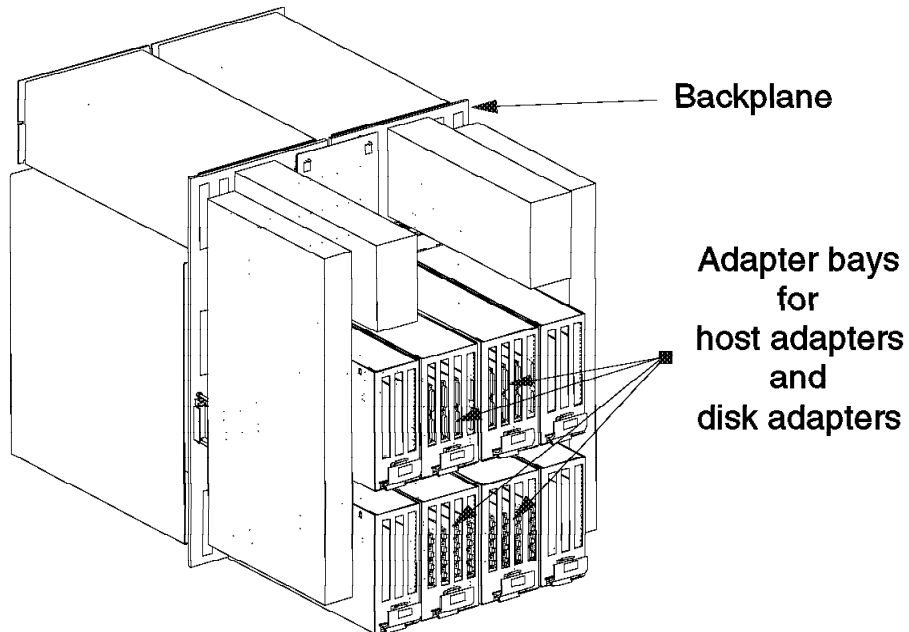
© IBM Corporation 1996

Storage Server (Front)

The foil shows a front view of the storage server showing the two SMP clusters installed. The DC power supplies for the storage servers are shown above each SMP cluster module.

Each cluster has its own motherboard that contains the SMPs, cache memory and other components (see "The Versatile Storage Server" on page 52 for a complete explanation of the various components of the storage server). In addition, each cluster has its own diskette drive and CD-ROM drive for loading microcode and diagnostics.

Storage Server (Rear)



© IBM Corporation 1998

Storage Server (Rear)

The foil shows a rear view of the storage server, revealing the adapter bays.

The four adapter bays are hot pluggable, hot removable bays used for connecting the host and drive adapters to the VSS storage servers. Each rack can be configured with a maximum of four adapters, giving a maximum of sixteen adapters. Of these, eight can be host adapters and eight can be drive adapters.

An internal backplane distributes DC power to the adapter bays and SMP clusters, as well as connecting the internal peripheral component interconnect (PCI) local bus between the bays and SMP clusters.

The Versatile Storage Server



- Symmetric multiprocessing (SMP) processors
- 332 MHz RISC CPUs
 - 0.5 micron technology ~5.1 million transistors
 - Low power: 2.5/5.0 V core, 3.3 V I/O
- CPU level 1 cache
 - 64 KB per CPU, 32 KB data, 32 KB instruction
- CPU level 2 cache
 - 256 KB per CPU external
- Snooping
 - Onboard, improves cache efficiency
- Packaging
 - Four CPUs per controller



© IBM Corporation 1996

The Versatile Storage Server

The Versatile Storage Server consists of two SMP-based processing clusters contained within the 2105-B09 rack. It uses leading-edge technologies to manage the interfaces to the connected hosts, disk drawers and drives, and cache.

Each SMP cluster uses a RISC planar, which contains a complete computer system in itself. It has all of the usual functional parts that a RISC has: central processing units (CPUs), level 2 cache, dynamic random access memory (DRAM), PCI bus, read-only storage (ROS), nonvolatile random access memory (NVRAM), an internal small computer systems interface (SCSI) bus, internal SCSI disk drive, flexible disk drive, compact disk read-only memory (CD-ROM), multidigit liquid crystal display (LCD), Ethernet, two serial ports, and a power supply.

Symmetric multiprocessing (SMP) processors

The CPUs run in an SMP configuration; that is, one common pool of memory contains a single copy of the operating system code. Each CPU accesses and executes the code independently of the other CPUs, at the same time preserving data integrity through the use of sophisticated locking mechanisms and private buses.

The 604e was designed to be used in both single-processor and multiprocessor applications. It contains these instructions specific to multiprocessing:

- Hardware-enforced cache coherence protocol for data cache (Modified/Exclusive/Shared/Invalid (MESI) protocol, explained below).
- Separate port into data cache tags for bus snooping. Cache tags are the memory addresses of the data that is currently in cache.
- Load and store with reservation instruction pair for atomic memory references and semaphores. In a multiprocessor environment, a single processor has to be able to load and store data in a cache or memory location without another processor attempting to do the same thing at the same instant. A load and store must be atomic; that is, it must complete in a single operation.

332 MHz RISC CPUs

Each SMP cluster uses four RISC 604e CPU chips, running at a clock frequency of 332 MHz. These chips are produced using 0.5 micron lithography technology. (the minimum width of a connection inside the chip is 0.5 micron). The 0.5 micron lithography process allows for efficient high density packaging; the 604e contains approximately 5.1 million transistors. The 604e also uses split voltages of 2.5 V DC for the core and 3.3 V DC for I/O. Split voltages greatly reduce power consumption—typical dissipation is 14 watts for the 604e.

CPU level 1 cache

Each CPU has 64 KB of level 1 (L1) cache internally. This cache is four-way set-associative split into 32 KB of instruction cache and 32 KB of data cache.

The instruction cache can provide up to four instructions in a single clock cycle. The RISC architecture defines special instructions for maintaining instruction cache coherence. The 604e implements these instructions.

The data cache is also four-way set-associative, and it is write-back with hardware support for reloading on cache misses. To ensure cache coherence in a multiprocessor configuration, the MESI protocol is implemented for each cache block. The four states indicate the state of the cache block:

- Modified — The cache block is modified with respect to system memory; that is, it is valid only in the cache and not in system memory.
- Exclusive — This cache block holds valid data that is the same as the data in system memory at this address. No other caches have this data.
- Shared — This cache block holds valid data that is the same as the data in system memory at this address, and at least one other cache.
- Invalid — This cache block does not hold any valid data.

CPU level 2 cache

Each CPU has 256 KB of level 2 (L2) cache externally. This cache consists of error correcting code (ECC) memory and is eight-way set-associative. L2 cache provides an extension of the L1 cache and improves overall performance by caching more data and instructions.

Snooping

The 604e provides for onboard snooping. It has a separate port into its data cache tags. Cache tags are the addresses of memory locations that are being held in cache. Snooping is the process whereby each CPU can access and examine another's L1 or L2 cache tags, determining whether to get the data it requires from another CPU or from main memory. As cache RAM is faster than conventional DRAM, snooping provides a significant improvement in cache hit efficiency.

Packaging

CPUs are packaged two per CPU board. Each SMP cluster has two CPU boards, or four CPUs per cluster, or eight CPUs in total.

More components of the VSS storage server are described on the next foil.

The Versatile Storage Server ...



- Internal SCSI adapter
 - 16-bit fast, controls internal disk and CD-ROM
- SCSI disk drive
 - Holds operating code, boot loader
- CD-ROM
 - Used for initial operating code load
 - Used for loading microcode, updates and diagnostics
- Diskette drive
 - Used for loading microcode and updates
- Ethernet
 - 10BaseT connection for intranet
 - Used for user configuration of the subsystem



© IBM Corporation 1998

The Versatile Storage Server ...

Internal SCSI adapter

Each SMP cluster has an internal 16-bit, fast SCSI adapter, used for controlling the internal SCSI disk drive and CD-ROM drive.

SCSI disk drive

Each SMP cluster contains an internal SCSI disk drive that is used to store its operating code and configuration data. It also contains the boot loader, diagnostics, and the reliability, availability, and serviceability (RAS) code. The cluster boots its code from the disk drive during its initialization sequence and uses it during operation for error logging and paging.

CD-ROM

Each SMP cluster contains an 8X speed CD-ROM that is used to initially load and install the operating code. It is also used to load diagnostics and microcode or operating code updates.

Diskette drive

A 1.44 MB 3.5 in. diskette drive is installed in each cluster. The diskette drive is used for loading microcode and updates, for when a diskette is the preferred distribution medium. It is also used for saving and loading configuration data; the diskette allows transfer of configurations between clusters and for offsite storage of configuration data.

Ethernet

A 10 Mbit/s Ethernet adapter is installed in each SMP cluster. The Ethernet adapter is used to connect to the customer's intranet (through 10BaseT) and is the primary interface for configuration of the subsystem.

The Versatile Storage Server ...



- Serial ports
 - One in each SMP Cluster has modem connection for "call home" and remote diagnostics
 - One in each cluster connects to RS-232 switch, to allow connection of ASCII terminal or CE laptop for maintenance purposes
- LCD display
 - Displays multi-digit status codes during boot and operation
- Service processor
 - Performs basic monitoring functions of the CPUs
 - Allows access when storage server microcode is not running
 - Can initiate "call home" when main CPUs are not running



© IBM Corporation 1998

The Versatile Storage Server ...

Serial ports

Two RS-232 serial ports per cluster are used for access by service personnel. One port is connected to a modem to enable remote diagnostic access and "call home." The "call home" feature enables the VSS to contact the IBM service center in the event of a failure or if the subsystem requires attention by a Customer Engineer (CE).

The second port is used by the CE for maintenance and repair actions to the subsystem.

LCD display

A multidigit liquid crystal display (LCD) is used to display status codes while the cluster boots and while the storage server is operational. Typically, the CE will use the status codes when performing maintenance and repair actions on the VSS.

Service processor

The service processor provides a means for low-level diagnosis of the storage server, even when the main CPUs are inoperable. The service processor allows access to cluster NVRAM, VPD, and error data. The service processor can also initiate a call home when the main CPUs are not running.

Storage Server Cache



- Error correcting code
 - Single bit correction, double bit detection
- Synchronous DRAM (SDRAM)
 - Bursting feature
 - Automatic precharge
 - Reduced number of access cycles
- Up to 6 GB of read/write cache (3 GB per storageserver)



© IBM Corporation 1998

Storage Server Cache

Each SMP cluster can be configured with up to 3 GB of ECC synchronous dynamic random access memory (SDRAM) as cache for temporary storage of read and write data from the subsystem disk drives. A combined total of up to 6 GB of cache memory can be installed in each subsystem.

Error correcting code

The ECC SDRAM provides single-bit error correction and double-bit error detection. When data is written to memory, the controller calculates check bits and stores those along with the data. On a read operation, the controller calculates the check bits again and compares them with the originals. Therefore, if a single bit of the memory location goes bad, the memory controller reconstructs the missing data. If two bits of the memory location go bad, the memory controller detects this and generates a check condition.

Synchronous DRAM (SDRAM)

SDRAM is the latest DRAM. When CPU speeds exceed 66 MHz, conventional DRAM is not fast enough to supply data to the CPU as well as perform its own refresh operations. Another limitation is that an external controller must supply each and every address that is to be accessed during a read or write operation—asynchronous operation. Many applications access memory in large sequential streams, or bursts. To accommodate access and provide better performance, SDRAM incorporates a bursting feature, which allows the DRAM

itself to provide the address of the next location being accessed once the first page access has occurred—synchronous operation.

SDRAM also incorporates automatic precharge, which eliminates the need for an external device to close a memory bank after a burst operation. In addition, it allows two row addresses to be open simultaneously. Accesses between two opened banks can be interleaved, hiding row precharge and first access delays.

Overall, SDRAM provides significant increases in performance. In CPU cycles, after an initial access of five cycles, SDRAM can provide subsequent bits at a rate of one per CPU cycle. This is twice the speed of extended data out (EDO) DRAM, or three times that of conventional fast page mode (FPM) DRAM.

Up to 6 GB of read/write cache (3 GB per storage server)

The basic VSS subsystem is configured with 256 MB of cache per cluster, or 512 MB in total. Total cache options are 512 MB, 1 GB, 2 GB, 4 GB, or 6 GB.

The cache is used as read and write cache but is considered volatile because it has no battery backup. The SSA disk adapters have nonvolatile storage (Fast Write Cache) and a write is not considered complete until the data has been transferred from the cluster cache into the disk adapter Fast Write Cache. As the disk adapter Fast Write Cache is battery protected, the host operating system is informed that the write is complete without having to wait for the data to be written to disk. This facility is known as *fast write*.

If an application or the host operating system requires the read-after-write function to verify data integrity, the read request is satisfied from the cluster cache, thus increasing the performance of this typically performance-degrading procedure.

Performance Enhancing Caching Algorithms

- Least recently used (LRU) destage
- Staging
 - Record
 - Partial track
 - Full track
- Adaptive cache algorithm
 - Predicts how to best serve a track of data
- Sequential prediction on read
 - Enhances read cache hits
- Fast Write Cache bypass on greater than Express Writes
 - Avoids flooding of adapter Fast Write Cache on writes



© IBM Corporation 1998

Performance Enhancing Caching Algorithms

A number of caching algorithms incorporated into the storage server microcode enhance performance of the subsystem. These are covered in full detail in Chapter 4, “Versatile Storage Server Data Flow” on page 109.

Least recently used (LRU) destage

Destage is the transfer of data from the disk adapter DRAM to physical disk media. When space is required for allocation in cluster cache or adapter Fast Write Cache, or the Fast Write Cache reaches 70% full, track slots that have not been accessed for the longest time are destaged from the DRAM.

The adapter attempts to destage a complete stripe if possible. This full-stripe destage improves performance by eliminating the RAID-5 overhead of having to read old data and parity, then write new data and new parity. Writing a whole stripe does not require two read operations and an extra write operation. Once an express write is in Fast Write Cache, it is destaged immediately, regardless of its LRU status.

If an express write of data is transferred from the cluster cache to the adapter in one operation, it is placed in the adapter memory (not Fast Write Cache) for calculation of parity data and then immediately written to disk without being transferred to Fast Write Cache.

If a whole stripe is not available when a record is scheduled for destage, all modified records of the particular track must be destaged.

If a particular track of data is constantly updated, it could remain at the bottom of the LRU list (that is, most recently used) and never be scheduled for destage. To circumvent this, an idle timer is implemented. After 5 s of idle time, a small percentage of the cache will be destaged on an LRU basis. Thus a constantly written block of data can be updated without being constantly destaged, and data integrity is maintained.

Staging

Staging is the term used for reading data from the physical disk media into the cluster cache. When the host accesses a track, if it is not in cache, a read “miss” occurs. Three types of staging can occur during the read access to the data:

- Record mode

Only the records asked for by the host are staged. Record mode is the default mode; that is, when a track is accessed for the first time, only the records asked for will be staged. If the track is being accessed frequently, it will be “promoted” to either partial or full-track staging (see below). Typically, data being serviced in record mode is being accessed so infrequently and randomly that the adaptive cache algorithm (see below) has decided it is better served in record mode.

- Partial track

The initial transfer of the data to cache is from the initial record accessed to the end of the track. If the initial record accessed is not from the index of the track, the cache image is a partial track.

- Full track

A full track is staged to the cluster cache if the initial record accessed is from the index record.

Adaptive cache algorithm

The adaptive cache algorithm is used to predict how much of a track should be staged on a read miss. For the purposes of this algorithm, disks are divided into bands. Various statistics about the usage patterns of the bands are maintained and used to determine whether to place a band into record mode, partial track, or full track staging.

Sequential prediction on read

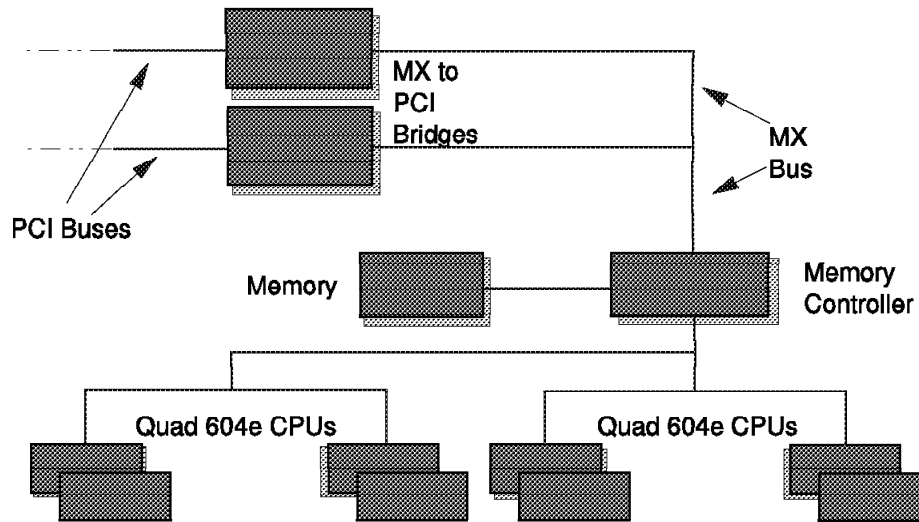
To enhance performance, a counter is kept for each track staged. When a track is staged, the previous track is looked up, and the counters are checked. If the previous track is in cache, the counter of the current track is incremented and checked against a preset value. If the counter shows more than the preset value, sequential access is predicted and a stage to the end of the sequential stage group is initiated. A sequential stage group is up to five RAID-5 stripes, depending on the configuration of the RAID rank.

Fast Write Cache bypass on greater than full-stripe writes

On write operations, if the amount of data being written is greater than a full RAID-5 stripe, the data is not backed up to the Fast Write Cache to prevent flooding it. The adapter still calculates the RAID-5 parity data but writes directly to disk from DRAM instead of mirroring to Fast Write Cache and then writing from DRAM.

For a full description of data flow and caching mechanisms, refer to Chapter 4, “Versatile Storage Server Data Flow” on page 109.

PCI Local Bus



© IBM Corporation 1996

PCI Local Bus

The PCI local bus was developed by IBM and others to create an industry standard local bus that could be used across various processor architectures as well as support the architectures as they evolved. PCI provides significantly improved performance over other common bus architectures. For example, it can complete a 4-byte write in as few as two bus clock cycles.

PCI is a clock synchronous bus that operates at 32 MHz. It provides a combination address and data bus that is multiplexed. The basic transfer operation for the PCI bus is a burst operation that enables a contiguous block of data to be transferred on the bus immediately after the address.

Each SMP cluster has two 4-byte PCI buses running at 32 MHz.

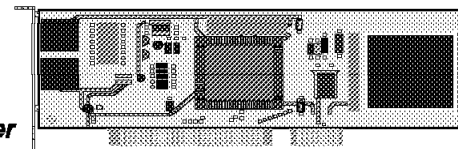
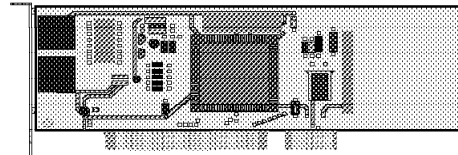
The foil shows a block diagram of the storage server memory, CPU, MX, and PCI buses. The M- to-PCI bus bridges are custom chips that allow any PCI adapter to access the memory bus or be accessed by the memory controller. Because the PCI buses are isolated from the system bus with the bridge chip, all CPUs have access to both PCI buses.

Adapters



- Host adapter
 - 32-bit PCI Ultra SCSI adapter
 - ▶ *Two host channels, 16 bit differential*
 - ▶ *SCSI-3 protocol and command set*

- Disk adapter
 - 32-bit PCI SSA adapter
 - ▶ *Similar to 6215 adapter with special firmware*
 - ▶ *Supports two loops*
 - ▶ *Fast write cache (NVS)*



**SSA
Adapter**



© IBM Corporation 1996

Adapters

The adapters installed in the VSS 2105-B09 rack connect to both hosts and disks. They are installed in the hot insertion and removal adapter bays, which are located in the rear of the 2105-B09 rack. Up to sixteen adapters in total can be installed: eight host adapters and eight disk adapters.

Host Adapter

The host adapter is a 32-bit PCI Ultra SCSI adapter. It supports the SCSI-3 protocol and command set and provides two 16-bit-wide, 40 MB/s differential channels, capable of reading and writing concurrently. Each channel of the host adapter can support up to 16 target IDs like any 16-bit SCSI bus, but each can support up to 64 logical unit numbers (LUNs) per target ID (in conformance with the SCSI-3 protocol).

Each of the channels can be attached to a single host, two homogeneous hosts, or two heterogeneous hosts. The hosts must attach either through an Ultra SCSI differential interface or a SCSI-2 fast and wide differential interface.

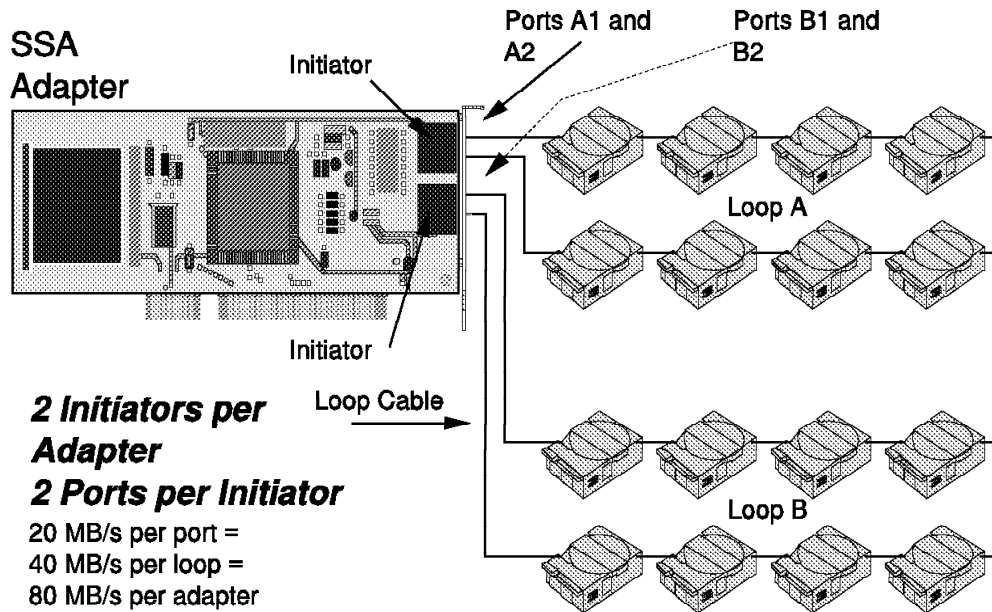
In the base VSS subsystem, two host adapters are included.

Disk Adapter

The disk adapter used in the VSS subsystem is a 32-bit PCI SSA RAID-5 adapter. This adapter is similar to the current 6218 SSA RAID-5 adapter used in the RS/6000 range of products but has a special firmware load for use in the VSS. Therefore, customers with existing 6218 adapters cannot migrate them to the VSS.

The VSS SSA adapter supports two separate loops of disks. As a general VSS configuration rule, 16 disks are supported per loop, because 16 disks are supported in one SSA drawer. In a maximum configuration of 18 drawers, however, two loops will support 32 drives in two drawers.

SSA Relationships



© IBM Corporation 1996

SSA Relationships

The foil shows a typical SSA adapter with dual loop configuration.

An SSA adapter contains two SSA initiators. An SSA initiator is an SSA node that is capable of initiating SSA commands. Each of the two initiators on the adapter is capable of functioning as a master. An SSA master can access a node by specifying the node's address in the SSA network. When an SSA network contains more than one initiator node, the initiator with the highest unique ID is the master.

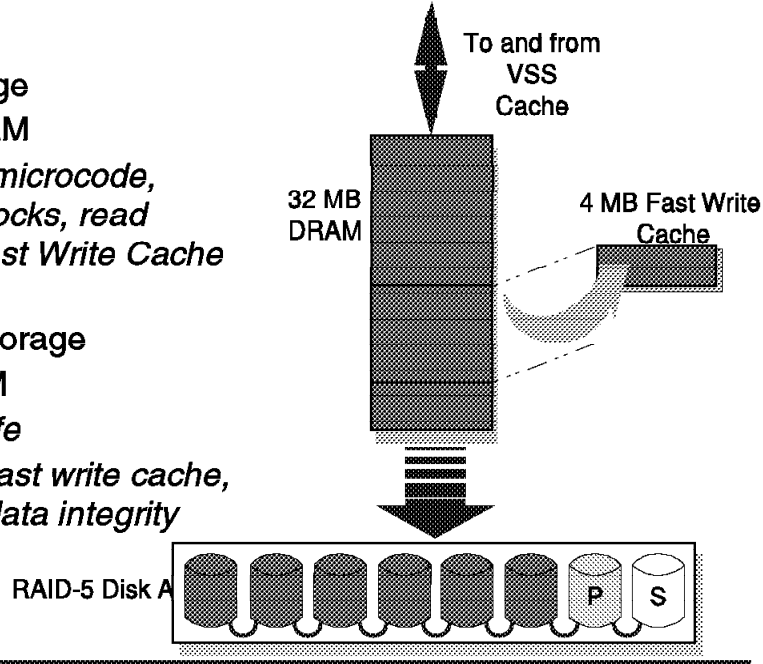
Each initiator controls two ports. When connected through a series of SSA disk drives, the ports form a loop.

Each port is capable of transmitting and receiving at 20 MB/s, giving an overall bandwidth of 40 MB/s per loop, or 80 MB/s for the adapter.

Disk Adapter Memory



- Volatile storage
 - 32 MB DRAM
 - *Used for microcode, control blocks, read cache, Fast Write Cache mirror*
- Nonvolatile storage
 - 4 MB SRAM
 - *10-year life*
 - *Used as fast write cache, ensures data integrity*



© IBM Corporation 1998

Disk Adapter Memory

The SSA adapter contains both volatile and nonvolatile memory to assist in caching, staging, destaging, and RAID parity calculations.

Volatile storage

The adapter contains 16 MB of DRAM, which is used for transferring data between the disk arrays and the SMP cluster cache. The adapter DRAM is the only place where RAID parity data is stored once it is staged from disk.

Parity data is never transferred to the VSS cache. It is cached in DRAM to assist in calculation of the new parity data if the data is modified. In addition, the DRAM is used for storing the adapter microcode, its code control blocks, and a mirror of the adapter Fast Write Cache. The DRAM is not partitioned in a physical sense; specific sizes are not set aside for read cache, fast write cache, adapter microcode, or control blocks. The size of the read cache depends on the amount of unused DRAM at any given time. The size of the fast write cache is limited to the size of the Fast Write Cache, 4 MB.

Nonvolatile storage

To prevent data loss in the event of power outages or cluster failure, the SSA adapter contains 4 MB of SRAM. The SRAM has a lithium battery backup and is designed to keep the data integral for 10 years. SRAM uses far less power than DRAM to maintain data integrity. When data is written from the SMP cluster cache to the adapter DRAM, it is copied into the Fast Write Cache to guard against data loss before the data is actually written to disk. As soon as the Fast Write Cache has a copy of the data, the adapter signals the SMP cluster that the write operation is complete, and in turn the SMP cluster signals the host that the write is complete. Fast Write provides a significant increase in performance over the usual practice of the host having to wait until the data is committed to disk. Without Fast Write Cache, it would not be possible to perform this fast write function and maintain data integrity.

In some circumstances, however, it may be undesirable to copy data into the Fast Write Cache. The following conditions cause data to bypass the Fast Write Cache:

- Full stripe write

If a full stripe of data is written from the VSS cache, it is beneficial to simply calculate parity and write straight to disk before signaling write completion.

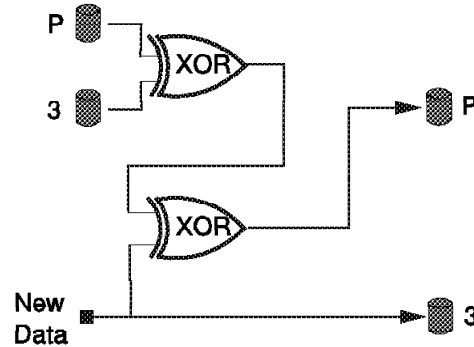
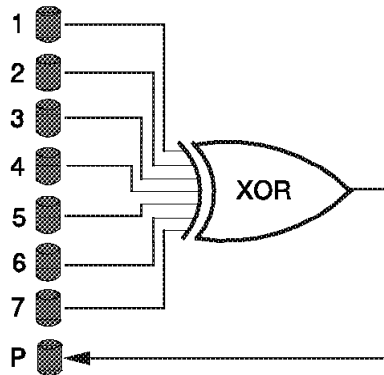
- Transfer size larger than full stripe or larger than Fast Write Cache size

If a write from VSS cache is larger than a full stripe, or larger than the Fast Write Cache size (4 MB), parity is calculated, and the data is written straight to disk to avoid flooding the Fast Write Cache.

Exclusive Or Function



- Parity calculation
 - Allows reconstruction of data
 - Minimizes I/Os required to write modified data



© IBM Corporation 1996

Exclusive-Or Function

Part of the function of RAID is the calculation of parity data that is used to reconstruct customer data in the event of a disk failure within the array. This function is easy to implement and can be implemented in either software or hardware. In the VSS subsystem, the Exclusive-Or (XOR) function for calculating the parity data is implemented in the SSA RAID adapter. Implementation of the XOR function in the adapter hardware results in faster parity calculation, which aids in write and data reconstruction performance.

Parity calculation

To maintain data availability if a drive in the array fails, when data is written an extra parity strip is written along with the data. Depending on the configuration of the array, a full stripe of RAID data on the VSS is either 224 KB for a six-drive-plus-parity array (6+P) or 256 KB for a seven-drive-plus-parity (7+P) array. From the numbers, we can see that a single strip of data is 32 KB. A strip is the amount of data written to a single drive in the array. In a 7+P array (eight drives in total), a full stripe is $8 * 32 \text{ KB} = 256 \text{ KB}$.

The parity data is calculated by XORing the seven 32 KB strips together. The foil shows a representation of the XOR function.

The XOR function is reversible in that, if you apply XOR to the parity data on any one of the seven data strips, the result is the XOR of the remaining six data strips. This reversibility provides two distinct advantages:

- If a single drive (strip) is lost from the array, the data can be reconstructed from the remaining data by applying XOR to the parity and comparing with the data from the other six good strips.
- If a single strip is updated, the parity is XORed with the *old* data, resulting in the XOR of the six unmodified strips. This result is then XORed with the new data to produce a new parity. Likewise, this process can be applied when more than one of the strips are updated. The foil shows a single strip update and parity calculation and update.

The XOR function minimizes the amount of I/Os that have to occur to calculate new parity data when a strip is updated. It minimizes the number of I/Os to four per write: read old data, read old parity, write new data, write new parity.

The Disk Drawer



- 7133 SSA disk drawer
 - 19 in. rack form factor
 - 16 hot insertion and removal slots for drives
 - 3 power supply and cooling modules
- 7133-010
 - Supported if customer already has these installed
- 7133-020
 - New improved features



© IBM Corporation 1997

The Disk Drawer

This foil discusses the types of disk drawers that are supported in the VSS subsystem.

7133 SSA disk drawer

The 7133 SSA disk drawer is a rack-mounted drawer that can contain up to 16 SSA disk drives. It fits into a standard 0.5 m (19 in.) EIA rack and takes up four EIA units vertically.

The 7133 contains four physically separate loops of slots for installing disk drives. The slots support hot insertion and removal of drives.

If disk drives are not installed in drive slots, a dummy drive must be installed to propagate the signals from adjoining slots. These four loops can be physically cabled in a variety of different configurations. In the VSS subsystem, however, only one loop of 8 or 16 drives per drawer is supported, so all four drive slot loops must be cabled together.

Slots 1 to 8 are physically located at the front of the 7133, numbered from left to right as you face the front. Slots 9 to 16 are physically located at the rear of the 7133, numbering from left to right as you face the rear of the 7133.

A 7133 has three separate slots for power supplies and cooling fans. In the VSS subsystem, all three power supply and cooling fan slots must be populated to ensure high availability.

7133-010

Two types of 7133 disk drawers are supported in the VSS subsystem: the 7133-010 and the 7133-020. Although the 7133-010 is no longer available, customers who already have the 7133-010 drawers can use them in the VSS subsystem.

There are some considerations when using 7133-010 drawers. For example, all disks within the drawers must be reformatted to the VSS specification. Thus the data cannot be migrated directly from existing RAID drawers to the VSS subsystem. Also, because the 7133-010 does not provide sufficient cooling air flow to support the 9.1 GB disk drives, it supports 4.5 GB disk drives only.

7133-020

The 7133-020 is the latest SSA disk drawer available from IBM. It is similar in design to the 7133-010, but has some upgraded features and, because of improved cooling air flow within the drawer, supports the 9.1 GB disk drives in addition to the 4.5 GB disk drives. Customers with existing 7133-020 drawers can also use them in the VSS. Differences between the 7133-010 and -020 are discussed under “7133 Model Differences” on page 76.

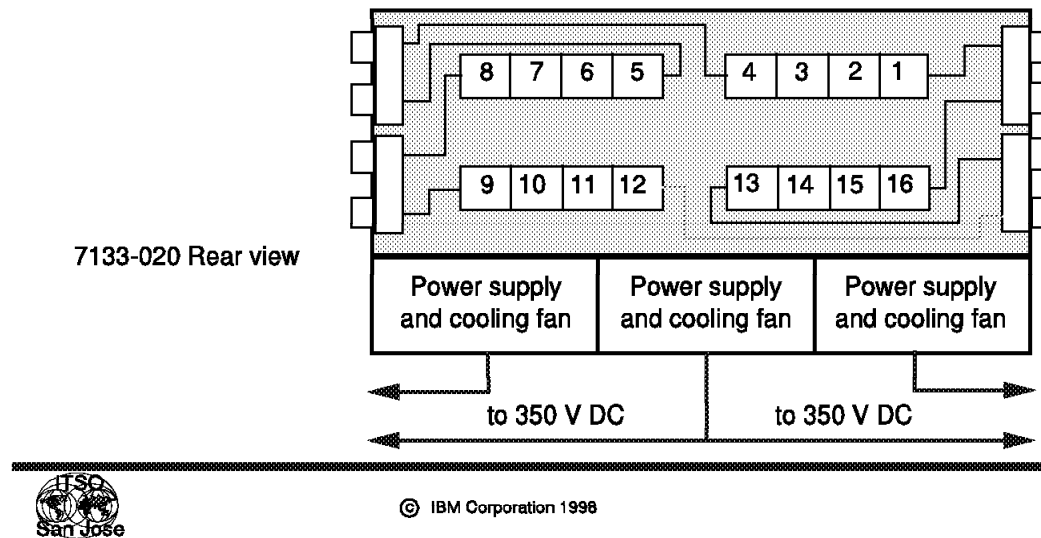
For more information about the VSS sector format, see “Versatile Storage Server 524-Byte Sector Format” on page 108.

For more information about migrating data with existing 7133s, see Chapter 7, “Migrating Data to the Versatile Storage Server” on page 221.

7133 Power and Cooling



- Three power supply modules
 - Drawer can run with only two operational
 - Input is 350 V DC in VSS
 - Improved cooling in 7133-020 supports 9.1 GB drive



7133 Power and Cooling

This foil shows a representation of the rear view of a 7133-020 disk drawer. The 7133-010 is similar, with respect to power supply and cooling modules.

Three power supply modules

Each 7133 drawer has slots for three power supplies. Earlier 7133-010 models had only two slots populated; the third power supply module was optional. Customers who have 7133-010 drawers must install the optional third module in all 7133-010s destined for use in the VSS subsystem. The 7133-020 has three modules installed in the base configuration.

Each power supply module also has a cooling fan assembly which not only cools the power supply module but also provides cooling to the 7133 drawer and the drive modules installed within. The 7133-020 has redesigned cooling air flow to support the 9.1 GB drive modules.

Each module can run off a 90 to 260 V AC line at a frequency of 47 to 64 Hz, or a 240 to 370 V DC line. In a VSS, the 7133s are powered from the two 350 V DC bulk power supply rails and do not use AC voltage. One module in the drawer is connected to both 350 V DC rails, whereas the other two connect individually to only one of the 350 V DC rails (refer to the foil). The connections are made this way to facilitate the redundancy built in to the power supply system. A 7133 can operate with only two of its three power supply modules operational, because two modules can supply all of the 7133 power requirements. If one of the 350 V

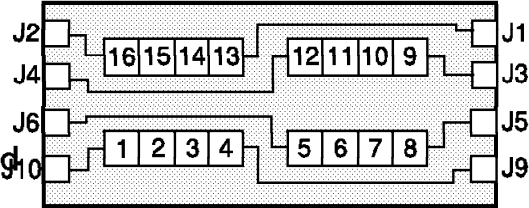
DC power supplies should fail, at least two of the power supply modules in the 7133 will have power and will therefore remain operational.

7133 Model Differences

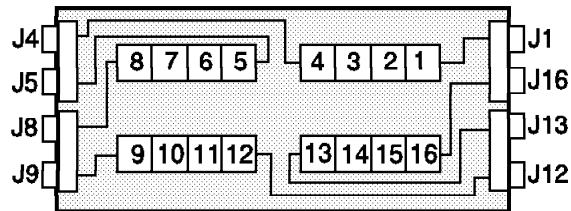


- Automatic host bypass circuits
- Third power supply module
- Redesigned front panel and on/off switch
- Different cable connector numbering

7133-010
Cabling



7133-020
Cabling
(rear view)



© IBM Corporation 1998

7133 Model Differences

The 7133-010 and 7133-020 differ slightly in design. The basic differences are these:

Automatic host bypass circuits

In a 7133-020, loop jumpers are not required to be fitted to maintain loop integrity in the event of a host, adapter, or cable loss. Automatic host bypass circuits also minimize the amount of intradrawer cabling required. The automatic host bypass circuits are discussed fully under "7133-020 Host Bypass Circuits" on page 78.

Third power supply module

A 7133-020 comes with the third power supply module in the base configuration. For a 7133-010, it is optional. However, 7133-010 drawers migrated to the VSS must be fitted with the third power supply.

Redesigned front cover and on-off switch

The 7133-020 has a redesigned front cover and on-off switch, which gives better visibility of the drive warning lights and access to the on-off switch.

Different cable connector numbering

In a 7133-010, the cable connectors for attaching the SSA cables are numbered in a way that makes no reference to the drive to which the cable connects. This makes cabling confusing and nonintuitive. The 7133-020 design has changed the numbering. The connector numbers are now more meaningful, as detailed below:

7133-010 Connectors

Connectors	Disk Drive Modules
J1 and J2	Back modules 13 through 16
J3 and J4	Back modules 9 through 12
J5 and J6	Front modules 5 through 8
J9 and J10	Front modules 1 through 4

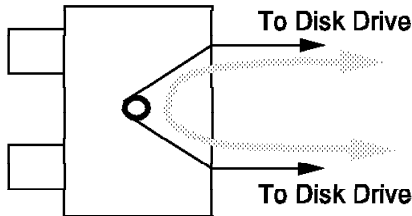
7133-020 Connectors

Connectors	Disk Drive Modules
J1 and J4	Front modules 1 through 4
J5 and J8	Front modules 5 through 8
J9 and J12	Back modules 9 through 12
J13 and J16	Back modules 13 through 16

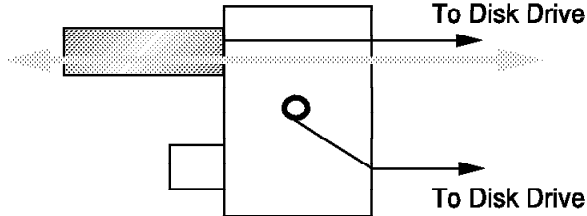
7133-020 Host Bypass Circuits



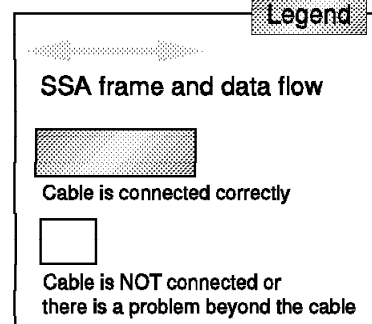
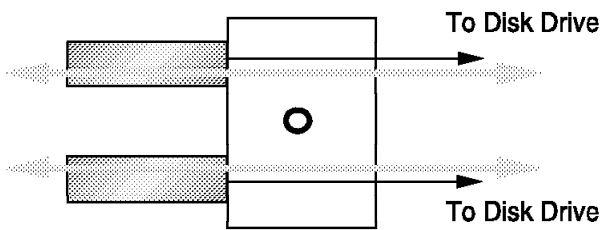
1. No SSA cable is connected or adapter fails.



2. One SSA cable is connected or adapter fails



3. Two SSA cables are connected, no adapter failure



© IBM Corporation 1996

7133-020 Host Bypass Circuits

In a 7133-010, the SSA loop connectors at the back of the unit are part of an SSA signal card, which connects two external SSA connectors to the backplane of the 7133. There are four cards per 7133 unit. When cabling the disk loops, for example, if you wanted to connect two loops of four into a single eight-disk loop, you would have to use loop jumper cables. In the event of a host or adapter failure, a loop jumper cable would have to be installed temporarily while the host was repaired or the adapter was replaced.

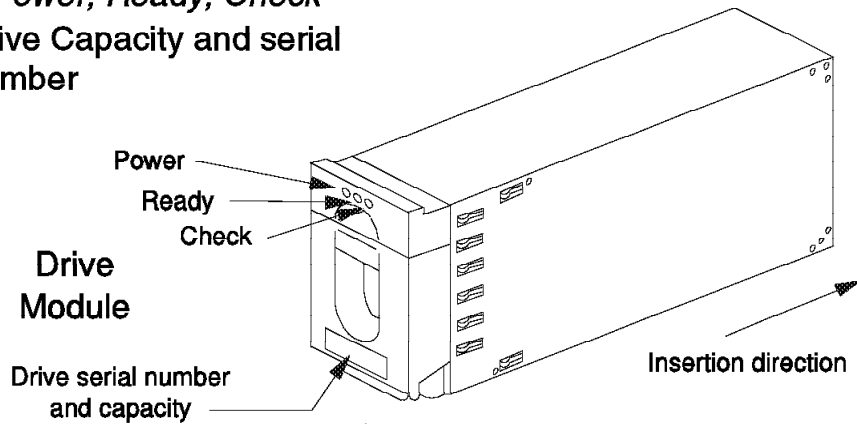
In a 7133-020, the SSA signal cards have been replaced by SSA bypass cards. The bypass cards can operate in two modes: bypass mode or forced inline mode. The bypass cards have active circuitry that can detect whether a cable is plugged into an external SSA connector. If a cable is not plugged in, the card operates in bypass mode, where the two connectors are looped together (Sequence 1). Bypass mode minimizes the amount of intradrawer cabling required and automatically connects a loop if the host or adapter fails. If the card detects that a cable is plugged in to one of the external connectors, it switches inline, connecting the internal SSA electronics to the external connector (Sequences 2 and 3).

The bypass circuitry can be disabled by an IBM service representative if a customer wants to run its 7133-020s in 7133-010 cabling mode. This is known as *forced inline mode*.

Drive Modules



- Disk drive module
 - Contains disk drive and supporting hardware
 - Status LEDs
 - *Power, Ready, Check*
 - Drive Capacity and serial number
- Dummy drive module
 - Contains connections for propagating SSA signals between adjoining slots



© IBM Corporation 1998

Drive Modules

Here we discuss the drive modules, which contain the actual disk drives. We discuss the disk drives themselves on the next foil.

Disk drive module

The disk drive module contains all of the hardware to physically connect the disk drive inside to the 7133 backplane. The module connects power and signals to the disk drive. The module is designed to be installed and removed while the 7133 is powered on and running. It consists of a metal container, which contains the disk drive, and a front panel assembly, which contains three drive status light-emitting diodes (LEDs) and a latching mechanism to secure the module in the 7133. Optionally, a locking mechanism that prevents the module from being removed inadvertently can be installed in the front panel of the disk drive module.

The LEDs provide the following status indicators:

- Power LED
 - This green LED, when lit, indicates that the drive module has its required power present.
- Ready LED

This green LED, when lit solidly, indicates that both SSA connections to the drive are good, and the drive is ready to accept commands from the host adapter.

When flashing slowly, the ready LED indicates that only one of the SSA connections is good.

When flickering, the ready LED indicates that the drive is being accessed or executing a command.

- Check LED

This LED is amber and when solidly lit indicates one of the following conditions:

- A failure in the module has been detected.
- Automatic self tests are being run.
- The drive is in service mode, meaning that the host system is not using the drive and service actions can be performed.

When the check LED is flashing, it indicates that the drive has been selected by the **Identify** option of the Set Service Mode service aid. This option is used to help confirm the identity of a physical drive module.

Dummy drive module

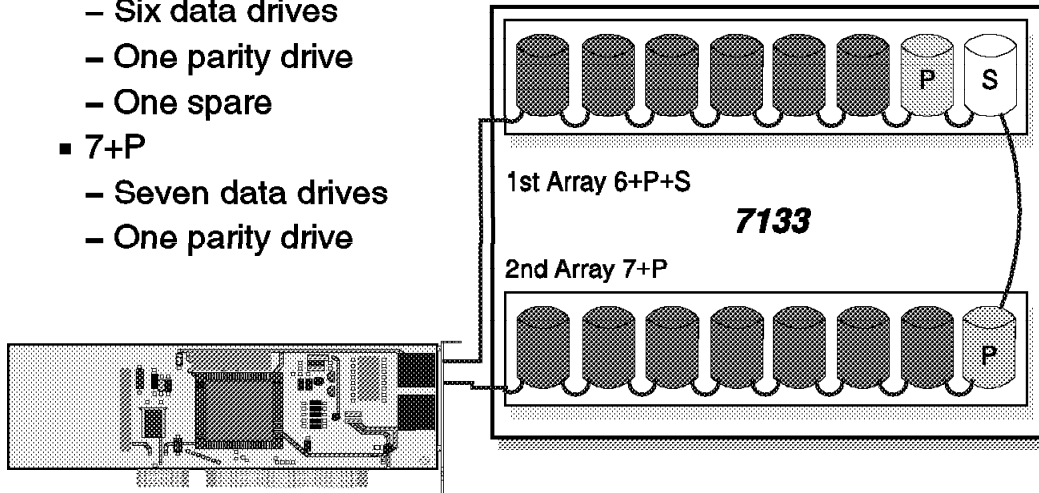
The dummy drive module is exactly the same size and shape as the disk drive module; however, it does not contain a disk drive or its associated electronics. The dummy module contains no active electronics, but it does contain passive connections to propagate SSA signals to and from adjoining slots. It is a necessary part of the 7133. The 7133 backplane does not complete the loop if a slot is empty. A dummy module must be installed in every slot that does not contain a drive.

As the dummy drive module does not contain any active circuitry, a limitation on the number of adjacent dummy modules is imposed. Only three dummy modules can be adjacent to each other, either in the same loop or adjoining loops. More than three dummy modules adjacent in a loop degrades signal quality enough to cause errors.

Array Definition



- 6+P+S
 - Six data drives
 - One parity drive
 - One spare
- 7+P
 - Seven data drives
 - One parity drive



© IBM Corporation 1998

Array Definition

An array is defined as a number of disks connected in a single loop that are part of a RAID configuration. In the VSS, two types of arrays are supported. Each array contains eight disks in a RAID-5 configuration, but the function of the disks varies slightly in the two configurations. The two types are the six-drives-plus-parity-plus-spare (6+P+S) array, and the seven-drives-plus-parity (7+P) array. The foil shows a typical configuration of a 7133 drawer and two arrays.

As a single loop of SSA disks in a VSS can contain up to 16 drives, one or two arrays are supported per loop. For each drawer, at least one array must be a 6+P+S array. The other can be another 6+P+S array, or a 7+p array.

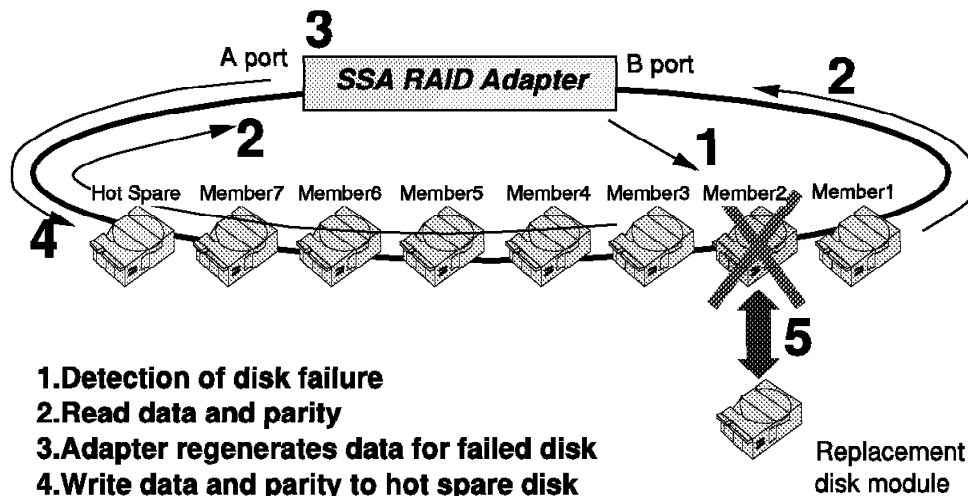
6 + P + S

The 6+P+S array contains six disks used to store customer data, one parity drive, and one spare drive. In a RAID-5 configuration, parity is not stored on one single drive, but for purposes of discussion, we assume that the parity is written on one drive. The spare drive is required to support a single loop in case of drive loss during normal operations. The adapter can detect the loss of a drive and perform an automatic reconstruction of the data for the lost disk. The reconstructed data is then written to the spare, and the bad disk is replaced as soon as possible. The spare drive must be the same size as the data and parity disks.

7+P

The 7+P array is used for storing customer and parity data and can only exist in a loop that already has a 6+P+S array.

Sparing



1. Detection of disk failure
2. Read data and parity
3. Adapter regenerates data for failed disk
4. Write data and parity to hot spare disk
5. Replaced disk is now configured as new hot spare disk



© IBM Corporation 1998

Sparing

The foil shows a representation of the automatic sparing procedure that is performed by the SSA RAID adapter in the event of a disk failure. The sparing function of the adapter provides a seamless method of taking the drive out of service for replacement.

In this case we assume that the array is configured as a 6+P+S array, that is 6 data drives, 1 parity drive, and 1 hot spare drive. A hot spare drive is one that is actually in the array and powered on, but idle; it is hot because it is at operating temperature. The built-in idle functions of the drive—disk sweeping and channel calibration—ensure that the drive is in good shape when required by the sparing operation.

Referring to the foil and following the sequence numbers, assume that Member 2 of the array above fails in some way. If the failure is complete (that is, the drive no longer responds to commands), the loop is broken and the SSA adapter is informed by the drives on either side of the failed drive in the loop (Members 1 and 3); see Sequence 1.

At this time, two things happen: first, any requests for the data from the failed disk are regenerated by the RAID adapters XOR function (see Sequence 3) by reading the parity data, and data from the other six drives (see Sequence 2). The increased number of reads required, and the reconstruction calculation in the adapter, cause a slight degradation in performance. Second, the hot spare

disk is now brought online into the array and the reconstruction of data from Member 2 is written to it (see Sequence 4). The reconstruction of data is performed sequentially across the disk, starting at the first LBA. To avoid unnecessary IOs, if requests are made for data not yet reconstructed (that is, out of sequence), the data is reconstructed by the adapter, passed to the requestor, and then written to the spare immediately to save having to recalculate it later in the process. Some degradation of performance may be noticed while the hot spare drive is rebuilt.

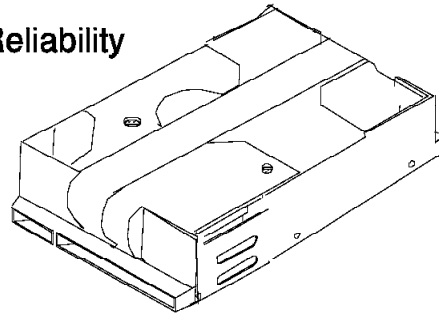
The hot spare disk is now no longer the spare, but a functional drive in the array, storing customer and parity data. The failed drive can now be replaced. Once it has been replaced, it then assumes the role of hot spare for the array (see Sequence 5).

The failed drive should be replaced as soon as possible, because as long as the failed drive remains in the array, should another drive fail (an unlikely event) no spare would be available. The adapter would then have to reconstruct data for every read request, resulting in decreased performance until drive replacement and data reconstruction had taken place. For more information on recovery procedures, see Chapter 10, "Subsystem Recovery" on page 333.

Disk Drives



- IBM Ultrastar 2XP
- Features
 - Thin film disk
 - Magnetoresistive heads
 - Partial response maximum likelihood
 - No-ID sector
 - Zoned recording
- Areal density
 - 829 Mbit/sq in.
 - 134,500 BPI
 - 6160 TPI
- Capacity
 - 4.5 GB
 - 9.1 GB
- Performance
 - 7200 rpm
 - < 8.5 ms access
 - 512 KB buffer
- Reliability



© IBM Corporation 1998

Disk Drives

Two types of disk drives are supported in the VSS subsystem. They are the 4.5 GB and 9.1 GB Ultrastar 2XP drives, in SSA configuration.

Both disk drives come in an industry-standard 3.5 in. factory package, which is compact and robust.

Many of the technologies used in the manufacture of the Ultrastar drives were developed at IBM's Almaden Research Center in San Jose, California, USA.

Features

The Ultrastar drives use the latest technology in disk drive manufacture from IBM. In recent years, many improvements in disk drive manufacture have led to increased capacity, performance, and reliability at decreased costs.

Thin film disk technology ensures uniformity in the recording surface. Magnetoresistive (MR) heads are small and easy to manufacture and allow increases in areal density. The partial response maximum likelihood (PRML) read channel uses a sampling algorithm to construct a data stream when reading data from the disk platter. PRML allows increased capacity and reliability. The No-ID sector feature allows drives to be formatted more efficiently, improving capacity and performance. Zoned recording makes maximum use of the outer tracks of a disk platter, increasing areal density and thus capacity.

Areal density

The combined features listed above give a recording density of 134,500 bits per inch (BPI), and a track density of 6160 tracks per inch (TPI). This equates to an outstanding areal density of 829 Mbit/sq in. maximum.

Capacity

The Ultrastar 2XP drives have capacities of 4.5 GB and 9.1 GB. The 4.5 GB model has five platters and nine MR heads mounted on a rotary voice-coil actuator. The 9.1 GB model has 9 platters and 18 MR heads mounted on a rotary voice-coil actuator.

Performance

The Ultrastar 2XP drive disks rotate at 7200 revolutions per minute (rpm), which reduces rotational latency to an average of 4.17 ms. Average access times are less than 8.5 ms, and track-to-track reads average 0.5 ms.

The Ultrastar 2XP drives use embedded sector servo technology, which stores servo (head positioning) data on the data areas of each platter, rather than on a dedicated servo platter. Embedded servo eliminates the repetitive thermal calibration routines required when dedicated servo platters are used, improving high-speed access to the data.

The Ultrastar 2XP drives can provide a data rate (stream of bits from the read head) of from 80.6 Mbit/s at the inner zones to 123.4 Mbit/s at the outer zones (for more information about zones, see “Zoned Bit Recording” on page 93).

To aid in read-ahead and faster write access, the Ultrastar 2XP drives contain a 512 KB buffer. The buffer is used for storage of write data on its way to the drive or as storage for data that has been read in advance in anticipation of being required by the adapter (read ahead).

Reliability

Recoverable read errors are less than 10 per 10^{13} bits read; nonrecoverable read errors are less than 10 per 10^{15} bits read. Seek errors are less than 10 per 10^8 seeks.

The Ultrastar 2XP drives have predictive failure analysis (PFA), which complies with the Self-Monitoring, Analysis, and Reporting Technology (SMART) industry standard.

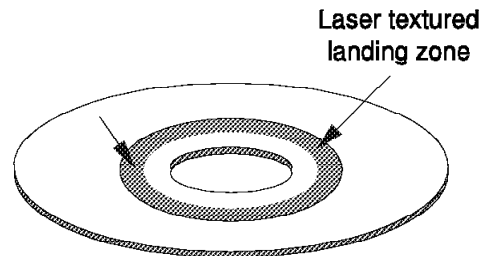
During idle periods, the drive collects various drive characteristics and writes them to a log area on the drive. If any of the characteristics of the drive exceed their thresholds, the drive notifies the host to which it is connected. The thresholds are set in such a way as to give at least 24 hours notice of a pending failure, which allows maintenance to be carried out before loss of data and possible downtime occur.

Thin Film Disk

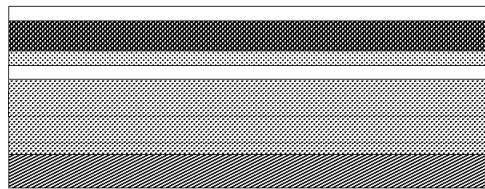


- Substrate
 - AlMg alloy
- Recording layer
 - CoPtCr alloy
- Protection layer
 - Carbon and lubricant
- Landing zone
 - Laser textured

Disk platter



Cross section



Lubricant
Carbon
CoPtCr
Cr
NiP
AlMg Substrate



© IBM Corporation 1997

Thin Film Disks

Thin film disks are the latest technology in disk platter manufacture. The process ensures uniform coating of the platter, which enhances reliability. Thin film disk platters are used in conjunction with MR heads, which greatly improve the recording and reading reliability (for more information about MR heads, see "MR Head Technology" on page 89).

Substrate

Thin film disk platters start with a substrate of aluminum and magnesium (AlMg) alloy, for lightness and strength. They are then coated with a nickel phosphorus (NiP) layer and a thin layer of chromium (Cr).

Recording layer

The recording layer is a thin layer of cobalt, platinum, and chromium alloy (CoPtCr) that is bonded to the chromium layer below it. The recording layer alloy provides high signal-to-noise ratio, which greatly enhances readability.

Protection layer

The recording layer is protected by an extremely hard carbon film and a microfilm of lubricant that enhances the mechanical stability of the head-to-disk interface.

Landing zone

Finally, a laser-textured head landing zone is placed at the inner edge of the disk. When the drive is powered down, the heads are moved to the center of the disk where they touch the platter. The textured landing zone allows the head to touch down without damaging itself or the platter and greatly reduces the likelihood of “head stiction”—the phenomenon where the head, once landed, sticks to the platter and prevents it from spinning when the drive is next powered up.

MR Head Technology



- Separate read and write elements
- Magnetic recording and reading process

Merged MR head



© IBM Corporation 1997

MR Head Technology

Magnetoresistive head technology is another IBM innovation that allows high bit densities and improved recording techniques.

Separate read and write elements

The basic design of the MR head consists of separate read and write elements formed over one another and sharing common material layers. The write element is a thin film inductive head. This optimized design is easier to build than inductive heads that must perform both read and write functions, because it requires fewer copper coils, material layers, photolithographic masking operations, and head-tolerance controls. As a result, IBM can achieve higher process yields, further reducing both manufacturing and end-user costs.

The read element consists of an alloy film, usually NiFe (nickel iron), that exhibits a change in resistance in the presence of a magnetic field—the MR effect. Shielding layers protect the MR elements from other magnetic fields. The second shield also functions as one pole of the inductive write head, thus giving rise to the term *merged MR head*.

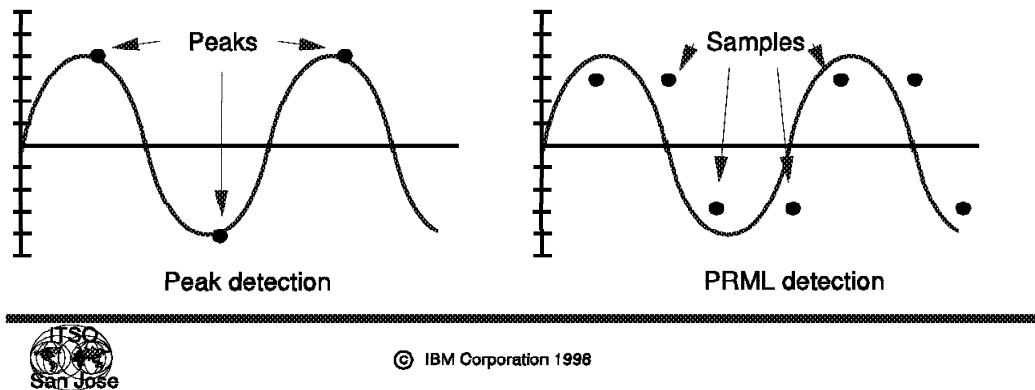
Magnetic recording and reading process

A merged MR head contains an inductive write element and a magnetoresistive read element, flying over a rotating disk. The inductive element writes information bits as magnetically biased regions within radially concentric areas, or tracks, which are subsequently read by the MR read element. The presence of a magnetic transition, or flux reversal, between bits causes the magnetization in the MR element to rotate. This rotation can be detected directly as a resistance change by a precision amplifier, which then produces the stronger signal that relays the information to the PRML read channel of the drive's electronics.

PRML Read Channel



- Peak detection
 - Uses the peak of the read head waveform to determine a 1 or 0
- Viterbi detection
 - Constantly samples read head waveform and compares with possible values



PRML Read Channel

The read channel of a disk drive is the main electronic device that determines how much data can be stored on any given area of a disk platter. It must provide the encoding and conversions needed to record data onto the platter and then read it back accurately. During a write operation, the digital data is serially converted into an analog current that is passed through the write head of the drive. The current causes a flux change in the magnetic material on the platter directly under the head. The platter stores the data bit as a magnetic flux change.

By storing data more densely, disk drives get smaller and less expensive. As data storage becomes more dense, however, data bits begin to interfere with their neighbors, which is known as *intersymbol interference* (ISI).

Peak detection

In a disk drive using a peak detection head, the read head senses the flux changes in the platter's magnetic material and generates an analog waveform that is passed to the read channel. The read channel detects the waveform peaks, each of which represents one bit of data. The data is then converted and deserialized back into digital data bytes.

Unfortunately, peak detection requires encoding techniques that minimize ISI by separating the signal peaks during reads. Ultimately data density is less on

drives using peak detection heads, which leads to less overall capacity and a lower data transfer rate.

Viterbi detection

The Viterbi algorithm is the key to the PRML read channel, named after Andrew Viterbi, the inventor of the algorithm. The PRML read channel uses analog signal processing to shape the read signal to the required frequency and equalize the incoming data to the partial response (PR) waveform. The PR circuits are tuned to match typical signals from the read head.

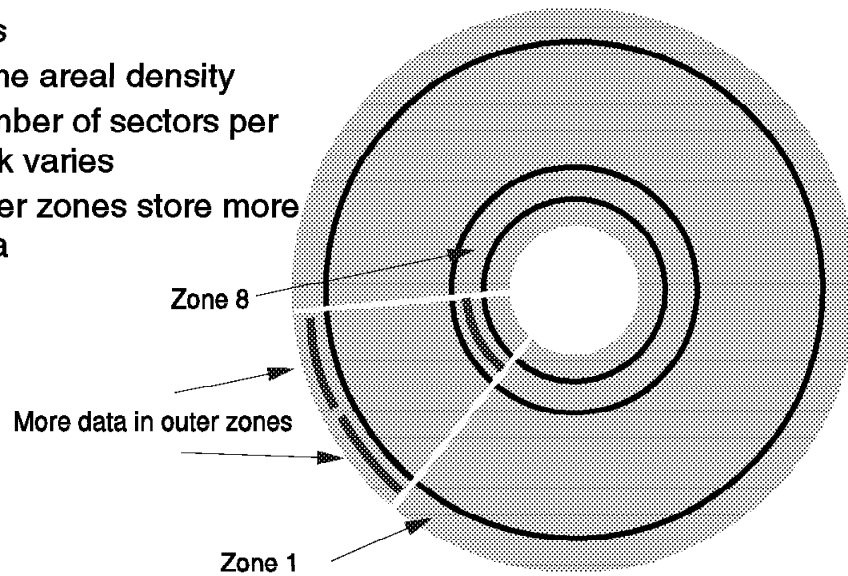
A Viterbi detector then processes the waveform, taking continuous samples along the waveform rather than just detecting the peaks. It does not decide immediately whether a sample represents a 1 or a 0. It compares the samples with sequences of possible values and uses the comparison to detect what a sequence of data bits should be—maximum likelihood (ML). The data is then converted back into digital data bytes.

Ultimately, PRML reduces error rates and allows higher capacities than drives with peak detection read channels.

Zoned Bit Recording



- Zones
 - Same areal density
 - Number of sectors per track varies
 - Outer zones store more data



© IBM Corporation 1998

Zoned Bit Recording

Because track lengths at the inner edge of a disk platter differ from those at the outer edge, before zoned bit recording the bit density was higher on the inner tracks than on the outer tracks. Sector lengths were the same at the inner and outer edges, so less data could be stored at the outer edge. Also, peak-detection read channels could not cope with the higher data rate at the outer edge (because of the higher angular velocity of the platter at the outer edge), thus limiting the amount of data that could be effectively stored.

Zoned bit recording takes advantage of the differing lengths of the tracks and keeps the areal density constant across the platter.

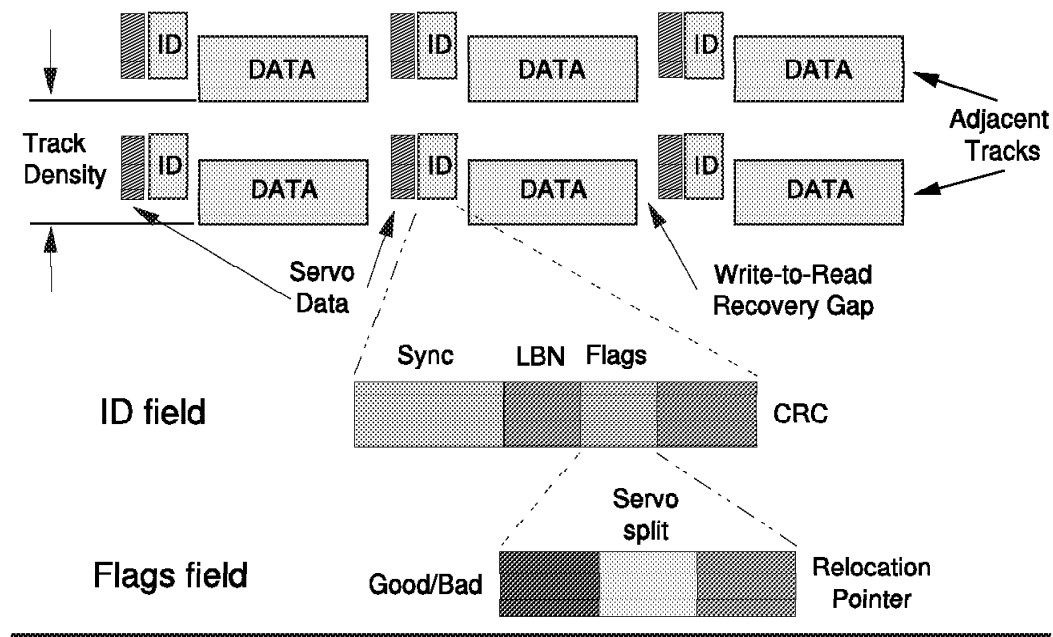
Zones

The Ultrastar 2XP platters are divided into eight zones. Since the areal density remains the same, the number of sectors per track increases toward the outer edge of the platter. Zone 1 is located at the outer edge, and Zone 8 is located at the inner edge.

The main advantage of using zoned bit recording is increased capacity without increased platter size. However, a side effect of having higher density at the outer edges is a higher data rate off the platter due to the higher angular velocity at the outer edge. The maximum data rate at the inner edge is 10.3 MB/s and at the outer edge, the maximum is 15.5 MB/s.

Also, more data is physically stored toward the outer edge; almost 60 % of all data resides in Zones 1 and 2.

ID Sector Format



© IBM Corporation 1996

ID Sector Format

There is overhead associated with storage of data on a disk drive. Each platter in the drive has to be formatted—the process whereby control information is written to the platter to enable the read/write heads to position correctly to read or write data. The term *format efficiency* refers to the amount of each track in a disk drive devoted to storing user data. Format efficiency can be improved by reducing the overhead. There are a number of contributors to overhead in the format of disk drives. Some of these, such as synchronization fields, are required for reading the data. Others, such as ECC and sector servo, offset their overhead by allowing the areal density to be increased. One contributor to the overhead that does not increase areal density is the header or ID field.

The foil illustrates the track layout of a typical fixed-block disk drive using embedded servo. Each track is divided into a number of data sectors and servo sectors. The servo fields contain the positioning information used to locate the head over a given track. The user data is stored in the data fields, each with an associated ID field. The ID fields contain information that identifies the data sector and other information, such as flags, to indicate defective sectors.

The majority of disk drives manufactured today use an addressing scheme where the data sectors are identified to the host system by a logical block number (LBN). In operation, the host computer sends a list of logical block numbers to be written or read. The disk drive converts these values into zone, cylinder, head, and sector (ZCHS) values. The servo system seeks for the

desired zone, cylinder, and head, and the disk drive begins reading ID fields until a match is found. Once the appropriate ID field has been read, the drive can then read or write the next data field.

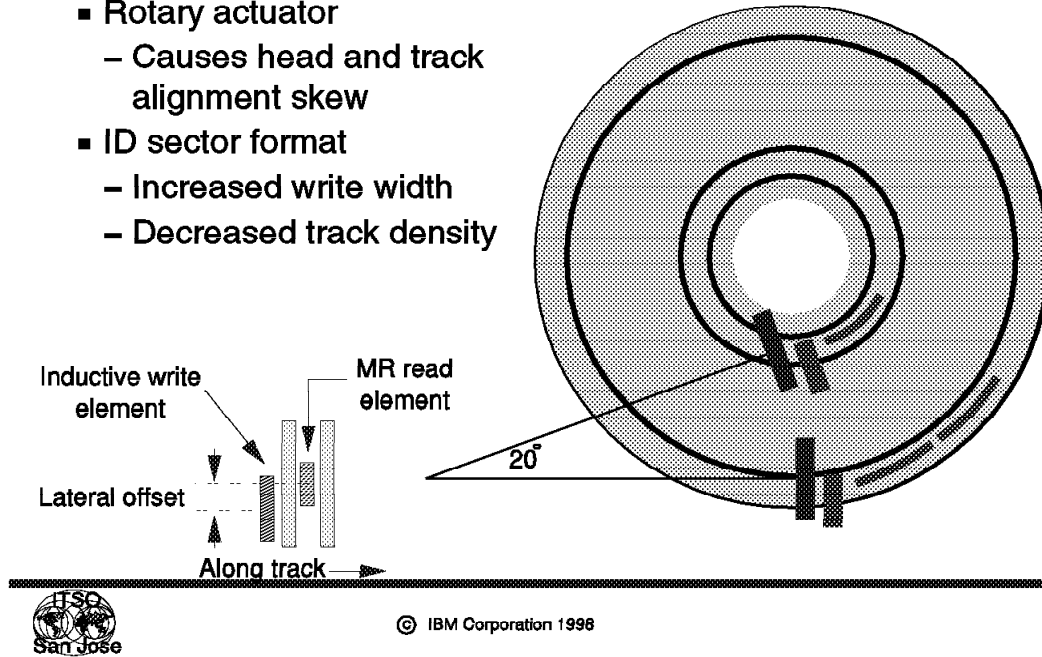
The use of ID fields allows great flexibility in the format and provides a simple mechanism for handling defects. However, substantial costs are associated with the use of ID fields. The ID fields themselves can occupy up to 10% of a track—space that would otherwise be used to store data. Further, because the disk drive must read the ID field for each sector before a read or write operation, additional space is required to allow for write-to-read recovery prior to each ID field (refer to the foil). The recovery gaps can occupy more than 5 % of a track.

Defect management is typically accomplished by reserving a fixed number of spare sectors at some chosen interval. If a sector is determined to be defective, the data is relocated to one of the spare sectors. The relocation process ranges from shifting all the sectors between the defect and the spare, to using a specific spare sector to replace the defective sector. Performance can be degraded if the sectors are not at their expected locations, and therefore require an additional seek operation. To reduce the likelihood of sector relocation, a large number of sectors are typically reserved as spares, which further reduces the format efficiency.

MR Head Geometry



- Rotary actuator
 - Causes head and track alignment skew
- ID sector format
 - Increased write width
 - Decreased track density



MR Head Geometry

The foil shows the basic geometry of an MR head, as seen from the disk surface. The head consists of a thin film inductive write element and an MR read element. The read element is typically narrower than the write element to improve the off-track performance. In practice, the longitudinal separation of the elements results in an offset between the centers of the read and write elements.

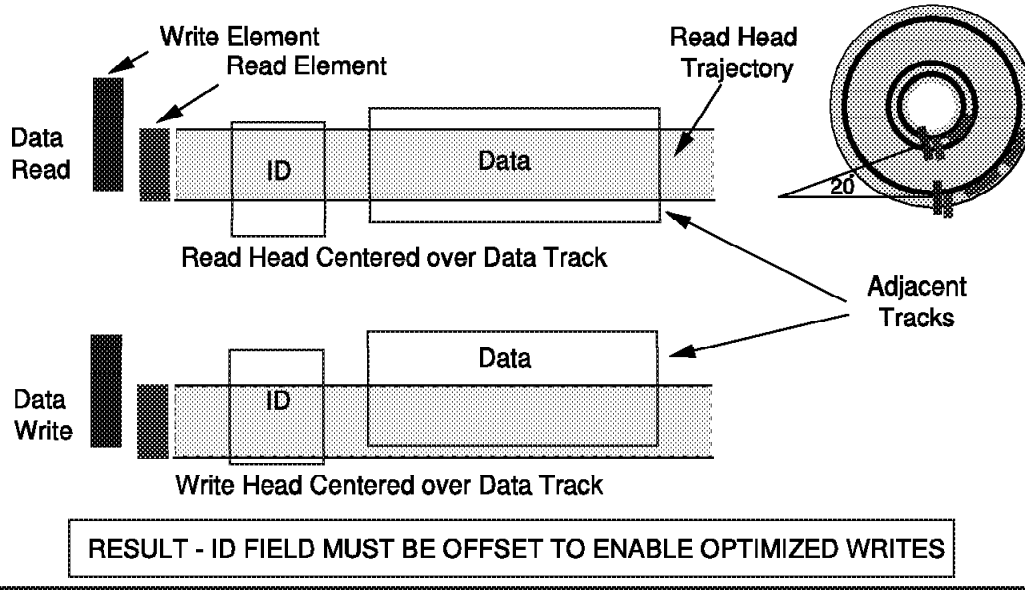
Rotary actuator

The Ultrastar drives use a rotary voice-coil actuator with approximately 20 degrees of angular movement. When used with a rotary actuator, the head is skewed with respect to the tracks as the actuator moves across the disk. The result is a lateral offset between the read and write head centerlines. Optimum performance is achieved by centering the read head over the data track for read operations and the write head over the data track for write operations. This operation causes the read head to be partially off-track during a write operation.

ID sector format

The read /write element offset presents a problem when ID fields are present, because they must be accurately read for both read and write operations. ID fields may be written partially off-track, requiring increased write width to ensure that the read head can reliably read the ID field. The increased write width imposes limits on the track density, ultimately decreasing the overall capacity of the drive.

MR Head Effect on ID Sector Format



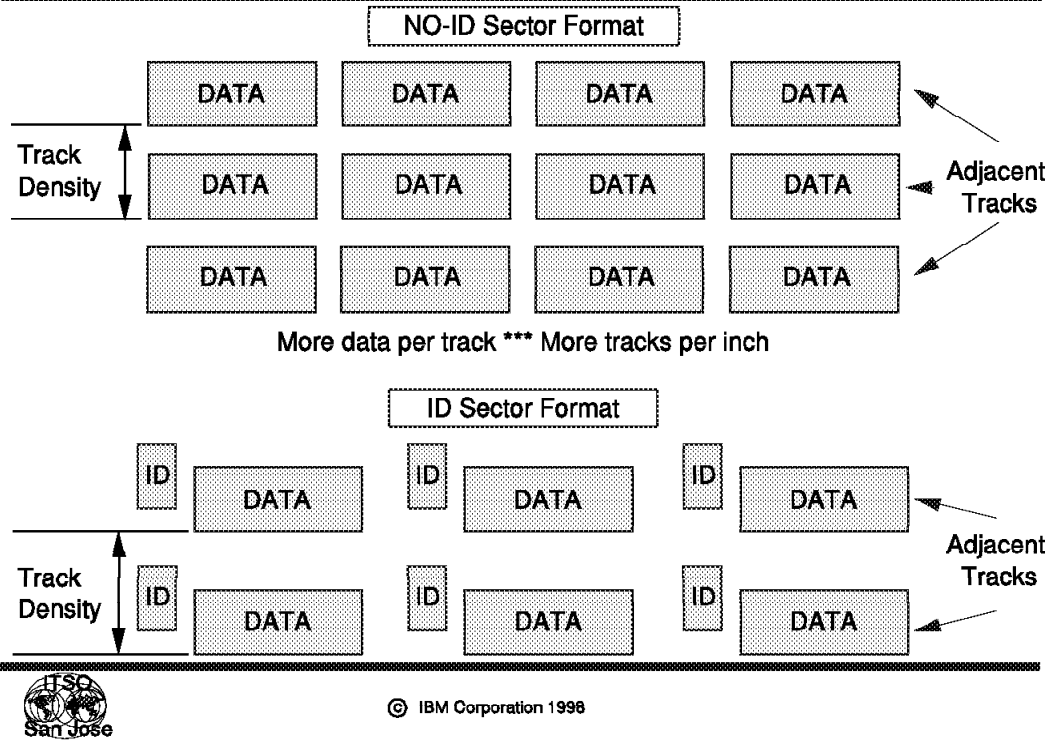
© IBM Corporation 1996

MR Head Effect on ID Sector Format

The ID field must be read before the corresponding data field can be read or written. This operation requires that the read head be positioned over the ID field before any read or write operation. For a read operation, the read head must be centered over the data track, and for a write operation the write head must be centered over the track. For both operations, the read head must be able to read the ID field.

The MR head elements are offset with respect to each other. When a write operation occurs, the write head must be positioned directly over the data field. However, the read head must be able to read the ID field immediately preceding. Because the offset of the read head with respect to the write head, the ID field has to be offset from the data field.

No-ID Sector Format



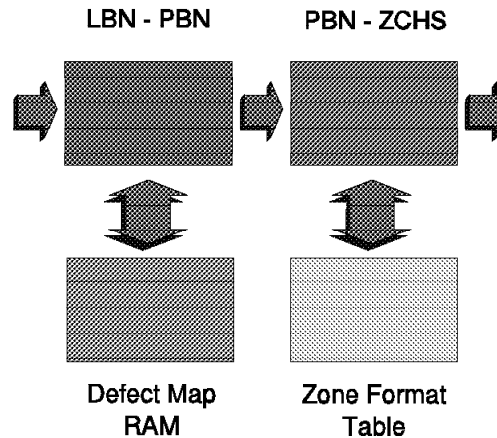
No-ID Sector Format

The No-ID sector format allows disk drives to be formatted more efficiently than drives using ID sector format, improving the capacity, reliability, and performance. The No-ID sector format eliminates the ID field and all of the information it contains from the track format. The ID field information is stored in solid state memory in the drive instead of on the disk surface. Storing the format in solid state memory results in the increased capacity of each track without impacting the linear density—approximately 5% to 15%. When combined with an MR head, the No-ID sector format dramatically increases the track density. The tracks can then be written closer together because the offset ID field has been removed. Further benefits are increased data throughput, and improved access time, defect management, and power management.

No-ID Sector Format ...



- RAM-based, servo-generated sector IDs
 - Defect map
 - Zone format
 - LBN to PBN conversion
 - Enhanced defect management and error recovery
- Other advantages
 - Increased capacity
 - Higher yield
 - Enhanced performance
 - Enhanced power management



© IBM Corporation 1996

No-ID Sector Format ...

Continuing from the previous foil, we discuss how the electronics and servo systems of the drive locate the desired sector.

RAM-based, servo-generated sector IDs

The No-ID sector format uses the drive's servo control system to locate physical sectors and uses a defect map stored in RAM to identify logical sectors. The drive electronics convert LBNs to physical block numbers (PBNs). The foil illustrates the sequence of operation.

The LBN is simply a number from 0 to the number of addressable blocks on the disk drive. The PBN is a number from 0 to the number of physical blocks on the disk drive, but with the defective and spare sectors mapped out. Once the PBN is computed, it may be converted to the exact ZCHS value for the sector. Because the defect information is known in advance, the proper logical block is guaranteed to be located at the computed ZHCS. The defect map is stored in a compressed format, optimized for small size and rapid lookup.

The servo system is used to locate the physical sector on the basis of knowledge of the track formats in each zone. This information includes the locations of any data field splits due to embedded servo, which are also stored in RAM.

The No-ID sector format enhances disk drive reliability because the header and data field split information is stored in RAM, not on the disk. Current disk drives

rely on cyclic redundancy check (CRC) or ECC to reduce the vulnerability to errors in the ID fields. However, if the drive cannot read an ID field, it may not be possible for it to recover the associated data sector. On a No-ID sector format drive, the sector ID is always known, so there is a greater chance that the data can be recovered from a defective sector.

Other advantages

No-ID sector format increases the capacity of disk drives by reducing the format overhead and allowing the MR head to be utilized to its fullest extent. Increasing track density gives up to a 15% increase in capacity, while increased linear bit density also gives up to a 15% increase in capacity, in the same space. Disk drive manufacturing yield is further enhanced by the advanced defect management capabilities. The performance is enhanced by the increased throughput (through reduced overhead) and by the knowledge of the absolute sector locations. Power management is enhanced because there is no need to supply current to the read electronics to read ID fields when searching for a sector.

Predictive Failure Analysis



- Overview
 - Two types of failures
 - ▶ *Unpredictable - cable, component*
 - ▶ *Gradual performance degradation*
 - Goal - provide 24 hour notice of pending failure
- Processes
 - Measurement driven
 - ▶ *Seven measurements*
 - ▶ *4-hour cycle*
 - ▶ *80 ms duration*
 - Symptom driven
 - ▶ *Error recovery logs*



© IBM Corporation 1998

Predictive Failure Analysis

Predictive failure analysis (PFA) is a process whereby the Ultrastar 2XP drive monitors itself to ensure that its components are functioning correctly. The drive can predict an imminent failure and notify the host computer in advance, ensuring the integrity of customer data.

Overview

With any electromechanical device, there are two basic failure types. First, there is the unpredictable catastrophic failure. A cable breaks, a component burns out, a solder connection fails. As assembly and component processes have improved, these defects have been reduced but not eliminated. PFA cannot provide warning for unpredictable failures. Then there is the gradual performance degradation of components. PFA has been developed to monitor performance of the disk drive, analyze data from periodic internal measurements, and recommend replacement when specific thresholds are exceeded. The thresholds have been determined by examining the history logs of disk drives that have failed in actual customer operation.

The design goal of PFA is to provide a minimum of 24 hours warning before a drive fails.

Processes

PFA monitors performance in two ways: a measurement-driven process and a symptom-driven process. The measurement-driven process is called *generalized error measurement* (GEM). At periodic intervals, the PFA GEM automatically performs a suite of self-diagnostic tests that measure changes in the disk drive component characteristics.

Seven measurements are taken over a 4-hour period, each taking approximately 80 ms. These seven measurements include various magnetic parameters of the head and disk, head fly height on all data surfaces, channel noise, signal coherence, signal amplitude, and writing parameters.

The symptom-driven process of PFA uses the output of data, nondata, and motor-start error-recovery logs. The analysis of the error log information is performed periodically during idle periods, along with the data collected by GEM. When PFA analysis detects a threshold-exceeded failure, the drive notifies the host system through the controlling adapter.

Predictive Failure Analysis ...



- **Error logs**
 - Saved to reserved areas
 - Approximately every half hour, when drive has been idle for 5 sec or more
- **Channel calibration**
 - Ensures optimum performance of read and write circuits
 - 20 ms duration, when drive has been idle for 5 s or more
- **Disk sweep**
 - Ensures low-use components do not become vulnerable to failure
 - Heads moved if drive is idle for more than 40 s
 - Heads moved again if in same location for more than 9 min



© IBM Corporation 1998

Predictive Failure Analysis ...

Error logs

The Ultrastar 2XP drive periodically saves data in error logs located in reserved areas of the disks. These reserved areas are not part of the usable customer-data area of the drive. Logs are written during idle times, usually only after the drive is idle for 5 s or more. The writing process takes about 30 ms, and the logs are saved about every half-hour.

Channel calibration

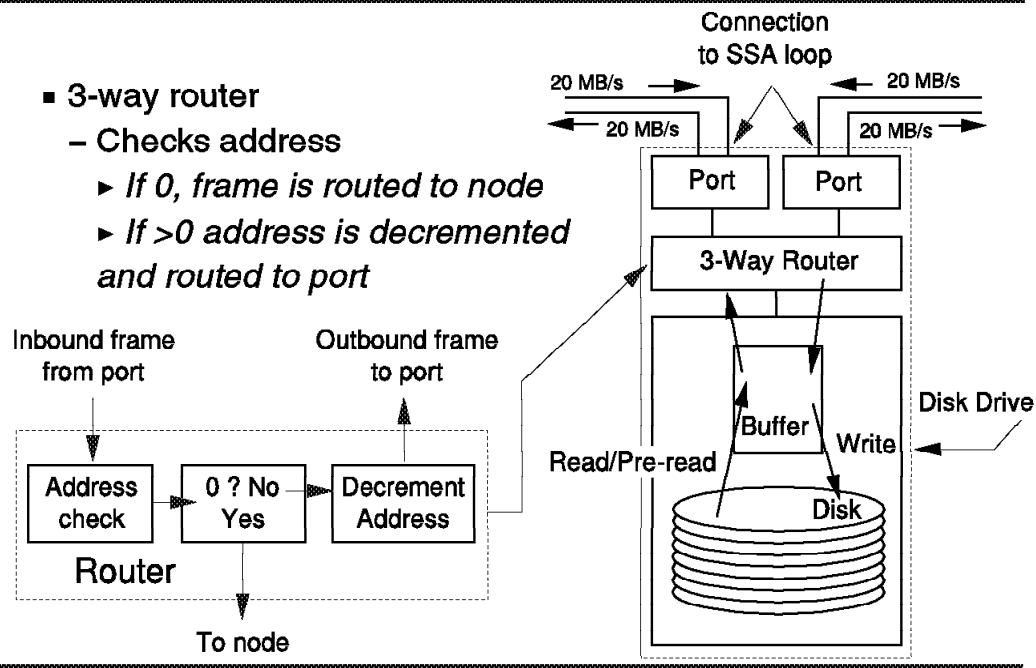
The Ultrastar 2XP periodically calibrates the read channel to ensure that the read and write circuitry is functioning optimally, thus reducing the likelihood of soft errors. Like the PFA functions, channel calibration is done during idle periods, every 4 hours. It takes about 20 ms to complete and requires the drive to have been idle for more than 5 s.

Disk sweep

Disk sweep is the process of ensuring that a drive does not remain idle for excessive periods. Components that remain idle for long periods become vulnerable to failure. If the drive has not processed a command for more than 40 s, it initiates a move of the heads to a random location. If the heads fly in the

same spot for more than 9 minutes, the heads are moved again, repeating for as long as the disk is idle.

Ultrastar 2XP in SSA Configuration



© IBM Corporation 1998

Ultrastar 2XP in SSA Configuration

When used in the VSS, the Ultrastar 2XP is configured for SSA operation with the addition of an assembly that contains two SSA bidirectional ports and a three-way router. The ports are capable of transmitting and receiving SSA frames similarly to the SSA disk adapter, except that they do not contain an initiator. In SSA terminology, the drive functions as an SSA “node.”

Three-way router

The three-way router is an important part of the SSA loop. The router takes inbound frames from either port, and checks the address field of the frame. The address field has two parts, the path and channel addresses. The router checks the path address. The path address is typically 1 byte but can be extended to up to 4 bytes in complex and switched webs. If the first byte of the path address is a zero, then the frame is assumed to be for the current node, and the frame is routed to the node (the drive). The channel address is used to route the frame within the node.

If the first byte of the path address is not zero, the frame is assumed to be for a device further along the loop. The router decrements the first byte of the path address, and transmits the frame out of the opposite port.

Thus, a frame travels around the loop only as far as it needs to, minimizing loop traffic. Minimization of traffic is known in SSA terms as *spatial reuse*.

Versatile Storage Server 524-Byte Sector



- ✓ 8 bytes of AS/400 header information
- ✓ 512 bytes of data
- ✓ 2-byte sequence number
- ✓ 2-byte longitudinal redundancy check

AS/400 Header	Data	SEQ #	LRC
8	512	2	2

524-Byte Sector



© IBM Corporation 1996

Versatile Storage Server 524-Byte Sector Format

Most fixed-block disk architectures use a fixed-byte sector of 512 bytes. Most UNIX systems, including AIX, use a fixed-byte sector of 512 bytes. When used in a VSS subsystem, a disk drive has a format of a fixed-byte sector of 524 bytes. The 524-byte sector format enables the VSS subsystem to connect to a wide range of host systems and share data between them.

At the start of the sector, 8 bytes are used by IBM AS/400 systems and are not used when the VSS is attached to UNIX hosts. The data portion of the sector remains at 512 bytes for all systems. A 2-byte sequence number and a 2-byte longitudinal redundancy check (LRC) increase the size of the sector to 524 bytes. The sequence number is a modulo-64K value of the LBA of this sector and is used as an extra method of ensuring that the correct block is being accessed. The LRC, generated on the SCSI host adapter, is calculated on the 520 data and header bytes and is used as an error-checking mechanism as the data progresses from the host, through the VSS storage server, into the SSA adapter, and on to the RAID array (see Chapter 4, "Versatile Storage Server Data Flow" on page 109 for a detailed description of data flow through the subsystem). The LRC is also used as an error-checking mechanism as the data is read from the array and passed up to the host adapter. The sequence number and LRC are never transferred to the host system.

Versatile Storage Server Data Flow



Versatile Storage Server Data Flow



© IBM Corporation 1998

In this chapter, we investigate the flow of I/O operations in the Versatile Storage Server.

Versatile Storage Server Data Flow Overview



- RAID-5 terminology
- Hardware
- Concepts
- Predictive cache algorithms
- Input/output operations (I/Os)



© IBM Corporation 1998

Versatile Storage Server Data Flow Overview

This chart describes the overall structure of the charts that follow in this section.

RAID-5 Terminology

First, we define the RAID terms that are used in this chapter.

Hardware

Next, we describe the cache and nonvolatile storage (Fast Write Cache) architecture of the Versatile Storage Server.

Concepts

Here we describe several important concepts unique to the VSS. It is important to understand these concepts in order to understand the algorithms the Versatile Storage Server uses to manage cache and access to the attached SSA disks.

Predictive cache algorithms

We discuss two key cache algorithms, sequential prestage and adaptive cache.

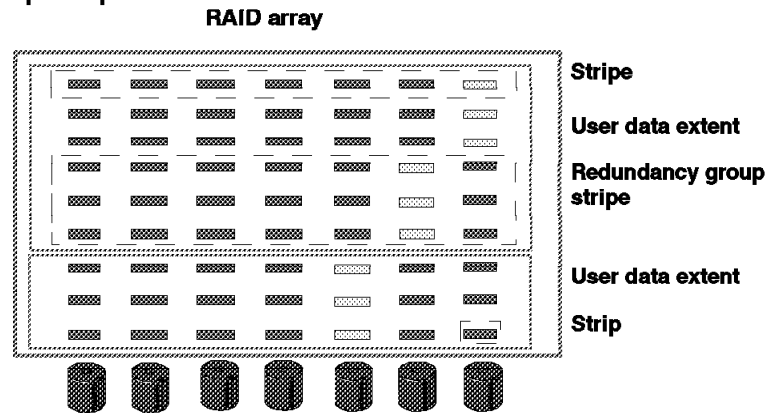
Input/output operations (I/Os)

Finally, we can describe the data flow for I/O operations. Foils describe different types of read and write operations.

RAID-5 Terminology



- RAID Advisory Board terms
- RAID array
- strip
- stripe
- redundancy group stripe
- virtual disk



© IBM Corporation 1998

RAID-5 Terminology

In this chart, we define the terms used to describe RAID-5 concepts in this chapter.

RAID Advisory Board terms

There are many terms used to describe the same RAID-5 concept. In this chapter, we primarily use RAID Advisory Board terms. In many cases, the RAID Advisory Board recognizes several terms for the same concept.

RAID array

The RAID Advisory Board defines a *RAID array* as “a disk array in which part of the physical storage capacity is used to store redundant information about user data stored on the remainder of the storage capacity.” We use the term *RAID array* to refer to the 6+P or 7+P group of member disks of the RAID-5 disk array in the VSS subsystem.

Strip

A *strip*, sometimes called a *stripe element*, is the unit of consecutively addressed blocks on a disk that are logically followed by a strip on the next data disk in the array. In the VSS, the strip size is 32 KB.

Stripe

A *stripe* is a set of positionally corresponding strips across all the member disks of an array. A stripe consists of n data strips and one parity strip.

In the VSS subsystem, there are always either seven or eight member disks in a RAID-5 array.

Redundancy group stripe

The *redundancy group stripe* is a group of stripes in which the RAID parity is stored on the same physical disk. In a RAID-5 implementation, parity is striped across all the disks in the array. The *redundancy group stripe depth* is the unit of striping of the parity across the member disks in the array.

In the VSS subsystem, the redundancy group stripe depth is four stripes for a 7+P array, and five stripes for a 6+P array.

Virtual disk

The total space in a disk array can be divided into virtual disks of different sizes. In the VSS subsystem, the virtual disks appear to the external SCSI host as a physical disk or hdisk.

A RAID array consists of strips of arbitrary but fixed size on each of the member disks of the array.

The total space in the RAID array can be subdivided into virtual disks.

A stripe is a 1 by n array of strips, where 1 strip is used to store parity and $n-1$ strips are used to store data.

The redundancy group stripe size must be small enough to effectively distribute the RAID parity update workload across all the member disks of the array. Otherwise, a RAID-5 array can have the parity “hot spot” characteristics of a RAID-4 array.

Versatile Storage Server Hardware Overview



- Storage Server cache
 - Improved performance for cache read hits
- SSA adapter nonvolatile storage (Fast Write Cache)
 - Protection of fast write data
- SSA adapter volatile cache
 - Protection of fast write data
 - Improved performance for RAID-5 update writes
- SSA disk buffer
 - Support for interleaved transmission of frames on SSA bus
 - Improved sequential read throughput
 - Improved sequential write throughput



© IBM Corporation 1998

Versatile Storage Server Hardware Overview

This foil is an overview chart, describing some of the topics to be covered in this section. The focus of the hardware overview is the cache and nonvolatile storage (Fast Write Cache) architecture of the VSS. The Fast Write Cache is a key component that ensures data integrity for fast write operations.

Storage server cache

The large cache in the VSS subsystem is part of the storage server. It is a volatile cache stored in DRAM. The primary purpose of this cache is to provide improved performance for read operations, since a read operation resulting in a cache hit need not access the backstore disk drives.

SSA adapter nonvolatile storage (Fast Write Cache)

The nonvolatile storage (Fast Write Cache) consists of static random access memory (SRAM) with a battery backup. The primary purpose of the Fast Write Cache is to protect fast write data not yet committed to disk. The Fast Write Cache protects the integrity of fast write data because a loss of power does not result in a loss of data updates. There is a 4 MB Fast Write Cache on each SSA adapter. If two SSA loops are attached to the adapter, they share the Fast Write Cache; it is not statically partitioned across the two loops.

The total Fast Write Cache in the Versatile Storage Server depends on the number of SSA adapters in the subsystem.

SSA adapter volatile cache

This cache resides in each SSA adapter, located in the storage server. Like the storage server cache, it is also volatile cache stored in DRAM. A key function of the SSA adapter volatile cache is to mirror the SSA adapter nonvolatile storage (Fast Write Cache). This protects data integrity because a failure of either SSA adapter DRAM or SSA adapter Fast Write Cache does not result in loss of data written but not yet committed to disk.

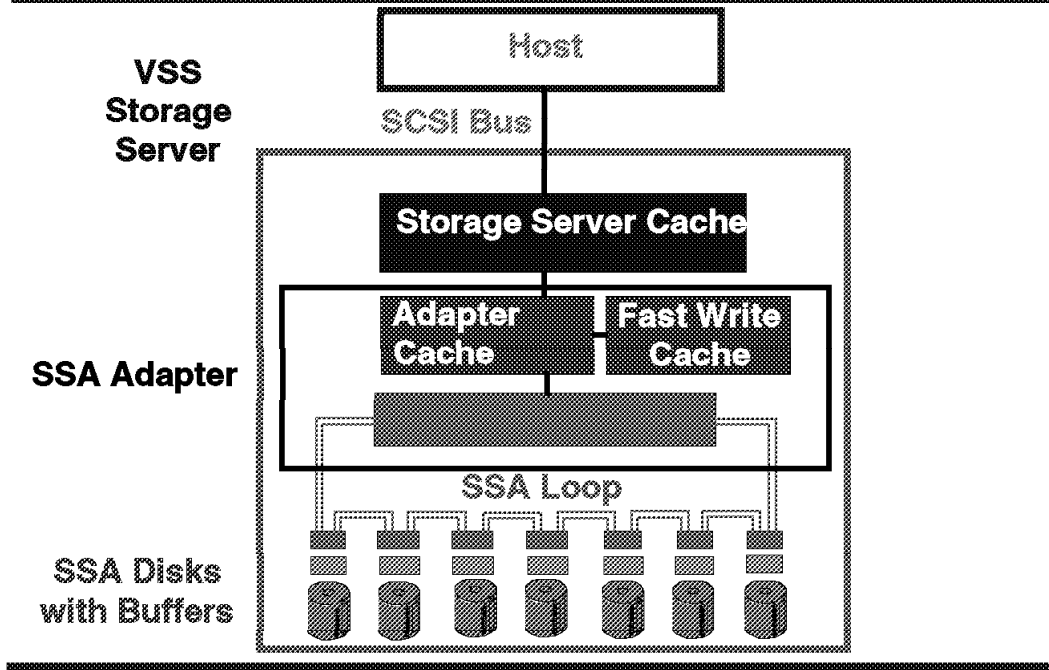
In addition to mirroring the Fast Write Cache, the primary purpose of the SSA adapter volatile cache is to improve performance for RAID-5 update writes by caching recently accessed data and RAID parity sectors.

The total adapter volatile cache in the VSS depends on the number of SSA adapters in the subsystem, usually one adapter per drawer.

SSA disk buffer

Each SSA disk has a disk buffer. The primary purpose of the disk buffer is to support the interleaved transmission of frames on the SSA bus. However, the buffer also allows readahead from the disk on read operations, which can improve sequential read throughput, and it improves sequential write throughput if there is a heavy sequential write load by accepting multiple write commands while waiting for disk latency.

Data Flow Architecture



© IBM Corporation 1998

Versatile Storage Server Data Flow Architecture

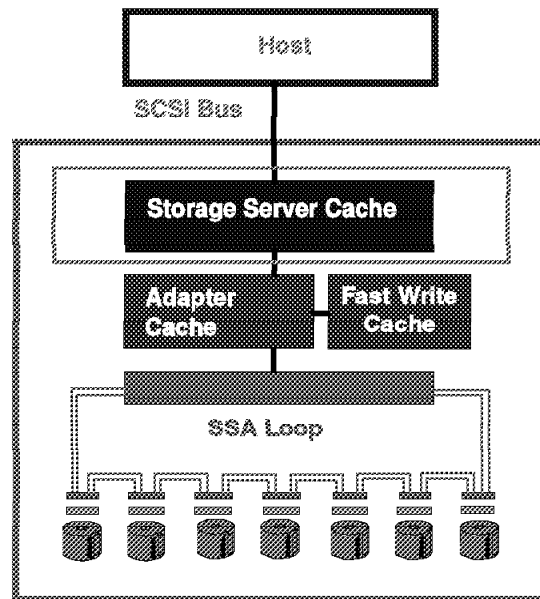
This foil shows the components of the Versatile Storage Server mentioned on the previous foil. It shows the path used by a single I/O to access data in a single RAID array.

There are two processing clusters in the storage server, and each can have access to multiple hosts through multiple SCSI ports and to multiple RAID arrays through multiple SSA adapters. For simplicity, this diagram, which will be used repeatedly in this section, shows only the path used by a single I/O. Typically, many concurrent I/Os will be processed by a single VSS subsystem, but this concurrency does not affect the data flow used in processing a single I/O operation.

Storage Server Cache



- Improved performance for cache read hits
- 512 MB to 6 GB volatile memory
- Managed by least recently used (LRU) algorithm
- Read and write cache
 - Data written is cached to facilitate cache hit if reread
- Data only
 - RAID parity is not stored in storage server cache



© IBM Corporation 1998

Storage server cache

The large cache in the VSS is called *storage server cache*. It is a volatile cache stored in DRAM.

512 MB to 6 GB volatile memory

Storage server cache is specified as the total size for the subsystem. The storage server cache is always divided symmetrically across the two storage servers.

The following cache size options are available:

- 512 MB
- 1024 MB (1 GB)
- 2048 MB (2 GB)
- 4096 MB (4 GB)
- 6144 MB (6 GB)

The choice of storage server cache size is dictated by two factors: the size of the disk backstore in gigabytes and the workload characteristics. Because UNIX systems typically provide very efficient disk caching in memory, larger cache sizes are most useful in VSS subsystems with a large disk backstore capacity or in environments where host disk caching is not highly effective.

Managed with LRU algorithm

As data is read or written, it is placed in cache. If room is needed to add more data to cache, the data least recently used (accessed) is purged from cache to make room. Reaccessing data already in cache causes the data to be moved to the top (most recently used end) of the access queue.

Read and write cache

The storage server cache is a read and write cache; it contains not only data that has been read but also data that has been written. Data written is cached to facilitate its reaccess by the same host or by a different host in a cluster environment.

The storage server cache is not involved in protection of fast write data. Fast write data is protected by SSA adapter cache and Fast Write Cache. Write data can be LRU destaged from storage server cache whether or not it has been written from Fast Write Cache to disk.

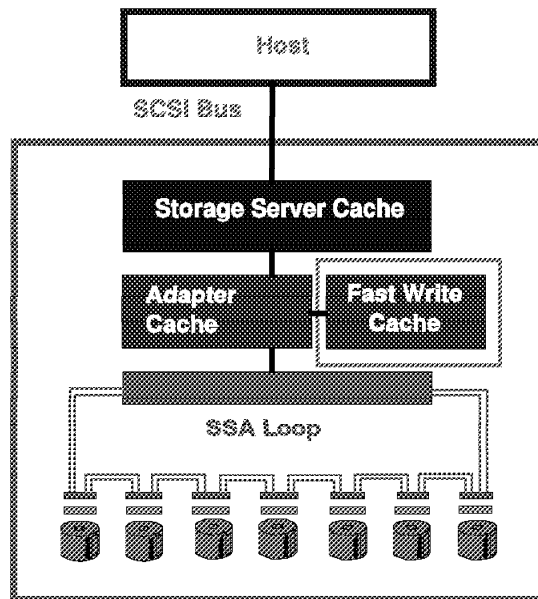
Data only

RAID parity is not stored in the storage server cache. RAID parity is never transferred from the SSA adapters to the storage server cache.

SSA Adapter Fast Write Cache



- Protection of fast write data
- 4 MB of static random access memory (SRAM)
 - Removable card with integral lithium battery
 - *Up to 10 year retention of data even if card is removed*
 - Fast Write Cache can be installed with data intact in replacement SSA adapter



© IBM Corporation 1998

SSA Adapter Nonvolatile Storage (Fast Write Cache)

The Fast Write Cache is a key component of the VSS subsystem that ensures integrity for fast write operations.

Protection of fast write data

The primary protection for fast write data is the use of nonvolatile storage or Fast Write Cache. Two copies of fast write data are kept until the data is successfully destaged to disk, one in Fast Write Cache and one in the SSA adapter cache. The Fast Write Cache copy ensures that the data is not lost in case of a power failure before it is committed to disk. The SSA adapter cache copy ensures that the data is not lost if there is a Fast Write Cache failure.

RAID parity data is not stored in Fast Write Cache, but an indication of required pending parity updates is. This is done to preserve Fast Write Cache space and because the SSA adapter can reread the data necessary to create parity if a volatile cache failure occurs.

4 MB of SRAM

The Fast Write Cache is SRAM, which requires less power to maintain the integrity of its contents than does DRAM. A battery can provide much longer data retention with SRAM than is possible with DRAM.

If two SSA loops are attached to the adapter, they share the Fast Write Cache. It is not statically partitioned across the two loops.

The Fast Write Cache is packaged as a removable card with an integral battery. The battery, which is a lithium battery and therefore not rechargeable, maintains the Fast Write Cache contents if the VSS is powered down for any reason. During a normal shutdown cycle, data in Fast Write Cache is destaged to disk before power is turned off to the subsystem.

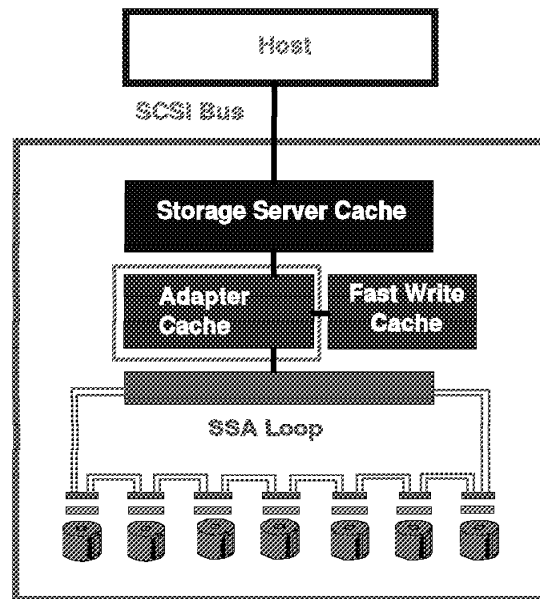
The battery is integrated into the card, so retention of data stored in Fast Write Cache is ensured even if the card is removed from the SSA adapter. The battery is designed to provide retention of data for up to 10 years.

A key advantage of this design is that in the unlikely occurrence of an SSA adapter failure, the Fast Write Cache can be removed from the failed adapter and installed in a replacement SSA adapter, without loss of the fast write data stored in Fast Write Cache.

SSA Adapter Volatile Cache



- Improved performance for RAID-5 update writes
- 32 MB of dynamic random access memory (DRAM)
 - 4 MB used to store fast write data mirrored in Fast Write Cache
 - 26 MB used as read and write cache
 - > *Data and parity*
 - 2 MB used for control blocks



© IBM Corporation 1998

SSA Adapter Volatile Cache

Improved performance for RAID-5 update writes

In addition to mirroring the Fast Write Cache, the SSA adapter cache improves performance for RAID-5 update writes by caching recently accessed data and RAID parity sectors.

32 MB of DRAM

Each SSA adapter contains 32 MB of DRAM, in two independent 16 MB chips. If there are two SSA loops attached to the SSA adapter, all cache on the adapter card is shared between the two loops. Least recently used (LRU) algorithms are used to manage adapter cache and Fast Write Cache across all arrays attached to both loops.

Data integrity is ensured by storing two copies of all write data on the SSA adapter card before task end or a write command is sent to the host. Of the SSA adapter DRAM, 4 MB are used to mirror the contents of Fast Write Cache so that data integrity is ensured in case of a Fast Write Cache failure.

When data is destaged to disk from Fast Write Cache, it is actually the volatile cache copy that is used. Technically speaking, the Fast Write Cache mirrors the 4 MB portion of SSA adapter volatile cache, not the other way around. However,

it is easier to speak of the Fast Write Cache and its volatile cache mirror, and doing so does not change the description of the protection of fast write data.

About 26 MB of the DRAM are used to store user data read from disk and parity data read from and written to disk. Although it is possible for the SSA adapter to process a read operation as a cache hit against the adapter cache, the primary purpose of the adapter cache is to enable cache hits when performing RAID-5 write processing. Because a RAID-5 update generates a read old data, read old parity, write new data, write new parity sequence, multiple updates to the same stripe result in a write to parity followed by a read of the data just written (the old parity for the second write). If the parity data written is still in the SSA adapter cache, the read of old parity data can be avoided.

While it is also possible to avoid the read of old data by a cache hit in the adapter cache if data in the same track is written twice, the write preempt capability of the Fast Write Cache management is likely to result in a single destage of the blocks. This is described in "Write preempts" on page 156.

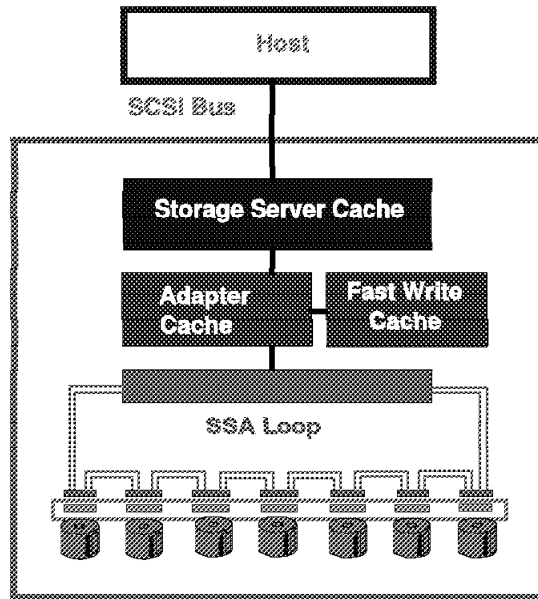
The SSA adapter uses a portion of the DRAM for control blocks. In this space, the adapter keeps track of information such as the SSA loop topology, which could change as a result of a disk or SSA cable failure.

SSA loop topology is not stored in Fast Write Cache since the topology can be determined again if necessary by the SSA adapter.

SSA Disk Buffer



- Improved sequential read throughput
- Improved sequential write throughput
- 512 KB buffer in each SSA disk
- Independent data transfer across SSA bus
- Interleaved data transfer on SSA bus



© IBM Corporation 1998

SSA Disk Buffer

Each 4.5 and 9.1 GB SSA disk has a 512 KB buffer, which is used for both read and write operations.

Improved sequential read throughput

All reads to an SSA disk result in the staging to the disk buffer of 64 KB of data beginning with the first sector requested. Only the data requested is transferred from the disk buffer to the SSA adapter. However, a subsequent read for additional sectors can result in a cache hit in the SSA disk buffer.

Improved sequential write throughput

When data is being written sequentially to the VSS, it will be written to disk in 32 KB tracks corresponding to a strip. The disk can accept up to three write operations before completing the first one, so multiple writes to the disk received in rapid succession can be written from the disk buffer to the disk surface in a single revolution of the disk.

512 KB buffer in each SSA disk

Each disk has a 512 KB buffer.

For reads, the disk stages into the buffer the sectors requested and following sectors up to a total of 64 KB. For writes, only the sectors written are placed in the buffer.

Independent data transfer across SSA bus

Data can be read from disk to buffer when the disk rotates to where the sectors to be read are under the heads and then transferred across the SSA bus when bus bandwidth is available. Similarly, data can be written to the buffer regardless of the rotational orientation of the disk, then destaged when the disk reaches the correct rotational orientation.

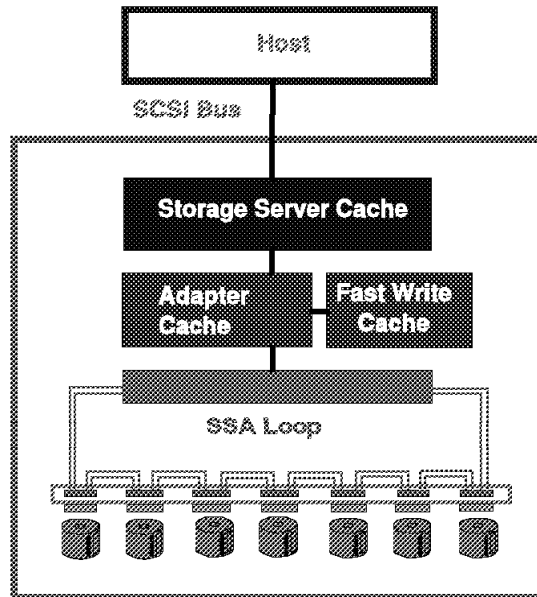
Interleaved data transfer on SSA bus

The SSA architecture allows data for different devices to be interleaved on the SSA bus. The key advantage of this approach is that a large data transfer from a device on the bus does not delay smaller data transfers from other devices. This interleaving of data from different devices requires that data transfer on the bus be asynchronous to the transfer of data from the disk surface to the bus, and this is accomplished by the use of the disk buffer.

SSA Frame Buffer



- Used for SSA frame management



© IBM Corporation 1998

SSA Frame Buffer

Used for SSA frame management

The frame buffer is used to pass SSA frames on the SSA bus. These frames can originate on this disk drive or from another drive and be passed around the loop. The frame buffer is large enough to handle the up to 128 byte frames moving on the SSA bus.

Versatile Storage Server Concepts



- 524 byte disk sector
- Cache concepts
- Cache transparency
- Front end and back end I/O



© IBM Corporation 1998

Versatile Storage Server Concepts

This foil introduces the concepts we discuss for the next several foils.

524 byte disk sector

The VSS stores 512 byte user sectors as part of a 524 byte sector on disk and in cache.

Cache concepts

We discuss the concept of a 32 KB track and how it relates to cache management.

Cache transparency

The VSS cache is transparent to SCSI applications.

Front end and back end I/O

A SCSI front end I/O will generate no back end I/O if it is a cache hit, or it may generate a number of I/Os if it reads or writes a lot of data. We'll talk about the relationship between SCSI front end I/O and SSA back end I/O.

524 Byte Disk Sector



- SSA disks formatted with 524 byte sectors
- System adds:
 - block sequence number
 - longitudinal redundancy check (LRC) bytes

AS/400 Header	Data	S E Q #	L R C
8	512	2	2

524-Byte Sector



© IBM Corporation 1998

524 Byte Disk Sector

All VSS disks use a 524 byte sector size. The extra bytes are added by the SCSI adapter on a write and removed by the SCSI adapter on a read. As a result, blocks stored in cache consist of 524 byte sectors.

SSA disks formatted with 524 byte sectors

The SSA disks used in the Versatile Storage Server are formatted with 524 byte sectors.

Of the 524 bytes, 512 is user data content.

System additions

The system adds a block sequence number and longitudinal redundancy check bytes to the user data. Space is reserved for an 8 byte AS/400 header whether the host is a UNIX host or an AS/400. AS/400 hosts always write data in 520 byte blocks.

The block sequence number is a modulo 64 KB LBA. This is used to double check that the disk is seeking the right cylinder. The LRC bytes are added as the data enters the system and remain with the data as it is written from cache to SSA adapter to disk and back. This ensures that any corruption of data inside the Versatile Storage Server is detected by the VSS.

Cache Concepts



- Cache segment
 - Storage Server cache managed in 4 KB segments
- Disk track
 - Unit of SCSI disk I/O is 512 byte sector
 - ▶ *Stored in the VSS as 524 byte sectors*
 - Unit of UNIX disk I/O is file system block
 - ▶ *UNIX file system block size defined by system administrator*
 - ▶ *Often 4 KB*
 - VSS track is 32 KB
 - ▶ *Really 64 524 byte sectors*
 - ▶ *RAID-5 strip size*
 - ▶ *Used by cache management algorithms*
 - VSS supports partial track images in cache



© IBM Corporation 1998

Cache Concepts

Several concepts used by the VSS cache are described here.

Cache segment

Cache is managed in 4 KB segments. These are segments of 4096 bytes, not segments of eight 524-byte blocks. This means that space in storage server cache will always be allocated as an integral number of 4 KB segments.

Disk track

The VSS uses the concept of a disk track, which is equivalent to the RAID-5 strip size. The disk track is used in cache management, but is not externalized to the host.

SCSI I/O is in units of 512 byte sectors for a UNIX host.

A system administrator will define a file system block size for a UNIX file system. This is both a unit of allocation for the file and a unit of disk I/O. The 4 KB is a common file system block size. While the SCSI commands used by a UNIX host can address 512 byte sectors, the UNIX host reads and writes fixed-length groups of sectors, or blocks.

The VSS uses a track size of 32 KB, which is really 64 sectors of 524 bytes each. The track size is the same as the RAID-5 strip size and is used by the cache management algorithms.

A set of contiguous blocks that are part of the same 32 KB VSS track will require from two to nine 4 KB cache segments, depending on the file system block and the number of blocks staged to cache.

Cache Transparency



- VSS Storage Server cache is managed by the subsystem
- No external control of caching



© IBM Corporation 1998

Cache Transparency

VSS cache managed by the subsystem

The Versatile Storage Server cache is managed by the predictive cache algorithms described in the next foil.

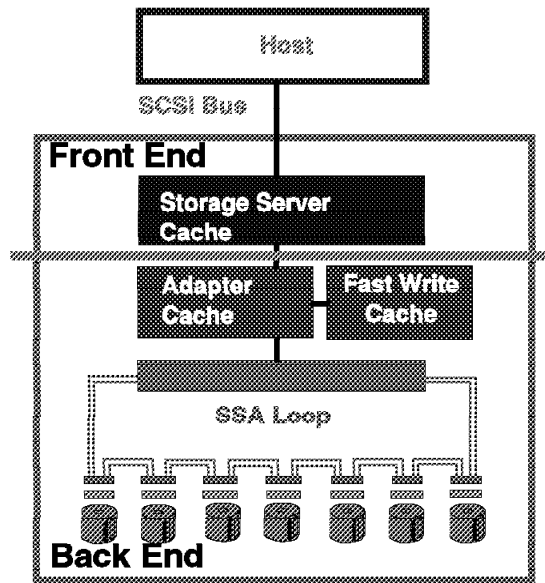
No external control of caching

The application has no control over how much data is staged to cache on a read, or which data is destaged when space in cache is required. There are also no installation parameters that modify the cache algorithms. The SCSI 10 byte read and write commands contain cache control bits that are ignored by the VSS. The SCSI-3 architecture defines cache control bits included in the 10 byte read and write commands. These cache control bits are ignored by the VSS.

Front End and Back End I/O



- **SCSI front end**
 - I/O as received from host
- **SSA back end**
 - Read cache miss only
 - Reads based on VSS algorithms
 - *Adaptive cache*
 - *Sequential prestage*
 - All writes
 - Writes based on size of write
 - *Partial 32 KB track*
 - *Full stripe*



© IBM Corporation 1998

Front End and Back End I/O

In this chart, we discuss the relationship of SCSI front end I/O to SSA back end I/O. Unlike the IBM 7190, which translates SCSI I/O into the equivalent SSA I/O, the VSS is a cache storage server that manages access to the SSA back end.

SCSI front end

The SCSI front end I/O is as sent by the host.

SSA back end

The SSA back end I/O, if any, is always generated by the VSS storage server.

Read cache hits don't require any SSA back end I/O.

Where SSA back end I/O is required, both how much data is read and how many I/Os are used to read the data are determined by the VSS.

With adaptive cache, the VSS may choose to read just the block requested, the block requested and the balance of the 32 KB track, or the entire 32 KB track.

In the case of a 4 KB block random read, either no back end I/O will be generated if the block is in storage server cache, or a single read of up to 32 KB will be generated.

Where the VSS detects a sequential pattern, sequential prestage is used. This generates 32 KB track back-end read I/Os and helps the VSS to try to stay ahead of the SCSI requests for data. Details of the sequential prestage are described in the next several foils.

Virtually all SCSI front end writes will generate SSA back end writes, except where a block is written twice before it has been destaged from Fast Write Cache. How data is destaged is controlled by the Fast Write Cache management algorithms, described later in this chapter.

How many SSA back end writes are generated from a SCSI write is a function of the size of the write and whether it generates a stripe write or an update write.

For update writes, as opposed to sequential writes, the VSS writes one or more blocks as part of a 32 KB track. Update writes require a parity update, which invokes the read data, read parity, write data, write parity sequence. Each of these steps consists of a block or a set of contiguous blocks that are part of the same 32 KB track.

In the case of sequential writes, the VSS uses stripe writes, avoiding the RAID-5 write penalty. Stripe writes are described later in this chapter.

Predictive Cache Algorithms



- Used for management of storage server cache
- Sequential predict and sequential prestage
 - Sequential access pattern detection based on residency of logically contiguous 32 KB tracks in cache
 - When sequential access detected, anticipatory staging of data begins
 - Improves sequential read throughput
- Adaptive cache
 - Dynamic selection of best random read caching algorithm based on data access patterns
 - Improves efficiency of cache management
 - Increased number of read cache hits in storage server cache



© IBM Corporation 1998

Predictive Cache Algorithms

Used for management of storage server cache

The algorithms described here are used to manage storage server cache, not SSA adapter cache. The SSA adapter volatile cache is managed by a simple LRU algorithm.

The algorithms are described in more detail on the next foil.

Sequential predict and sequential prestage

The VSS detects a sequential access pattern and begins reading data ahead of the data being requested in order to improve the sustained data rate of the sequential read.

Since SCSI I/O gives no explicit signal that a sequential I/O pattern is under way, the VSS detects this based on the residence in cache of logically contiguous 32 KB tracks in cache. It is important to emphasize that these are logically contiguous not physically contiguous tracks, and that logically contiguous 32 KB tracks are stored on different member disks in the array,.

Once the VSS has determined that a sequential read operation is under way, it begins anticipatory staging of data.

The result of sequential prestage is an improvement in the sustained data rate achievable for the sequential read stream.

Adaptive cache

For random reads, the VSS performs adaptive caching.

Depending on the access patterns for data, the VSS will either stage just data read to cache, or the data read and the balance of the 32 KB track, or the entire 32 KB track. The choice of caching algorithm is adaptive, so if different data is stored on the disk, the caching algorithm will adapt to the access pattern of the new data. Similarly, if there is a change in access pattern of data, the algorithm will adapt.

The advantage of adaptive caching is improved efficiency of cache management. Adaptive caching doesn't put more data in cache; it puts more of the right data in cache.

A higher percentage of cache hits may be experienced in a given cache size with adaptive caching. Data with a high locality of reference will tend to be staged with partial or full track staging. Data with a low locality of reference will tend to be staged with block mode staging. This helps ensure that the data in cache is data likely to produce read cache hits.

Sequential Predict and Sequential Prestage



- Used for management of storage server cache
- Sophisticated sequential access pattern detection
- Anticipatory staging of data
- Storage Server prestages a sequential staging unit
 - 5 stripes if 6+P RAID rank (30 tracks)
 - 4 stripes if 7+P RAID rank (28 tracks)
- SSA back end reads are of 32 KB tracks
- Parallel access increases sequential bandwidth



© IBM Corporation 1998

Sequential Predict and Sequential Prestage

Used for management of storage server cache

The algorithms described here are used to manage storage server cache, not SSA adapter cache. The SSA adapter volatile cache is managed by a simple LRU algorithm.

Sophisticated sequential access pattern detection

The sequential prediction algorithm recognizes a sequential pattern in the SCSI access even if there are intervening read or write I/Os to other portions of the virtual disk or RAID array. I/Os to another file in the same file system do not terminate a sequential prestage that is under way.

Anticipatory staging of data

Sequential prestage begins when a sequential access pattern has been recognized. When a read request is received, if multiple 32 KB tracks preceding the one containing the requested blocks are in storage server cache, the VSS begins sequential prestage operations. Prestage continues as long as read requests are received and the preceding 32 KB tracks for the requested block are in storage server cache.

Because the sequential prediction algorithm cannot predict when the sequential retrieval will stop (it has no knowledge of the file system using the logical disk),

it is likely that a partial or full sequential striping unit will be read to storage server cache but not accessed by the sequential read stream. This small amount of unnecessary disk I/O is an unavoidable side effect of sequential prestage. Since sequential prestaging does not begin until multiple consecutive 32 KB tracks are in storage server cache, every small burst of sequential reads does not initiate sequential prestaging.

Storage server prestages a sequential staging unit

The unit of prestage is a sequential staging unit:

- For a 6+P array, this is five stripes, or five 32 KB tracks from each of the six data disks in the RAID stripe (30 tracks total).
- For a 7+P array, this is four stripes, or four 32 KB tracks from each of the seven data disks in the RAID stripe (28 tracks total).

SSA back end reads are of 32 KB tracks

A sequential prestage operation always results in issuing multiple SSA back end reads of 32 KB each.

Parallel access increases sequential bandwidth

Because each disk in the RAID array is independent, each can be processing a read request in parallel with one or more others. Subject to the bandwidth limitations of the SSA loop, the SSA adapter, and the host SCSI adapter, the parallel access to the RAID-5 array allows the array to perform like a RAID-3 array when performing large sequential transfers.

The SSA disks used in the VSS will read the data requested and the following sectors up to 64 KB. The 32 KB reads issued by the VSS will result in 64 KB being read into the SSA disk buffer. The VSS will issue reads for four or five 32 KB tracks on each disk (depending on whether the array is 6+P or 7+P) in rapid succession. The SSA disks can accept up to three I/O operations before the first is performed.

This design allows the VSS to read up to four 32 KB blocks in a single rotation, yet still receive a task-complete indication from the SSA back end when data from the first stripe is available. Data being read sequentially is returned to the SCSI host as soon as it is requested and as soon as it is in the storage server cache. The storage server requires a task-complete indication from the SSA back end to know that data from a stripe is available in cache to be transmitted to the host.

Sequential Prestage Cache Management



- Sequential prestage synchronized with host access
- Sequential data LRU-destaged (preferentially)



© IBM Corporation 1998

Sequential Prestage Cache Management

In this foil, we talk about some of the details of the effects of sequential prestage on VSS storage server cache management.

Sequential prestage synchronized with host access

The VSS begins the prestage of the next sequential staging unit when the SCSI front end read stream reaches the middle of the current sequential staging unit. This allows the VSS to attempt to stay ahead of the SCSI front end read stream without allowing a single sequential prestaging operation to flood the storage server cache.

Sequential data LRU-destaged

Data read as part of a sequential prestage operation is expected to be less likely to generate a later read cache hit than data that is read randomly. As a result, the data from a sequential prestage operation is preferentially but not immediately destaged. Read cache hits from random reads or another sequential read stream are possible, although the cache residence time of the data will be less than for data read or written as part of a random I/O.

Adaptive Cache Concepts



- Management of storage server cache
- Dynamic selection of best caching algorithm
- Partial and full track staging
- Adaptive caching managed in track-bands
 - Algorithm selected for a 1920 track band
 - Bands do not span virtual disks



© IBM Corporation 1998

Adaptive Cache Concepts

In this foil, we investigate the concepts used by the adaptive cache algorithms.

Management of storage server cache

Adaptive cache manages the VSS storage server cache, not the SSA adapter caches.

Dynamic selection of best caching algorithm

The objective of adaptive caching is the selection of a random read caching algorithm best suited to the data access patterns for particular data.

If many reads are issued for data that would have been staged had a different cache management algorithm been used, the VSS may change to a more appropriate cache management algorithm for that data. The default is to cache blocks read.

Partial and full track staging

If a random read is issued for data in the same 32 KB track as a block already in cache, this is noted by the VSS.

With partial track staging, the block requested and the balance of the 32 KB track are staged into cache.

With full track staging, the entire 32 KB track containing the block requested is staged into cache.

Adaptive caching managed in track bands

Adaptive cache manages access in bands of 1920 32 KB tracks.

The same caching algorithm is used for all data in a band. The choice of algorithm is dynamic, but if it changes, it changes for access to all data in a band.

Bands do not span virtual disks. A virtual disk will be divided into bands of 1920 tracks, except that the last band will probably have fewer than 1920 tracks. A new virtual disk will always begin a new 1920-track band.

Adaptive Cache Algorithm



- Storage Server generates cache statistics
 - Subsequent miss
 - ▶ *Track is in cache but block requested is not*
 - Back access
 - ▶ *Requested block follows block that caused track stage*
 - Front access
 - ▶ *Requested block precedes block that caused track stage*
 - Block access mode
 - partial track mode
 - full track mode
- Algorithm adapts to changing data access patterns
- Algorithm is not biased by sequential access



© IBM Corporation 1998

Adaptive Cache Algorithm

In this foil, more detail is given about the adaptive caching algorithm.

Storage server generates cache statistics

Statistics are kept for all full or partial 32 KB tracks in cache, except those staged as a result of a sequential prestage operation. This prevents a sequential read from changing the access pattern used for random processing of data. Statistics for the 1920-track band determine use of block access mode, partial track mode, or full track mode for all reads to tracks in the band.

A subsequent miss occurs if part of a 32 KB track is in cache but the block requested is not. A subsequent miss is an indication that partial track or full track staging would be beneficial for this data. A low subsequent miss rate is an indication that block access mode should be resumed for a band currently performing partial track or full track staging.

A back access is a miss where the block requested follows the block that caused the track to be in cache. A back access is an indication that partial track staging would have been beneficial for this data.

A front access is a miss where the block requested precedes the block that caused the track to be in cache. A front access is an indication that full track staging would have been beneficial for this data.

The Versatile Storage Server periodically examines these statistics to determine if the access mode for the band should be changed.

Algorithm adapts to changing data access patterns

The algorithm is dynamic: it adapts to changing data access patterns, either because the data stored on a virtual disk has changed, or because the access pattern to data already stored has changed.

Algorithm not biased by sequential access

The algorithm is not biased by sequential access to begin partial or full track staging for data that would not benefit from this caching when accessed randomly (for example, data backup).

The choice of caching algorithm for a band does not affect sequential prestage for data in that band.

Input/Output Operations



- Random reads
 - Cache hit
 - Cache miss
- Sequential reads
- Writes
 - Fast write
 - RAID-5 write penalty
 - Fast Write Cache management
 - Stripe writes
 - Writes that are not fast writes



© IBM Corporation 1998

Input/Output Operations

We finally have enough background on the concepts and operation of the VSS to begin a discussion of how the VSS processes specific I/Os.

This is an introductory chart that describes I/O data flow. For this discussion, I/O operations are characterized as random reads, sequential reads, and writes. The specific flow for each of these I/O types is described in the following foils.

Random reads

Read operations reading a small amount of data (usually one or two blocks) from a disk location, then seeking a new location before reading again, are called *random reads*. Random reads are typical of transaction-based systems and are usually response-time sensitive. We describe the processing flow of a random read that is a cache hit in the storage server cache and a random read that results in a cache miss in the storage server cache.

Sequential reads

Read operations reading large contiguous blocks of data, such as file system backups and database scans used in decision support applications are called *sequential reads*. Sequential read and write operations are usually throughput-sensitive rather than response-time sensitive; the sustained throughput (in megabytes per second) is usually more important than the response time of an individual read operation.

The sequential detect and sequential prestaging algorithms, discussed in “Sequential Predict and Sequential Prestage” on page 135, are designed to increase sequential read throughput.

The discussion of sequential reads and writes pertains to reads and writes in LBA sequence. A UNIX file system will perform sequential readahead, and the file system knows

whether data for a given file is stored in more than one extent on the logical volume. A UNIX file system considers sequential access to be sequential relative to the file system, the VSS considers sequential access to be sequential relative to the VSS virtual disk. Except in very badly fragmented file systems, a UNIX sequential read of a file 10 MB or larger will generate at least some I/O considered sequential by the VSS.

Writes

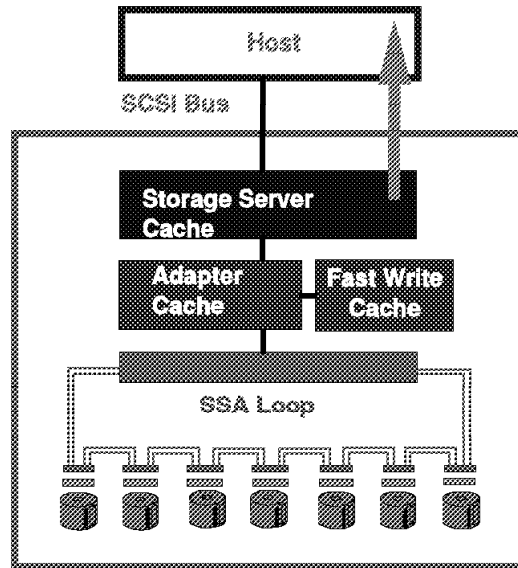
Because much of the processing of random writes and sequential writes is the same, they are covered in a single discussion.

The discussion of write operations includes a description of how data is written to Fast Write Cache and cache in the SSA adapter, and how the adapter manages the destage of write data to disk. Virtually all writes to the VSS are fast writes. Exceptions are noted later in this chapter.

Random Read Cache Hit

IBM Corporation
© 1998
All rights reserved.
IBM, the IBM logo, and
Versatile Storage Server
are trademarks of International
Business Machines Corporation.
Other names may be the
trademarks of their respective
owners.

- Cache directory searched
- Transfer from storage server cache to host



© IBM Corporation 1998

Random Read Cache Hit

This chart describes the processing of a SCSI read command that results in a cache hit in the Versatile Storage Server cache.

Cache directory searched

For all read commands, the first thing the VSS does is to search the cache directory to determine if the read can be satisfied from the cache.

Transfer from Storage Server cache to host

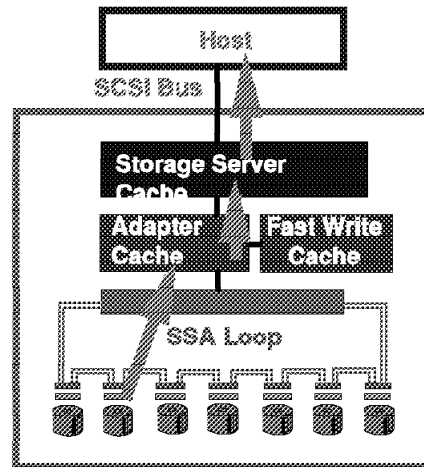
If the requested data is already in the cache, no I/O to the SSA disks is required. The requested data is transferred from the storage server cache to the host. The 32 KB track containing the data read is placed at the top (most recently used end) of the queue used for cache management.

If the read request can be partially satisfied from cache but also requires some sectors not in the cache, then the read request is treated as a cache miss. Only those sectors not already in the cache are read from the SSA disks. There is no reason to reread data from disk that is already in the cache. Once all the data requested is in the storage server cache, it can be returned to the SCSI host.

Random Read Cache Miss



- Cache directory searched
- Storage Server determines what data to read from disk
- Read from disk
 - Disk seeks, if necessary
 - Latency
 - Data read from disk surface to disk buffer
 - Data transferred from disk buffer to adapter cache and controller cache
 - Blocks requested sent to host



© IBM Corporation 1998

Random Read Cache Miss

This chart describes the processing of a SCSI read command that asks for at least some data not in the VSS cache.

Cache directory searched

For all read commands, the first thing the VSS storage server does is to search the cache directory to determine if the read can be satisfied from the cache.

If at least some of the data requested is not in the cache, it must be read from the SSA disks. Only the data not in the cache is read from disk.

Read from SSA disk

A read is issued by the storage server to one or more SSA disks. Most read requests result in a read to a single SSA disk. The VSS has a 32 KB strip size, which means that logically contiguous blocks of 32 KB are written across the member disks in an array. Access to more than 32 KB in a single read results in access to more than one member disk in the array, and will require more than one SSA read I/O.

Depending on the location of the heads, a seek operation may be necessary.

It is unlikely that a host program will know whether a seek will be necessary to perform a read. In the VSS subsystem, what is seen by the host as a physical

disk is not really a single hard disk, but rather an extent of a RAID array. The RAID array may contain many virtual disks. The read of a single block will be directed to a single member disk, moving the heads on that disk only. In addition, because a read can be satisfied from the storage server cache, disk access can be avoided for an I/O, which means that some host-initiated I/Os result in no SSA I/O and therefore do not move the disk heads.

A host application attempting to manage arm movement on VSS logical disks may be relying on incorrect assumptions. In most cases, this renders such programs ineffective but not harmful to application performance.

The primary use of the disk buffer is to allow data transfer on the SSA loop asynchronously to the rotational position of the disk surface in relation to the heads. This allows the interleaved transmission of frames from many disks on the SSA bus.

Data is read from the disk surface to the disk buffer, and can then be placed on the SSA loop as frames are available and as dictated by the SSA SAT protocol.

Data read from the disk is transferred on the SSA loop to the SSA adapter where it is placed in the cache. It is then also transferred to the storage server cache. The queue used for cache management is updated for both the adapter cache and for the storage server cache.

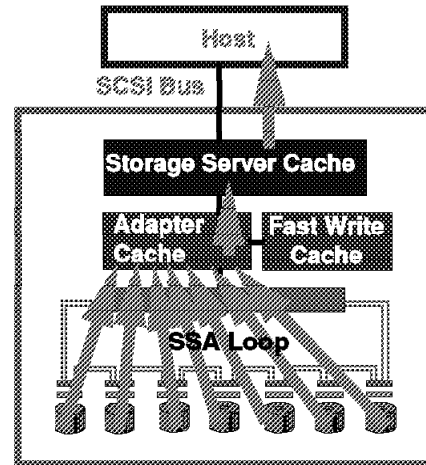
When all data necessary to satisfy the host request is in storage server cache, it can be transferred across the SCSI bus to the host.

In a case where the adaptive caching algorithm in the VSS storage server has determined that full or partial track staging should be used, the blocks requested will not be available for transfer to the host until the completion of the staging operation.

Sequential Read



- Cache hit possible if data recently written
- Data prestaging begins when sequential access is detected
- Storage Server overlaps prestage with host access
- Data read sequentially preferentially destaged from storage server cache



© IBM Corporation 1998

Sequential Read

Cache hit possible if data recently written

While sequential access reads usually result in a cache miss, an entire sequential read operation can be processed as a cache hit if the data is recently written, or recently read. Because the VSS caches data written as well as data read in the storage server cache, a file written to disk and then reread by the same or a different host may still be in cache, and could be returned without access to the backstore disks.

Data prestaging begins when sequential access is detected

The first few reads in a sequential pattern are processed as random reads. This is done so that every small sequential read pattern does not invoke sequential prestage. Once multiple 32 KB tracks of data are in the cache, the VSS recognizes a sequential access pattern and begins sequential prestaging. If sequential prestaging prestages data faster than the host is requesting it, it appears to the host as though the data was already in storage server cache.

Even if the host is requesting data faster than the sequential prestage can provide it, sequential throughput is still maximized because the VSS is issuing back end I/Os designed to exploit the capabilities of the SSA disks and the SSA bus.

Storage server overlaps prestage with host access

When the 32 KB track in the middle of the sequential staging unit group is read by the host, the stage of the next sequential staging unit is begun. In this manner, the VSS attempts to prestage data in anticipation of a host read but without unnecessarily flooding cache with data from a single read stream.

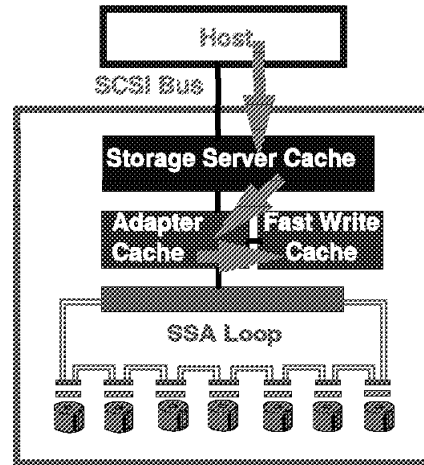
Data read sequentially preferentially destaged from storage server cache

Data read as part of a sequential prestage operation is preferentially destaged from storage server cache. Data read as part of a sequential prestage operation is expected to be less likely to generate a later read cache hit than data read randomly. As a result, the data from a sequential prestage operation is preferentially but not immediately destaged. Read cache hits from random reads or another sequential read stream are possible, although the cache residence time of the data will be less than for data read or written as part of a random I/O.

Fast Write



- Fast Write bypass when appropriate
- Optimum data availability – three copies of data
- Data destaged asynchronously from Fast Write Cache (logically)



© IBM Corporation 1998

Fast Write

Fast Write bypass when appropriate

A write operation that includes one or more full stripes will not be processed as a fast write. The portions of the write that are not part of full stripes will be written to SSA adapter cache and Fast Write Cache and processed as fast writes. Any full stripes will be written to the adapter as a single I/O for the stripe. This data will be written to adapter cache but not Fast Write Cache. Parity will be generated in the adapter cache and a stripe write will be issued. The task-complete signal will not be sent until fast write data is in adapter cache and Fast Write Cache and any full stripes and the parity strip are written to disk.

This is done because the RAID-5 write penalty can be avoided through the stripe write without the use of Fast Write Cache for a large write.

Optimum data availability – three copies of data

Fast write data is written to storage server cache and then, in one or more write operations, to SSA adapter cache and Fast Write Cache. The number of SSA back end I/O operations depends on the number of 32 KB tracks spanned by the data being written.

Task completion is signaled once data is safely stored in the SSA adapter cache and Fast Write Cache. The storage server cache copy is not required to ensure fast write integrity.

Data integrity is ensured because there are two copies of the data stored in the SSA adapter, one in adapter volatile cache, and one in Fast Write Cache. In the unlikely event of a failure in the SSA adapter, the Fast Write Cache copy ensures integrity. In the even less likely event of a failure of the Fast Write Cache component, the SSA adapter cache copy will be used to perform writes.

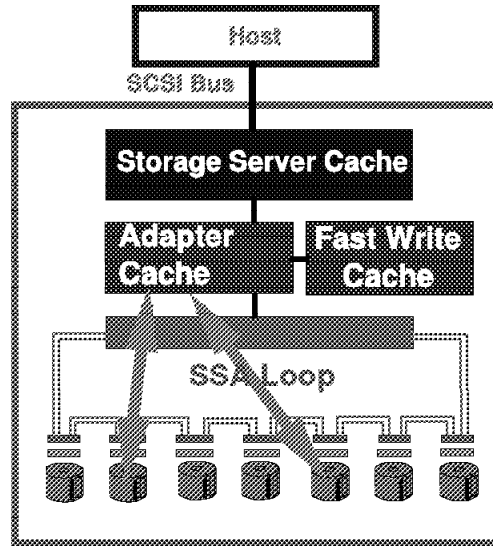
Data destaged asynchronously from Fast Write Cache

Destage of data from Fast Write Cache is not immediate. Fast Write Cache is managed by threshold, which will be described in more detail in a following foil. The word logically is placed on the foil to show that conceptually, data is destaged from the Fast Write Cache, but in reality, it is destaged from the Adapter Cache.

RAID-5 Write Penalty



- Update writes subject to RAID-5 write penalty
 - Read old data
 - Read old parity
 - Write new data
 - Write new parity
- Fast write masks the write penalty on update writes
 - SCSI task complete is signaled as soon as data is in SSA adapter cache and Fast Write Cache
 - Resulting I/O occurs after SCSI task complete sent



© IBM Corporation 1998

RAID-5 Write Penalty

In this foil, we describe the RAID-5 write penalty.

Update writes subject to RAID-5 write penalty

In order to complete a RAID-5 update write, parity must be updated. It would be possible to update parity by reading the corresponding blocks from all the disks not being written, performing an exclusive- or (XOR) against that data and the data to be written, and rewriting the new parity. This approach, while conceptually simple, requires an I/O to each of the member disks of an array.

A better algorithm relies on the ability to reverse an XOR operation by performing it again. If I have a block that contains the XOR of blocks A, B, C, D and E and I apply XOR to block A with this parity block, the result is the same as for an XOR of B, C, D, and E without A. Because of this, the sequence

- Read old data
- Read old parity

XOR old data with old parity, which yields the XOR of all the other member disks in the array.

XOR new data with the result of the above XOR to get the new parity data.

- Write new data
- Write new parity

can be used for an array with any number of disks to perform an update write with four I/Os, two each to two of the member disks.

Fast write masks the write penalty on update writes

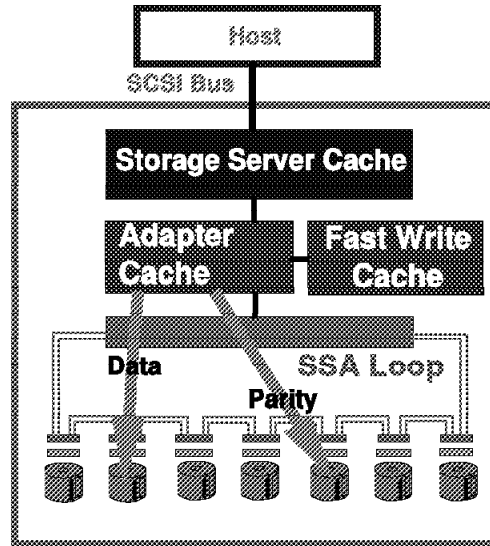
The RAID-5 write penalty can result in a long service time for an update write to a RAID-5 array, since it can result in four I/O operations. Fast write masks the write penalty, because SCSI task complete is signaled as soon as the data is in the SSA adapter cache and Fast Write Cache. The destage of the data written to Fast Write Cache occurs after the SCSI task is complete.

There is still a bandwidth limitation to update writes that can be performed to a RAID array in the VSS or any other RAID-5 disk subsystem. If this bandwidth limit is not approached, fast write effectively masks the overhead of the RAID-5 write penalty. If the bandwidth limit is approached or exceeded, elongated service times may still be experienced for RAID-5 update writes. The VSS design ensures that the RAID-5 write penalty is masked whenever possible.

Data Integrity for RAID-5 Parity



- RAID-5 update write integrity challenge
- Fast Write Cache processing assists parity integrity
- Parity regenerated if SSA adapter cache malfunctions
 - No RAID-5 update write integrity exposure



© IBM Corporation 1998

Data Integrity for RAID-5 Parity

RAID-5 update write integrity challenge

One of the challenges facing the designers of a RAID-5 array subsystem is that update writes that result in a write of data and a write of parity must act like a single operation to ensure data integrity. If a data block is updated but the associated parity is not updated, the update would be lost if data regeneration were ever required for that block. The RAID Advisory Board refers to this issue as the RAID-5 *write hole*.

Fast Write Cache assists parity integrity

The VSS does not store parity in Fast Write Cache, but it does store an indication of pending parity updates in Fast Write Cache until that update is complete. The indication that a parity update is required is placed when data is written. In normal processing, parity is generated as part of the destage operation of data from Fast Write Cache to disk. The Fast Write Cache indication is not removed until the write of parity to disk is complete.

Parity regenerated if SSA adapter cache malfunctions

In case of SSA adapter cache failure, the Fast Write Cache indication ensures that the parity update will occur. All data required to regenerate the parity can be read from disk.

This approach minimizes the Fast Write Cache space used to ensure data integrity for RAID-5 parity. In the unlikely event of an SSA adapter failure, the data required to regenerate the parity is read from disk.

Fast Write Cache Management



- Asynchronous destage to disk
- Fast Write Cache managed by threshold to
 - Stripe writes
 - ▶ *RAID-5 write penalty avoided by writing an entire stripe including its parity*
 - Write preempts
 - ▶ *Only one destage necessary if block is rewritten while still in Fast Write Cache*
 - Write blocking
 - ▶ *Adjacent blocks are destaged together to minimize arm movement and latency*



© IBM Corporation 1998

Fast Write Cache Management

Asynchronous destage to disk

Data is not destaged immediately from Fast Write Cache. While this could be done, data integrity is ensured as soon as the data is stored in adapter cache and Fast Write Cache, so there is no data integrity exposure in implementing a more sophisticated destage management algorithm.

Fast Write Cache managed by threshold

A write to a RAID-5 array incurs the RAID-5 write penalty (read old data, read old parity, write new data, write new parity). A single RAID-5 write can therefore generate four I/O operations to disk.

A significant increase in sequential write throughput can be achieved by using stripe writes. If all the data for a full 6+P or 7+P stripe is in the SSA adapter Fast Write Cache, the parity can be calculated in adapter cache and the data and parity for the stripe can be written with seven (6+P) or eight (7+P) writes. Otherwise, up to 28 I/Os would be required.

Stripe writes increase the sequential write throughput of the VSS by reducing the number of disk I/Os required to write data. By holding data sequentially written in the Fast Write Cache until a full stripe is collected, the VSS can realize this throughput improvement regardless of the host applications write blocking.

I/Os containing a full stripe are written by the VSS storage server so that any stripes are written individually in a single SSA write. Such stripes bypass Fast Write Cache. Where I/Os do not include full stripes, stripe writes are collected in Fast Write Cache.

Write preempts

Write preempts occur when a block still in Fast Write Cache is updated again. The VSS allows the update to occur in Fast Write Cache without the block first being written to disk. This in turn allows a block that is frequently written, such as a file system superblock, to be written to the SSA adapter cache and Fast Write Cache but destaged fewer times than it was written, all while maintaining full data integrity protection.

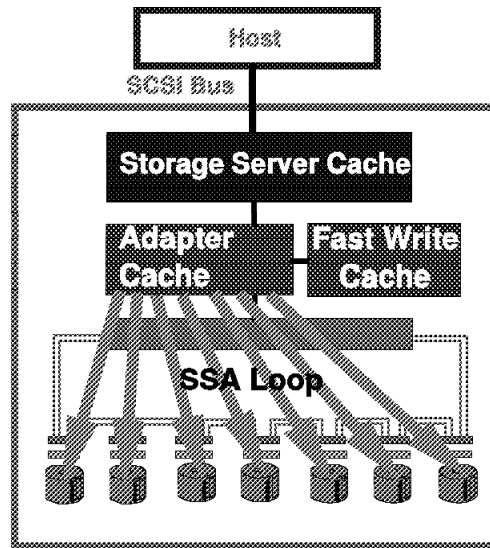
Adjacent blocks are destaged together to minimize arm movement and latency. When a block that has reached the bottom (least recently used end) of the cache management queue is selected to be destaged to disk, any other blocks adjacent to it are destaged also. This allows more effective utilization of the target disk drive and the drive containing associated parity.

Note that storage server cache is managed by 32 KB tracks, so that any blocks in a 32 KB track will be destaged from storage server cache together. Fast Write Cache is managed by blocks. Blocks that are part of the same 32 KB track are destaged independently unless they are adjacent.

Stripe Writes



- RAID-5 write penalty avoidance
 - Write data to six or seven strips
 - Write parity
- Sequential write throughput



© IBM Corporation 1998

Stripe Writes

In this foil, we describe in more detail the processing of stripe writes.

RAID-5 write penalty avoidance

Stripe writes can be used only when all of the data in a stripe is written to disk. In the VSS, this occurs whenever the data in a stripe is written sequentially, regardless of whether this is in a single I/O or in several I/Os. It is theoretically possible that the data in a stripe could be LRU destaged from Fast Write Cache before the write of all data in the stripe is complete. However, since most sequential write I/O is rapid, this is unlikely to occur in real environments.

Parity is generated by XOR of the data to be written. Data is written in 32 KB tracks, which is a VSS strip, to each member disk of the array containing data within this redundancy stripe.

A 32 KB strip is written to the member disk containing parity for this redundancy stripe.

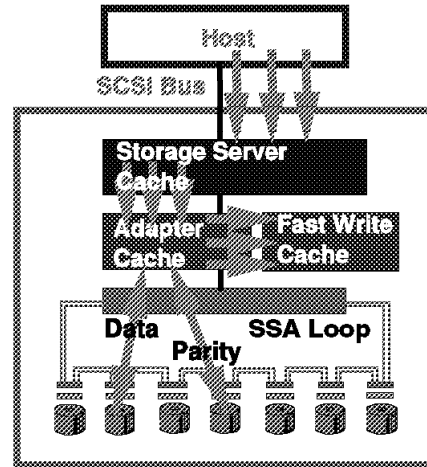
Sequential write throughput

Stripe writes not only avoid the RAID-5 write penalty by writing to multiple member disks in an array concurrently, stripe write operations exploit the parallel processing capabilities of a RAID-5 array, making the write operation RAID-3-like. Write throughput with stripe writes can exceed the write throughput possible to a single disk.

Write Preempts



- Update write to a particular block
- Fast Write "hits"
- Multiple updates processed by single destage



© IBM Corporation 1998

Write Preempts

Write preempts are described in more detail.

Update write to a particular block

A write preempt begins with an update write to a particular block. The block is placed at the top (most recently used end) of the LRU queue used for Fast Write Cache management.

Fast Write Hits

A write preempt occurs if additional update writes to the same block occur before the block is destaged from Fast Write Cache. The VSS allows the second or subsequent write to occur in Fast Write Cache without destaging the data from the first write. The block is again placed at the top (most recently used end) of the LRU queue used for Fast Write Cache management.

Multiple updates processed by single destage

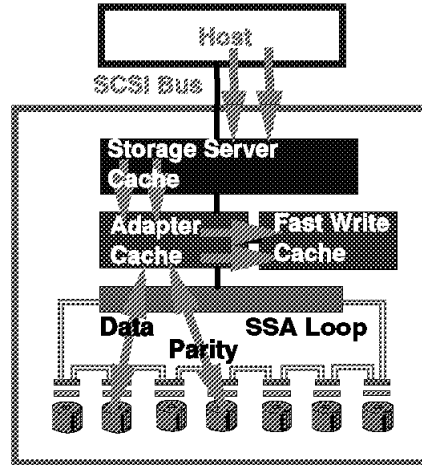
Only one destage from Fast Write Cache is required even though two or more update writes have occurred. Since each destage would probably result in a RAID-5 write penalty, avoiding a destage avoids four I/O operations.

This foil depicts three update writes to the same block. All three are written to SSA adapter cache and Fast Write Cache. Only the last write is destaged to disk, which involves the read data, read parity, write data, write parity sequence.

Write Blocking

IBM Corporation
© 1998 IBM Corporation
All rights reserved.
IBM, the IBM logo, and
Versatile Storage Server
are trademarks of International
Business Machines Corporation.

- Adjacent blocks destaged together



© IBM Corporation 1998

Write Blocking

More detail is provided about write blocking.

Adjacent blocks destaged together

Fast Write Cache LRU management is on a block, not 32 KB track, basis. However, when a block is chosen to be LRU destaged, the VSS will determine if an adjacent block is also in Fast Write Cache. If it is, they are destaged together to minimize the RAID-5 write penalty overhead.

Write blocking will occur whether the blocks were written in the same or different SCSI I/Os. Write blocking occurs if adjacent blocks are in Fast Write Cache and one is chosen to be destaged.

This foil depicts two SCSI update writes that write adjacent blocks in the same 32 KB track. Both are written to SSA adapter cache and Fast Write Cache. When the first write is destaged to disk, which involves the read data, read parity, write data, write parity sequence, the second block is also destaged. In this case, "read data" reads two adjacent blocks, as does "read parity." The two write operations also write two adjacent blocks. Both blocks are updated with what is effectively a single RAID-5 write penalty sequence.

Fast Write Cache Destage



- Fast Write Cache destage triggered by threshold
- Other destage triggers:
 - Once a stripe has been collected in the Fast Write Cache, destage is immediate
 - During periods of inactivity, data is slowly LRU destaged
 - During normal shutdown of the VSS, all data is destaged from Fast Write Cache to disk
- Destage from SSA adapter cache



© IBM Corporation 1998

Fast Write Cache Destage

Fast Write Cache destage triggered by threshold

Fast Write Cache destage is triggered when Fast Write Cache reaches 70% full. The 70% threshold is specified in the VSS Licensed Internal Code and cannot be overridden by an installation. If the Fast Write Cache is less than 70% full, destage will not occur, with the following exceptions:

Other destage triggers

Because the most significant performance benefit of stripe writes is achieved once a single stripe is written, data is destaged as soon as a full stripe is collected in Fast Write Cache. While additional throughput improvements could be realized by writing more than one complete stripe in rapid succession, the additional Fast Write Cache utilization does not justify this.

Because writes to a disk are written first to the disk buffer, the disk itself can manage writing up to three 32 KB tracks in one operation if the subsequent 32 KB tracks are written to the disk before the first can be written from the disk buffer to the media. This can occur if writes occur at a high rate as they can during a file system restore or the creation of a logical volume copy.

If there has been no activity to the SSA adapter for 5 s, a small amount of data is LRU destaged from Fast Write Cache even if the Fast Write Cache is not at the

70% full threshold. If the period of inactivity lasts, all data will slowly be Fast Write Cache destaged.

During normal shutdown of the VSS, the VSS issues a command to the SSA adapter to destage all data from Fast Write Cache to disk. The VSS waits for all data to be destaged before shutting down.

This is not required to ensure the integrity of fast write data, since data would remain in Fast Write Cache until the machine was again powered up. It is done so that system administrators who want all data destaged to disk on normal shutdown will be satisfied.

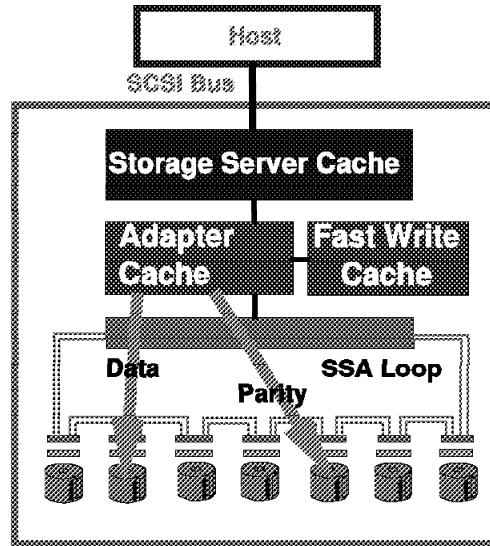
Destage from SSA adapter cache

Throughout this chapter, we speak of destaging data from Fast Write Cache. The actual destage is from the SSA adapter DRAM mirror of Fast Write Cache. The distinction is insignificant, but is mentioned for completeness.

Writes from Fast Write Cache to Disk



- Destage uses hierarchy of VSS storage
- Data removed from Fast Write Cache following destage
- SSA adapter cache copy retained to improve hit rates
 - Data and parity blocks
 - May provide cache hit for future update write



© IBM Corporation 1998

Writes from Fast Write Cache to Disk

Destage uses hierarchy of VSS storage

When data is destaged from Fast Write Cache to disk, it is written from the Fast Write Cache to the disk buffer. This means that writes can occur regardless of the rotational position of the disk relative to the blocks to be written, and regardless of whether data being written on the SSA bus to many drives requires that frames be interleaved on the bus.

Theoretically, data being read and being written could be moving in the same direction on the SSA bus. In the VSS, this will not occur, since the SSA adapter sends data to a disk by the shortest route.

Data removed from Fast Write Cache following destage

Data is not removed from the SSA adapter cache or Fast Write Cache until the data has been written from the disk buffer to the disk. This ensures data integrity by ensuring that a write is committed to disk before the data is destaged from Fast Write Cache. In case of a transient or permanent error in writing to the disk, the write is recoverable by the VSS even if the data must be held in Fast Write Cache until a repair action is complete.

It is also essential that the parity associated with the data destaged to disk be updated. The VSS ensures that this write will also occur. This was described in the foil entitled "Data Integrity for RAID-5 Parity" on page 153.

SSA adapter cache copy retained to improve hit rates

While the Fast Write Cache copy is kept only until the data has been written to disk, the SSA adapter cache copy is kept until it is demoted by LRU processing. This is done to enable a possible SSA adapter cache hit on a future update write when the data written as new parity or new data could be read as old parity or old data for the subsequent write.

Transfers from Storage Server Cache to SSA Adapter

- VSS data transfer
 - Segmented on 32 KB track boundaries
 - Except for stripe write which is sent to SSA adapter as a single write of six or seven 32 KB tracks
- Fast writes
 - Except for stripe writes, which are written to adapter cache (so that parity can be generated) and then directly to disk, bypassing Fast Write Cache
 - ▶ *Performance advantage of stripe write without using Fast Write Cache*
 - All writes to VSS are fast writes
 - ▶ *Except single writes incorporating one or more full stripes*



© IBM Corporation 1998

Transfers from Storage Server Cache to SSA Adapter

In this foil, we present more detail about transfers from the VSS storage server cache to the SSA adapter.

VSS data transfer

As we have seen, a single read or write can generate no back end I/O, a single read or write to the SSA adapter, or multiple reads or writes to the SSA adapter. How much data is read from the SSA back end is controlled by the adaptive cache of the VSS.

As a general rule, SSA back end I/O is segmented on 32 KB track boundaries. An I/O that writes 64 KB that spans three 32 KB tracks will generate three SSA back end writes. An I/O that reads 64 KB that spans three 32 KB tracks will generate up to three SSA back end reads, depending on what data is already in cache.

The exception to the general statement that I/O is segmented on 32 KB track boundaries is a stripe write. If a write includes one or more full stripes, each full stripe will be written to the SSA back end in a single I/O. This single I/O stripe write will bypass Fast Write Cache and will be processed in SSA adapter cache only.

Fast Writes

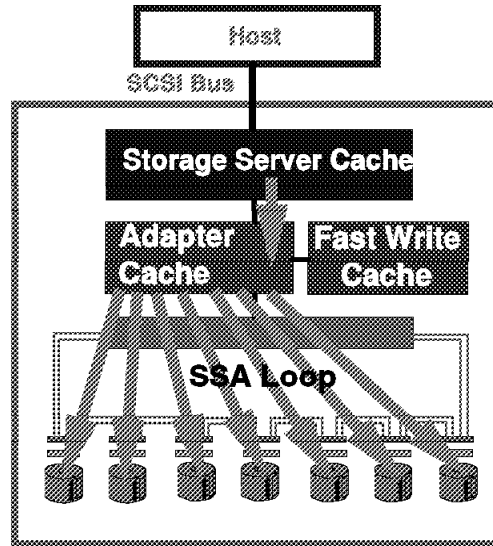
Except for SSA back end writes of a full stripe, all writes to the SSA back end are processed as fast writes by the SSA adapter.

As a result, any write that does not cause the VSS to generate a stripe write as a single I/O will be processed as a fast write. The VSS will generate a stripe write as a single I/O for any write including one or more full stripes.

Stripe Write to SSA Adapter



- Bypass Fast Write Cache
- Data is written from storage server cache to adapter cache
 - Parity generation is done in the adapter cache
- Data and parity written to disk buffers and to disks
- Transfer complete when data written to physical disk



© IBM Corporation 1998

Stripe Write to SSA Adapter

The processing of a stripe write written to the SSA adapter as a single I/O is described in more detail.

Bypass Fast Write Cache

Fast Write Cache is bypassed only when a stripe write is presented to the SSA adapter as a single write. This is the only time when Fast Write Cache is bypassed in the VSS. While the SSA RAID adapter cards used by the VSS have other instances in which Fast Write Cache will be bypassed, these conditions never arise with the I/O generated by the VSS storage server.

Data is written from storage server cache to adapter cache

Data for a stripe write I/O is written from storage server cache to adapter cache, but not to Fast Write Cache.

The strips in the stripe are XOR'd together to create the parity strip. Note that parity generation is always performed in adapter cache in the VSS.

Data and parity written to disk buffers

The data and parity strips are written to the appropriate disks, depending on the redundancy group.

Task complete when data written to physical disk

When the data write to disk media from the disk cache is complete, a transfer complete signal is sent by the SSA adapter to the storage server. The storage server then sends the formal task-complete signal to the host.

Versatile Storage Server Data Flow Summary

- Storage Server cache
 - Improved performance for read cache hits
- Adaptive cache
 - Increased number of read cache hits in storage server cache
- Sequential prediction and sequential prestage
 - Improved sequential read throughput
- Fast writes
 - Masks RAID-5 write penalty
- SSA adapter nonvolatile storage
 - Protection of fast write data



© IBM Corporation 1998

Versatile Storage Server Data Flow Summary

Storage server cache

Storage server cache provides read cache hits, either for data that is reread by the same or a different host, or for data that is staged by the adaptive cache algorithms. A read cache hit is a fast I/O.

Adaptive cache

The adaptive cache algorithms of the VSS increase the number of read cache hits for a given cache size. The adaptive cache algorithms dynamically observe the locality of reference of data stored on the VSS, and choose a cache staging algorithm based on the locality of reference observed. Data with high locality of reference will have more data loaded into cache for each cache miss read. Data with low locality of reference will load just the data read into cache, preserving cache for other uses.

The algorithms adapt to changing data access patterns.

Sequential prediction and sequential prestage

The sequential prediction and sequential prestage capabilities complement the sequential preload of the UNIX host. The sequential predict capability determines when sequential prestage should be used and sequential prestage increases the sustained data rate of sequential read streams.

Fast writes

Fast write masks the RAID-5 write penalty. While the RAID-5 entails an unavoidable overhead for update writes in a RAID-5 disk subsystem, the effects of the RAID-5 write penalty can be masked by fast write.

SSA adapter nonvolatile storage

Fast write wouldn't be very attractive if the integrity of data written to the disk subsystem were vulnerable until the data was committed to disk. The Fast Write Cache architecture of the SSA adapter used in the VSS protects the integrity of data written as soon as a task- complete indication is sent to the host in response to the SCSI write command.

Chapter 5. Drawer Design

Drawer Design



- Drawer design
 - Component
 - Drawer
 - Host bypass circuit
 - Disk drive
 - SMART/PFA
 - Disk drive reliability
 - Configuration
 - Sparing
 - Sparing procedure
 - Component interaction

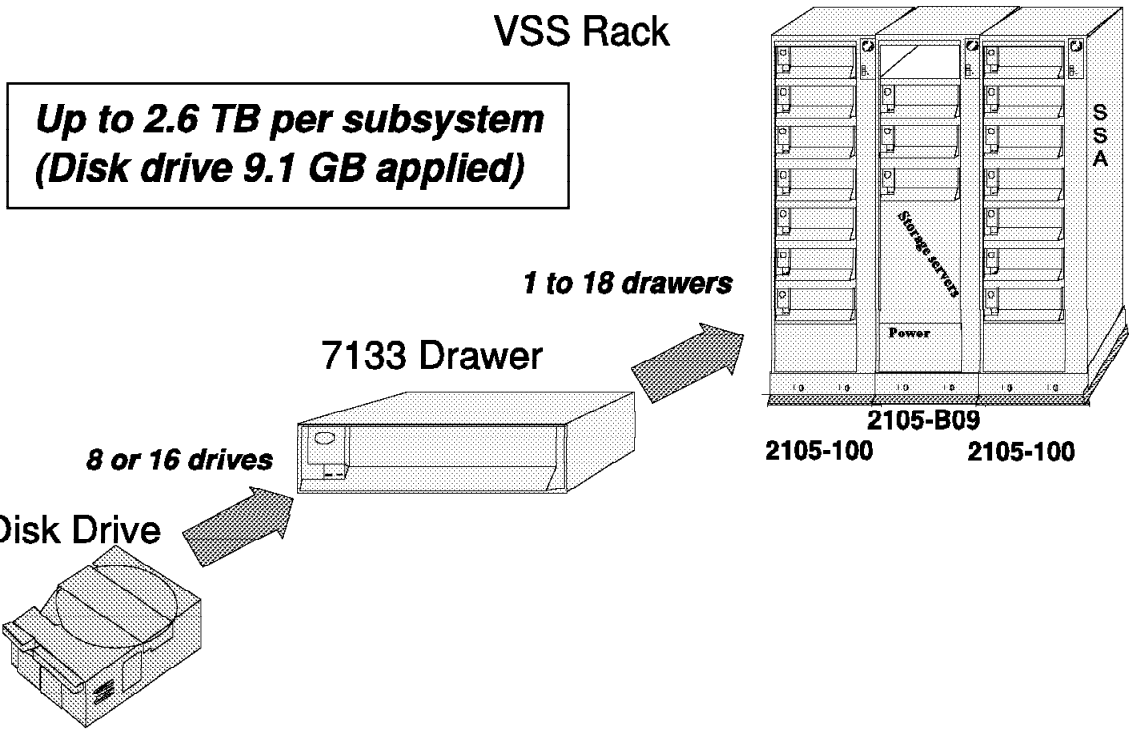


© IBM Corporation 1998

Drawer Design

This foil shows the topics we discuss in this chapter. In this chapter, we present the drawer design of VSS. The drawer contains 8 or 16 disk drives and carries the internal SSA path that connects the disks installed in the drawer. The drawer plays a key role in the SSA loop configurations, and is the physical interface between the SSA RAID adapter and the disks. In addition, each drawer has three power supplies and cooling fan modules so that the drawer can continue operation in case one of the power supplies, or a cooling fan, fails. We examine the hardware functions and, finally, explain the component interaction around the drawer.

Component

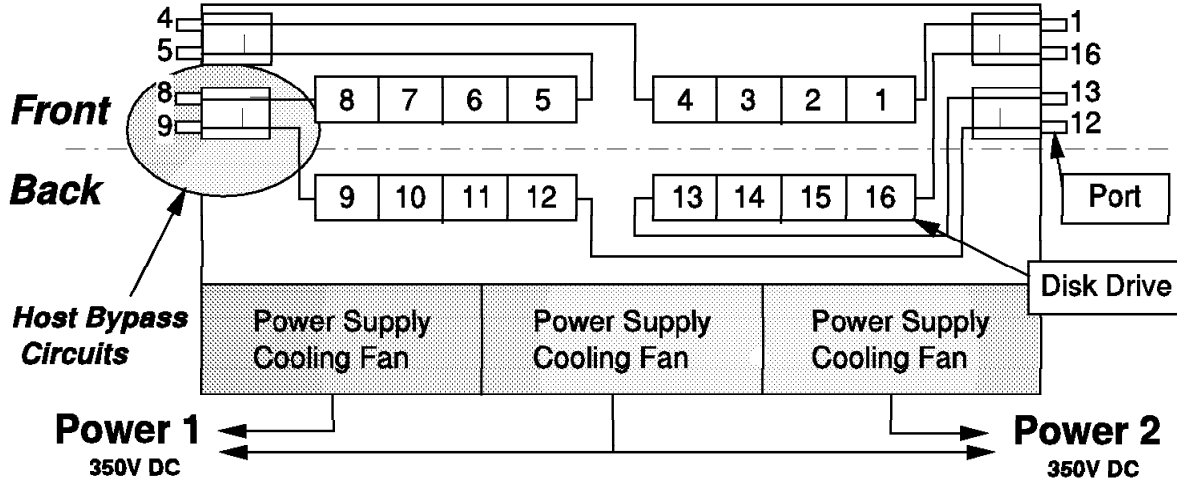


© IBM Corporation 1998

Component

This foil shows the components of the disk drive, drawer, and VSS racks. Either 8 or 16 drives can be installed in one VSS drawer and from 2 to 18 IBM 7133 drawers can be installed in one VSS subsystem. There are two types of rack: 2105-B09 and 2105-100. The 2105-B09 includes the basic facilities to operate the VSS, such as the storage servers, SSA RAID adapters, disk drives (sixteen 9.1 GB disks), and power supplies. Up to four drawers can be installed in the 2105-B09 rack. If more than four drawers are needed, the 2105-100 expansion rack can be installed in the VSS subsystem. Up to seven drawers can be installed in the 2105-100 expansion rack, and the maximum configuration of VSS is one 2105-B09 and two 2105-100. So the maximum usable capacity that can be installed in one VSS is 2.6 TB. This capacity includes the area required to store the parity of the RAID-5 implementation and the hot spare disk drives. If all RAID drawers are configured as a 6+P+S array and a 7+P array (sharing the spare between the two arrays) for 16 drives * 18 drawers, the amount of usable application data storage could be 2.129 TB. (This doesn't include either the parity or the hot spare disk drives.) If all RAID arrays are configured as 6+P+S for 16 drives * 18 drawers, the amount of usable data storage could be 1.965 TB. (This does not include either the parity or the hot spare disk drives.)

Drawer



© IBM Corporation 1998

Drawer

This foil shows the internal SSA connection path and the power supply and cooling fan unit in the VSS drawer.

Internal SSA path

The drawer provides SSA loop path and redundant power supplies and cooling fans for 8 or 16 drives. The foil shows 16 drives installed in one drawer, numbered from 1 through 16. In addition, there are eight ports per drawer, also numbered. Port 1 is connected to Disk 1, Port 4 is connected to Disk 4, and the other ports (5,8,9,12,13 and 16) follow the same rule. This rule simplifies loop connection design.

The connection between two drawers or between the host and the drawer is the SSA cable, which attaches to the port indicated in the foil. The ports of the drawer include a function called *host bypass circuit* (This function is explained in the next foil.)

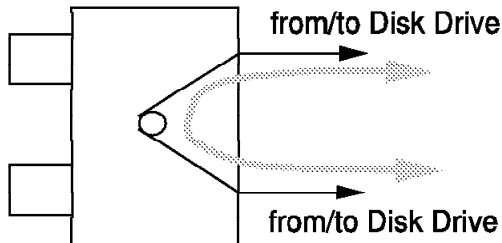
Redundant power supply and cooling fan

Each drawer has three power supplies and cooling fans. The power supply and cooling fan are together in one module. If one of the three power supplies in a drawer fails, no VSS operational outage occurs. In addition, VSS itself has two power supplies that meet all the power demands in the VSS rack, and each drawer takes power from both power supplies in the VSS rack. The power supply module of the VSS rack incorporates redundancy, as do the drawer's power supply and cooling fan modules. The power supply and cooling fan unit can be replaced without shutting down the power to the drawer or any other components. The drawer contains multiple disks, whose grouping helps to configure the SSA loops, along with the power and cooling modules.

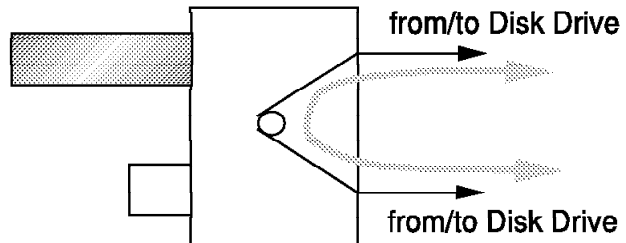
Host Bypass Circuit



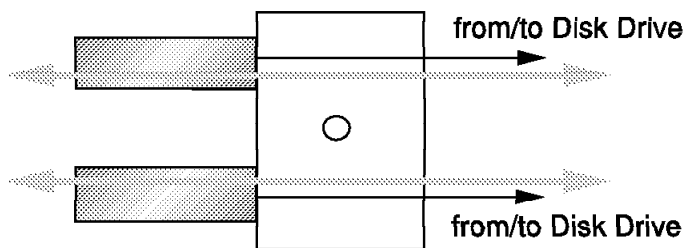
1. No SSA cable is connected or adapter fails.



2. One SSA cable is connected adapter fails



3. Two SSA cable is connected and no failure on SSA opposite side of connected SSA cable



Examples

Electric flow/SSA frame flow



Cable is connected correctly and no problem beyond the cable



Cable is NOT connected or there is a problem beyond the cable



© IBM Corporation 1998

Host Bypass Circuit

This foil shows the function of the host bypass circuit, one of the key functions of the drawer in which it is installed. The host bypass circuit works as follows:

- Pairs the Ports 1 and 16, 4 and 5, 8 and 9, 12 and 13.
- If one or both ports of the pair detect any abnormal electronic state on (or beyond) an attached cable, then each port can establish the connection internally (Patterns 1 and 2 on the above foil).
- If the SSA cables are attached to both ports of the pair and the device beyond the cable is working correctly, then each port will connect with the initiator or another drawer through the cable. (Pattern 3 on the above foil).

This bypass is an efficient remedy when the cable, the initiator, or the target fails. The circuit also provides easy cabling for the SSA loop of VSS. To configure 16 drives installed in one drawer on one loop, the cabling should connect Ports 1 and 16 with the two ports on the SSA RAID adapter. Then all other ports in the drawer function as the bypass circuit, so the SSA topology is configured as one loop.

Disk Drive

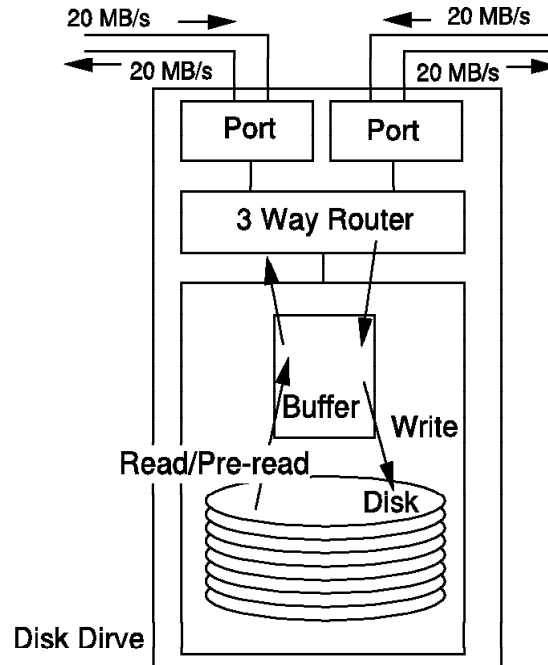


Ultrastar 2XP 9.10 GB DISK Drive

- ▶ Rotational speed: 7200 rpm
- ▶ Sustained data transfer rate: 6.5 - 10.0 MB/s
- ▶ Average read seek time: 8.5 ms
- ▶ Buffer (multi segmented): 512 KB
- ▶ Form factor: 3.5" standard
- ▶ Number of disks/heads: 9/18
- ▶ Dimensions(H x W x D):1.63" x 4.0" x 5.75"

Ultrastar 2XP 4.51 GB DISK Drive

- ▶ Rotational speed: 7200 rpm
- ▶ Sustained data transfer rate: 6.5 - 10.0 MB/s
- ▶ Average read seek time: 7.5 ms
- ▶ Buffer(multi segmented): 512 KB
- ▶ Form factor: 3.5" low profile
- ▶ Number of disks/heads: 5/9
- ▶ Dimensions(H x W x D):1.0" x 4.0" x 5.75"



© IBM Corporation 1998

Disk Drive

This foil shows the internal structure and detailed information of the Ultrastar 2XP 9.1 GB or 4.5 GB disk drive.

Ultrastar 2XP disk drive

There are 8 or 16 disk drives installed in one drawer. Two types of disk drives can be installed in the VSS drawer: the 4.5 GB disk drive and the 9.1 GB disk drive. All drives in the same drawer must have the same capacity, either 9.1 GB or 4.5 GB. The characteristics of each disk drive are shown in the above foil.

Buffer

The disk drive has a 512 KB buffer that can improve the I/O performance of the disk drive and VSS. The data to be written or read is stored in this buffer and then physically written to the disk surface or sent to the initiator. The initiator here means the SSA RAID adapter and the target means the disk drive. In the write operation, the data to be written can wait in this buffer for access and seeking. In the sequential read operation, this buffer improves performance by prefetching the data from the contiguous block.

4.5 GB disk or 9.1 GB disk

In view of the I/O performance, two 4.5 GB disks may achieve better performance than one 9.1 GB disk. Because each disk has 512 KB of buffer whatever the disk capacity, the amount of buffer for two 4.5 GB disks is greater than for one 9.1 GB disk. In addition, both disk drives have the same I/O performance per disk drive, so two lower capacity disks may carry out more I/O operations than one higher capacity disk.

In terms of cost, one 9.1 GB disk drive is less expensive than two 4.5 GB disk drives. Also, the maximum capacity that one VSS can be configured is 2.0 TB using 9.1 GB disk drives and 1.0 TB using 4.5 GB disk drives if all RAID arrays are configured as 6+P+S.

Routing function

Each disk drive is recognized as a node by the initiator. The initiator means the pair of ports on the SSA RAID adapter. Each SSA RAID adapter has two initiators because each adapter has four SSA ports. Because there are multiple disks (up to 48) on one loop and the loop path is shared by all nodes, the control command or data may have to pass through the other disks before arriving at the destination disk. Each disk is assigned an individual address on the loop when the configuration or reconfiguration is executed on the VSS storage server.

The SSA RAID adapter sends the data or command with the destination address of the disk on the loop. If a disk receives the data, a command comes as the SSA frame through the SSA cable, and the three-way router checks the address header contained in the SSA frame. If the address is zero, then the three-way router recognizes this frame is sent for it and the data or command is sent to the disk drive. If not, the three-way router sends it to a contiguous node (most often a disk) after decrementing the address value. All initiators or targets on the SSA loop work in the same way as the “router” described and it functions in compliance with the SSA protocol.

For details of the disk drive technology, such as the disk format, see Chapter 3, “Versatile Storage Server Technology” on page 43.

SMART/PFA



Ultrastar 2XP disk drive PFA function

- **GEM monitors:**

- Head fly height
- Cannel noise
- Signal coherence
- signal amplitude
- writing parameters
- and more

Notify to the host if a specified threshold is exceeded

SMART Protocol

Host

- **Symptom driven uses:**

- output of data
- output of non-data
- motor start error recovery logs

SMART=Self-Monitoring Analysis and Report Technology
PFA=Predictive Failure Analysis



© IBM Corporation 1998

SMART/PFA

This foils lists the functions of the predictive failure analysis (PFA) monitors and the self-monitoring analysis and report technology (SMART) protocol.

PFA

PFA monitors key device indicators for change over time or exceeding specified limits. The device notifies the system when an indicator surpasses a predetermined threshold.

PFA is an attractive solution to disk drive maintenance. It can minimize the probability of loss of data access, and at a much lower cost than data redundancy. PFA not only provides a new level of data protection, it allows scheduled replacement of the disk drives.

How the PFA works

There are two types of disk drive failure: on/off failure and gradual performance degradation. The on/off type of failure occurs if a cable breaks, a component fails, or an unpredictable catastrophic event occurs. These types of failures have been reduced recently, but not eliminated. PFA cannot provide any warning for on/off unpredictable failures.

Gradual performance degradation can be predicted using PFA algorithms. PFA is designed to monitor the performance of the disk drive, analyze the data from periodic measurements, and recommend part replacement when a specified threshold is exceeded. The thresholds have been determined by examining the history logs of the disk drives that have failed in actual customer operation.

Monitoring

PFA monitors the performance of the disk drive in two ways: a measurement-driven process, and a symptom-driven process. The measurement-driven process is based on IBM's exclusive generalized error-measurement feature. At periodic intervals, PFA's generalized error measurement (GEM) automatically performs a suite of self-diagnostic tests that measure changes of the characteristics of the disk drives. The GEM circuit monitors the head fly height on all data surfaces, channel noise, signal coherence, signal amplitude, writing parameters, and more. From those measurements and equipment history, PFA then recognizes the specific mechanisms that can cause the disk drive failure.

The symptom-driven process uses the output of the data, nondata, and motor-start error-recovery logs. The analysis of the error logs is performed periodically during idle time. When the PFA detects a measurement that exceeds the threshold, the host system is notified.

SMART

The reports or warnings generated by the activities of the PFA comply with industry-standard self-monitoring analysis and reporting technology (SMART).

Disk Drive Reliability



PFA and SMART

- ▶ Analyses drive parameters
- ▶ Predicts imminent drive failures
- ▶ Idle-time function

Channel calibration

- ▶ Calibrates read and write circuits
- ▶ Idle-time function

Ultrastar 2XP Disk Drive Reliability

Log Recording

- ▶ Saves periodic data on reserved area of the disk
- ▶ Idle-time function
- ▶ Data is used for failure analysis

Disk Sweep

- ▶ Avoids errors cause by low-use of the disk drive components
- ▶ Idle-time function



© IBM Corporation 1998

Disk Drive Reliability

This foil lists the functions performed in Ultrastar 2XP disk drive to measure the disk drive parameters and predict whether component failures are imminent.

PFA and SMART

Predictive failure analysis is an idle-time function that consists of seven measurements taken for each head on the disk. The idle-time functions are the least intrusive way of testing for potential component malfunctions, although productive I/O processing can preempt the idle-time functions. The seven PFA measurements are taken for each head over a 4-hr period, and each takes approximately 80 ms.

The PFA function complies with the SMART industry standard. The PFA functions invoked when the disk drive is idle measure the conditions of each head and save the data to a log in a reserved area of the disk. This log data can be retrieved and analyzed.

Channel calibration

The Ultrastar 2XP disk drive periodically calibrates the disk drive channel to ensure that the read and write circuits function optimally. This reduces the likelihood of soft errors. This function is performed every 4 hr and typically completes within 20 ms. The calibration exercise starts only if the disk drive has been idle for 5 s, and there is no interference with the production I/O processing.

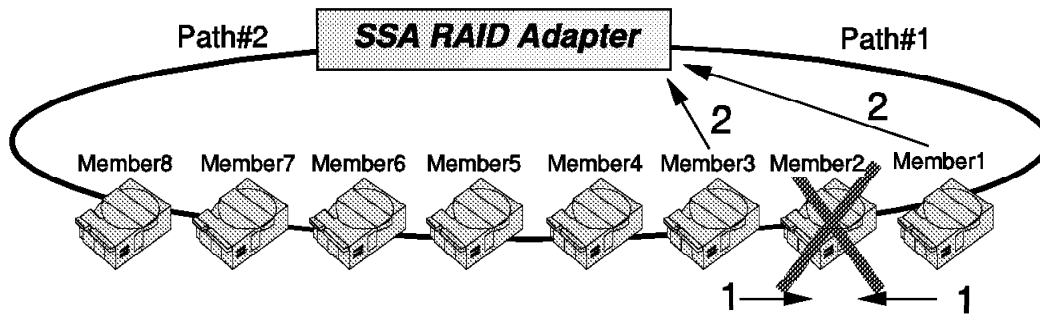
Log recording

The Ultrastar 2XP disk drive periodically saves the data in the logs located in a reserved area on the disks. These reserved areas on the disks are not included in the area used for storing customer data. The log recording is an idle-time function that is invoked only if the disk drive has been idle for 5 s or more. The log recording takes approximately 30 ms, and the logs are saved about every half-hour. The logs are used for failure analysis.

Disk sweep

If the Ultrastar 2XP disk drive has not processed a SCSI command for at least 40 s, the disk drive executes the disk sweep, another of the idle-time functions. The disk sweep exercises the heads by moving them to another area of the disk, and it initiates a second movement of the heads if they fly in the same point for 9 min. This function ensures that low-use disk drive components do not become vulnerable to errors through inactivity.

Configuration



One of the following conditions trigger the SSA loop configuration.

- ▶ SSA loop is broken because of disk failure or cabling problem.
- ▶ Disk drive is removed or added.
- ▶ SSA topology has been changed.
- ▶ By operator instruction



© IBM Corporation 1998

Configuration

This foil shows the conditions that trigger the reconfiguration of the SSA loop. Reconfiguration is required when a component in the SSA loop fails, VSS is installed, or any disks are added to or removed from VSS. Here, we look at the timing of triggering the SSA loop configuration.

Automatic SSA loop configuration

Because the RAID array must consist of eight disks (7+P or 6+P+S) in VSS, if you plan to increase the capacity, the number of disks to be added must be a multiple of eight. You can add the disk drives to the existing SSA loop without shutting power off. In addition, if a disk fails, the failed disk can be removed from the drawer and a repaired or new disk can be added to the drawer without shutting power off. When a disk is removed from the drawer, the adjacent SSA nodes on the SSA loop automatically detect it within an interval. These nodes then alert the SSA RAID adapter. The VSS storage server and SSA RAID adapter check the SSA loop status, including the disk status and the cable connection status. The above procedure can be performed on the SSA loop when

- The SSA loop is broken because of disk failure or a cabling problem.

- A disk drive is removed or added.
- The SSA topology is changed.
- By operator instruction (from the VSS storage server)

We explain an example of loop reconfiguration here.

An example of the loop reconfiguration

In this foil, Member 2 disk fails. The error detection and the SSA loop reconfiguration sequence would be as follows:

1. When Member 2 disk fails, either Member 1 or Member 3 detects the loss of Member 2.
2. Member1 or Member 3 signals the event (loss of Member 2) to the SSA RAID adapter. The SSA RAID adapter and the SSA device driver reconfigure the SSA loop.

Usually Members 1 through 4 send and receive commands and data through SSA Path 1. However, the loss of Member 2 breaks the path between Member 1 and Member 2. By reconfiguration, the SSA loop changes the route as follows:

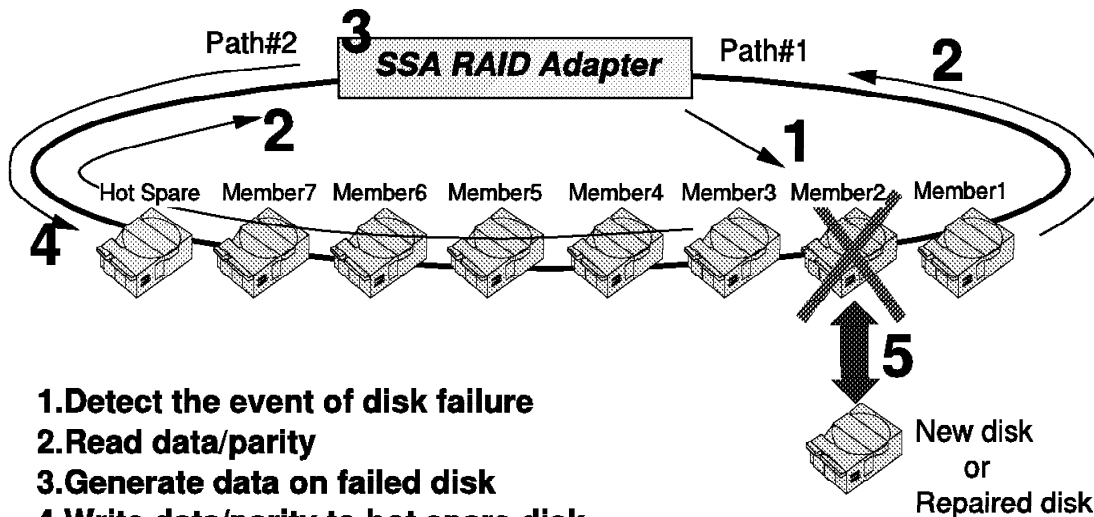
- Member 1 uses SSA Path 1.
- Members 3 through 8 use SSA Path 2.

New disk installation

As mentioned, when you add a new RAID array to VSS, the number of the disk drives to be added should be eight or a multiple of eight. In this case, either a new SSA loop is configured or the disk drives are added to the existing SSA loop. You can add the disks in a way similar to the above.

Thus the event of the single point of failure on the SSA loop is recovered.

Sparing



- 1. Detect the event of disk failure**
- 2. Read data/parity**
- 3. Generate data on failed disk**
- 4. Write data/parity to hot spare disk**
- 5. Replaced new disk can be configured as hot spare disk**



© IBM Corporation 1998

Sparing

This foil shows how the sparing event works in the VSS drawer. In this case, we assume there is at least one hot spare disk in the SSA loop. Sparing is performed as follows:

1. Disk Member 2 in the array fails. The SSA RAID adapter receives a report of the failure of that disk from Disk Member 1 or 3. In response, the adapter begins to reconstruct the data stored on Member 2 to the hot spare disk. (The error detection is performed as in the previous example.)
2. To regenerate the data stored on Member 2, read processing occurs for Member 1 and Members 3 through 7 to regenerate the data stored on the missing disk, Member 2. In the event of disk failure, SSA loop topology must change because the path of the loop is blocked by the failed disk. However, the SSA RAID adapter or target disks have already changed the routing table of the loop, so data on Member 1 is sent through Path 1 and data on other disks is sent through Path 2 in the example above.
3. The data stored on the missing disk is regenerated on the SSA RAID adapter by parity processing. This function is performed on the SSA RAID adapter.

4. The regenerated data is written to the hot spare disk. This data reconstruction continues until all of the data stored on the Member 2 is reconstructed on the hot spare disk.
5. Either during or after the reconstruction of the Member 2 data, you can remove the missing disk and then plug a repaired or new disk into the SSA loop without shutting power off or interrupting the other processing or access to the disk. After the sparing processing is completed, the data in the array resumes the redundancy provided by the RAID-5.

By this process, access to data is allowed when a disk in the RAID array fails. However, the failed disk should be replaced with a repaired or new disk as soon as possible. The repaired or new disk can then become the hot spare disk. The previous hot spare disk is now the data or parity disk replacing Member 2. That is, the hot spare disk becomes the member of the RAID array.

Sparing Procedure



Sparing with Hot Spare Disk

- ▶ Disk failure occur
- ▶ SSA RAID adapter detects the event of disk failure
- ▶ The missing disk mark as "damaged"
- ▶ SSA RAID adapter instructs to reconstruction data on hot spare disk
- ▶ Data reconstruction processing starts

Sparing without Hot Spare Disk

- ▶ Disk failure occur
- ▶ SSA RAID adapter detects the event of disk failure
- ▶ The missing disk mark as "damaged"

▶ Remove missing disk and put repair or new disk to drawer
▶ Instruct to reconstruct data on failed disk onto repair or new disk from operator terminal.

- ▶ Data reconstruction processing starts

Operator Intervention is needed



© IBM Corporation 1998

Sparing Procedure

This foil shows the sparing procedure both when there are hot spares available in the loop and when there are no hot spares.

Sparing with hot spare disk

As already shown in the example, when one disk in the array fails, the SSA RAID adapter or disks detect the failure and report to the SSA RAID adapter. The SSA RAID adapter then begins to reconstruct the data from the failed disk to the hot spare disk, if available. It is done automatically and no operator intervention is required to execute the data reconstruction on the hot spare disk.

Sparing without hot spare disk

If there is no hot spare disk in the SSA loop, the sparing function is still available, although some operator intervention is required to trigger the sparing. Suppose one disk in the RAID array fails:

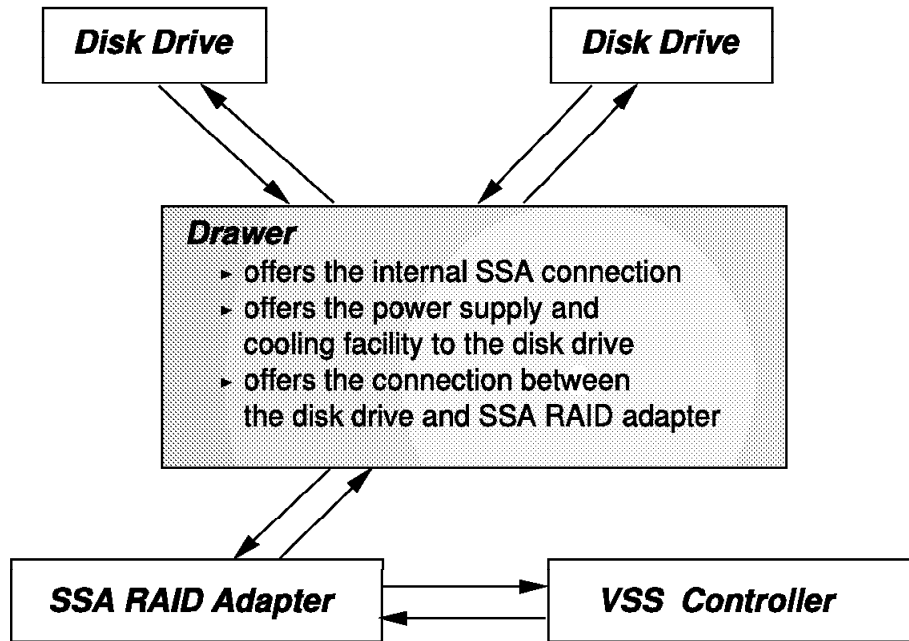
1. The SSA RAID adapter or a disk detects the disk failure.
2. The routing table of the SSA loop is then changed to access all other disks. Requests for data on the failed disk are met by the data regeneration

function performed by the SSA RAID adapter. However, data reconstruction is delayed until a new disk is installed.

3. Remove the failed disk and plug in the repaired or new disk
4. When you replace the failed disk with a functioning one, the SSA RAID adapter detects the replacement and reconfigures the SSA loop.
5. The operator enters an instruction from the VSS terminal that triggers a shift of the sparing function of the missing disk to the replaced disk.

Thus, although there is no available hot spare disk, the sparing function is performed on VSS. However, no further redundancy is provided before the sparing processing is successfully completed. The missing disk should be replaced with a new or repaired one as soon as possible.

Component Interaction



© IBM Corporation 1998

Component Interaction

This foil shows the component interaction around the drawer. The drawer interacts with the components that make up the VSS.

SSA connection

The drawer offers the SSA loop connection between the disk drives. This facility eliminates the cabling between the disk drives installed in the same drawer.

Power supply and cooling fan

The drawer has three power supply and cooling fan modules that supply the electrical power to all disk drives installed in the drawer and cool the disks. Three modules offer enough redundancy to minimize the likelihood of a total loss of power and cooling to the disks.


Interface between the SSA RAID adapter and the disk drive

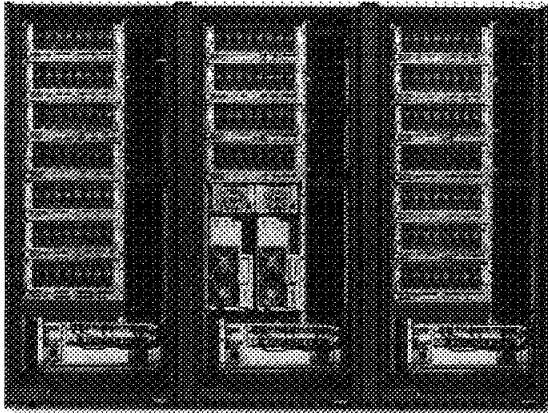
The drawer provides an interface between the SSA RAID adapter and the disk drives by way of the ports. Each drawer has eight ports that constitute the host bypass circuit. The drawer simplifies the SSA cabling and availability of the disk drives.


Chapter 6. Versatile Storage Server Configuration Options

Configuration Options

- Host connectivity
 - supported hosts
 - adapters
- Storage Server
 - four-way
 - cache size
- Disks
 - adapters
 - array size
 - loops
- Racks
- Power supplies
- Configuration
 - management
 - performance and availability


© IBM Corporation 1998




San Jose

© IBM Corporation 1998

Versatile Storage Server Configuration Options

In this chapter we discuss the possible configurations of the VSS components and the factors that influence configuration decisions.

Host connectivity

We discuss supported hosts, adapter options and factors that influence connectivity choices.

Storage server

The storage server includes four-way SMP clusters. Cache size can be a big factor in influencing system performance, the minimum size is 512 MB and the maximum is 6 GB.

Disks

The main factors affecting the overall size of the system are the total amount of data to be stored and the workload characteristics, in particular the I/O rate. These influence the number of disk arrays and how arrays should be presented to the attached hosts in terms of number and size of logical disks.

Racks

In many ways the number of racks is an easy decision; it is based mainly on the number of disk drawers to be installed. Where a customer wishes to incorporate existing racks of 7202 or 7015-R00 SSA disk drawers, there are factors to consider.

Power supplies

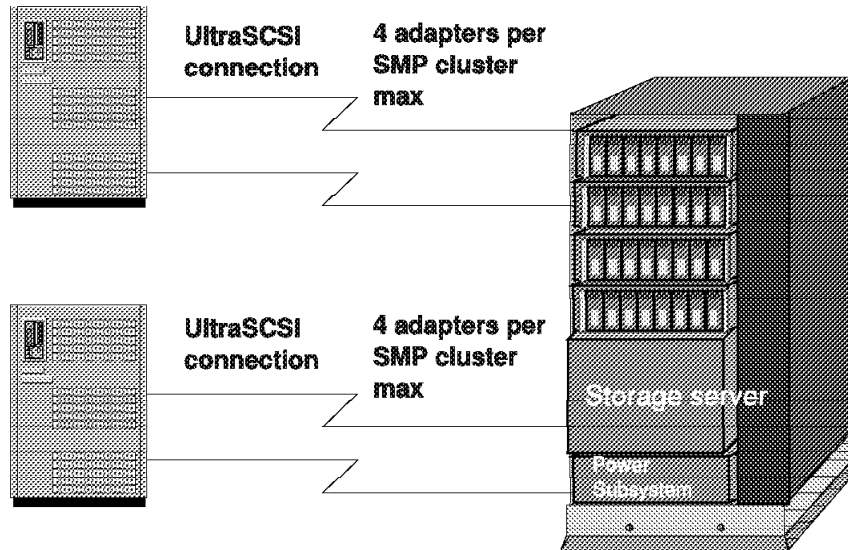
An optional power supply battery backup can be considered.

Configuration

We discuss the configuration methods and factors that influence the configuration such as availability and performance considerations.

It must always be remembered that budget constraints are an influence on system design and configuration. Some compromises in the final system configuration are inevitable. It is essential that the impacts of compromises are clearly understood and are communicated to all concerned.

Host Connectivity



©IBM Corporation 1998

Host Connectivity

Each host interface adapter contains a dual-ported Ultra-SCSI host attachment, which enables up to two hosts to be attached. Two attachments are included in the base machine. Up to six additional host adapters, making eight in total, can be ordered. This means that up to 16 hosts can be attached to the VSS.

Host attachment cables

Cables must be ordered in pairs for each host adapter. Each host adapter provides two host connections (ports). The maximum distance that the host can be from the VSS is 20 meters (20 m).

Up to four homogeneous host servers can be attached to a single Ultra-SCSI attachment. The hosts are linked in sequence (daisy-chained) along the SCSI bus. The number of attachments is dependent on the support available within the host operating system. AIX supports four hosts per bus, HP/UX supports two hosts per bus, but Solaris supports only a single host per bus. You can have as many as 64 hosts attached to a single VSS.

Connectivity Considerations



- Number of hosts
- Volume of data
- Data rate
- I/O rate
- Availability requirements
- Backup requirements



©IBM Corporation 1998

Connectivity Considerations

In this foil we discuss the considerations that affect the choice and number of host connection kits and cables.

Number of hosts

Each host will require at least one connection to the VSS. Each host adapter has to be configured so that it is "owned" by one of the two SMP clusters. This means that the adapter can access only disks that are attached to its cluster. If the host needs to access disks that are attached to the other cluster, then a second host connection to the VSS is required. In practice, in order to balance the workload across the two clusters, in most situations, a host's disks will be spread across the two clusters; therefore, at least two connections per host are required.

Volume of data

Each Ultra-SCSI adapter in a host can support up to 15 targets with 64 LUNs per target. In VSS, we can configure the maximum LUN size to be 32 GB.

Data rate

The maximum data rate that an Ultra-SCSI adapter can support is 40 MB/s and a SCSI-2 adapter can support a rate of 20 MB/s. These are peak rates and realistic operating rates would be 25 MB/s and 15 MB/s respectively.

I/O rate

Very often in the early stages of designing a new system, the actual application data and I/O rates are not known. Some guesstimate based on previous knowledge has to be made. Where assumptions have been made, they should be documented and validated as the project progresses.

Availability

The loss of a host-to-VSS connection means that the host cannot access the data that is normally available by that route. If this loss of data availability would seriously affect the organization's business operations, then an alternative path to the data should be configured using a second adapter on the host and VSS. Availability can be further improved by configuring a second host that uses HACMP-type software to provide protection in the event of primary host failure.

Note Not all host operating systems will support adapter failover.

Backup requirements

With the increasing use of around-the-clock operations, the time available for system backups is decreasing. It is important to consider the demands that system backups make on the storage system when doing system design and configuration. When online backups are carried out, an even greater workload is placed on the system. The backup requirement may change the data rate and I/O rate requirement but is less likely to affect the storage volume unless backups are going to be made from a mirrored copy of the data.

Summary

There are four factors that influence the number of host-to-VSS connections required:

- Volume of data
- Data rate
- I/O rate
- Availability requirement

Normally one of these factors will be the bottleneck and determine how many connections are required.

Host Support



- Hewlett Packard
 - HP 9000 800, D,E,G,H,I, K, T Series, EPS
 - HP/UX 10.01, 10.10, 10.20, 10.30
- Sun Microsystems
 - Sun Ultra 1000, 1000E, 2000, 2000E, 3000, 4000, 5000, 6000
 - Sun Solaris 2.5.1, 2.6
- Compaq
 - ProLiant 3000, 5000, 5500, 6500 and 7000
 - *NT 4.0*
- DG
 - AViON 4900 and 5000
 - *DG/UX 4.2*
- IBM
 - RS/6000 (both servers and SP)
 - *AIX 4.1.5, 4.2.1, 4.3 +*
 - AS/400 (9406)
 - *OS/400 V3R1, V3R2, V3R6, V3R7, V4R1 and V4R2*
 - NetFinity PC (325, 704, 3500, 5500 and 7000)
 - *NT 4.0*



© IBM Corporation 1997

Host Support

VSS includes the necessary support to operate in a UNIX, AS/400 and some PC environments.

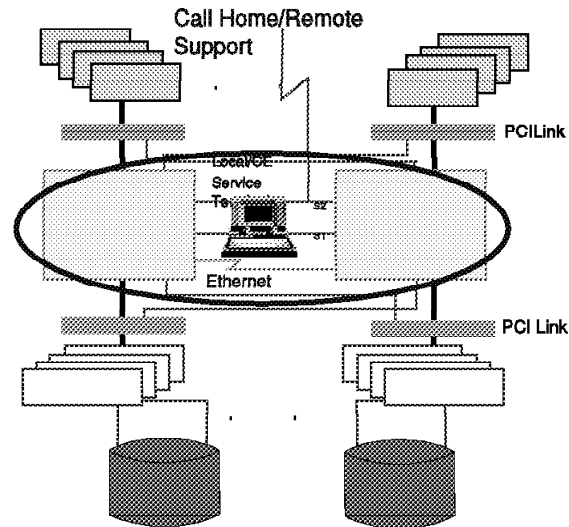
Advanced software

VSS provides the capability to share data as part of the base system. No special features are required to utilize data sharing. Data sharing is under the control of application code to regulate data access and prevent corruption.

SMP Cluster Options



- Processors
 - Four-way SMP high-performance (four-way on each side)
- Read cache
 - 512 MB minimum
 - 6 GB maximum
- Standard features
 - RS-232 port
 - Ethernet port
 - CD-ROM, diskette drive, and internal hard disk



©IBM Corporation 1998

SMP Cluster Options

As standard, each cluster has two SMP processor boards, each of which contains two 604e CPUs. This means that the VSS is configured with 4 four-way engines in each cluster.

VSS as standard comes with 512 MB of cache. This is split across both of the SMP clusters. The cache can be upgraded from 512 MB to a maximum of 6 GB (3 GB per cluster).

Processors

The type of processors required in each cluster is a function of the I/O rate that the cluster must process. I/O rate has more effect on the processors than the data rate.

Read cache size

The size of the cache will depend on the nature of the workload, whether it is sequential or random, and the expected cache hit ratio. See Chapter 8, "Versatile Storage Server Performance" on page 251. The minimum read cache size is 512 MB and the maximum is 6 GB.

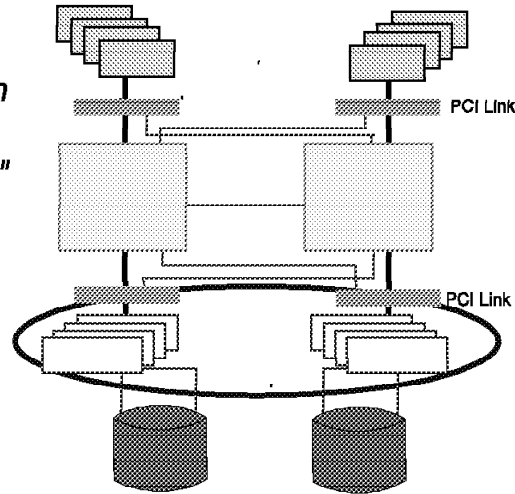
Standard features

As standard, each cluster has RS-232 ports for service support, an Ethernet connection for system configuration, and a CD-ROM and diskette drive for microcode loading.

Disk Adapters



- Number of adapters
 - Two to eight adapters
 - max of two loops per adapter
 - Fast write cache 8 MB
 - ▶ 4 MB used as write through cache
 - ▶ 4 MB used as "permastore"
- Guidelines
 - minimize cost
 - ▶ two loops per adapter
 - ▶ two arrays per drawer
 - maximize write cache
 - ▶ one loop per adapter
 - ▶ one array per drawer



©IBM Corporation 1998

Disk Adapters

The number of SSA disk adapters must be specified. Each adapter will support two SSA loops. Adapters are "owned" by the SMP cluster to which they are attached. You cannot configure the adapter to be "owned" or accessed by the other cluster. The only exception is when one cluster fails and the remaining storage server accesses all the disk adapters.

Number of adapters

As standard, the VSS comes with one SSA disk adapter. The maximum number of adapters that can be installed is eight. The number of adapters required is a function of the volume of disk storage.

Guidelines

There is one SSA loop per disk drawer, and each drawer will house either one or two RAID arrays of eight SSA disks. Unless the VSS has more than 16 disk drawers, the standard method of connecting adapters to disk drawers is to have one SSA loop per disk drawer. Each disk adapter will therefore be supporting up to two disk drawers. If more than 16 disk drawers are required, then there will be two adapters supporting two disk drawers (that is, 32 disks).

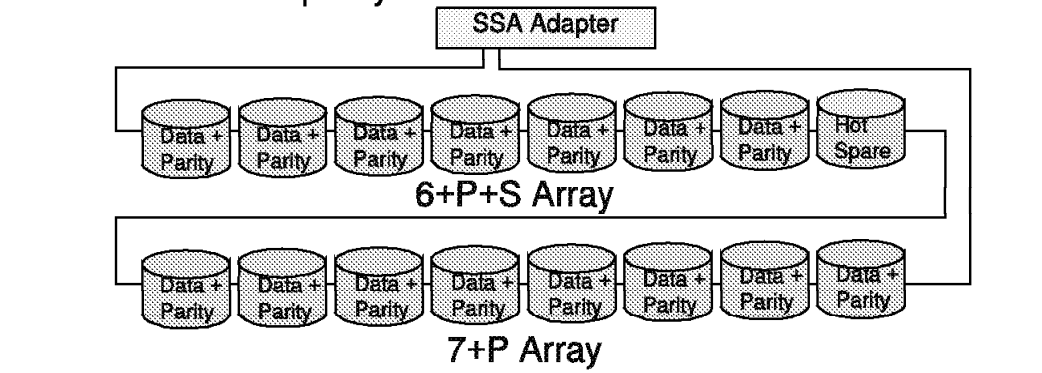
Configuration rules are based on the need for maximum performance. The base configuration includes two SSA adapters and two SSA disk drawers (32 disk drives). As disk drawers (7133s) are added to the configuration, additional SSA

adapters are added. For example, having a total of eight disk drawers (regardless of the number of disk drives in the drawers) requires eight SSA adapters. When Drawers 9 through 16 are added, they are connected to the second loop on each adapter. Thus, 16 drawers are configured as 16 loops (maximum of 16 disk drives per drawer), with each SSA adapter supporting two loops. If Drawer 17 or 18 is added, they are added to existing loops. These two loops support up to 32 drawers.

RAID Arrays



- 6+P+S
 - One per loop, first array in the loop
 - Spare can float across all arrays on the loop
- 7+P
 - All drives in array contain data and parity. Maximizes useable capacity



©IBM Corporation 1998

RAID Arrays

There are two options as to how to configure RAID arrays:

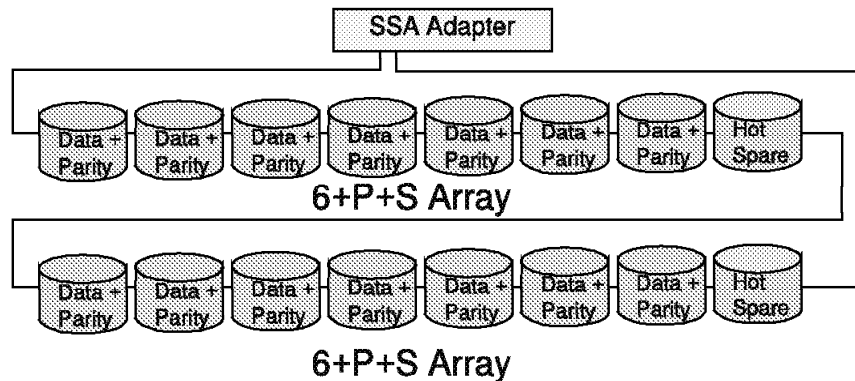
1. To have a 6+P+S configuration
2. To have a 7+P configuration

The first array on a loop must be of the 6+P+S type, that is six data disks with one parity drive and one spare. The second array can be of either type. To maximize the available storage, configure an array of each type on the same loop. This configuration gives a total of 13 data drives with two parity drives and one spare. All drives in a loop must have the same capacity, either 4.5 GB or 9.1 GB.

Maximum Availability



- Two 6+P+S RAID arrays
 - One per loop, both arrays in the loop
 - Two spares can float across all arrays on the loop
 - Maximizes availability if two disks should fail



©IBM Corporation 1998

Maximum Availability

To maximize availability in the unlikely event that two drives in a loop fail together, then configure two arrays of the first type. That is, both arrays would be of the 6+P+S, giving 12 data drives with two parity drives and two hot spares. All drives in a loop must have the same capacity, either 4.5 GB or 9.1 GB.

Disk Sizing

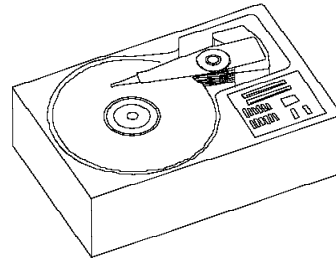


Two capacities available

- 4.5 GB
- 9.1 GB

State-of-the-art technology

- PRML
- 3rd generation MR head technology



High performance

- 7200 RPM
- 6.9 to 7.8ms average access
- up to 15.4 MB/s media transfer rate



©IBM Corporation 1998

Disk Sizing

There are two possible disk sizes that can be selected, 4.5 GB or 9.1 GB. The 9.1 GB drives will give the largest capacity in the VSS and, for a given volume of storage, will be cheapest as fewer disk drawers and expansion racks will be needed. In storage devices, smaller disk drives offer the customer more disk actuators per megabyte of storage, giving better performance for applications with high I/O rates. In VSS, however, the use of read and write caches means that for many applications the data will be read from or written to the cache. The use of high-speed cache in front of the high-performance SSA disk subsystem and high-performance IBM hard drives means that the 9.1 GB drives are the best option for most applications. For more details on this, see Chapter 8, "Versatile Storage Server Performance" on page 251.

State-of-the-art technology

IBM disk drives are acknowledged as leaders in the field of disk-drive technology. They have many advanced features, such as PRML and third-generation MR head technology (see Chapter 3, "Versatile Storage Server Technology" on page 43).

High performance

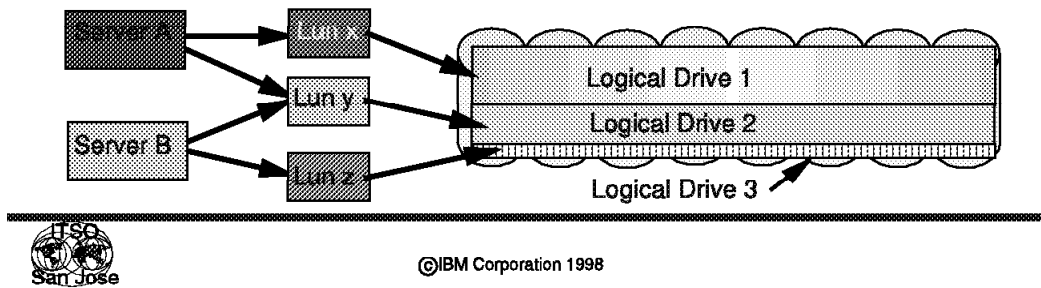
IBM disk drives are leaders in high performance. They offer the customer fast access speeds and high data rates (see Chapter 3, “Versatile Storage Server Technology” on page 43).

Logical Volume Allocation



- Server
 - VSS presents logical view of storage to the server.
 - Variable LUN size
 - Sharing of LUNs

- VSS
 - Controller manages the logical/physical relationship
 - Data is striped across the RAID-5 array



Logical Volume Allocation

The application servers attached to VSS do not “see” the actual disk drives or RAID arrays but a logical representation or virtual disk.

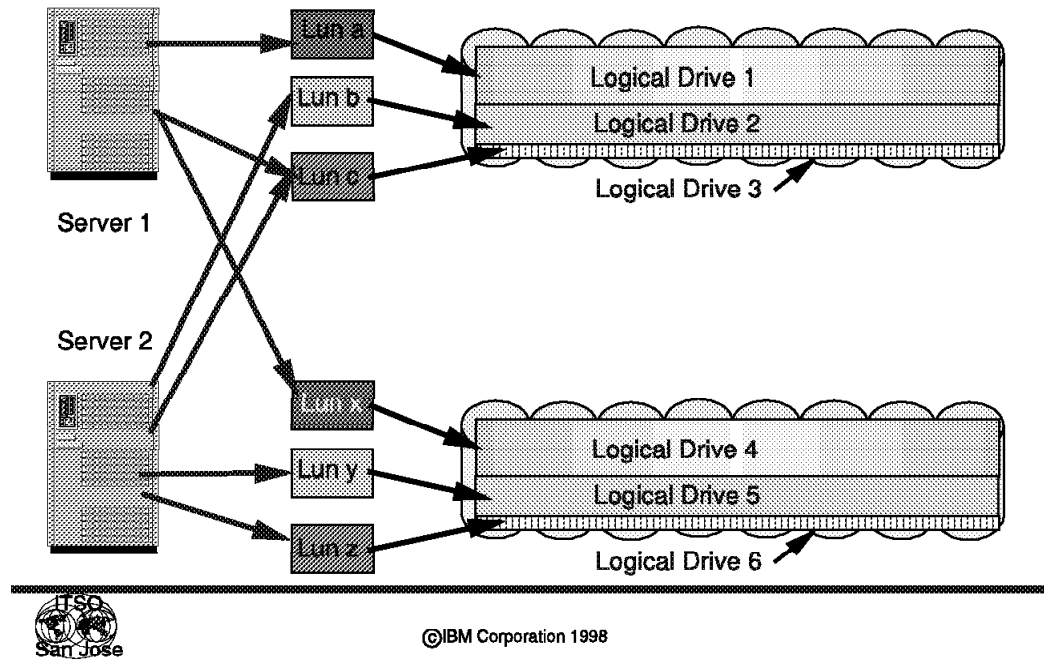
Server

A separate storage pool can be configured for each server attached to the VSS. The number of LUNs and their size will depend on two main factors: first, the total volume of storage required by the server and, second, the characteristics of the workload. For example, if the work involves the use of large sequential I/O, then a small number of large LUNs is appropriate, while for random workload with very high I/O rates, having more small LUNs spread over several disk arrays is more appropriate, see Chapter 8, “Versatile Storage Server Performance” on page 251. Note that a single LUN cannot span more than one disk array.

VSS

The configuration of the actual disk arrays into a logical representation is done using the IBM StorWatch Versatile Storage Specialist (VS Specialist), a web-browser-based management tool.

Multiple Access



Multiple access

When configuring logical disks and deciding the host allocation, account must be taken of the workloads of all the hosts. Performance will be reduced, for example, if all of the heavily used virtual disks from multiple hosts are assigned to the same physical RAID array. In the example shown, if LUNa attached to Server 1 and LUNb attached to Server 2 are both heavily used, then the I/O performance of both servers will suffer, as logical drive 1 and logical drive 2 are both on the same physical RAID array. It would be better to configure LUNb to logical drive 5, which is on a different physical RAID array.

2105-B09 Storage Server Rack



Storage Server

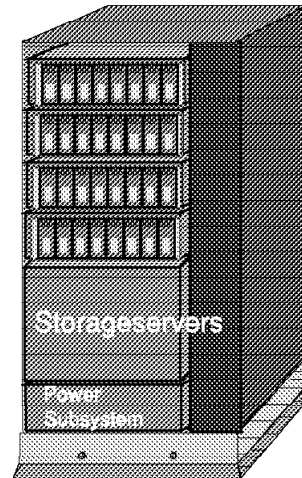
- Four-way as standard
- 512 MB cache
- Two host adapters
- Two SSA disk adapters

Power supply

- Single or three phase
- Battery backup option

Disk storage

- Two RAID arrays both 6+P+S
- Internal SSA cabling



©IBM Corporation 1998

2105-B09 Storage Server Rack

VSS can be configured as a single rack or a multiple rack complex. The 2105-B09 (storage server rack) is always the first rack and includes the electronics complex, disk drawers, and a fully redundant power control subsystem. The 2105-B09 consists of a 1.8 m rack and the following standard features:

- Dual SMP clusters with 4 four-way SMP processors.
- 512 MB storage server memory, which can be upgraded to 6 GB
- Two Ultra-SCSI host adapter cards
- Two dual-loop SSA RAID-5 disk adapters
- 8 MB of Fast Write cache (located on dual-loop SSA adapters)
- 230 GB of RAID-5 protected storage (32 9.1 GB drives, arranged in four RAID-5 arrays). Two additional 7133s can be placed in the rack.
- Redundant power control subsystem (single phase or three phase power) with two AC power cords.

Disk storage

Included in the base 2105-B09 rack are four RAID-5 arrays, (32 9.1 GB drives). All VSSs are RAID-5 protected. The first RAID array in an SSA loop must include a hot spare drive. All drives in a loop must have the same capacity, either 4.5 GB or 9.1 GB. There is room in the 2105-B09 rack for two additional 7133 drawers. Each drawer can support up to two RAID-5 arrays. New or existing 7133s (Model 10 or 20) can be placed in the rack. Use 7133 Feature 2105 when

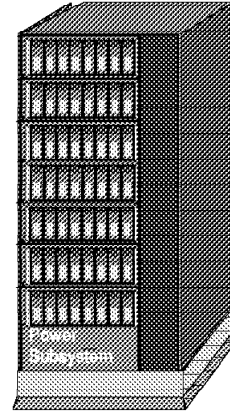
ordering new 7133s to be placed in new VSS racks (2105-B09 or 2105-100). Use 7133 Feature 2106 when ordering new 7133s to be placed in existing VSS racks (2105-B09 or 2105-100).

2105-100 Expansion Rack



- **Power supply**
 - single or three phase
 - battery backup option

- **Disk Drawers**
 - none as standard
 - space for seven drawers
 - new or existing 7133



©IBM Corporation 1998

2105-100 expansion rack

The 2105-100 expansion rack includes space for up to seven disk drawers and always comes with a fully redundant power control subsystem.

Power supply

The power supply and control system is fully redundant. Power is supplied through two power cords with either 50 ampere (single phase) or 50/60 ampere (three phase) connectors with each cord capable of providing 100% of the power requirements. The type of power supply must be specified. The power unit in the 2105-100 is controlled by the main power control unit in the 2105-B09 storage server rack.

An optional battery backup system is also available. The battery is designed to assist with system shutdown in the event of catastrophic power failure. It will also provide power during a temporary loss such as a brownout. If the optional battery is installed in the 2105-B09 rack, then we recommend that it also be installed in all expansion racks.

Disk drawers

Each 2105-100 can house up to seven 7133-020 drawers. All VSS storage is RAID-5 protected. The 7133 can accommodate two RAID-5 arrays of SSA disks. The first RAID array in an SSA loop must include a hot spare drive. All drives in a loop must have the same capacity, either 4.5 GB or 9.1 GB. There is room in the 2105-100 rack for seven 7133 drawers. New 7133-020 drawers must be ordered with Feature 2105 as part of the order. A 7133-020 with Feature 2105 is not shipped with power cords, as power cords for the 7133s are in VSS racks.

SSA cables

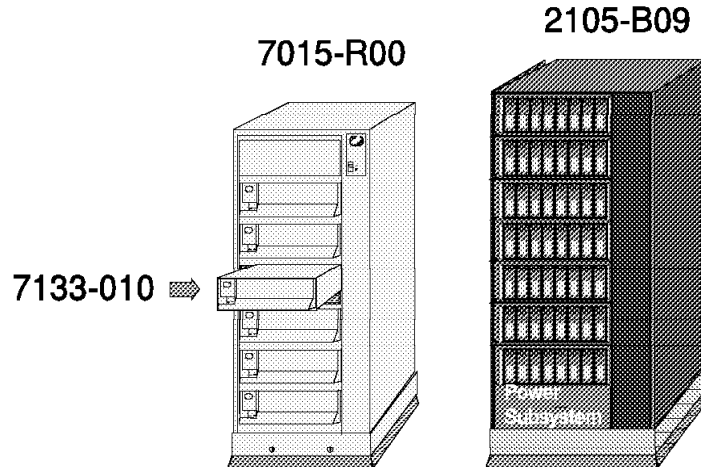
7133 drawers shipped with the 2105 racks (using 7133 Feature 2105) do not require SSA cables. These cables are provided with the rack. If additional new 7133s (using 7133 Feature 2106) are ordered, then SSA cables need to be ordered with the 7133. Two 5 m cables are standard. If longer cables are required, please order the 10 m or 25 m lengths.

If existing 7133s are placed in the 2105-1000 rack, you may need new cables, depending on the length. Cables of 5 m or longer are required.

Investment Protection



7015-R00 and 2105-100



©IBM Corporation 1998

Investment Protection

Existing 7133-010 and 7133-020 drawers can be used. They can be installed in the 2105-B09 storage server rack or the 2105-100 expansion rack. Alternatively, an existing 7133 can be integrated in its existing 7105-R00 rack.

Note If existing racks are used, the power control is not integrated into the power storage server in the 2015-B09 control rack.

In either case, the disk drives must be reformatted to 524 byte sectors before they can be used in VSS.

Maximum Configuration



2105-B09

- 2 high-performance 4-way SMP controllers
- 6 GB read/write cache
- 8 UltraSCSI adapters
- 8 disk adapters
- 4 disk drawers
- 9.1GB disk drives

2105-100

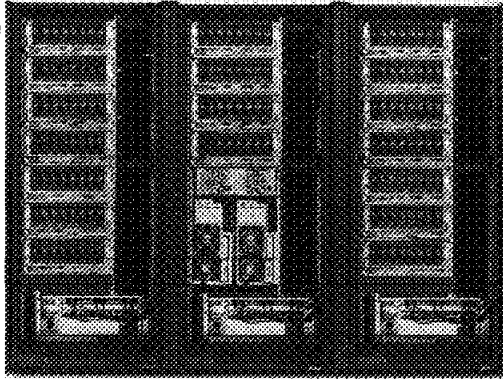
- 7 disk drawers
- 9.1GB disk drives

2105-100

- 7 disk drawers
- 9.1GB disk drives

Total useable storage

- 2 TB protected storage



© IBM Corporation 1998

Versatile Storage Server Maximum Configuration

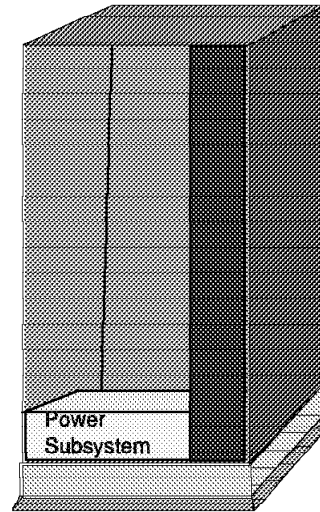
Up to two 2105-100 expansion racks can be attached to a 2105-B09. The configuration considerations for the second 2105-100 expansion rack are the same as for the first expansion rack. The maximum configuration has:

- Two high-performance four-way SMP storage servers
- 6 GB of read and write cache
- Eight Ultra-SCSI host adapters with potential to connect to up to 16 hosts (64 host connections are possible with daisy-chaining, four hosts per SCSI bus)
- Eight SSA disk adapters
- Eighteen drawers of SSA disks, each drawer containing 16 SSA disks configured as two RAID-5 arrays.
- 2 TB of protected storage

Power Options



- Redundant Power
 - Dual line cords
 - Provides 350V DC to all components
- Disk drawer power cords are provided
- Optional battery - powers entire rack
- Several minutes of run time
 - Provides power for short term outage
- Access from front and rear for service
- Service required and power indicators



©IBM Corporation 1998

Power Options

In this foil we discuss the power options available in the VSS.

Redundant power

All power supplies are fully redundant. Power is supplied through two power cords with either 50 ampere (single phase) or 50/60 ampere (three phase) connectors with each cord capable of providing 100% of the power requirements. This flexibility allows VSS to be installed into existing environments with a minimum of disruption. The type of power supply must be specified.

Disk drawer power cords

The disk drawer power cords are supplied as an integral feature of the racks.

Optional battery

An optional battery backup system is also available. The battery is designed to assist with system shutdown in the event of catastrophic power failure. It will also provide power during a temporary loss, such as a brownout. If the optional battery is installed, we recommend that it be installed in all of the VSS racks. The battery can provide power for several minutes.

Access

Access from the front and rear of the rack is required, to work on the power supplies.

Service indicators

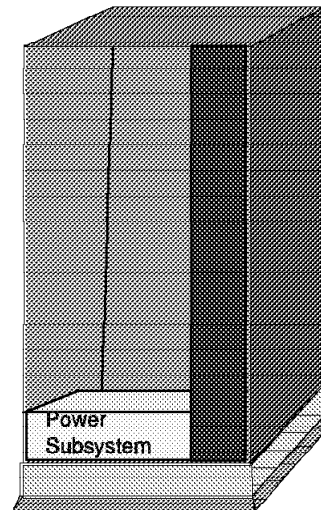
There are indicators for when service is required and for power availability.

Note If existing 7202, 7014-S00 and 7015-R00 racks are used, they are not integrated into the 2015-B09 power control unit.

Versatile Storage Server Enclosure



- Physical Characteristics
 - 36 EIA units
 - Height: 70 inches (1.8 meter)
 - Width: 33 inches (0.8 meter)
 - Depth: 51 inches (1.3 meter)
 - Footprint: 11.67 square feet (1.038 square meter)
 - ▶ 2 TB (3 racks) provides 56.5 GB per sq. ft. (5 GB per sq. m)
 - Weight: about 1600 pounds (fully loaded)



©IBM Corporation 1998

VSS Enclosure

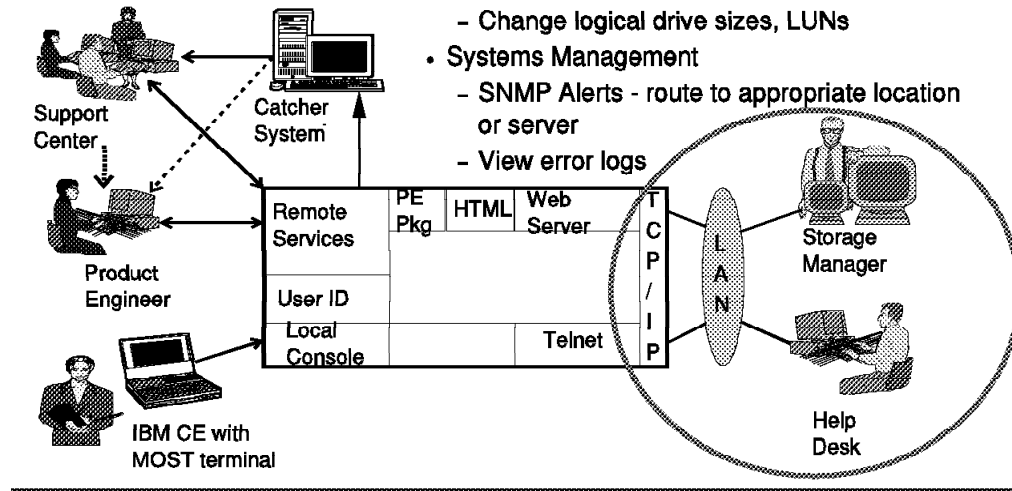
The physical dimensions of the VSS are:

- Height 70 in. (1.8 m)
- 36 EIA units
- Width 33 in.(0.8 m)
- Depth 51 in.(1.3 m)
- Footprint 11.67 sq ft (1.036 m²)
- 2 TB in three racks provides 56.5 GB per sq ft
- Weight 1640 lb (738 kg) for the 2105-B09, with each rack fully configured.
- Weight 1696 lb (753 kg) for the 2105-100, with each rack fully configured.

Configuration Management



- Intranet access
 - Connect via ethernet, use router for TR access
- Web interface available at designated locations
 - Configuration Manager
 - Add, remove logical drives, physical arrays
 - Change logical drive sizes, LUNs
 - Systems Management
 - SNMP Alerts - route to appropriate location or server
 - View error logs



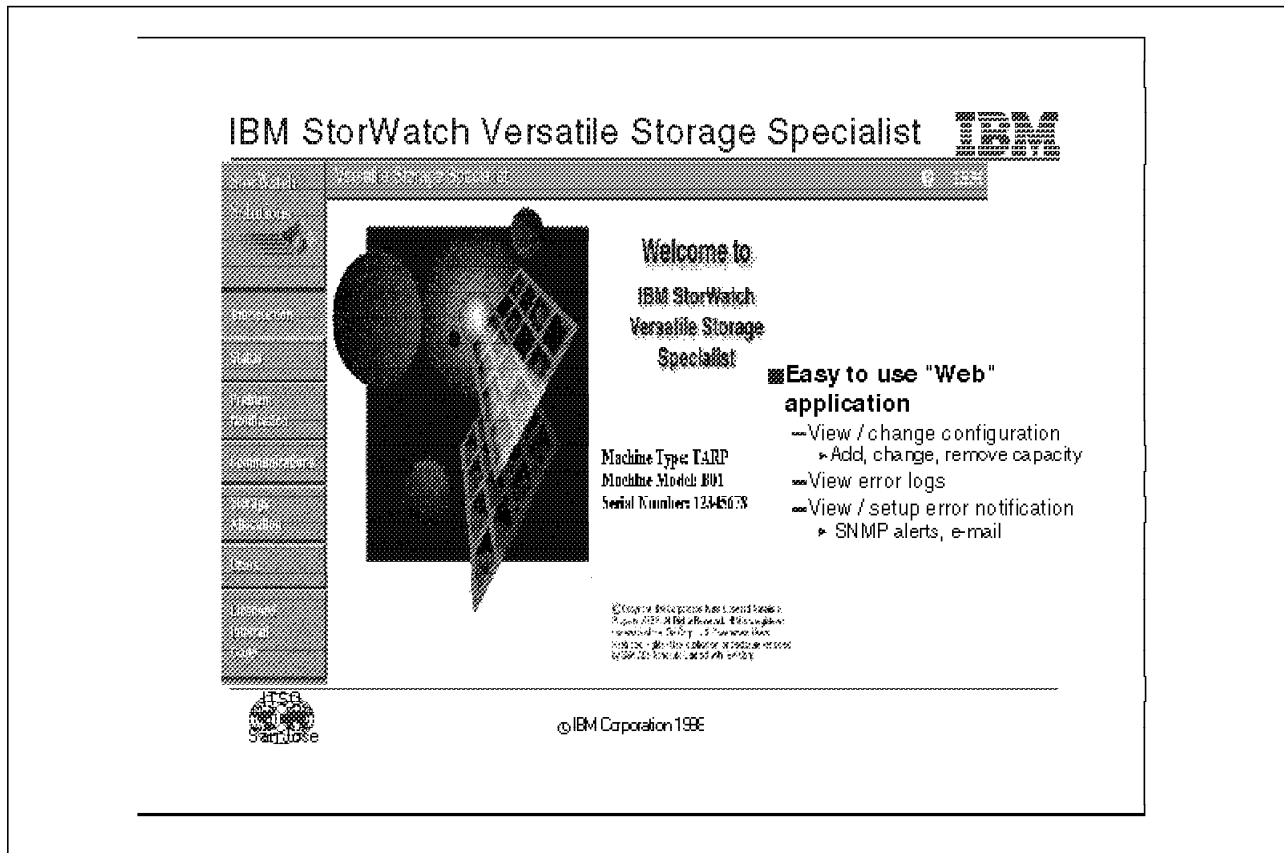
©IBM Corporation 1998

Configuration management

Both the user and service technician can configure the VSS, assigning storage to individual hosts. The configuration manager can be accessed from the subsystem interface.

Intranet access

Configuration access is usually through a customer-controlled network, or intranet. Using an existing customer intranet allows the user to manage the complex from his or her usual place of work and usual desktop. The connection on VSS is Ethernet; if token ring access is required, it must then be made via a router.



Web interface

The configuration manager is a web based GUI application called the IBM StorWatch Versatile Storage Specialist (VS Specialist), providing an easy to use interface into the VSS. Using this interface the storage manager add and/or remove physical disks and arrays and see the status of the disks in the system. A typical management screen is shown. The area on the left shows the typical tasks that the system administrator can perform. The host's view of virtual disks can be changed, the access path to the virtual disks and the size can be altered from 0.5 GB to 32 GB.

Systems management tasks can also be carried out using this interface. SNMP alerts can be routed to the appropriate location or server. Error logs can be viewed and appropriate remedial action taken. The microcode level can be viewed and new levels applied.

Configuration Considerations



- Number of hosts
- Volume of data
- Data rate
- I/O rate
- Availability



©IBM Corporation 1998

Configuration Considerations

There are many factors to consider when configuring the VSS.

Number of hosts

This is the number of hosts that will be attached to the VSS.

Volume of data

This is usually a relatively easy figure to obtain. The ease with which VSS can be upgraded online means that if some of the initial sizing assumptions are wrong, it is relatively easy to rectify them later.

Data rate

The data rate affects two areas of the subsystem:

1. Host to VSS connection
2. Disk to VSS connection

Care must be taken when designing a system that the overall configuration is balanced and that all the parts of the configuration are synchronized. The host to VSS connection is affected only by the data rates of the applications running on the individual host. The disk-to- VSS connection data rate is the sum of the data rates of all the LUNs that are configured on that adapter. (See Chapter 8, "Versatile Storage Server Performance" on page 251.

I/O rate.)

The effect that the I/O rate has on system configuration is similar to that of the data rate: that is, two areas are affected. The first is the host-to-VSS connection and the second is the disk-to-VSS connection. One additional factor that must be considered is that each logical I/O from the host to VSS will, in the case of write operations, result in four actual I/Os. Although the performance impact of the RAID-5 write penalty is masked from the application by the use of a write cache, the effect on the disk activity must be considered (see Chapter 8, “Versatile Storage Server Performance” on page 251).

Availability

VSS has been designed to have zero exposure to loss of data integrity and so all maintenance can be carried out online and without need to power down attached host systems. There can be failures when the host cannot access its data. In such cases, consideration should, if the host operating system supports adapter failover, be given to providing alternative paths to the same LUN. If this is not the case, then disk mirroring should be considered. Mirroring across LUNs on different disk adapters should be considered if disk adapter failure seems likely.

Chapter 7. Migrating Data to the Versatile Storage Server

Migration

Migration



© IBM Corporation 1998

Overview



Overview

A large, empty rectangular box with a thick black border is centered on the page. Inside the box, the word "Overview" is written in a large, black, sans-serif font.

© IBM Corporation 1998

Overview

In this chapter we discuss migrating data to the VSS storage subsystem. We discuss issues facing system administrators needing to migrate data from existing storage subsystems, including supported servers, software prerequisites, how the host system views a VSS subsystem, the various methods of transferring data, the impact on availability and performance, and the use of existing 7133 SSA disk drawers in a new VSS subsystem.

Issues



- Connection
 - Hosts
 - Power
 - Network
- Replacement for existing storage ?
- Reformat of existing drives
- Method of transfer
 - Volume management software
 - Direct copy with cpio -p
 - AIX backup and restore
 - Unix dump and restore
 - Impact
- Use of existing 7133's and drives
 - Order features
 - Restrictions on drive sizes and numbers
- One or two spare drives per loop ?

Sector format

AS/400 Header	Data	S E Q #	L R C
8	512	2	2



© IBM Corporation 1998

Issues

Typically, the VSS will be used in large homogeneous and heterogeneous environments where there are large servers that need access to centralized shared data storage subsystems. There are a number of issues that must be taken into account when installing and implementing a VSS subsystem in this environment.

Connection

As with any storage subsystem, the VSS subsystem must be physically connected to any host that wishes to use it to store data. While making a physical connection is typically a quick and easy operation, it requires that the host system be shut down and powered off to plug the cable into the host. If the VSS is to be accessed by a number of different hosts, connecting to each host should be coordinated in order to minimize the necessary downtime.

The VSS cabinets, along with the 2105-B09 and 2105-100 racks, have dual power line cords to facilitate their high availability. Customers should ensure that they have enough power outlets to accommodate the number of racks being installed.

Configuration of the VSS subsystem is performed using a TCP/IP-based HTML browser network client or through the CE interface. Using the network client requires that an intranet connection be available for connection to each VSS 2105-B09 cabinet. Intranet connection is through 10baseT twisted-pair.

Replacement for existing storage

Is the VSS being installed as a replacement for existing storage? If so, either the VSS must be partitioned so that its virtual disks are similar in configuration to the drives it is replacing, or new configurations must be large enough to accommodate the existing data.

Reformat of existing drives

Although existing disk drives can be used in the VSS, they cannot simply be plugged in to the VSS. The track format is different from that of standard SSA disks used in AIX RS/6000, HP, SUN, or other UNIX systems. The AIX Logical Volume Manager (LVM) uses a fixed-byte sector of 512 bytes, whereas when used in a VSS subsystem, the format is a fixed-byte sector of 524 bytes. The data portion of the sector remains at 512 bytes. An 8 additional bytes are used for AS/400 headers; a 2-byte sequence number and a 2-byte longitudinal redundancy check (LRC) complete the extra bytes. The sequence number and LRC are used by the VSS microcode and are not transferred to the host system.

Method of transferring data

No special tools or methods are required for moving data to or from the VSS. That is, no extra software is required. Migration or movement of data is done at the host operating system level, as the host sees the VSS as one or more physical disk drives. For UNIX hosts, there are a number of varied methods of copying or moving data from one disk drive to another. These include (but are not limited to):

- Volume management software — AIX, Solaris and HP-UX all have volume management software that directly controls the disks and storage devices attached to the system. It provides the system administrator with tools for configuring and managing disk drives, file systems, paging and swap spaces, as well as providing the operating system interface to the data.

Volume management software provides specific tools for wholesale movement of data. Beginning with “AIX Logical Volume Manager” on page 235 we discuss some methods and provide examples.

- Direct copy with **cpio -p** — The **cpio** command is a standard UNIX command for archiving and copying data. The **-p** (pass) option allows data to be copied between file systems without the creation of an intermediate archive. The **cpio** command reads a list of files to copy from its standard input. It is typically used with the UNIX **find** command to copy complete directory structures.
- AIX **backup** and **restore** commands — These commands are commonly used on AIX systems to archive and restore data. However, they do not support a direct disk-to-disk copy facility, requiring an intermediate device such as a tape drive or spare disk drive to store the archive created by the **backup** command.
- **dump** and **restore** commands — These commands are similar in function to the AIX **backup** and **restore** commands, and are found on most other forms of UNIX. They too require an intermediate device.
- Others — There are other UNIX commands that provide archival facilities, such as the **tar** command, which can be used to transport data. Again, these commands require that an intermediate archive be created, usually on a tape drive or spare disk device.

Typically, the method chosen is the best compromise between efficiency and least impact on the users of the system. Using volume management software is typically the first choice as it provides simple robust methods that can generally be used during production without disturbing users. The AIX LVM provides methods which can be used at any time without disrupting access to the data by users. A small performance degradation may be noticed, but this is better than having to shut down databases or require users to log off the system. Methods using backup and restore procedures will generally have the most impact on the system usage, requiring databases and file systems to be in quiescent states to ensure a valid snapshot of the data. Ultimately, the system administrator will choose the method that offers the best compromise.

Migration of existing 7133s and drives

If existing 7133 and SSA disk drives are to be used in the VSS, there are features that have to be ordered in order for the existing 7133s to be mounted in the VSS racks. Also, there are restrictions on disk sizes, numbers, and spares that need to be taken in to account. See “Use of Existing 7133 Drawers with Versatile Storage Server” on page 246 for more information.

One or two spare drives per loop

Before data can be placed on the VSS, the physical drives in each array must be configured. Part of that configuration process is the designation of spare drives. The spare drive is required for reconstruction of data should one drive in the array fail. At least one spare drive is required per loop and, in most applications, one will suffice. However, if availability is more important than overall capacity, it may be valuable to configure two spare drives in a single loop (one per array). For more information regarding spare drive configuration, see Chapter 6, “Versatile Storage Server Configuration Options” on page 191.

Supported Servers and Prerequisites



- IBM RS/6000
 - All models including SP
 - ▶ *Require AIX 4.1.5, 4.2.1, 4.3 and higher*
 - ▶ *VSS will ship with installation package which will contain VPD data to populate ODM*
- IBM AS/400
 - Models 9406 and Advanced Series 300, 310, 320
 - ▶ *Require OS/400 release V3R1 or V3R2*
 - Advanced Series 500, 510, 530
 - ▶ *Require OS/400 release V3R6, V3R7, V4R1 or V4R2*
 - e Server S20, S30, S40, 620, 640, 650
 - ▶ *requires OS/400 release V4R1, or V4R2*



© IBM Corporation 1998

Supported Servers and Prerequisites

The VSS is supported on various platforms from IBM, Sun Microsystems, and Hewlett-Packard. The foil discusses the supporting systems and required operating system levels for each of the supporting systems.

All host platforms must have, at the very least, a differential SCSI-2 fast and wide adapter, although an Ultra SCSI adapter is the preferred host adapter, in order to take advantage of the extra capabilities of Ultra SCSI in the VSS.

IBM RS/6000

All IBM RS/6000 servers support VSS, including the SP.

Operating system levels required to support the VSS are 4.1.5, 4.2.1, 4.3 and 4.3.1. The VSS ships with a small installation package, similar to that of other external IBM storage subsystems; it contains the vital product data (VPD) necessary to populate the AIX ODM database. The VPD allows AIX to see the VSS as a VSS, not as a “generic SCSI disk.”

From the next release of AIX, the VPD for the VSS will be included in the base releases.

IBM AS/400

The AS/400 model 9406, Advanced Series 300, 310, 320 systems and the e Series S20, S30, S40, 620, 640, and 650 support the VSS. These systems require OS/400 release V3R1 or V3R2. AS/400 Advanced Series 500, 510 and 530 systems support the VSS. These systems require OS/400 release V3R6 or V3R7. The AS/400e Series require OS/400 V4R1, or V4R2.

Supported Servers and Prerequisites ...



- IBM PC
 - 325, 704, and Netfinity 3500, 5500, 7000
 - ▶ *NT 4.0*
- Sun Microsystems
 - Ultra servers 1000, 1000E, 2000, 2000E, 3000, 4000, 5000, 6000
 - ▶ *Solaris 2.5.1, 2.6*
- Hewlett Packard
 - 9000 series 800, D, E, G, H, I, K and T series and EPS series.
 - ▶ *HP-UX releases 10.01, 10.10, 10.20, 10.30*
- Compaq
 - ProLiant 3000, 5000, 5500, 6500, 7000
 - ▶ *Windows NT 4.0*
- Data General
 - AViiON 4900, 5000
 - ▶ *DG/UX 4.2*



© IBM Corporation 1998

Supported Servers and Prerequisites ...

Sun Microsystems

Sun systems supporting the VSS are the Ultraserver Models 1000, 1000E, 2000, 2000E, 3000, 4000, 5000, 6000. They require Solaris versions 2.5.1 or 2.6. No other special software is required.

Hewlett-Packard

The VSS is supported on HP 9000-800 series, D, E, G, H, I and K series, T series and EPS systems. HP-UX levels 10.01, 10.10, 10.20 or 10.30 are required. No other special software is required.

Compaq

The VSS is supported on Compaq ProLiant Models 3000, 5000, 5500, 6500, and 7000. Compaq requires Windows NT 4.0. No other special software is required.

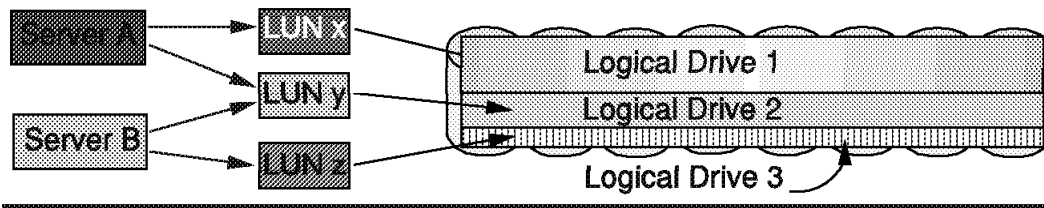
Data General

The VSS is supported on Data General AViiON Models 4900 and 5000. DG/UX 4.2 is required. No other special software is required.

Host View of the Versatile Storage Server



- Unix systems
 - Generic SCSI disk driver
 - Sees VSS as SCSI disk(s)
 - AIX sees VSS as *physical volume - hdisk*
 - Up to 15 targets
 - Up to 64 LUNs
 - Capacity of 0.5 GB to 32 GB
- AS/400 systems
 - VSS emulates 9337 disk subsystems
 - Logical 4 GB drives emulate 9337-480
 - Logical 9 GB drives emulate 9337-590



© IBM Corporation 1998

Host View of the Versatile Storage Server

VSS supports connection to host systems through an Ultra SCSI adapter. A variety of UNIX-based systems and AS/400 systems support the VSS as a peripheral device. This foil discusses how the systems view the VSS.

UNIX systems

For UNIX-based systems, the VSS emulates a SCSI disk drive. The host system will access the VSSs virtual drives as if they were a generic SCSI disk drive. The AIX operating system will contain entries in its ODM database to identify the VSS properly, but will access it using its generic SCSI disk driver. The VSS will appear to be a standard physical volume or *hdisk* to AIX. The VSS will appear similarly to Solaris and HP-UX systems.

When using Ultra or wide SCSI interfaces, a total of 16 devices are supported connected to the bus. The initiator (a device capable of initiating a SCSI command—usually an adapter) uses one address, leaving a total of 15 address for devices, typically called *targets*. Each target can have a total of 64 LUNs for Ultra SCSI or 32 LUNs for SCSI-2 wide. A VSS can be configured to appear as up to 64 LUNs per SCSI interface, each LUN having a capacity of 0.5 GB up to 32 GB (valid LUN sizes are: 0.5, 1, 2, 4, 8, 12, 16, 20, 24, 28, and 32 GB).

The number of LUNs per target depends on the operating system.

There does not have to be any correlation between physical drives in the VSS subsystem and the virtual LUNs seen by the UNIX server.

AS/400 systems

The VSS emulates 9337 subsystems when attached to an AS/400. As AS/400s require 9337 subsystems to have a minimum of four drives and a maximum of eight, the VSS must be configured to support a four- to eight-drive subsystem. The 9337-580 emulation requires logical 4 GB drives, while the 9337-590 emulation requires logical 8.5 GB drives. There is no relationship between the physical disk drives and the logical volumes assigned to the AS/400. The AS/400 expects to see a separate device address for each disk drive in the subsystem, so the VSS will report unique addresses for each virtual drive defined to the AS/400.

For more information regarding configuration of the VSS, see Chapter 6, “Versatile Storage Server Configuration Options” on page 191.

Transferring Data - Unix Systems



- Volume management software
 - Logical volume or file system copy
 - Migrate whole disk
 - Create mirror, split mirror
- Direct copy
 - `cpio -p`
 - ▶ *Can copy complete directory structures*
 - ▶ *Can only do on a per file system basis*
- Backup and restore
 - Different methods available
 - Slow
 - May be the only way with databases using raw file systems



© IBM Corporation 1998

Transferring data—UNIX Systems

Migrating data to a VSS can be done using standard host utilities, as the host sees the VSS as one or more generic disk drives. The method chosen depends on a number of factors, including the amount of data to be migrated, the amount of time available, the availability of tape devices or spare disks to facilitate temporary storage, and the format of the data itself. This foil discusses some of the methods available.

A VSS in minimum configuration comes with four arrays of eight disk drives. It is possible that some part of the array can be used as temporary storage for archives instead of using a slower tape device.

Volume management software

AIX and HP-UX 10.XX ship with volume management software as part of the base operating system. AIX is built around its LVM, which provides complete control over all disks and file systems that exist on an AIX system. HP-UX has similar volume management software.

There is a basic volume management product for Solaris, called *Solstice*, which is available from Sun Microsystems. Similarly, the Veritas Volume Manager (VxVM) and Veritas File System (VxFS) can be purchased as optional products for Solaris. All of these volume managers provide tools and utilities to move data—at a logical volume level, at a physical volume (disk drive) level, or by selective use of the mirroring features of the volume managers.

Beginning with “AIX Logical Volume Manager” on page 235 we discuss methods and examples pertaining to AIX.

Direct copy

If the data to be migrated resides as individual files on UNIX file systems, and no volume management software is available, the next easiest method of moving the data is to use a utility that supports a direct copy feature, such as **cpio** with the **-p** (pass) option.

Command **cpio** is available on all of the UNIX operating systems that support the VSS, and is easy to use. The command **cpio** reads a list of files to copy from its standard input. The easiest way to produce the list of files is to use the UNIX **find** command and pipe its standard output to the standard input of the **cpio** command. The following output shows a typical example of the use of **cpio** to move data.

```
# mount /dev/lv00 /mnt
# cd /data
# find . -print | cpio -pdmuv /mnt
.
.
.
# umount /mnt
# umount /data
# mount /dev/lv00 /data
```

In the example, we first assume that a file system has been made on the `/dev/lv00` logical volume, which is how the AIX LVM views part or all of the virtual disk made available to the system by the VSS. We mount the logical volume on a temporary mount point, in this case `/mnt`. We then change directories to the directory at the top of the file system we wish to move (`cd /data`). Using the **find** command, we produce a list of file names which is passed to the **cpio** command by a pipe (`|`). Once finished we unmount both file systems, and mount the new file system over the original mount point directory.

Backup and restore

In some cases, the only method available to transfer data is to back it up to a tape device and restore it to the new disk device. This method is obviously slower, as tape devices are essentially slow devices. However, if disks are being removed before the VSS is installed, the only way to move the data is via a tape device.

There are a number of different utilities on UNIX systems with which to archive data. The **cpio** command mentioned above can also create and read archives on tape devices. Using the **-o** (output) option creates archives and the **-i** (input) option reads and extracts data from archives.

AIX provides a pair of commands **backup** and **restore**. The **backup** command has two different methods of creating archives, either by file name or by inode. The **restore** command must be used to read any archive created by **backup**. Solaris and HP-UX provide the **dump** and **restore** commands, which provide backup and restoration by inode.

The **tar** command is available on all UNIX systems that support the VSS, and is another form of creating and extracting tape archives. The **c** option is used to create archives, while the **x** option is used to extract files from archives.

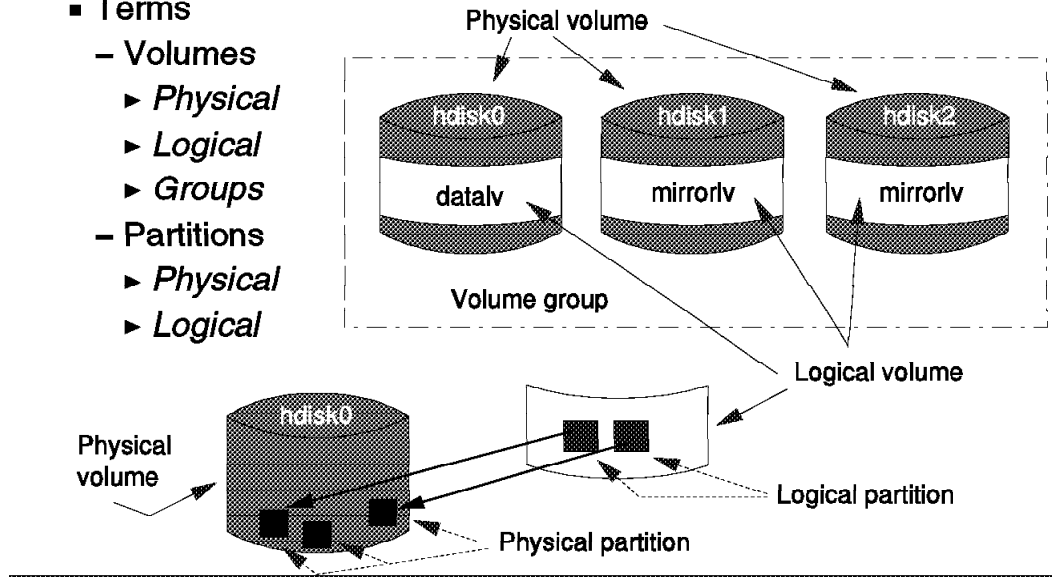
For databases that use raw file systems, logical volumes, or methods other than a UNIX file system, it may not be possible to use the volume management methods of migrating data, especially if the database uses volume serial numbers in its licensing code, or validity checking. If databases use licensing methods or validity checking, it may only be possible to export the database from its old locations, and import the database to its new location. It is up to the database software to provide some mechanism to move the data. This may take the form of a standard database backup and restore if it does not have any specific tools for movement of data.

AIX Logical Volume Manager



■ Terms

- Volumes
 - ▶ Physical
 - ▶ Logical
 - ▶ Groups
- Partitions
 - ▶ Physical
 - ▶ Logical



© IBM Corporation 1998

AIX Logical Volume Manager

The AIX LVM provides useful tools and utilities for migration of data as part of the AIX base operating system release. These tools can be used to move data to and from the VSS as they would on any other disk drive connected to an AIX system. This foil discusses some of the methods that can be used to migrate data, and shows some example commands.

For most purposes of data migration, these methods are preferred, as they work below the file system level—they don't care what sort of data resides on the logical volume, be it a UNIX file system or a raw database.

Terms

For the purposes of this discussion, the following terms refer to elements of, or definitions used by, the AIX LVM:

- Volume — a collection of data
- Physical volume — a disk drive, upon which a collection of data is stored.
- Logical volume — how AIX views collections of data. Logical volumes reside on physical volumes, but there is not necessarily any correlation between the size or position of logical volumes and the physical volumes underneath them. That is, a logical volume may span one or more physical volumes, or there may be more than one logical volume on a physical volume.

- Volume group — a group of physical volumes upon which logical volumes reside.
- Partition — the smallest piece of data known to the LVM.
- Physical partition — the smallest piece of a disk drive. A physical volume is divided into physical partitions when it is introduced into a volume group. A physical volume can have up to 1016 physical partitions. A physical partition can range from 4 MB in size to 256 MB, depending on the size of the physical volume. All physical volumes in a volume group must have the physical partitions of the same size.
- Logical partition — a logical volume is made up of logical partitions. A logical partition maps or points to one to three physical partitions, providing the logical to physical volume relationship. When data is written to a logical partition, the physical partitions it points to will have data written to them. When a logical partition points to more than one physical partition, the logical volume is being mirrored. Logical partitions are the same size as the physical partitions they point to; a volume group with physical partitions of 8 MB will have logical partitions of 8 MB. A logical partition is the smallest amount of data that can be allocated to a logical volume.

For more information on the AIX LVM, see the ITSO Redbook, *AIX Storage Management* (GG24-4484).

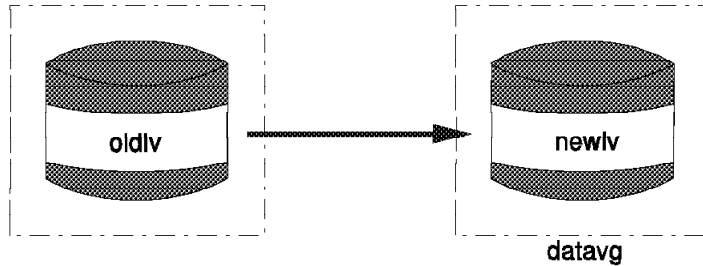
Now that we have discussed some basic terms regarding the AIX LVM, let's discuss how to copy and migrate data using the facilities provided by the LVM.

Complete Logical Volume Copy



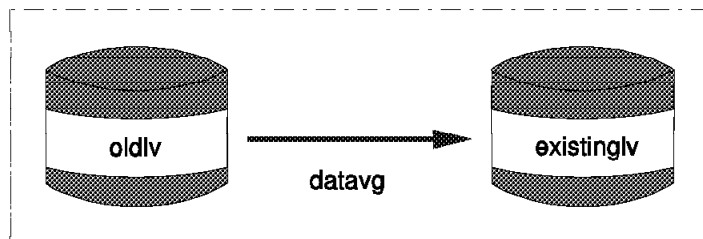
```
cplv -v datavg -y newlv oldlv
```

Copies a logical volume to another volume group, creating new logical volume.



```
cplv -e existinglv oldlv
```

Copies a logical volume over an existing logical volume in the same volume group.



© IBM Corporation 1998

Complete Logical Volume Copy

The AIX LVM provides the **cplv** command for copying logical volumes within volume groups or to different volume groups. It can be used to create a new logical volume while running the command or it can overwrite an existing logical volume. With reference to the foil, here are some examples:

```
# cplv -v datavg -y newlv oldlv  
# cplv -e existinglv oldlv
```

In the first example, the **cplv** command is copying data from the existing logical volume *oldlv* creating a new logical volume called *newlv* (-y) in the volume group *datavg* (-v). If the -v option is omitted, the volume group to which the existing logical volume belongs will receive the new logical volume. The **cplv** command, when creating a new logical volume for you, will create the new logical volume with exactly the same characteristics as the existing logical volume.

The second example shows the use of **cplv** to copy the data from the existing logical volume *oldlv* to the existing logical volume *existinglv* (-e). When using the -e option, the existing target logical volume will be overwritten with the data from the source logical volume. When using the -e option, the characteristics of

the existing target logical volume are maintained. Care should be taken when using this option.

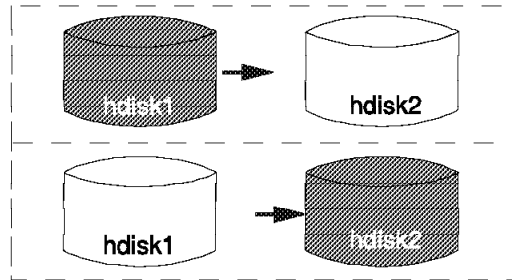
The command **cplv** is a good method for copying or migrating a single logical volume. Sometimes, however, it may be necessary to migrate all the data from a physical volume. The next foil discusses **migratepv**

Migrate Physical Volume



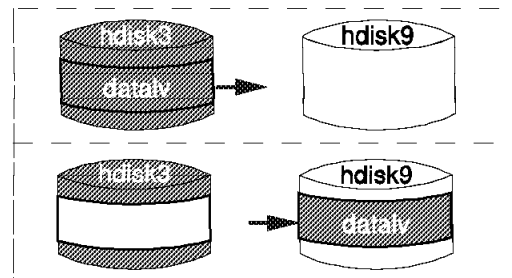
```
migratepv hdisk1 hdisk2
```

Migrates all data from one physical volume to another physical volume, in the same volume group.



```
migratepv -l data1v hdisk3 hdisk9
```

Migrates all data in one logical volume from one physical volume to another physical volume, in the same volume group.



© IBM Corporation 1998

Migrate Physical Volume

The AIX LVM provides the **migratepv** command to migrate complete physical volume data from one to another. It also provides options to migrate only portions (at the logical volume level) of data from one physical volume to another. Referring to the foil, here are some examples:

```
# migratepv hdisk1 hdisk2  
# migratepv -l data1v hdisk3 hdisk9
```

In the first example, we are simply migrating all data from *hdisk1* to *hdisk2*. The **migratepv** command updates all LVM references so that from the time of completion of the command, the LVM no longer references *hdisk1* to access data that was previously stored there. As the data is physically moved, the target physical volume must have enough spare physical volumes to accommodate the source physical volumes data. The source physical volume can then be removed from the volume group.

In the second example, we are moving any data in the logical volume *data1v* which physically resides on *hdisk3* to *hdisk9*. If any other data resides on *hdisk3* it is not moved.

The **migratepv** command can be used while the system is active, without disturbing users.

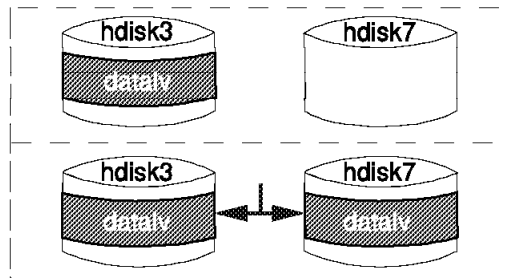
The **migratepv** command works by creating a mirror of the logical volumes being moved, then synchronizing the logical volumes. The original logical volume is then removed. The next example describes two ways of using mirroring to duplicate and migrate data.

Mirror Migration



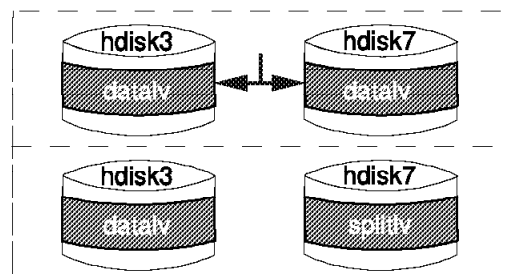
```
mkivcopy -e m -s y -k datalv 2 hdisk3 hdisk7
```

Creates a mirrored copy of logical volume, within the same volume group.



```
splitivcopy -y splitlv datalv 1
```

Splits a mirrored logical volume, into two separate logical volumes within the same volume group.



© IBM Corporation 1998

Mirror Migration

One of the facilities of the AIX LVM is to provide RAID level 1 data mirroring, in software. Mirroring can be used to create copies or move data, similar to **cplv** and **migratepv**.

The **mkivcopy** command creates a second copy of the data within the logical volume. Command **mkivcopy** is all that is needed to start mirroring of data in a logical volume. The *mkivcopy* process can be used on an active system without disturbing users. Some performance degradation may be noticed while the new data copy is being synchronized with the existing copy.

The **splitivcopy** command is used to break a mirrored logical volume into separate logical volumes using the copies of the data that are available. It will create new logical volumes leaving one intact with its original name. Command **splitivcopy** can be used on an active logical volume, but it is not recommended, as corruption can occur if the data is being actively updated while the split operation is being run.

The **rmlvcopy** command can be used to delete a copy of a mirrored logical volume. The space currently being used by the copy that is removed is freed up to be used by other logical volumes in the volume group. In this way, data can be moved from one physical volume to another. To create a copy of the data, use **mkivcopy**. Once the synchronization of the new copy is completed, the old copy can be removed, and the data has effectively been moved. This method

can be performed without affecting users. Referring to the foil, here is an example of the above method:

```
# mklvcopy -e m -s y -k data1v 2 hdisk3 hdisk7
.
.
.
# splitlvcopy -y splitlv data1v 1
```

The example shows how to create a mirror copy in a logical volume. The options specify to use minimum interdisk allocation policy (**-e m**), strictly allocate mirror copies on separate physical volumes (**-s y**) and synchronize new copies immediately (**-k**). Here, *data1v* is the name of the logical volume we wish to start mirroring, two is the number of copies that we wish to make of the data, in total (this can be a maximum of three). Disks *hdisk3* and *hdisk7* are the physical volumes upon which the logical volume will reside. Disk *hdisk3* is the physical volume already holding the data of *data1v*. Disk *hdisk7* is the physical volume that will hold the mirror copy, and where we want to move the data.

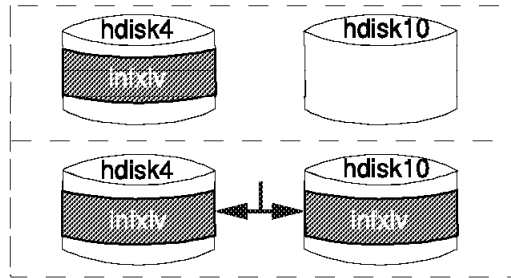
The **splitlvcopy** command creates a new logical volume called *splitlv* with the **-y** option. The source logical volume is *data1v*. The number 1 specifies how many copies are to remain in the logical volume. Logical volume *data1v* will still exist as it did before, residing on *hdisk3*, and the new *splitlv* will reside on *hdisk7*.

Mirror Migration ...



```
mklvcopy -e m -s y -k inxlv 2 hdisk4 hdisk10
```

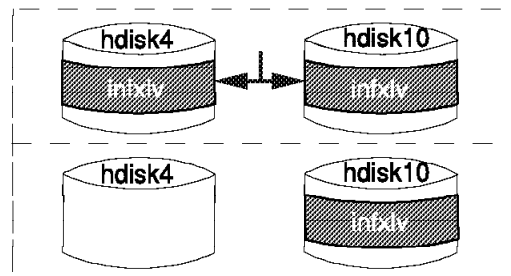
Creates a mirrored copy of logical volume, within the same volume group.



```
rmlvcopy inxlv 1 hdisk4
```

Removes a mirrored logical volume copy.

Effectively, the logical volume has been moved physically. Logically, it is the same.



© IBM Corporation 1998

Mirror Migration ...

The second example we discuss here shows a similar **mklvcopy** command to the previous example to start mirroring the *inxlv* logical volume.

```
# mklvcopy -e m -s y -k inxlv 2 hdisk4 hdisk10
.
.
.
# rmlvcopy inxlv 1 hdisk4
```

The disk *hdisk4* is where the data for *inxlv* already exists, while *hdisk10* is where the mirror copy will reside, and where we want to ultimately move the data.

The **rmlvcopy** command, as shown, specifies that we wish to remove a copy from *inxlv*, leave one remaining copy of the data, and remove the existing copy from *hdisk4*. The disk *hdisk10* will contain the remaining data; effectively the data has been moved from *hdisk4* to *hdisk10*.

Impact on Availability



- Volume management methods
 - **cplv**
 - ▶ *Best performed on quiesced logical volumes*
 - ▶ *Requires /etc/filesystems update and fsck run*
 - **migratepv**
 - **mklvcopy, splitlvcopy**
 - ▶ *splitlvcopy should be run on closed logical volumes*
 - ▶ *Requires /etc/filesystems update and fsck run*
 - **mklvcopy, rmlvcopy**
 - Direct copy
 - **cpio -p** should be run when file systems quiesced
 - Backup and restore
 - **slow, requires file systems quiesced**
-



© IBM Corporation 1998

Impact on Availability

Unfortunately, most methods of data migration have some sort of effect on the everyday operational aspects of a computer system. When data is moved, it must be in a known state, typically requiring that updates or changes cease while the movement occurs. Depending on the amount of data to be moved and the method chosen, the data could be unavailable for a long time.

Other factors like creation of new logical volumes or file systems, modification of configuration files, data integrity checks, and testing all contribute to the unavailability of data being migrated. This foil discusses some of the aspects that must be considered when choosing a method to use for migrating data.

Volume management methods

Command **cplv** is used to make entire copies of logical volumes. The command can operate without disrupting access to data contained within the logical volume it is copying. If the data is actively being updated, however, the resulting copy may not look anything like the logical volume it started as. Some updates may make it to the copy, others may not. Inconsistency is a drawback of using **cplv** on active logical volumes. We recommend that any logical volumes that are going to be copied using **cplv** be closed beforehand.

Closing of a logical volume requires either that the file system built upon it be unmounted, or that any database that has the raw logical volume open should be shut down. After running **cplv**, the file system configuration file `/etc/filesystems`

should be updated to include the relevant configuration data regarding the new logical volume and file system. The file system integrity checker *fsck* should then be run to ensure data consistency within the new logical volume.

The command **migratepv** is one of the better methods for moving data without disrupting users. Command **migratepv** can be run on an active system, as it first creates a mirror of each logical volume contained within the physical volume, and synchronizes both copies. Once the synchronization is complete, the original copy is removed, leaving the new copy active and available to users. Users may notice some performance degradation due to the atomic nature of creation and synchronization of the mirror, as each physical partition is written separately, locked from being accessed by any other process. This can slow down access to the data, but ensures data integrity.

The *mkivcopy*, *splitivcopy* method is ideal for creating a copy of logical volumes, and then running with both copies, one in a production environment and one in a test environment. The **mkivcopy** command ensures data integrity by creating a mirror copy of the data and synchronizing it atomically. Command **splitivcopy**, however, should not be run on an active logical volume for the same reasons that **cplv** should not be run on an active logical volume. If processes are updating the data while the split is taking place, the consistency of the data on both copies can not be guaranteed.

After running **splitivcopy**, the file system configuration file */etc/filesystems* should be updated to include the relevant configuration data regarding the new logical volume and file system. The file system integrity checker *fsck* should then be run to ensure data consistency within the new logical volume.

The *mkivcopy*, *rmivcopy* method is similar to the method used by *migratepv*. Mirrors are created and removed atomically, ensuring data integrity. This method is ideal for migrating active logical volumes, where a slight performance loss is acceptable.

Direct copy method

This method suffers from the same drawback as do using *cplv* and *splitivcopy* on logical volumes; it can be used on active file systems, but data consistency between the original and new copies cannot be guaranteed. Command **cpio -p** should be used only on file systems that are in a quiescent state, to guarantee consistency between the copy and original. This in turn mandates some disruption of service to users.

Backup and restore methods

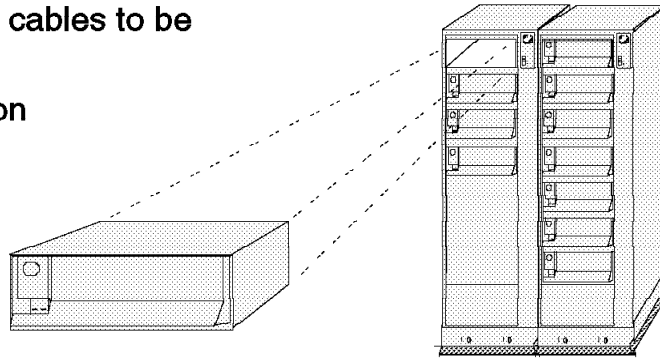
All backup and restore methods require that the file systems or logical volumes being backed up be in a quiescent state. Typically, backups are run after hours when there is minimal use of the system and databases can be shut down. Typically, backup and restore methods are slow, time consuming, and ultimately inefficient methods of migrating data.

Some databases require that to move data between file systems or logical volumes, a database export must occur, followed by an import of the data on to the new file system or logical volume. This is inefficient and time consuming but often necessary. Some reconfiguration of the database may have to occur to point at the new data location.

Use of Existing 7133 Drawers with VSS



- Existing 7133s can be migrated to new racks
 - 7133 MES may be required
 - ▶ *Old power cables not used*
 - ▶ *Front cover is removed*
- Existing 7133s can remain in 7014, 7015 or 7202 racks
 - Only requires loop cables to be connected to VSS
 - No power protection



© IBM Corporation 1998

Use of Existing 7133 Drawers with Versatile Storage Server

7133s can be migrated to new racks

To protect a customer's investment in SSA technology, use of existing 7133-010 or 7133-020 drawers is supported in the VSS subsystem. All data on drives being migrated should be backed up before commencing movement of the drawers. The physical movement of the drawers or racks must be performed by an IBM CE.

Existing 7133s can be migrated to new racks

Existing 7133s can be moved into new 2105-B09 or 2105-100 racks. The VSS racks are different from other IBM racks.

Although 7133s in a VSS rack run on 350 V DC, no special procedures need to be done in order to facilitate this—the 7133 has auto-sensing power supplies. The power cords from the original enclosure or rack are not used, however, as the VSS rack already has power cables installed. Because of configuration restrictions (see "Use of Existing Disk Drives in VSS" on page 248), existing 7133-010 drawers may require extra loop jumper cables to complete loop connection.

Existing 7133s can remain in 7105 or 7202 racks

If a customer has a number of 7133s already installed in existing 7105 or 7202 racks, and does not wish to physically move the drawers into VSS racks, the existing racks can still be used. No MES kit is required, as the original power cabling remains within the existing racks. However, the 7105 and 7202 racks do not provide any power supply redundancy or protection against outages as the VSS racks do.

SSA cabling of the drawers is to the SSA adapters in the 2105-B09 adapter bays. Configuration restrictions (see “Use of Existing Disk Drives in VSS” on page 248) mean that existing 7133-010 drawers may require extra loop jumper cables to complete loop connection.

Use of Existing Disk Drives in VSS



- **Strict drive requirements**
 - Only 4.5 GB and 9.1 GB drives supported
 - 7133-010 only supports 4.5 GB drives
- **All drives in array and drawer must be same size**
 - 8 or 16 drives per drawer
 - ▶ *Arrays must be either 6+P+S or 7+P*
 - To allow for spare to cover whole drawer
- **Drives must be reformatted**
 - New 524-byte sector
 - Data should be backed up to temporary storage or migrated to new location prior to physical movement to VSS



© IBM Corporation 1998

Use of Existing Disk Drives in VSS

As there are almost infinite cabling and disk drive options for use in SSA subsystems, it is likely that customers will have disk configurations different from those supported by the VSS. This foil discusses the requirements to move existing SSA disk drives into VSS subsystems. Because of these requirements, the customer may have to purchase extra drives or loop jumper cables to accommodate the new configurations supported by the VSS.

Strict drive requirements

Only 4.5 GB and 9.1 GB disk drives are supported in the VSS. Customers with 1.1 GB or 2.2 GB drives will not be able to migrate them into the VSS. Because of its different design, the 7133-010 is unable to provide adequate cooling for the 9.1 GB drives. For this reason, the 7133-010 can support only 4.5 GB drives in a VSS subsystem.

All drives in an array and drawer must be the same size

Referring to “Array Definition” on page 81, an array is defined as eight drives, used in either a 6+P+S or 7+P configuration. A drawer, in the VSS can contain either 8 or 16 drives—one or two arrays. Because of the way RAID-5 treats the drives, drives must be the same size to fully utilize the capacity of each drive; the usable capacity of each drive in the array is limited by the size of the smallest drive.

The spare drive is another factor in the requirement for restricting a drawer to same-size drives. The spare drive is able to provide backup for both arrays in the drawer. If the two arrays were made up of different- sized drives, it is conceivable that a 4.5 GB drive could become a data drive in the 9.1 GB-drive array if a 9.1 GB drive failed. This presents a problem as there would not be enough room on the 4.5 GB drive to accommodate data reconstruction from the 9.1 GB array. Similarly, if the spare was a 9.1 GB drive and a 4.5 GB drive failed, half of the 9.1 GB drive would remain unused in the 4.5 GB-drive array.

Drives must be reformatted

To provide extra security checking of data and to allow different operating systems to use the VSS, the VSS disks are formatted to a 524-byte sector rather than the traditional 512-byte sector of fixed-block-architecture disk drives. The 524-byte sector format mandates that all disks being migrated to VSS subsystems be reformatted. As reformatting will erase any data already contained on the drives, the data must be backed up to some form of temporary storage.

As all VSS subsystems come with two arrays of eight drives, it is possible that these drives can be used for either direct migration of data, or as temporary storage while the existing drives are moved. If this is not possible, a removable media device such as a tape drive would have to be used to temporarily store the data while the drives are reformatted.

“524-byte Sector Format” on page 250 explains the new format in detail.

524-Byte Sector Format



- ✓ 8 bytes of AS/400 header information
- ✓ 512 bytes of data
- ✓ 2-byte sequence number
- ✓ 2-byte longitudinal redundancy check

AS/400 Header	Data	SEQ #	LRC
8	512	2	2

524-Byte Sector



© IBM Corporation 1998

524-byte Sector Format

Most fixed-block disk architectures use a fixed-byte sector of 512 bytes. This includes most UNIX systems, including AIX. When used in a VSS subsystem, a disk drive has a format of a fixed-byte sector of 524 bytes. The 524-byte sector format enables the VSS subsystem to connect to a wide range of host systems and share data between them.

The 8 bytes at the start of the sector are used by IBM AS/400 systems, and are not used when the VSS is attached to UNIX hosts. The data portion of the sector remains at 512 bytes, for all systems. A 2-byte sequence number and a 2-byte LRC increase the size of the sector to 524 bytes. The sequence number is a modulo 64k value of the LBA of this sector and is used as an extra method of ensuring that the correct block is being accessed.

The LRC, generated by the SCSI host adapter, is calculated on the 520 data and header bytes and is used as an error-checking mechanism as the data progresses from the host, through the VSS storage server, into the SSA adapter, and on to the RAID array (see Chapter 4, "Versatile Storage Server Data Flow" on page 109 for a detailed description of data flow through the subsystem). The LRC is also used as an error-checking mechanism as the data is read from the array and passed up to the host adapter. The sequence number and LRC are never transferred to the host system.

Chapter 8. Versatile Storage Server Performance

Performance



Performance



© IBM Corporation 1998

In this chapter, we examine the options in configuring the Versatile Storage Server for a given capacity, and the performance implications of these options.

Performance Overview



- VSS Performance Highlights
- Performance Planning Information
- Total System Performance
- VSS I/O Operations
- Guidelines for VSS Configuration
- Workload Characterization
- Other Performance Considerations
- Summary



© IBM Corporation 1998

Versatile Storage Server Performance Overview

The overview foil introduces the topics we cover.

VSS Performance Highlights

First, we give a short list of the key performance capabilities of the Versatile Storage Server.

Performance Planning Information

We define the types of performance guidelines presented in this chapter. This chart helps ensure that the rules of thumb included in this chapter are not misconstrued.(There is no such heading.SWH)

Total System Performance

Reminds the reader that disk subsystem performance is just part of overall system performance.

VSS I/O Operations

Gives a quick review of how SCSI front end I/O operations relate to SSA back end I/O operations in the VSS.

Guidelines for VSS Configuration

In this set of foils, we look at the actual configuration options that can be varied independently of subsystem capacity. They are:

- Storage server cache size
- Number of UltraSCSI ports per host
- Choice of 4.5 or 9.1 GB disk drives
- Number of RAID-5 arrays per SSA loop

Workload Characterization

In this set of foils, we look at several aspects of a workload that may affect the configuration of the VSS. They are:

- Read to write ratio
- Synchronous write I/O content
- Sequential I/O content
- Caching characteristics of data

Other Performance Considerations

In this set of foils, we look at several miscellaneous factors that may affect the configuration of the VSS. They are:

- Race conditions
- Migrating data from RAID-1 file systems to RAID-5.
- Use of parallel query processing in database management systems

Summary

In these foils, we first repeat the key performance capabilities of the VSS in review, then summarize the key points presented in this chapter.

Performance Highlights



- Storage Server cache
 - Improved performance for read cache hits
- Adaptive cache
 - Increased number of read cache hits in storageserver cache
- Sequential prediction and sequential prestage
 - Improved sequential read throughput
- Fast write
 - Masks RAID-5 write penalty
- SSA adapter nonvolatile storage
 - Protection of fast write data



© IBM Corporation 1998

Performance Highlights

In this foil, we present a very high-level look at the performance capabilities of the VSS.

Storage server cache

The VSS can be configured with a large storage server cache. The primary advantage of this disk subsystem cache is to provide read cache hits for reads not satisfied from the host memory.

Adaptive cache

The VSS storage server cache is managed by an adaptive caching algorithm which determines how much data should be staged to storage server cache when data is read from the disk back store. The storage server can either stage just the block (or blocks) read, the blocks read and the balance of the 32 KB track, or the entire 32 KB track containing the requested blocks.

The key advantage of this adaptive caching is that it intelligently manages which data is placed in cache. Adaptive caching can increase the number of read cache hits given the same size storage server cache.

Sequential prediction and sequential prestage

The VSS detects sequential access to the disk in logical block address (LBA) sequence, and will begin read-ahead prestaging once sequential access is detected. With sequential prestaging, the sustained sequential data transfer rate for sequential reads is increased.

Fast write

Its fast write capability allows the VSS to return a SCSI task-complete indication when data has safely been stored in both cache and nonvolatile storage (Fast Write Cache) in the SSA adapter. The fast write capability allows the VSS to mask the overhead associated with RAID-5 update writes.

SSA adapter nonvolatile storage

Fast write wouldn't be very attractive if the integrity of data written to the disk subsystem were exposed until the data was committed to disk. The Fast Write Cache architecture of the SSA adapter used in the VSS protects the integrity of data written as soon as a task- complete indication is sent to the host in response to the SCSI write command.

Total System Performance



- Disk subsystem performance part of overall system performance
- System performance considerations
 - Processor resources
 - Memory resources
 - Network resources
 - I/O resources



© IBM Corporation 1998

Total System Performance

Disk subsystem performance as part of overall system performance

This foil serves as a reminder that the subject of this chapter, VSS performance as a disk subsystem, is only one part of overall system performance.

System performance considerations

Factors affecting system performance include:

- Processor resources

How powerful is the host processor? Are host processor resources overcommitted?
- Memory resources

How much memory is available to the host processor? Are any memory processes constrained?
- Network resources

How do transactions reach the host? Are the network resources constrained?
- I/O resources

What are the response time and throughput capabilities of the disks or disk subsystems, tape drives or tape subsystems, and other I/O devices?

System performance can be improved by adding resources where there are bottlenecks. Adding additional resources where there is no bottleneck can have little or no effect in improving overall system performance.

An excellent reference on the subject of system performance for RS/6000 hosts is the Redbook *Understanding IBM RS/6000 Performance and Sizing (SG24-4810)*.

I/O Operations



- Front end I/O
 - SCSI I/O from hosts
- Back end I/O
 - SSA I/O from VSS controller to disk
 - Assume uniform distribution across member disks in an array
 - Random read cache misses
 - ▶ *Reads x (1 - cache hit ratio)*
 - Random writes (including RAID-5 write penalty)
 - ▶ *Writes x 4*
 - Sequential reads
 - ▶ *Bytes read / 32 KB*
 - Sequential writes
 - ▶ *(Bytes written / 32 KB) x 1.15*



© IBM Corporation 1998

I/O Operations

Use this foil to differentiate between the SCSI front end I/O sent to the VSS and the resulting SSA back end I/O. It is not necessary to make these calculations in order to properly configure the VSS for most workloads. This information is presented so that the relationship between front end and back end I/O can be understood and investigated further if necessary.

Front end I/O

The front end SCSI I/O received by the VSS is whatever I/O load and mix are generated by the host applications. SCSI reads will be issued by a UNIX file system whenever data requested is not in the host cache. SCSI writes will be issued for a UNIX file system by the synchronization daemon, either by interval or when explicitly requested by a user or program. Information about this I/O load can be gathered from the AIX IOSTAT command, or equivalent.

Back end I/O

Back end I/O is the I/O generated by the VSS in accessing the SSA disk back store. The relationship between the front end I/O load and the back end I/O load is primarily a function of the read cache hit ratio and the read-to-write ratio of the I/O workload.

You can assume that the access patterns for the member disks in a 6+P or 7+P array will be reasonably uniform. Since this is a RAID-5 disk subsystem,

there is no single parity disk in the array. Some parity data is stored on all disks in the array.

Access skew smoothing is an indirect benefit of RAID-5 disk subsystems when compared with non-RAID storage. A 7+P RAID-5 disk array has the same total capacity as seven JBOD (“just a bunch of disks”) disks. You would expect to see a skew in the access pattern for seven JBOD disks, with some disks busier than others. Migrating that workload to a 7+P RAID-5 array has the effect of smoothing the workload across the eight disks in the RAID array.

While the RAID array has an extra disk, the smoothing of the access pattern isn’t because of the addition of the extra disk, it’s because of striping the workload of each of the seven JBOD disks across the RAID array. A “hot spot” RAID array is less likely to occur than a “hot spot” JBOD disk.

Not all reads directed to the VSS result in access to the SSA disk back store. Only cache read misses require access to the back store. A VSS subsystem with a high cache hit ratio can have much less back end I/O activity than front end I/O activity.

To estimate the back end I/O load being generated for a given workload of random reads, use the formula $reads \times (1 - cache\ hit\ ratio)$.

On the other hand, random writes typically generate more back end I/O load than front end I/O load, because of the RAID-5 write penalty. While the VSS will cache both data and parity strips in the SSA adapter cache, which can result in cache hits during parity generation for a random write, a conservative assumption is that all random writes will generate four I/Os to the back store SSA disks. Use the formula $writes \times 4$ to calculate the write I/O load that will be generated from a given random write workload.

By *sequential reads*, we mean reads that are in LBA sequence for the virtual disk. Sequential processing in a UNIX file system can refer to processing all blocks in a file, which may not be stored together on the disk, depending on the structure of the logical volume definition and whether the disk is fragmented.

When the UNIX file system detects sequential processing, it will begin read-ahead into the host cache. The sequential detect and prefetch of the VSS complement this sequential read-ahead by increasing the read throughput from the VSS.

When performing sequential prestaging, the VSS will issue reads for 32 KB tracks to the SSA disk back store, regardless of the size of the SCSI reads being received. So we use the formula $bytes\ read / 32\ KB$ to calculate the number of read operations generated against the SSA disk back store as a result of sequential read I/O operations.

Data written sequentially is written in stripe writes by the VSS. As a result, there is no RAID-5 write penalty for sequential writes, although a parity strip must be written for every stripe. Since the parity overhead of a 6+P RAID array is not very different from the parity overhead of a 7+P RAID array, use the formula $(bytes\ written / 32\ KB) \times 1.15$ to calculate the number of write I/Os generated to the SSA back store for a given sequential write workload.

Guidelines for Configuration--Overview



- Storage Server cache size
- 4-way symmetric multiprocessors (SMPs)
- Number of SCSI ports per host
- Choice of 4.5 or 9 GB disk drives
- Number of RAID-5 arrays per SSA loop



© IBM Corporation 1998

Guidelines for Configuration—Overview

This foil introduces the configuration choices for the VSS that can be varied independently of the target capacity. Each of these will be discussed in more detail in following foils.

Storage server cache size

The total storage server cache must be distributed symmetrically across the two storage servers, but any cache size can be configured for any back store capacity. We will see that the rules of thumb for configuring storage server cache state that the storage server cache should scale with the capacity of the disk back store.

Number of SCSI ports per host

Each host must be attached to an UltraSCSI port owned by an SMP cluster side in order to be able to access RAID-5 arrays owned by that cluster side. It is also possible to define VSS virtual disks owned by the same cluster side as being accessible by different SCSI ports attached to the same host. We will see that the throughput requirements of the workload from a given host will determine the number of SCSI ports per side that should be configured.

Choice of 4.5 or 9.1 GB disk drives

The VSS supports the use of either 4.5 or 9 GB disk drives in a RAID-5 array; all disks in a RAID-5 array must have the same capacity. The frequency of back end access to SSA disks determines whether 9 GB disk drives are appropriate for the workload.

Number of RAID-5 arrays per SSA loop

For VSS subsystems with up to eight drawers, the rule of thumb is one SSA adapter for every 16-disk drawer, with both RAID-5 arrays in the drawer on the same SSA loop so that they can share a spare disk. We recommend that the second SSA loop on the adapter not be used.

Storage Server Cache Size



- Storage Server cache size determined by two factors
 - Effectiveness of host disk caching
 - Storage capacity of VSS subsystem



© IBM Corporation 1998

Storage Server Cache Size

This is the first of several foils that address choosing the right cache size for a given workload and capacity.

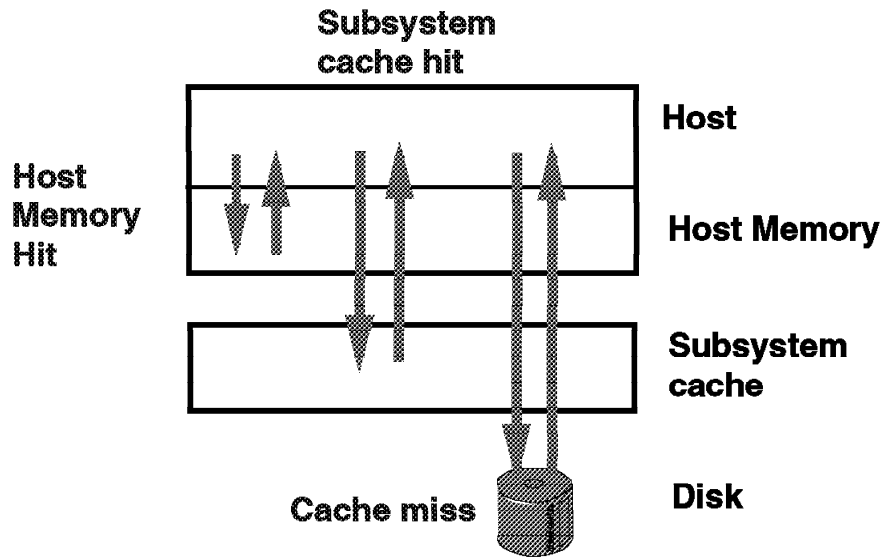
Storage server cache size determined by two factors

There are two factors that are used to determine the proper storage server cache size: the effectiveness of host caching, and the storage capacity of the VSS subsystem. These factors are examined more closely in the next several foils.

Depending on many factors, including workload, application and available host memory, host caching can be highly effective or only somewhat effective. The effectiveness of an external disk subsystem cache depends largely on whether cache hits are already effectively being serviced from the host cache.

Given a certain effectiveness of host caching, the amount of VSS storage server cache should scale with the size of the SSA disk back store.

Host Memory Caching



© IBM Corporation 1998

Host Caching

This foil depicts three possible outcomes for a read request by a UNIX application. The read request can be satisfied from the host cache, in which case it is a disk cache hit. The read request can result in a host cache miss and can be sent to an external disk subsystem where it can result in a subsystem cache hit. Or, the read request can result in both a host cache miss and a subsystem cache miss, in which case the data must be retrieved from the disk.

From this foil, it's easy to see why the effectiveness of a disk subsystem cache depends on the effectiveness of host caching. Host cache hits do not result in I/O to the disk subsystem. Host cache misses that result in I/O to the disk subsystem are unlikely to be in an external disk subsystem cache, unless the disk subsystem cache is considerably larger than the host cache, since the disk subsystem cache is likely to contain much of the same data as the host cache.

Effective Use of Host Memory



- Effectiveness of host caching depends on several factors
 - How much memory is available in the host for caching?
 - Is the host software exploiting it?
 - Is the same data accessed for update by multiple hosts without shared memory access?
 - ▶ *Integrity protection usually involves invalidating cache copy in other hosts if data is updated*



© IBM Corporation 1998

Effective Use of Host Cache

In this foil, we investigate how to determine if host caching is effective. No metric of host cache effectiveness is provided, but a binary indication of “effective host caching” versus “less effective host caching” will be used in the following foils to determine which rule of thumb for cache sizing to apply.

Effectiveness of host caching depends on several factors

Since the data being accessed could be a UNIX file system or a raw disk managed by a database management system, there are several factors that can affect the effectiveness of host caching.

The first question is, how much memory is available in the host for caching? An AIX system will use all memory not committed to other purposes as a disk cache. Other UNIX implementations may allow the system administrator to specify the size of the disk cache. Database management systems typically require an explicit specification of the amount of memory that should be used for disk buffers.

Data stored in a UNIX file system will automatically be cached in a host cache. Database data may be stored in either a UNIX file system or, more commonly, in raw disk. The host may have plenty of memory, but unless a database management system is using it for buffers, the host caching of the database data is not going to be highly effective.

The data integrity management of data accessed for update by multiple hosts introduces a requirement that other hosts know whether or not a block that may be in their cache has been updated by another host. This is usually addressed by having a host perform an update notifying each of the sharing hosts while the sharing hosts invalidate the copy of data in their cache, forcing a reread of the data from the disk subsystem if it is accessed again. This is known as the *ping effect*.

Host caching will be less effective in such a shared disk environment, depending on the percent of total data being updated, and the likelihood that updated data will be retrieved, since recently accessed data is flushed from cache in all systems but the one that has performed the update.

Host Caching Environments



- **UNIX file systems**
 - Automatically exploit host caching
 - Effectiveness of caching depends on host memory resources
- **Database Management Systems**
 - Typically must explicitly specify buffer sizes
 - Effectiveness of caching depends on host memory resources, exploitation of host memory resources
- **DB2 Parallel Edition**
 - Work routed to processor owning data
 - Allows effective host caching with processor clusters
- **Oracle Parallel Server**
 - Exploitation of Virtual Shared Disk (VSD) in RS/6000 Scalable Parallel (SP)



© IBM Corporation 1998

Host Caching Environments

In this foil, we examine specific application environments to understand the effectiveness of host caching for each.

UNIX file systems

UNIX file systems automatically exploit host caching, which is managed globally by UNIX for all file systems in the host. AIX uses all memory not in use for another purpose as a host cache. Other UNIX implementations may allow the system administrator to specify how much memory should be used as a disk cache for file systems.

If the workload to be placed on the VSS consists primarily of data that is part of a UNIX file system, the effectiveness of host disk caching largely depends on the amount of memory available for the host cache. Systems that are memory constrained will have less effective host caching, where systems with abundant memory should perform very effective host caching, assuming that the cache size has not been severely limited by the system administrator.

Database management systems

Most database management systems access data stored on raw disks, since the overhead is lower for this than for most UNIX file systems. Database management systems typically require the explicit specification of the memory to be used for host caching, since the database management system is not likely to be aware of other applications running on the same host.

Because the database management system is dependent on the proper specification of host cache size, the effectiveness of host disk caching for a database application is a function not only of the host memory resources but of how they are being exploited.

DB2 parallel edition

DB2 Parallel Edition is a special case of a database management system, since it is designed to be used in a cluster environment. In DB2 Parallel Edition, data is owned by a given processor and DB2 routes work to the processor owning the data.

This architecture allows the effective use of host caching, because data needs to be stored in only one host cache, the cache managed by the processor owning the data. DB2 Parallel Edition requires an explicit specification of disk cache size, but there are no special considerations for the effectiveness of host disk caching for a DB Parallel Edition configuration, since data will be in only one host cache.

Oracle Parallel Server

Oracle Parallel Server, when running on an RS/6000 SP system, exploits an AIX capability known as *Virtual Shared Disk* (VSD). In the VSD architecture, a process runs on a single node in the SP system and owns specific AIX logical volumes. There may be VSD processes running on several nodes, each controlling its own set of AIX logical volumes. Accesses to these logical disks are routed by AIX to the appropriate VSD process. Since access to data is from a single node, the VSD process running in that system can effectively use host caching.

A new AIX capability, the Generalized Parallel File System (GPFS), exploits the VSD capability to provide access to data in an AIX file system.

Subsystem Effect of Host Caching



- Where host caching is highly effective
 - Most cache hits that could occur in disk subsystem cache occur in host disk cache
 - Typically results in low cache hit ratio on disk subsystem cache
 - Very large subsystem cache can provide additional cache hits
 - ▶ *Rule of thumb: Disk subsystem cache four times or more the size of host disk cache*
- Where host caching is not highly effective
 - Benefit from cache in disk subsystem
 - Consider larger VSS cache than with effective host caching



© IBM Corporation 1998

Subsystem Effect of Host Caching

In this foil, we examine the effects on an external disk subsystem of highly effective and not highly effective host caching.

Where host caching is highly effective

The term *highly effective* is subjective, but we can differentiate between configurations in which host caching is highly effective and those in which it is not.

Where host caching is effective, the cache hits that could occur in the external disk subsystem cache do not because I/O is never generated to the external disk subsystem. As an extreme example, if there is only one host, all data is stored in one external disk subsystem, and both host and disk have 1 GB of memory in use as a disk cache, the two disk caches are likely to contain much of the same data. The LRU chains in the two caches would be different, because the external disk subsystem cache would be unaware of rereads of the same blocks, but the last n blocks read into the host cache would also be the last n blocks read into the disk subsystem cache.

Since most cache hits occur in the host cache, the cache hit ratio for reads in the external disk subsystem cache is typically lower. This helps explain why workloads that would be considered “cache unfriendly” in an S/390 mainframe environment are the norm in a UNIX environment.

It is still possible to attain reasonably high cache hit ratios in an external disk subsystem cache, even with highly effective host disk caching, but doing so requires a very large external disk subsystem cache.

The rule of thumb states that a disk subsystem cache four or more times the size of the host cache can provide significant performance benefit through additional read cache hits.

In the case of a cluster environment, the rule of thumb would state that the disk subsystem cache should be four times or more the size of the total storage used for host caching in all hosts in the cluster.

Where host caching is not highly effective

The terms *not highly effective* and *less effective*, used interchangeably, are also subjective, but again we can differentiate between configurations in which host caching is highly effective and those in which it is not.

As a general rule, in a configuration where host caching is less effective, there will be a performance benefit from read cache hits in the VSS subsystem. This is because fewer read cache hits are satisfied in the host cache.

In a configuration with effective host caching, performance benefit begins to be realized only when the external disk cache size is four times or more the host cache size. Where host caching is less effective, performance benefits can be realized as disk subsystem cache scales up from minimum size to larger size caches. Since there is benefit from each cache size increment, consideration should be given to investing in a larger disk subsystem cache in these environments. This is reflected in the rules of thumb presented later in this chapter.

Storage Server Cache Size Guidelines-Part I



- Disk subsystem cache critical to performance
 - ▶ *Consider large VSS cache*
- VSS cache should scale with disk backstore
- Rule of thumb: 2 GB cache per 450 GB disk

Cache	Number of drawers with	
	9 GB disk	4.5 GB disk
512 MB	1	2
1 GB	2	4
2 GB	4	7
4 GB	7	14
6 GB	11	18



© IBM Corporation 1998

Storage Server Cache Size Guidelines—Part I

Host caching

This is the first of three VSS storage server cache size guidelines, to be used where host caching is considered “less effective,” or unknown.

In a configuration with less effective host caching, disk subsystem cache can be more critical to overall system performance. Read requests not satisfied in the host cache are being sent as SCSI I/O to the external disk subsystem. A disk subsystem cache hit is much faster than a cache miss, and therefore subsystem cache hits are beneficial to performance.

The reason to consider a large VSS cache in a configuration with less effective host caching is that each increment in VSS storage server cache can give measurable improvement to I/O performance.

The greater the capacity of the SSA disk back store in gigabytes, the more cache. Again, it is capacity in gigabytes, not number of disks or drawers that matters. Therefore, 450 GB of 9 GB disks is equivalent to 450 GB of 4.5 GB disks when sizing VSS storage server cache.

This rule of thumb states that in a configuration with less effective host caching, twice as much VSS storage server cache should be configured as in a configuration with effective host caching.

In the table, information is shown defining how many 9 GB disk drawers and 4.5 GB disk drawers can be installed for each available cache size. Both columns assume that all drawers are of the disk capacity indicated. The table does not contain new information; it applies the rule of thumb to the available VSS storage server cache sizes.

Note that, in an environment with less effective host caching, the largest storage server cache sizes available for the VSS would be indicated for a VSS subsystem of less than the maximum number of drawers of 9 GB disks.

Storage Server Cache Size Guidelines-Part II



- Where host caching is effective
 - VSS cache should scale with disk backstore
 - Rule of thumb: 1 GB cache per 450 GB disk

Cache	Number of drawers with	
	9 GB disk	4.5 GB disk
512 MB	2	4
1 GB	4	8
2 GB	8	16
4 GB	16	18
6 GB	18	18



© IBM Corporation 1998

Storage Server Cache Size Guidelines—Part II

Where host caching is effective

This is the second of three VSS storage server cache size guidelines, to be used where host caching is considered “effective.”

The cache should be configured based on the capacity of the SSA disk backstore. Capacity in gigabytes is what matters, not capacity in disks or drawers. Therefore, 450 GB of 9 GB disks is equivalent to 450 GB of 4.5 GB disks when sizing VSS storage server cache.

The usable capacity of a VSS drawer containing a 6+P RAID array, a 7+P RAID array, and a spare disk is 57 GB with 4.5 GB disks and 115 GB with 9 GB disks. The rules of thumb for VSS storage server cache size are stated in terms of 450 GB of disk, which is the capacity, in round numbers, of four drawers of 9 GB disks or eight drawers of 4.5 GB disks.

The foil includes a table showing how many 9 GB disk drawers and 4.5 GB disk drawers can be installed for each available cache size. Both columns assume that all drawers are of the disk capacity indicated. The table does not contain new information; it applies the rule of thumb to the available VSS storage server cache sizes.

In an environment with effective host caching, however, the largest storage-server cache sizes available for the VSS would not be indicated by this rule of thumb.

Storage Server Cache Size Guidelines-Part III

- Where host caching is effective
 - VSS cache should scale with host cache
 - Rule of thumb: VSS cache should be four or more times the host memory used for caching
 - ▶ *Total of all systems in a cluster environment*

Host Disk Cache	VSS Cache
256 MB	1 GB
512 MB	2 GB
1 GB	4 GB
1.5 GB	6 GB



© IBM Corporation 1998

Storage Server Cache Size Guidelines—Part III

Where host caching is effective

This is the third of three VSS storage server cache size guidelines, to be used where host caching is considered “effective” and where you desire additional performance benefit from VSS subsystem cache.

In this case, it is the host cache, not the capacity of the disk back store, that is the scaling factor.

The rule of thumb states that for an external disk subsystem cache to be effective, it must be four or more times the size of host memory used for disk caching. In a cluster environment, this is the total of all systems in the cluster environment.

There is no upper limit to this rule of thumb. An external disk subsystem cache eight times the size of the host memory used for disk caching is not unreasonable if the objective is to improve performance by providing subsystem cache read hits.

Note that if you seek an increase in performance where host caching is already effective, you have a choice between increasing the memory available for host caching and using a large cache in the VSS storage server. Many factors may affect this decision, such as the age of the UNIX hosts, their upgradability, and

the ease of installation of a large VSS storage server cache. Since these factors are not directly performance related, the rule of thumb given does not attempt to recommend one versus the other.

In the table, information is shown defining what VSS storage-server cache sizes should be used to increase performance for a given host cache size.

4-way SMPs



- Both storageservers are 4-way SMPs
- High-performance 4-way
- Rules of thumb:
 - Select high-performance 4-way SMPs where I/O rate is 10,000 operations per second or more
 - Select high-performance 4-way SMPs where capacity is greater than 450 GB
 - ▶ 4 drawers of 9 GB disks
 - ▶ 8 drawers of 4.5 GB disks
- Consider storage server capability during failover
 - 4-way SMPs may provide additional I/O throughput capability during failover???
 - ▶ Consider 4-way SMPs for higher availability ?



© IBM Corporation 1998

Four-Way SMP

Where both storage servers are four-way SMPs

A VSS is configured with high-performance four-way SMPs on each storage server. Both storage servers must be configured the same way.

Rules of thumb

The first rule of thumb is to select high-performance four-way SMPs where SCSI front end I/O reaches or exceeds 10,000 operations per second.

Where the SCSI front end I/O rate is not known, a reasonable assumption can be made by applying a typical access density to the size of the SSA disk back store. Assuming an access density of about three I/O per second per GB, a 450 GB VSS subsystem would be expected to receive about 1500 I/O per second.

This capacity would be realized with either

- 4 drawers of 9 GB disks, or
- 8 drawers of 4.5 GB disks

The second rule of thumb, select high-performance four-way SMPs where capacity is greater than 450 GB, is just an application of the first rule of thumb.

It is SCSI front end I/O rate, not SSA disk back store capacity, that drives the storage server processor utilization.

In case of a failure of a storage server processor or memory, the VSS will shift the workload of that cluster side to the other cluster side (failover) until a repair action can be performed against the failed component. In cases of failover, a single cluster side will be performing more work than under normal operation.

Consider Storage server capability during failover

High-performance four-way SMPs, even for VSS subsystems of less than 450 GB, will provide additional throughput capability during failover. While such failover is unlikely, the additional processing power provided by high-performance four-way SMPs could mask the performance effects of a cluster side failure.

Note also that a failover procedure is invoked during nondisruptive activation of a new level of Licensed Internal Code (LIC).

In an installation with critical availability requirements, the choice of four-way SMPs for all VSS subsystems provides an availability advantage in addition to any performance gain.

Versatile Storage Server SCSI Ports



- Emulates multiple LUNs
 - LUNs may be addressed across more than one target ID
 - On a SCSI bus with a VSS SCSI port and a single host, there will be no SCSI bus arbitration even if more than one target ID is used
- Supports multiple SCSI initiators
 - Multiple host attachments on a VSS SCSI port should be avoided since this introduces SCSI bus arbitration
- UltraSCSI adapters
 - Can also support attachment to SCSI-2 host adapter
- Throughput considerations



© IBM Corporation 1998

Versatile Storage Server SCSI Ports

The VSS supports the configuration of up to eight dual-port UltraSCSI adapters.

Emulates multiple LUNs

On a given SCSI port, the VSS will provide addressability to multiple virtual disks as logical units (LUNs).

Up to 64 LUNs can be addressed by a single SCSI target ID. In some cases, the UNIX operating system used will limit the number of LUNs per target ID to a smaller number. In any case, the VSS can emulate multiple LUNs and multiple target IDs on a single SCSI port.

Where a SCSI bus is dedicated to a VSS SCSI port and a single host, there will be no SCSI bus arbitration overhead. This is true even if the LUNs are addressed as more than one target ID.

The absence of SCSI bus arbitration significantly improves the throughput of a single SCSI connection to the VSS.

Supporting multiple SCSI initiators

The SCSI bus attaching the VSS to a host is a standard SCSI bus and up to four SCSI initiators (hosts) can be supported in addition to the VSS initiator. The VSS requires that the host types be homogeneous, which means that all must be RS/6000 AIX, or all must be HP systems running HP-UX, or all must be Sun systems running Solaris. AIX supports up to four initiators, while HP supports only two and Sun supports only one. Different hardware and different software releases on the same bus are supported as long as the host types are homogeneous.

From a performance viewpoint, dedicating a SCSI port to a single host should be considered. Where more than one host attaches to a VSS SCSI port, SCSI bus arbitration will be required, which will have performance impacts beyond the sharing of the SCSI bandwidth.

UltraSCSI adapters

VSS adapters are UltraSCSI adapters, which operate at a burst speed of 40 MB/s. As is true for any UltraSCSI adapter, backward compatibility with SCSI-2 is provided, so that the VSS UltraSCSI adapter can be attached to a host SCSI-2 adapter. In this instance, the VSS SCSI port will operate at the 20 MB/s burst speed supported by SCSI-2 when transferring data.

Throughput considerations

The throughput of a VSS SCSI bus depends not just on the VSS SCSI adapter but also on the host SCSI adapter. In addition to whether the host SCSI adapter is UltraSCSI or SCSI-2, various SCSI adapters have different limitations in terms of maximum I/O rate. The maximum I/O rate of a VSS SCSI port may be determined by the host SCSI adapter limitation.

Number of SCSI Ports Per Host



- VSS SCSI adapter throughput
 - 80-120 GB per port
- Consider high-availability SCSI connection
- Multiple SCSI attachment options
- Consider virtual disk partitioning
 - For very high I/O rates
 - ▶ *May be further limited by host SCSI adapter*
 - High sequential throughput requirements



© IBM Corporation 1998

Number of SCSI Ports per Host

VSS SCSI adapter throughput

The sustained throughput of the VSS UltraSCSI adapter for sequential reads is as follows:

- Up to 25 MB/s UltraSCSI
- Up to 15 MB/s SCSI-2

One reason these figures are lower than the 40 and 20 MB/s burst rates of UltraSCSI and SCSI-2, respectively, is that all command and control transfer on a SCSI bus occurs at 10 MB/s for compatibility purposes, as required by the SCSI-3 architecture. Only data transfer occurs at the 20 or 40 MB/s speeds.

Consider high-availability SCSI connection

A VSS SCSI attachment is used by only one cluster side at a time. While a SCSI port can be reassigned to the other cluster side during failover, one SCSI attachment per cluster side per host is required during normal operation if the host has access to VSS virtual disks owned by both cluster sides.

Multiple SCSI attachment options

More than one SCSI attachment per host per storage server side can be configured if dictated by performance requirements. In this case, each will access particular virtual disks owned by the storage server side.

A virtual disk can be shared between two SCSI attachments that are used by two different hosts, but a virtual disk cannot be shared between two SCSI attachments used by the same host.

Consider virtual disk partitioning

In general, a single SCSI attachment from a VSS cluster side to a host will be adequate (AS/400 configurations require one SCSI port for every eight logical volumes). However, partitioning the virtual disks used by a host across several SCSI attachments should be considered if the I/O rate to the virtual disks accessed exceeds 1000 I/Os per second, which may occur if the virtual disks used by a host that are owned by one cluster side exceed 350 GB in capacity.

The capacity of a VSS – host SCSI attachment may be further limited by the host SCSI adapter, especially if it is SCSI-2.

High sequential throughput requirements, above 25 MB/s for all virtual disks sharing a VSS – host UltraSCSI attachment, or 15 MB/s for a SCSI-2 attachment, would dictate partitioning virtual disks across multiple SCSI attachments.

Disk Capacity Selection



- Disk specifications
- Access density



© IBM Corporation 1998

Disk Capacity Selection

The choice of 4.5 GB versus 9 GB capacity disk drives may be a key determinant of VSS subsystem performance. In selecting one capacity or the other, attention must be paid not just to the relative cost of the drives, but the relative cost of a configured VSS subsystem containing the drives. Comparing a VSS subsystem containing 4.5 GB drives to one of the same subsystem capacity containing 9 GB drives, the 4.5 GB drive subsystem will have twice as many drawers, and potentially twice as many racks.

We will discuss the following topics in the next two foils.

Disk specifications

The next foil shows a comparison of the specifications of the 4.5 GB and 9 GB SSA disks used in the VSS.

Access density

Access density is measured in I/Os per second per GB, which is a key metric in choosing between 4.5 GB and 9 GB drives.

Disk Specifications

IBM Corporation
San Jose, CA
© IBM Corporation 1998

	9 GB disk	4.5 GB disk
Average seek (read)	7.5 msec	7.5 msec
Average latency	4.2 msec	4.2 msec
Media data rate	10.2 to 15.4 MB/sec	10.2 to 15.4 MB/sec

- Disk can perform 50 I/Os per second
– Regardless of disk capacity



© IBM Corporation 1998

Disk Specifications

The specifications of the 4.5 GB and 9 GB drives used in the VSS are given in this foil. The major difference between the two drives is that the 9 GB drive has more platters and more heads. The capacity of a single disk platter is the same for both drives. Both drives have the same number of cylinders and the same number of bytes per track on corresponding tracks.

Because there are more heads on the actuator arm in the 9 GB drive, the average seek time for the 9 GB drive is slightly longer than for the 4.5 GB drive. Rotational speed, and therefore average latency, are the same.

Disk can perform 50 I/Os per second

As a rule of thumb, either the 4.5 GB or the 9 GB disk can perform 50 I/Os per second at a reasonable response time. Higher I/O rates are possible, but demand increased response time per I/O.

Access Density



- I/Os per second per GB
- Selection of capacity depends on access density
 - Read cache misses
 - Writes including RAID-5 write penalty
 - All virtual disks on a RAID array
- Rule of thumb: 50 I/O operations per second
 - 350 I/Os per second for a 6+P RAID array
 - 400 I/Os per second for a 7+P RAID array
- Select 4.5 GB disk drives where appropriate



© IBM Corporation 1998

Access Density

Access density is an important metric in the choice of 4.5 versus 9 GB capacity disk drives.

I/Os per second per gigabyte

Access density is the number of I/Os per second per gigabyte. Since the disk drives used in the VSS are both capable of about 50 I/Os per second per drive, the access density that can be supported by the 4.5 GB drive is twice that of the 9 GB drive.

Selection of capacity depends on access density

The access density that must be considered in selecting 4.5 GB versus 9 GB drives is VSS back end I/O. This I/O consists of:

- read cache misses, and
- writes including the RAID-5 write penalty

for all virtual disks in a RAID array.

Calculation of the SSA back end I/O load for a given SCSI front end I/O load was previously discussed in "I/O Operations" on page 258.

Rule of thumb

As a rule of thumb for access density, use 50 I/O operations per second.

Since SSA disks in the VSS are always part of a RAID array, this can be stated as:

- 350 I/Os per second for a 6+P RAID array
- 400 I/Os per second for a 7+P RAID array

Use 7 and 8 as multipliers for the 6+P and 7+P RAID arrays, respectively, since in a RAID-5 configuration, both data and parity are stored on all disks in the array. The I/O per second capability of the array, therefore, includes all disks in the array.

For sequential workloads, the sustained transfer rate is not limited by I/Os per second to each disk in the array. A rule of thumb for sequential read workloads can therefore be specified as 25 MB/s for a RAID array, regardless of the disk capacity used in the array.

For mixed sequential and random workloads, the sequential I/O load will expend some of the SSA disk I/O capacity and therefore some of the RAID array I/O capacity.

Select 4.5 G disk drives where appropriate

Since 9 GB drives offer a lower price per megabyte than 4.5 GB drives, the 9 GB drive should be the default choice except where access density requirements dictate the use of 4.5 GB drives.

Migrating data from RAID-1 subsystems may present an exception to this rule of thumb.

Number of RAID Arrays per SSA Loop



- SSA adapter can support one or two loops
 - Adapter bandwidth can be used by I/Os to either loop
 - Adapter cache and Fast Write Cache can be used for data on disks attached to either loop
- Rules of thumb:
 - Use one SSA adapter per drawer for up to 8 drawers
 - ▶ *For high performance configurations, consider using only one SSA loop on adapter*
 - ▶ *Regardless of disk capacity if disk capacity selection driven by I/O rate*
 - Use one SSA adapter per two drawers for 9 to 16 drawers
 - 18 drawer subsystems must have at least two loops that span drawers



© IBM Corporation 1998

Number of RAID Arrays per SSA Loop

Since the SSA adapters used in the VSS can support one or two SSA loops, the number of RAID arrays per SSA loop and therefore the number of RAID arrays per SSA adapter has several effects:

- Each adapter contains 4 MB of Fast Write Cache, which is shared by all RAID arrays on both loops of the adapter. Increasing the number of SSA adapters for a VSS configuration of a given capacity increases the number of 4 MB Fast Write Caches.
- Each adapter, which manages the RAID-5 storage of the attached RAID arrays, has bandwidth limitations, especially for RAID-5 update writes. Increasing the number of SSA adapters provides more RAID-5 update write bandwidth.
- Since each adapter can support one or two SSA loops, there is an SSA bandwidth associated with the adapter. Increasing the number of SSA adapters increases the SSA bandwidth.

SSA adapter supporting one or two loops

Each VSS SSA adapter can support one or two loops.

Adapter bandwidth, except for SSA loop transfer bandwidth, is shared by all RAID arrays attached to either loop.

The cache and Fast Write Cache on the SSA adapter is managed globally for data attached to all RAID arrays attached to either loop. Cache space (and Fast Write Cache space) is not statically partitioned across RAID arrays.

Rules of thumb

The relevant rules of thumb are stated in terms of numbers of drawers in the VSS subsystem regardless of disk drive capacity. It is assumed that 9 GB disks are chosen unless the access density requirements dictate the use of 4.5 GB disks. Since the rules of thumb are driven by the SSA back end I/O workload, the workload is expected to be the same for either 4.5 GB or 9 GB disks.

For configurations containing up to eight drawers, use one SSA adapter per drawer. Configure both RAID arrays (one 6+P and one 7+P) on the same loop so they can share a spare drive.

Where 4.5 GB drives are chosen for reasons other than access density, such as where existing drives are used from an installed 7133 rack, it may be possible to use both loops on an SSA adapter to support two drawers from a single SSA adapter. Even if the rationale for using 4.5 GB drives is unrelated to performance requirements, it is important to determine that performance requirements would not have dictated the use of 4.5 GB drives before configuring more than two RAID arrays per SSA adapter.

Since there can be only eight SSA adapters in a VSS subsystem, configurations using 9 to 15 drawers require using both loops on some SSA adapters. Configurations with 16 drawers require using both loops on all SSA adapters. Where multiple drawers share an SSA adapter, the two RAID arrays in a drawer should be configured in one loop, and the two RAID arrays in the second drawer should be configured in the other loop.

Note that this means that VSS configurations of nine drawers or more will have less SSA back-end bandwidth per gigabyte for those drawers sharing an SSA adapter. As an example, in a nine drawer system, there would be two drawers sharing an SSA adapter, and seven drawers configured with one drawer per SSA adapter. For many workloads, there will be a measurable performance difference between the RAID arrays in drawers with one drawer per SSA adapter and the RAID arrays in drawers with two drawers per adapter.

Since there can be only eight SSA adapters in a VSS subsystem, configurations using 17 or 18 drawers will require two of the eight SSA adapters to attach the RAID arrays in three drawers each.

Workload Characterization Overview



- Read to write ratio
- Synchronous write I/O content
- Sequential I/O content
- Caching characteristics of data



© IBM Corporation 1998

Workload Characterization Overview

In this overview foil, we introduce the topics that will be investigated as part of the workload characterization section of this chapter.

Read to write ratio

In a RAID-5 disk subsystem, because an update write can generate four I/O operations to disk, it is important to understand the read to write ratio of the workload when configuring the subsystem. It is generally true for RAID-5. disk subsystems that the read throughput is greater than the write throughput.

Synchronous write I/O content

The fast write capability of the VSS gives a significant performance boost to applications using synchronous writes, such as the database redo logs of database management systems.

Sequential I/O content

The sequential prestage capability of the VSS in processing sequential reads and the stripe write capability in processing sequential writes significantly improves the throughput of the VSS when processing sequential workloads. Since sequential workloads also tend to use more SCSI port resources, the sequential I/O content of the expected workload should be understood for effective configuration.

Caching characteristics of data

Finally, the caching characteristics of the data, which are application-dependent, is a consideration in proper configuration. No cache management algorithm is going to effectively predict who is next in line to use an automated teller machine, whereas the data associated with many order processing systems displays excellent locality of reference. Where known, these factors should not be overlooked.

Read to Write Ratio



- Significant in RAID disk subsystems
- Random writes
 - RAID-5 write penalty
 - ▶ *Masked by fast write cache in VSS*
 - Except for sectors rewritten while still in Fast Write Cache, all SCSI writes generate SSA disk I/O
 - For high random write content applications
 - ▶ *Consider configuring virtual disks across multiple RAID arrays and combining these disks through a UNIX logical volume manager*
- Sequential writes
 - No RAID-5 write penalty due to VSS stripe writes
 - ▶ *Stripe collected in Fast Write Cache*



© IBM Corporation 1998

Read to Write Ratio

Significant in RAID disk subsystems

In a RAID-5 disk subsystem, an update can generate four I/O operations to disk. The read to write ratio of the workload is important to understand when configuring the subsystem. It is also significant in RAID-1 disk subsystems where writes must be written to both copies.

Random writes

For random writes in a RAID-5. disk subsystem, there is the RAID-5 write penalty—read data, read parity, write data, write parity. Since a single front end write I/O can generate four back end I/Os, the effect of the RAID-5 write penalty on the back end I/O load can be considerable.

In the VSS., the RAID-5 write penalty is masked by fast write cache. Note that it is *masked*, not eliminated.. With fast write cache, a SCSI task-complete indication is sent as soon as the data written is safely stored in SSA adapter cache and Fast Write Cache. As long as the arrival rate of update writes does not exceed the capability of the SSA adapter to destage these writes to disk, the RAID-5 write penalty is effectively eliminated. During write bursts, it may be necessary for SCSI write operations to wait for Fast Write Cache space to be cleared before being processed as fast writes.

The Fast Write Cache management algorithms work to ensure that destage I/O is as efficient as possible. See the chapter Chapter 4, “Versatile Storage Server Data Flow” on page 109 for more information about Fast Write Cache management.

If a block is rewritten before it has been destaged from Fast Write Cache, the VSS will update the SSA adapter cache and Fast Write Cache copy and only one destage will occur. For planning purposes, a conservative assumption would be that all SCSI writes generate SSA disk write I/O.

If the RAID-5 update write bandwidth requirement will be very great, consider configuring multiple VSS virtual disks in different RAID arrays, and then combining these virtual disks using the logical volume manager of the UNIX implementation.

Sequential writes

The stripe write capability of the VSS allows RAID-3-like processing for sequential writes. This can speed database loads and file system recoveries from backups. In a stripe write, data is written to each of the data strips in the stripe and parity is generated from that data and written to the parity strip. No read before write is required. A sequential write not only is not subject to the RAID-5 write penalty, but the striping across multiple drives increases the write throughput beyond what would be possible with a single unstriped disk.

Stripes are collected in Fast Write Cache. Stripe writes in the VSS are not dependent on having SCSI write operations write data in exact stripe increments. Where a SCSI write contains a complete stripe, the VSS storage server recognizes this and passes the data to the SSA adapter accordingly. Where SCSI writes are smaller, the collection of the strips in a stripe occurs in Fast Write Cache and full stripes are destaged to disk as soon as they are collected.

Synchronous Write I/O Content



- Used by database management systems
- Response time sensitive
- Benefit significantly from fast write
 - Task complete sent to the host as soon as data is written to SSA adapter cache and Fast Write Cache
 - VSS accepts responsibility ensuring that the write to disk is complete once task complete is sent to the host
- Fast Write Cache usage



© IBM Corporation 1998

Synchronous Write I/O Content

A workload that uses synchronous writes, as opposed to writes to host cache later written by the synchronization daemon, should particularly benefit from the Fast Write Cache architecture of the VSS.

Used by database management systems

Synchronous writes are used by database management systems to ensure recoverability of data. Writes to database redo logs are usually synchronous writes, while database update writes may or may not be, depending on the environment.

Response time sensitive

Unlike writes to host cache that are written to disk by the synchronization daemon, which are largely throughput- but not response-time sensitive; synchronous writes usually are response-time sensitive. A database management system may wait for a synchronous write to complete before indicating successful completion of an update transaction.

Benefit significantly from fast write

Since fast write returns SCSI task complete as soon as data is safely written to the SSA adapter cache and Fast Write Cache, fast writes appear to complete almost instantaneously.

The database management system wouldn't wait for the completion of a synchronous write if the integrity of the write were not important, often for recoverability. The VSS subsystem accepts responsibility for ensuring that the data is successfully written to disk when the task-complete signal is sent in response to the SCSI request. Storage of two copies in the SSA adapter, one in volatile cache and one in Fast Write Cache, ensures that write integrity can be ensured even if there is a single component failure. This write integrity includes ensuring that associated parity data will also be written to complete a RAID-5 update write.

Fast Write Cache usage

Synchronous writes are typically response-time sensitive, but since they generally occur in smaller, more frequent bursts than the write bursts associated with a synchronization daemon, they are less likely to experience any Fast Write Cache constraints.

Sequential I/O Content



- Sequential I/O
 - Typically large amounts of data
 - ▶ *megabytes to hundreds of megabytes in a stream*
 - Throughput sensitive more than response time sensitive
- Sequential detect
 - Sequential prestage does not need large subsystem cache to be effective
 - Data sequentially prestaged preferentially LRU destaged to avoid flooding cache
- SCSI host attachment utilization
 - Ensure that adapters configured support required bandwidth
 - ▶ *especially if SCSI-2 rather than UltraSCSI*



© IBM Corporation 1998

Sequential I/O Content

The sequential versus random I/O content of your workload should be understood whenever possible.

Sequential I/O

Sequential I/O is characterized by the transfer of large amounts of data. From a disk subsystem perspective, sequential I/O refers to retrieval of data in LBA sequence. Where a UNIX host does sequential read-ahead, its read-ahead is based on the logical structure of the files in the file system, which may be stored in a fragmented manner on disk and which may be striped across several logical volumes.

Sequential I/O is typically throughput sensitive over many I/O operations rather than response time sensitive for each I/O. We often speak of the *sustained data rate* in discussing sequential throughput, which underscores the throughput versus response time focus. A fast sequential I/O is of little value if it cannot be followed immediately by another fast sequential I/O.

Sequential detect

The VSS has a sequential detect algorithm that detects sequential access patterns even when there are intervening I/Os for other data from the same host. Sequential prestage, when invoked, attempts to read the disks in the RAID array in parallel to increase the sustained data rate achieved in the sequential retrieval.

Sequential prestage does not require a large subsystem cache to be effective. Sequential prestage attempts to stay up to 1 MB ahead of the SCSI read requests driving the sequential prestage.

A requirement for sequential processing alone will not dictate a need for a large VSS subsystem cache.

Data that has been read through sequential prestage is destaged more quickly than data staged to cache through random reads and writes, which avoids flooding the cache with data read sequentially.

SCSI host attachment utilization

On the other hand, sequential I/O by its nature can generate a high utilization of SCSI host attachments. A VSS virtual disk that is often scanned sequentially as part of a data mining application can effectively use all the bandwidth of the SCSI port and bus.

Review requirements for sequential processing, including data mining, decision support applications, database and file system backups, and large sort tasks. Ensure that the virtual disks owned by a VSS storage server side are partitioned across enough SCSI ports attached to the required host (or hosts) to support the sequential bandwidth. Remember that a SCSI-2 host adapter will see significantly lower sequential throughput than an UltraSCSI host adapter.

Caching Characteristics of Data



- VSS adaptive caching
 - Selects best caching algorithm for data
 - ▶ *Record caching*
 - ▶ *Partial 32 KB track staging*
 - ▶ *Full 32 KB track staging*
 - Improves effectiveness of storage server cache
- VSS caching of data written
 - Enables read cache hit if data is reread
- Dependent on access pattern



© IBM Corporation 1998

Caching Characteristics of Data

Ultimately, the caching characteristics of data are application-dependent. However, the VSS has several capabilities designed to ensure that storage server caching is effective as possible.

VSS adaptive caching

The adaptive caching capability of the VSS is designed to ensure that the most effective caching algorithm is used, based on the access patterns to the data. The algorithm will select one of three staging algorithms for staging data to storage server cache.

Only the blocks read in a SCSI I/O request are placed in storage server cache. This uses the least room in cache and therefore allows data to remain in cache for as long as possible to enable a hit if the data is reread by the same or a different host.

Or, the blocks read and the rest of the 32 KB track they reside in are placed in storage server cache. This choice is best where the data exhibits locality of reference and blocks adjacent to one already read are often later referred to.

In some cases, the locality of reference is not just forward in LBA sequence but to blocks clustered around the blocks read. Or, where indicated, the VSS will read not just the block requested but the entire 32 KB track into storage server cache.

Adaptive caching improves the effectiveness of storage server cache because it intelligently chooses a staging algorithm to fit the access pattern that has been observed for that data.

For more information about VSS adaptive caching, see Chapter 4, “Versatile Storage Server Data Flow” on page 109.

VSS caching of data written

The VSS caches not only data that is read, but also data that is written. This is done to enable a read cache hit if the data is reread after it is written, either by the same or a different host.

Dependent on the access pattern

No caching algorithm is best for all access patterns; however, for all caching algorithms there will be a worst case access pattern. By dynamically selecting among three caching algorithms, the VSS adapts to the access pattern of the data stored and attempts to maximize the effectiveness of the disk subsystem cache for your data.

Other Performance Considerations



- Race conditions
- Migrating data from RAID-1 file systems to RAID-5
- Parallel query processing



© IBM Corporation 1998

Other Performance Considerations

In this section of the chapter, we will discuss some miscellaneous considerations related to VSS performance.

Race conditions

Race condition is the term used to describe contention for resources where the contenders are unequal in speed. We will see that race conditions resulting from the attachment of hosts of different processor speeds, and with different SCSI attachment capabilities, do not affect the VSS.

Migrating data from RAID-1 file systems to RAID-5

There are special considerations when migrating data from a RAID-1 file system to RAID-5 storage. The RAID-1 “write both, read either” capability delivers significant read bandwidth. If this read bandwidth is required for data stored in a VSS, the use of 4.5 GB disks should be considered.

Parallel query processing

Considerations relative to the definition of virtual disks in the VSS, when used with parallel query processing are discussed.

Race Conditions



- Heterogeneous environments
 - Multiple hosts with a variety of processor speeds share a VSS subsystem, or
 - Mixture of UltraSCSI and SCSI-2 attachments to VSS
- Shared SCSI bus environments
 - Each SCSI port receives I/O requests
 - I/O requests are queued
 - Queued I/O request are processed



© IBM Corporation 1998

Race Conditions

Race condition is the term used to describe contention for resources where the contenders are unequal in speed.

As we will see, race conditions are not an issue for the VSS. This information is included to address concerns about VSS performance in heterogeneous environments.

Heterogeneous environments

Heterogeneous environments include:

- Multiple hosts with a variety of processor speeds sharing a VSS subsystem
- A mixture of UltraSCSI and SCSI-2 attachments to VSS

Shared SCSI bus environments

Such heterogeneous environments do not require special planning as long as each VSS SCSI port is used by a single host only. If there are multiple hosts contending for the SCSI bus through SCSI bus arbitration, race conditions may indeed occur.

If configured as suggested, each VSS SCSI port is in communication with one and only one host, and will receive SCSI I/Os in whatever sequence they are presented by that host.

I/O requests are queued in the VSS. Any SCSI port can add requests to the queue; there is no race condition between VSS SCSI ports.

I/O requests are selected and processed by an available request server, in the order they were placed on the queue.

While an UltraSCSI attachment will outperform a SCSI-2 attachment, it will not monopolize I/O request processing. Therefore, while a VSS subsystem may have a mixed environment of UltraSCSI and SCSI-2 attachments, I/O requests from all hosts will be serviced.

Migrating Data from RAID-1 File Systems



- RAID-1 (software mirroring or in hardware)
 - Both copies updated when written
 - Either copy read when read
 - ▶ *Significant read bandwidth*
- RAID-5 provides a level of fault tolerance comparable to RAID-1
- Consider read bandwidth
- Consider 4.5 GB disks



© IBM Corporation 1998

Migrating Data from RAID-1 File Systems

There are special considerations when migrating data from a RAID-1 file system to RAID-5 storage. The RAID-1 “write both, read either” capability delivers significant read bandwidth. If this read bandwidth is required for data stored in a VSS, the use of 4.5 GB disks should be considered.

RAID-1 (software mirroring or in hardware)

This may be an issue whether the RAID-1 is implemented as logical volume mirroring or in a hardware RAID-1 implementation.

In a RAID-1 implementation, an update write must update both copies of data. In a read, however, either copy may be read.

This provides significant read bandwidth, both for random reads and for sequential reads.

Consider read bandwidth

When migrating from RAID-1 storage, consideration should be given to the read bandwidth requirements for both random and sequential reads to ensure that adequate bandwidth can be provided by the VSS. In many cases, the read bandwidth provided by RAID-1 is not exploited, but it is important to know whether or not it has been in a given installation.

Consider 4.5 GB disks

When migrating from RAID-1 storage, consider using 4.5 GB capacity disks to increase the back-end disk bandwidth.

Use of Parallel Query Processing



- Database management parallel query processing
- A VSS virtual disk is defined as a physical disk to the host system
- A VSS virtual disk is a partition of a RAID-5 array
- A VSS RAID array equivalent to physical disk
- Plan mapping of virtual disks to RAID arrays



© IBM Corporation 1998

Use of Parallel Query Processing

Support for parallel query processing by a database management system is easily provided by the VSS, provided that consideration is given to the definition and placement of virtual volumes on the VSS.

Database management parallel query processing

Database management systems performing parallel query processing assume knowledge of physical disk drives capable of independent operations in order to exploit the read bandwidth capabilities of the disk storage used.

A VSS virtual disk defined as a physical disk to the host system

A virtual disk, or partition of a RAID array, is defined as a physical disk to the host system.

A VSS virtual disk as a partition of a RAID-5 array

The virtual disk that the host system sees as a physical disk is really a partition of a RAID array. More than one virtual disk can share a RAID array.

A VSS RAID array equivalent to physical disk

For the purposes of the database management system in exploiting read bandwidth, the closest equivalent in a VSS subsystem to a physical disk is a RAID array.

Plan mapping of virtual disks to RAID arrays

As long as only one virtual disk on a RAID array is used for parallel query data storage, the query optimization of the database management system will be effective. Placing two virtual disks on the same RAID array will limit the sequential throughput of both to about the same as a single virtual disk, which may cause the database management systems optimization to underperform.

It is acceptable to define additional virtual disks on the same RAID array as a virtual disk used for parallel query processing. The situation to avoid is when two virtual disks are sharing the same RAID array but are both defined to the host as independent physical disks, causing the parallel query software to believe that they can effectively be read in parallel with no loss of bandwidth.

Performance Review



- Storage server cache
 - Improved performance for read cache hits
- Adaptive cache
 - Increased number of read cache hits in storage server cache
- Sequential prediction and sequential prestage
 - Improved sequential read throughput
- Fast write
 - Masks RAID-5 write penalty
- SSA adapter Fast Write Cache
 - Protection of fast write data



© IBM Corporation 1998

Performance Review

In this foil, begin to summarize this chapter by reviewing at a high level the key performance capabilities of the VSS. The text here is repeated from “Performance Highlights” on page 254.

Storage server cache

The VSS can be configured with a large storage server cache. The primary advantage of this disk subsystem cache is to provide read cache hits for reads not satisfied from the host cache.

Adaptive cache

The VSS storage server cache is managed by an adaptive caching algorithm which determines how much data should be staged to storage server cache when data is read from the disk backstore. The storage server can either stage just the block or blocks read, the block or blocks read and the balance of the 32 KB track, or the entire 32 KB track containing the requested blocks.

The key advantage of this adaptive caching is that it intelligently manages which data is placed in cache. Adaptive caching can increase the number of read cache hits given the same size storage server cache.

Sequential prediction and sequential prestage

The VSS detects sequential access to the disk in logical block address (LBA) sequence, and will begin read-ahead prestaging once sequential access is detected. With sequential prestaging, the sustained sequential data transfer rate for sequential reads is increased.

Fast write

The fast write capability of the VSS allows it to return a SCSI task-complete indication when data has safely been stored in both cache and nonvolatile storage (Fast Write Cache) in the SSA adapter. The fast write capability allows the VSS to mask the overhead associated with RAID-5 update writes.

SSA adapter Fast Write Cache

Fast write wouldn't be very attractive if the integrity of data written to the disk subsystem were exposed and vulnerable until the data was committed to disk. The Fast Write Cache architecture of the SSA adapter used in the VSS protects the integrity of data written as soon as a task- complete indication is sent to the host in response to the SCSI write command.

Summary



- Overall system performance
- Disk subsystem cache
- Improved performance through Fast Write
- Configuration options:
 - Storage Server cache size
 - 4-way symmetric multiprocessors (SMPs) in the storageserver
 - 4.5 or 9 GB capacity disk drives
 - Bandwidth of SCSI attachments per host
 - Disks per SSA loop and SSA adapter
- Configuration flexibility



© IBM Corporation 1998

Summary

Finally, we review the key points made in this chapter.

Overall system performance

Disk performance is an important part of overall system performance, but it is only part of overall system performance. A system performance bottleneck will not be addressed by a faster disk subsystem if the bottleneck is not related to disk I/O.

Disk subsystem cache

Even when there is effective host caching, a large external disk subsystem cache can be used to improve I/O performance. Where host caching is less effective, which may be for a variety of reasons, an external disk subsystem cache can be highly beneficial to disk I/O performance.

Improved performance through Fast Write

The fast write capability of the VSS will improve performance for many workloads, especially database transaction systems which use synchronous writes.

Configuration options

- Storage server cache size
- 4.5 or 9 GB capacity disk drives
- Bandwidth of SCSI attachments per host
- Disks per SSA loop and SSA adapter

Configuration flexibility

The VSS has a flexible architecture that can be scaled in subsystem capacity and can be configured independently of subsystem capacity to meet your performance needs.

Chapter 9. Versatile Storage Server Maintenance

Maintenance



Maintenance



© IBM Corporation 1998

Overview



- Philosophy
 - Concurrent
 - Disruptive
- Repair Actions
 - Customer
 - CE
 - PE
- Interfaces
 - HTML browser
 - ▶ *TCP/IP based client PC or workstation*
 - ASCII terminal CE maintenance interface
 - Remote access



© IBM Corporation 1998

Overview

In this chapter, we discuss maintenance of the VSS subsystem. This foil provides an overview of the topics we discuss in greater detail throughout the chapter.

Philosophy

One of the main driving forces behind the VSS was the need to provide uninterrupted service to the owner or user of the storage subsystem. Typically, maintenance on most components of a computer system causes some sort of interruption to service as an engineer replaces a part or a system administrator reconfigures around a degraded peripheral or component.

The VSS is designed to cause as little disruption as possible when maintenance is being performed. Being able to replace components while the subsystem is running is known as *concurrent* maintenance.

Components that are most likely to fail can be replaced without disruption of service. The components most likely to fail in the VSS are the disk drives. The 7133 disk drawer is designed to facilitate hot replacement and hot insertion of drives and its power supplies. The RAID-5 adapter can detect a failed disk drive and initiate *sparing*—the process whereby a good spare drive is automatically brought online into an array to take over from a failed drive.

Although the VSS is designed to be easy to maintain and to minimize disruption of service while maintenance is performed, customer repairs are limited to field disk drive modules, 7133 power supplies, and limited application of code engineering changes (ECs). A qualified IBM CE or PE is required to perform all other hardware and software maintenance of the VSS.

Repair Actions

There are three distinct levels of repairs and maintenance actions that can be performed on the VSS. These are as follows:

- Customer repairs. Repair actions for the customer are limited.
- CE repairs. Typically, most failures that occur on a VSS will be repaired by the CE.
- PE repairs. The PE will perform code (EC) functions, and provide high-level support.

Repair actions are explained in detail beginning with “Repair Actions – Customer” on page 327.

Interfaces

A number of interfaces are provided to allow access for maintenance purposes:

- Web browser. A web-based HTML browser is the primary interface for configuring the VSS.
- ASCII terminal. A serial port is provided for each storage server to enable the CE and PE to access error logs, configure error reporting, run diagnostics, and upgrade microcode.
- Remote support interface. A serial port is provided for each storage server to allow attachment of a modem to enable remote access by an IBM support representative. The IBM support representative can examine error logs, configure error reporting, and run diagnostics without needing access to customer data.

Overview ...



- Reporting
 - Error log
 - Email
 - SNMP
 - Call home
 - Sparing
 - Hot spare available
 - No hot spare available
 - Upgrades
 - Hardware components
 - Code EC Management
 - EC process
 - Interfaces
 - Distribution
-



© IBM Corporation 1998

Overview ...

Reporting

Depending on the configuration of the customer's site and network, the VSS provides different levels of error reporting to accommodate any situation:

- Error log. The VSS constantly monitors itself and logs any errors and unusual occurrences in its error log.
- SNMP. If the customer has a large site that uses the SNMP protocol for network management, the VSS can be configured to send SNMP alerts to the network management stations.
- E-mail. When a problem is reported, the VSS can e-mail a list of users to inform them of the problem.
- Call home. Through the modem connected for remote access, the VSS can automatically dial and connect to the local IBM support office and log a call for service. The appropriate IBM CE is then dispatched with a replacement part, or a support representative can dial in to examine the error logs and make recommendations.

Sparing

The SSA RAID adapter has built-in sparing capability. Sparing is the process whereby the adapter detects a failed disk drive, and automatically brings online a good spare that has been configured in the array. Missing data from the failed drive is reconstructed using the XOR facility of the RAID-5 adapter and written to the spare disk. The call-home facility alerts IBM and a CE is dispatched with a new drive module, which replaces the failed drive and becomes the new spare for the array. Procedures are also available for when no hot spare exists in the array—typically because the spare has replaced a failed drive that has not yet been replaced.

Upgrades

Upgrade of the components of the VSS is made easy by the concurrent maintenance features of the subsystem. All components of the VSS that are designed to be replaced while the subsystem is running can be upgraded while the subsystem is running: disk modules, disk drawer power supplies, the storage servers, clusters, adapters and rack power supplies. In addition, system microcode can be upgraded while components are running, through the code EC management process.

As when components fail, some degradation of subsystem capabilities may be noticed when components are taken off line for upgrade.

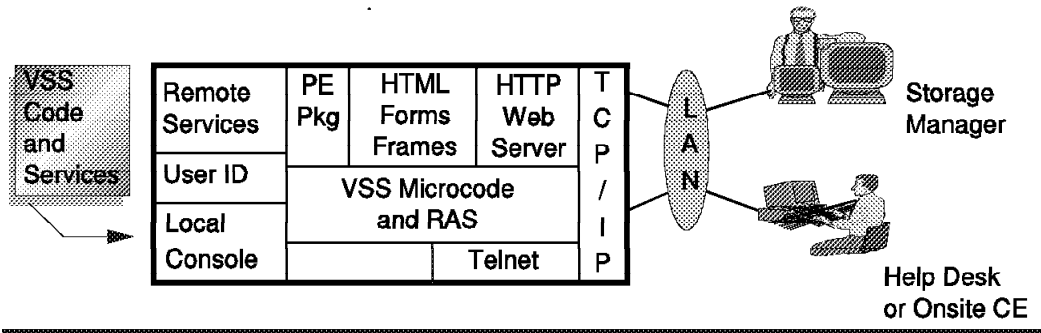
Code EC management

The code EC management process exists to provide concurrent management of licensed internal code (LIC). It allows the interrogation, modification, and maintenance of the installed ECs. The EC process is available through the standard interfaces: web-browser, CE maintenance interface, and remote support interface. ECs are distributed on CD-ROM, diskette, and network.

VS Specialist



- HTML browser
 - Text markup, user interface
 - CGI binary utilities
 - TCP/IP based client
 - Requires forms and frames
- Status Screen
 - Graphical array display
 - Interactive drive table
 - Faulty disk in red
 - Sparring procedure
 - Error reporting options



© IBM Corporation 1998

VS Specialist

On this foil we discuss the web-based interface available to access the configuration and maintenance options of the VSS. This interface is typically used by both the customer and IBM support representatives.

HTML browser

The standard interface to the configuration of the VSS is via a web-based HTML browser, so called because it uses the World-Wide Web style of hypertext and information browsing currently available on the Internet and intranets. HTML is the text markup language that is the front end of all web applications. The VSS configuration menus, written in HTML, provide the user interface. The data provided by the user is passed through the Common Gateway Interface (CGI) to binary utilities that perform the actual configuration and repair actions of the VSS.

The HTML browser is a TCP/IP based application that runs on a client PC or workstation. To fully utilize the configuration menus and screens, the browser chosen to perform the configuration and maintenance tasks must support HTML forms and frames. For this reason, character-based browsers such as Lynx are not supported.

Status screen

The main part of the display shows a graphical representation of the arrays configured in the VSS. If a drive is faulty, it will flash red in this display, immediately alerting the viewer to the condition. When a user chooses a drive, the interactive drive information table shows the drive serial number, location, status, and a host-by-host listing of which volumes are on the drive. It also contains a procedure to “spare the drive out,” if the drive needs to be replaced and has not already been spared by the SSA adapter.

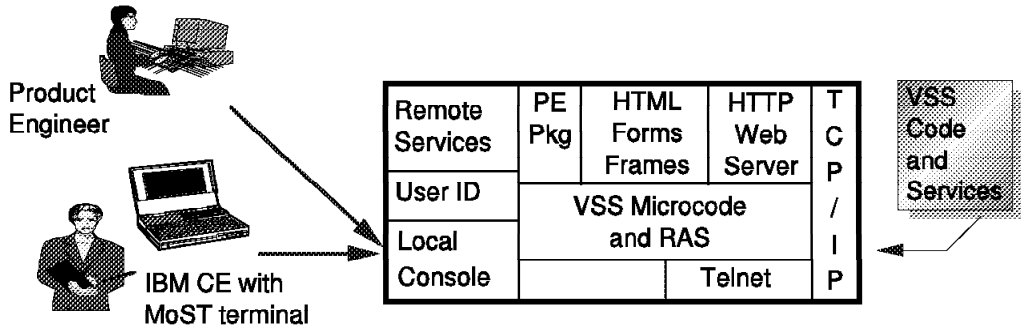
The maintenance options for the VSS are available from the system status screen. From here, the following options are available:

- Change error reporting
- Allow IBM service access to your system
- Allow remote configuration
- View error logs

On-site Maintenance Interface



- ASCII Terminal
 - Customer supplied
 - CE or PE laptop with MoST terminal
 - Connects through switch to both controllers
- Character-based
 - Perform service procedures
 - Initial installation and setup



© IBM Corporation 1998

Onsite Maintenance Interface

This foil illustrates the onsite maintenance or CE interface, which uses the RS-232 serial ports of each storage server. The interface provides lower-level access to maintenance and repair functions.

ASCII terminal

The services provided through this interface are designed to be displayed on character-based asynchronous ASCII terminals, such as the IBM 3151 or DEC VT100. Typically, the CE or PE will use this interface while performing onsite maintenance. The ASCII terminal can be supplied by the customer, or the CE or PE can plug in a laptop containing the mobile solutions terminal (MoST) emulator. The terminal connects through a null-modem cable to an RS-232 switch that is in turn connected to a RS-232 serial port on each storage server cluster. Either storage server cluster can be selected from the switch, without having to physically unplug the terminal cable. The switch is located in the rear of the 2105-B09 cabinet.

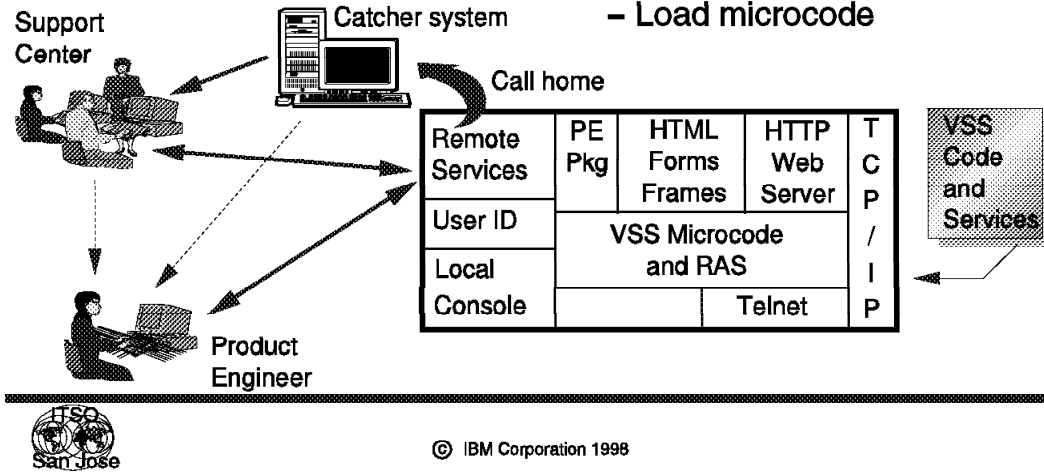
Character based

The character-based menus are similar to the AIX SMIT menus, and allow the CE to run service procedures such as diagnostics and component swapping. The character interface is also used to install the VSS operating microcode and for initial setup of the VSS. The PE can also install microcode updates from the CE interface.

Remote Support Interface



- Call home
 - Async modem dial out
 - *Automatic notification of problem or potential failure to IBM*
- Support center
 - CE or PE
 - Examine error logs
 - Run diagnostics
 - Reconfigure
 - Load microcode



© IBM Corporation 1998

Remote Support Interface

The remote support interface provides dial-in access to the maintenance services of the VSS, and dial-out services for the VSS to “call home.” The remote support interface uses another of the RS-232 serial ports of each storage server cluster. A modem is connected to each cluster of the VSS, allowing the cluster to call home, and allowing service personnel to dial in. The modems are not powered by the VSS. They connect through a standard RS-232 asynchronous modem cable. The remote interface is character-based, providing the same menus and screens as the onsite CE maintenance interface.

Call home

The call-home facility provides automatic notification of problems or potential problems to IBM, as well as notifying the customer by e-mail or SNMP alert. The VSS RAS code of each cluster constantly monitors and logs errors or events that occur. The logs are analyzed on a regular basis by a separate RAS process, and any immediate problems or trends that may cause problems are noted. If necessary, another RAS process dials out to the closest IBM support center, where it communicates with a “catcher system” that logs the data passed to it by the VSS. The catcher system then informs support center personnel of the problem. The support center will then dial back in to the VSS and examine the error logs, making the necessary recommendations. If needed, the support center notifies the PE, who also dials in to examine the problem, or the support center can dispatch the appropriate CE with a replacement for the failing part.

Support Center

Through the remote support interface, the support center can perform a set of tasks similar to those of an onsite CE or PE. CE functions such as error log examination and running of diagnostics do not require or allow access to customer data, maintaining system security. On the other hand, some of the PE functions such as reconfiguration and microcode loading or updates will require that the PE have the root password of the storage server to which he or she has dialed in. The customer can configure access for both CE and PE. In the event the PE requires access to the root password, the customer can enable the access and have it automatically revoked after 8 hours.

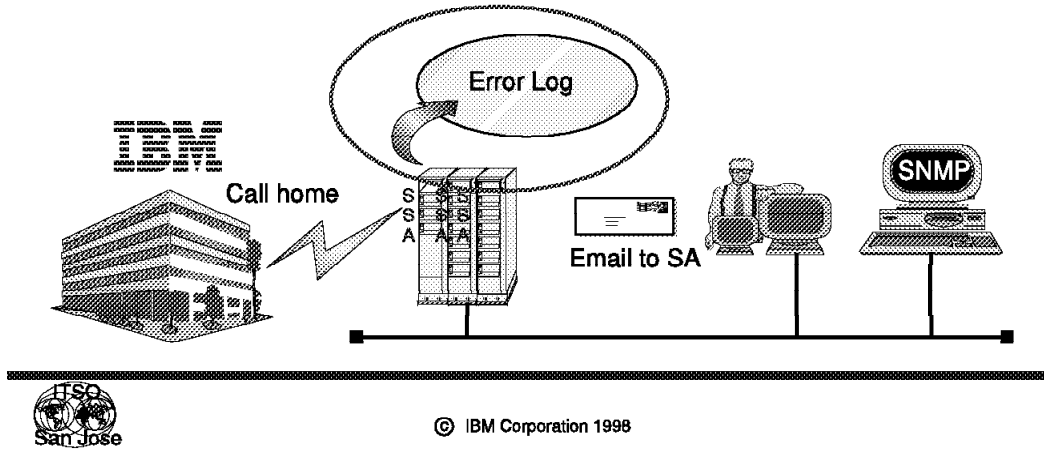
Limited support for a storage-server-down situation is provided through the storage-server-service processor. The storage server must still be functional at the service processor level—typically this would be in the event of a microcode failure. Access via the service processor is limited to the storage server NVRAM, VPD, and error data.

For code ECs, the remote support interface supports only emergency microcode fixes, because of the limited data rate of asynchronous modems.

Reporting - Error Log

IBM
VERSATILE STORAGE SERVER
SERIES 3500
SERIES 3600
SERIES 3700
SERIES 3800
SERIES 3900
SERIES 4000
SERIES 4100
SERIES 4200
SERIES 4300
SERIES 4400
SERIES 4500
SERIES 4600
SERIES 4700
SERIES 4800
SERIES 4900
SERIES 5000
SERIES 5100
SERIES 5200
SERIES 5300
SERIES 5400
SERIES 5500
SERIES 5600
SERIES 5700
SERIES 5800
SERIES 5900
SERIES 6000
SERIES 6100
SERIES 6200
SERIES 6300
SERIES 6400
SERIES 6500
SERIES 6600
SERIES 6700
SERIES 6800
SERIES 6900
SERIES 7000
SERIES 7100
SERIES 7200
SERIES 7300
SERIES 7400
SERIES 7500
SERIES 7600
SERIES 7700
SERIES 7800
SERIES 7900
SERIES 8000
SERIES 8100
SERIES 8200
SERIES 8300
SERIES 8400
SERIES 8500
SERIES 8600
SERIES 8700
SERIES 8800
SERIES 8900
SERIES 9000
SERIES 9100
SERIES 9200
SERIES 9300
SERIES 9400
SERIES 9500
SERIES 9600
SERIES 9700
SERIES 9800
SERIES 9900
SERIES 10000

- Error log
 - Tracks all subsystem errors
- Error log analysis
 - Determines if a service action is required
- Problem record generation
 - Problem record generated
 - *Contains relevant data about the problem*
- Service alert



Reporting – Error Log

The VSS has a number of different notification methods for error reporting, providing flexibility to the customer. The customer can elect to use all or some of the methods depending on individual configurations and site requirements. The error-log facility is the basis upon which all reporting and maintenance is built.

Error log

The VSS has code for logging, tracking, and managing errors and events that occur. Each storage server runs a copy of the same code and keeps track of its own errors, both hardware and software (microcode). In addition, the two storage servers of the VSS monitor each other across the Ethernet interface using TCP/IP and UDP/IP protocols. Configuration options include the log file name, maximum size of the log, and the maximum memory buffer size. The memory buffer is a small (8192 bytes by default) circular buffer in the VSS operating kernel to which errors are written. A separate process monitors the kernel buffer, extracts entries that are written to it, and logs them. The next step in the process is to analyze the logs.

Error log analysis

The error log analysis process examines the error log to determine if a service action is required. Typically, the following conditions require service actions:

- A hardware resource has a permanent failure.
- A threshold for recoverable hardware errors has been reached.
- A threshold for recoverable firmware or microcode errors has been reached.
- A data exception condition exists where data associated with a logical volume has been permanently lost. (Data exception conditions are reported and handled by the customer, and do not involve a CE action unless the customer specifically requests assistance from the CE.)

When a service action is required, a problem record is created and a service alert is generated.

Problem record generation

Problem record generation executes when it is determined by the error log analysis that a service action is required. First, the process determines whether a problem record already exists for the problem in the problem log. If not, a new problem record is generated and logged. Only one problem record exists even if the failure condition results in multiple error log entries in the error logs of one or both storage server clusters. The problem record contains the following data:

- A summary of the error data that caused the problem record to be generated
- Data required for presenting the service alert
- Problem state and age information
- Data required by the CE to perform maintenance procedures
- Actions to resolve the problem
- Additional data for the PE for resolution of problems that cannot be solved by the CE.

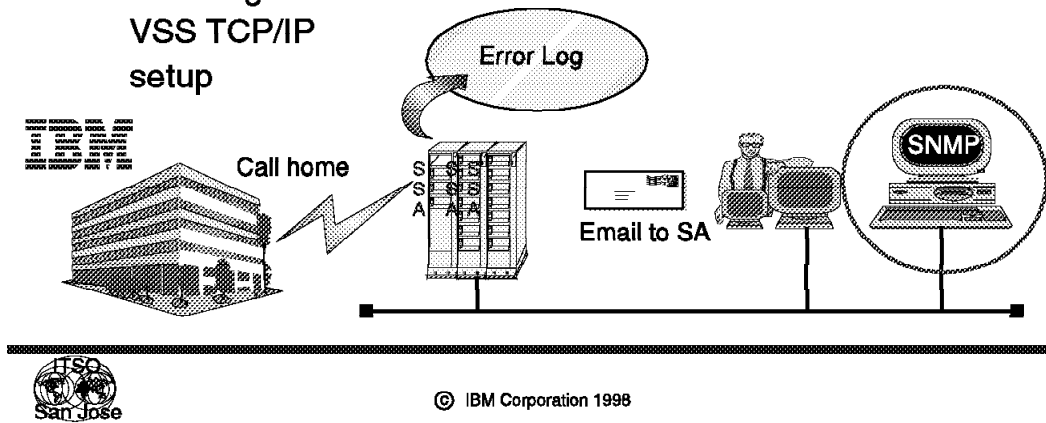
Problem records are available for viewing by the customer, CE, and PE through all interfaces.

Once the problem record has been generated, a service alert is sent, according to one of the methods discussed on the next three foils.

Reporting - SNMP

IBM Corporation
San Jose, CA
© IBM Corporation 1998

- **SNMP**
 - VSS can alert a network management station upon error or threshold exceeded
 - Requires IP address of NMS registered in VSS TCP/IP setup
- **Two MIBs**
 - Configuration information
 - Pending problems
 - *SRNs, FRUs, ESCs*
 - *Previous two traps*



© IBM Corporation 1998

Reporting – SNMP

If the customer's intranet is managed by, or has an SNMP-capable network management station, the VSS can be configured to send SNMP traps to the station upon detection of an error or a threshold reached.

Simple network management protocol

SNMP is a common network management protocol, typically used in large networks, both homogeneous and heterogeneous, to manage all network entities from a single point (typically a network management station). SNMP traps are the primary method of notification of errors from the VSS. The customer's network management station must have its IP address registered in the TCP/IP configuration of the VSS in order to receive SNMP traps from the VSS. The network management station will show that there is an error condition on the VSS requiring intervention.

Two management information bases

There are two MIBs for the VSS. An MIB is simply a collection of objects that describe an SNMP-manageable entity. The first MIB contains configuration information, allowing a user of an SNMP network management station to display the configuration of the VSS.

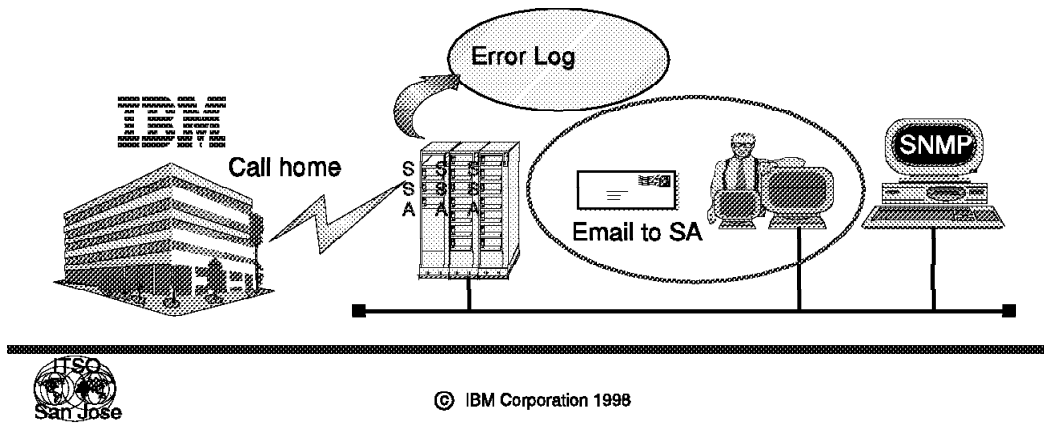
The second MIB contains information about pending faults. The information allows the user of the network management station to display any pending

service alerts, and any service request numbers (SRNs), field replaceable units (FRUs), and exception symptom codes (ESCs) associated with the pending service alerts. The user can also display data for the last two trap events.

Reporting - Email

Email notification

- Optional
- Sent on service alert
- Requires SMTP gateway
- Standard internet address - **username@host.domain**



© IBM Corporation 1998

Reporting – E-mail

The VSS RAS code can be configured to send e-mail to a specified user or users, or list of users when a service alert is generated. E-mail configuration is part of the **Change Error Reporting** option of the **System Status** screen. Use of the e-mail facility requires a simple mail transmission protocol (SMTP) mail gateway or server on the local intranet to handle forwarding of the mail sent from the VSS. E-mail addresses are standard internet "@" format—*username@host.domain*.

When the storage server microcode is not running, but the service processor is still functional, the service processor can initiate a call home. The data provided is only that which is available in the storage server nonvolatile RAM. The NVRAM data consists of:

- Customer ID
- The VSS callback password
- Time stamp
- Callback telephone number (to the VSS)
- Account telephone number (to the customer)
- VSS serial number
- Subsystem LCD code (error or halt code)
- Minimal problem description if available (up to 64 characters)
- RETAIN login ID number (assigned during RETAIN registration)

The call is screened by the support center and the appropriate IBM representative is contacted to provide support. This support may take the form of a CE arriving on site with a replacement for the failing part, or a support representative dialing in to examine error logs and subsystem status. The call-home facility is configured from the **Configure Error Reporting** option from the **System Status** menu.

Repair Actions - Customer



- Console specification
- Logical configuration
 - Disk arrays and LUNs
- Limited physical configuration
 - Drives
 - Host connections
 - TCP/IP, users and passwords
- Limited repair
 - Problems associated with drives
- Subsystem code EC management
 - Query and install



© IBM Corporation 1998

Repair Actions – Customer

Soon after general availability of the VSS, limited repair actions of the VSS are available to the customer. This foil discusses the requirements for the customer to perform repairs and the actions that the customer can take.

Console specification

To perform repairs on the VSS, the customer must have a console running a web browser capable of handling HTML forms and frames. It attaches to the customer intranet, and communicates with the hypertext transfer protocol (HTTP) server running on the VSS. If the customer wishes to install code ECs distributed on diskette and CD-ROM, the console must have the appropriate drive installed. If the customer plans to install code ECs over the network, an Internet or IBM-NET connection is required.

Logical configuration

The customer is able to perform all logical configuration tasks on the VSS. Logical configuration is defining resources to the host systems that are accessing the VSS. The customer is able to perform the following logical configuration tasks:

- View, define, change, and remove disk arrays.
- View, create, change, and remove LUNs.
- View, define, change, and remove configuration data.

Limited physical configuration

The customer is able to perform some limited physical configuration functions. Physical configuration is configuration of the actual physical resources of the VSS. The actions that the customer is able to perform are as follows:

- View VSS storage server configuration.
- Add or remove disk drives.
- View, define, change and remove host connections.
- View, define, change, and remove TCP/IP configuration, users, and passwords.

Limited repair

The customer is able to perform some limited repair actions of the RAID disk arrays in the VSS.

All actions apart from the following are performed by the CE or PE:

- Display problem information associated with disk drives.
- Determine disk drive problems that are customer repairable.
- Request that a failed disk be conditioned for repair.
- Upon completing a disk repair, identify the repaired drive.
- Return the repaired drive to the RAID array.
- Close the associated problem.

Subsystem code EC management

The customer can perform subsystem code EC management functions through the web interface. The functions supported are as follows:

- Query subsystem code EC levels.
- Determine whether EC is concurrent or not.
- Copy code EC to VSS from distribution media—diskette, CD-ROM or network.
- Install and activate new EC code.
- Restore previous code EC level.
- Code EC installation recovery if problems are encountered.

Repair Actions - CE and PE



- Service procedures
 - Service processor menus
 - ▶ *Communicate with service processor only*
 - SMS menus
 - ▶ *Controller firmware based*
 - Online menus
 - ▶ *Configuration and repair menu*
 - ▶ *Diagnostics and Service Aids menu*
 - Hardware tests
 - POST
 - Isolation tests
 - Machine checkout tests
 - Repair verification tests
-



© IBM Corporation 1998

Repair Actions – CE and PE

The CE and PE are able to perform various repair and maintenance functions in addition to those the customer can perform. The primary interface for the CE and PE is through the CE ASCII terminal interface. Many of the CE repair actions are designed to be performed concurrently. The PE typically provides higher-level support and code EC management tasks.

Service procedures

The service procedures that the CE is required to perform comprise a number of different “levels.” The first of these is access to the service processor of the storage server, typically when the storage server microcode is not running (either after a fatal error or prior to booting the storage server). Access to the service processor provides limited access to the VPD, NVRAM, and some error data. Typically, access to the service processor is warranted only after an internal error causes the microcode to halt.

The second level of service procedures are those performed through the system management services (SMS) of the storage server firmware. Access to the SMS is when the storage server microcode is not running. The SMS can be used to provide access to the storage server bootlist (the list of devices that the storage server attempts to boot from, in order), extended memory tests, and update storage server firmware.

The third level of service procedures is through the online menus that are accessible when the storage server microcode is running. There are two types of menus, the configuration and repair menus and the diagnostics and service aids menus. The menus are similar to the AIX SMIT menus. From the configuration and repair menus, the CE is able to install the storage server microcode and perform initial configuration, logical and physical subsystem configuration, all subsystem repairs, as well as manage code EC, copy error and problem logs to diskette, and view or modify subsystem VPD.

The diagnostic and service aids menus are similar to those of the RS/6000 online diagnostic and services menus. They allow the CE to run online diagnostics on the storage server, and access low-level functions of the subsystem, such as viewing and setting the storage server bootlist, and viewing and setting NVRAM.

Code EC Management



- Supported interfaces
 - Web and CE
 - Remote support
 - Code EC process
 - View subsystem LIC levels
 - Copy update to disk for future installation
 - Install and activate LIC update
 - Restore LIC to previous level
 - Clean up after failed or interrupted update
 - Release process and media types
 - CD-ROM
 - Diskette
 - Network
-



© IBM Corporation 1998

Code EC Management

Code EC management involves the ability to view and update the various microcode levels of subsystem components. Many code EC updates are concurrent; that is, they can be performed without completely removing access to customer data. In some cases, a slight degradation of performance may be noticed while the update is performed.

Supported interfaces

The web-browser interface discussed under “VS Specialist” on page 314 and the CE onsite maintenance interface discussed under “Onsite Maintenance Interface” on page 316 are the primary supported methods for code EC management. Through the web-based interface, the customer has limited access to EC management. The CE interface supports only ECs from CD-ROM, diskette, and from previously loaded file sets.

The remote support interface provides limited support for code EC management. Typically, asynchronous modems do not provide enough bandwidth to upload large file sets, so only emergency code fixes will be supported through the remote support interface.

Code EC process

Five separate functions are available as part of the code EC management process. These are as follows:

- View subsystem LIC levels — This function provides the capability to display EC levels of subsystem components, relevant to EC management and for support purposes.
- Copy an update to disk for future installation — This function provides the ability to load ECs to the storage server's internal hard disk drive from removable media or network, without installing them. Installation can be scheduled at a later date and time, or the file sets can then be transferred to the other storage server for future installation.
- Install and activate LIC update — This process allows the installation of an EC previously loaded to hard disk, or from removable media. The EC is first installed and then applied. During the installation process, the data being updated is saved to allow restoration in the event of failure of the new EC. The application process activates the new code.
- Restore LIC to previous level — As discussed above, if a new EC fails after installation (that is, the problem is not solved or, worse, a different problem is created), the EC can be backed out and the previous level restored.
- Clean up after failed or interrupted update — In the event of an interrupted or failed update, it is possible that partially installed ECs may cause problems. The clean-up option is available to restore the subsystem to a known working state.

Release process and media types

CD-ROM is the primary distribution method for code ECs. CD-ROMs can accommodate large distributions and are required for sites that do not have a connection to the EC distribution network. A CD-ROM drive is provided as part of each storage server. Network distribution is the secondary method of distribution for ECs. WWW, Internet, and IBM-NET connections are supported. The customer has limited access to ECs, although all EC information is provided. Network distribution ensures access to the very latest fixes as soon as they are available. ECs are distributed on diskette where the fix is small enough to fit—typically firmware updates.

Chapter 10. Subsystem Recovery

Subsystem Recovery



- **Types of failure**
 - **host connection adapter**
 - **storage server**
 - **disk or disk adapter**
 - **power system**
- **Data integrity**
- **Data availability**
- **Concurrent maintenance**



© IBM Corporation 1998

Subsystem Recovery

In this chapter we discuss how to recover from a failure within the VSS. From a host-system perspective VSS is “just another disk storage system.” However, resilience, availability, and serviceability (RAS) are built into VSS, so that most types of possible failure are masked from the host operating system and allow normal data processing to continue. If a failure should occur, the RAS systems will ensure that repair actions can take place while the system is online.

Data integrity is key to VSS; in all circumstances, once the host has been signaled that the write is complete, the integrity of the data is assured.

However, the most common cause of data loss and corruption is human error. As VSS cannot protect data from this, it is vital to have backup and recovery plans and procedures in place, and to test them. Therefore, all the normal rules governing data backup, restore, and disaster recovery plans still apply.

The VSS has built-in diagnostic and error warning systems. It can return error codes relating to failing components back to the attached hosts, send out an SNMP alert, and dial out and send messages to remote maintenance personnel.

Types of Failure

Within the VSS storage subsystem, there are four main categories of possible failure:

- Failure of the host adapters in the PCI bridge
- Failure of the storage servers
- Failure of the disks or disk adapters
- Failure of the power system

We discuss each of these types of failure in detail, how to recover from them, and the effect on system availability.

Data integrity

VSS has been designed so that in any set of circumstances, no single failure will result in any loss of data integrity. Data integrity is assured for any completed I/O operation.

Data availability

In most circumstances, data availability is always maintained. When repair action is taken, it may be necessary to restrict data access to the failed area within VSS while concurrent maintenance is taking place.

Concurrent maintenance

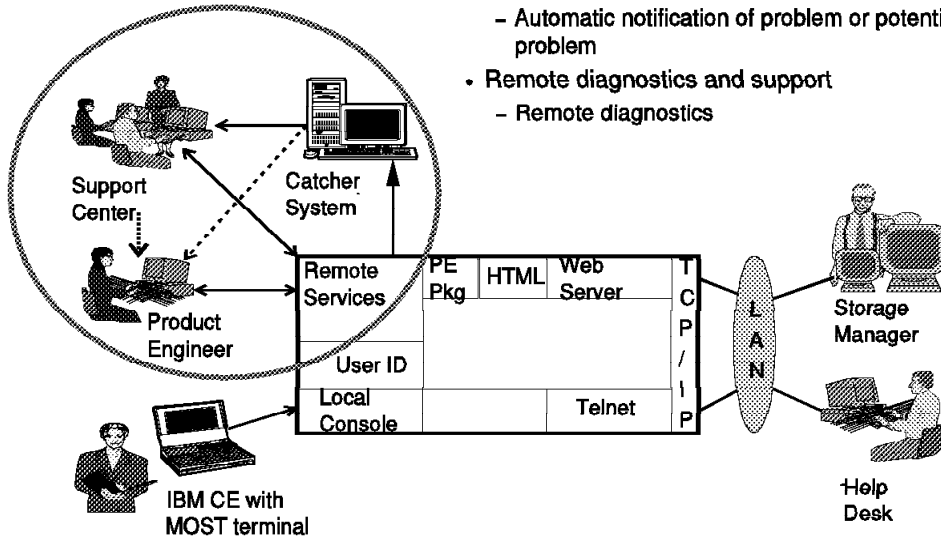
VSS has been designed so that nearly all maintenance can be carried out without having to power the system down or take it off line.

Remote Services



- Access through phone line and modem

- Call home support
 - Automatic notification of problem or potential problem
- Remote diagnostics and support
 - Remote diagnostics



© IBM Corporation 1998

Remote Services

VSS is not designed for customer setup or maintenance. However, it is equipped with sophisticated and easy-to-use interfaces to detect, monitor, and rectify faults.

Access through phone line and modem

VSS provides an RS232 connection which can be used to connect a phone line. The phone line does not have to be dedicated, although a dedicated line is recommended as it ensures that fault conditions are reported as soon as they are detected.

Call-home support

This is a continuous self-monitoring system that initiates a call to the service provider if a failure or potential failure is detected. With this feature, a service technician can automatically be sent out to replace the failed or failing component and can bring along the required replacement parts. Repair time is thus minimized.

Remote diagnostics and support

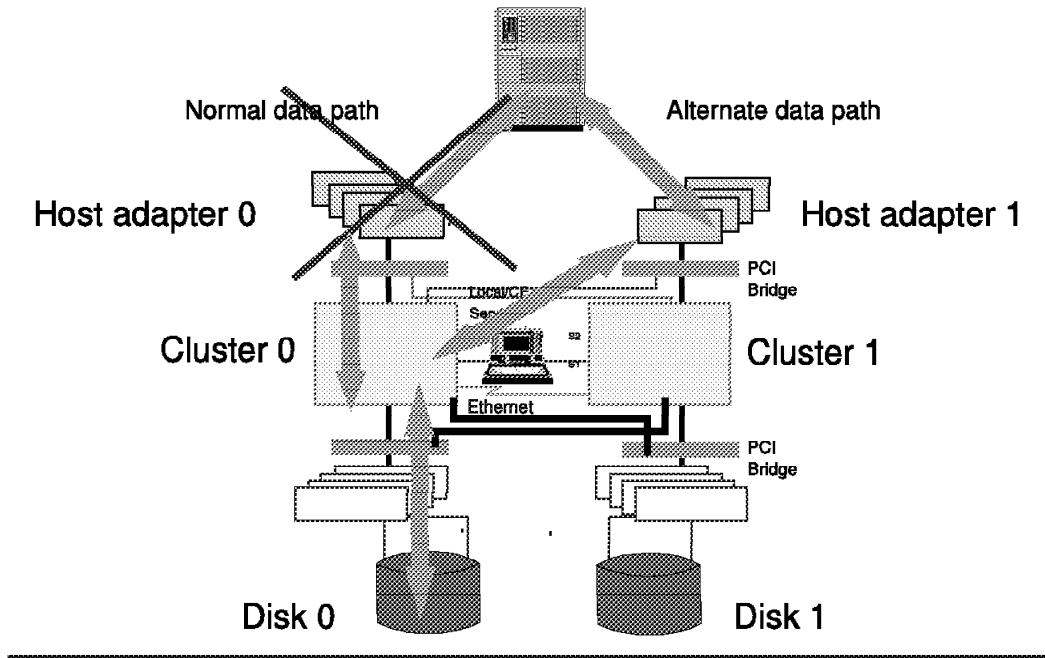
Using the RS232 connection, the service provider can initiate a call to the VSS storage server and remotely analyze potential failures and determine the type of failure and the proper fix. This is particularly important when microcode problems are suspected. Using the remote support facility, IBM can initiate a call to the VSS storage server and correct many types of microcode problems. When a microcode problem is reported, a fix can be created and then applied from a remote location.

It is important to stress that this remote access does not compromise security of the data stored on disk.

There are two levels of support that can access the system via the RS-232 port. The normal customer service engineer (CE) support does not have access to any customer data. The product engineer (PE), the highest level of support, may require root access to the system, which will give him or her access to the data. This level of access must be specifically unlocked by the customer on site; it relocks automatically after 8 hours. For all support levels, the customer has to authorize the use of remote support. If remote support is activated, then the customer is notified by e-mail. Call-back modems and other security procedures (such as passwords) should be used to prevent unauthorized access to the system.

Both the user and the CE can configure the VSS, assigning storage to individual hosts. The configuration manager can be accessed from the subsystem interface, the RS232 port, or through a customer-controlled network (Ethernet based). Using an existing customer network (intranet) allows the user to manage the complex from the normal desktop. The configuration manager is a web-based GUI application, providing an easy-to-use interface into the VSS.

Host Connection failure



© IBM Corporation 1998

Host Connection Failure

There are three possible causes of a host connection failure:

- Failure of the adapter within the host
- Failure of the connecting cable
- Failure of the host connection adapter in VSS

In all cases, the result is the same: the host is not able to read or write data to the storage subsystem. There is no issue of data integrity, as any in-flight transactions that are taking place when the failure occurs will abort. The application running on the host will not receive an affirmative return code from VSS and so will take its normal recovery procedures. In order to ensure continuous data availability, two paths using different adapters on the host and VSS can be used.

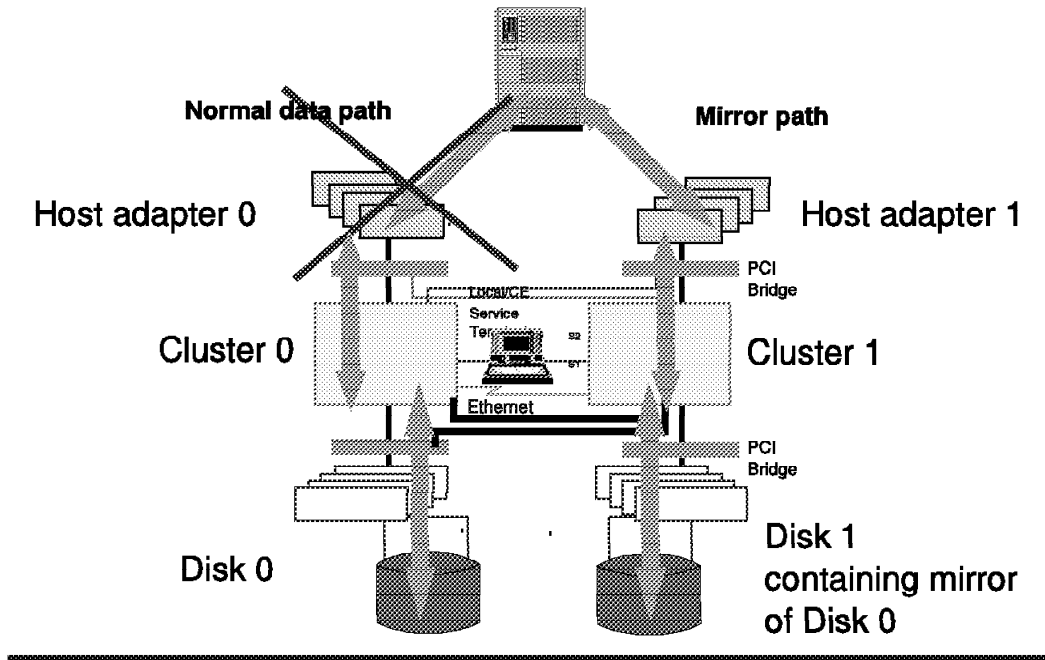
Note This requires the host operating system to have adapter failover capability or some other method of defining two different routes to the same piece of data.

VSS has been designed so that concurrent maintenance can take place. This means that in the event of host connection adapter failure, the failing component can be replaced without the need to power VSS down. Maintenance procedures for this are fully described in Chapter 9, "Versatile Storage Server Maintenance" on page 309. It must be understood by the operators that while VSS has been designed for concurrent maintenance and hot pluggability of components, host

operating system requirements may require that user and application access be restricted while maintenance is under way.

Disk mirroring

IBM
CORPORATION
San Jose, CA
© 1998

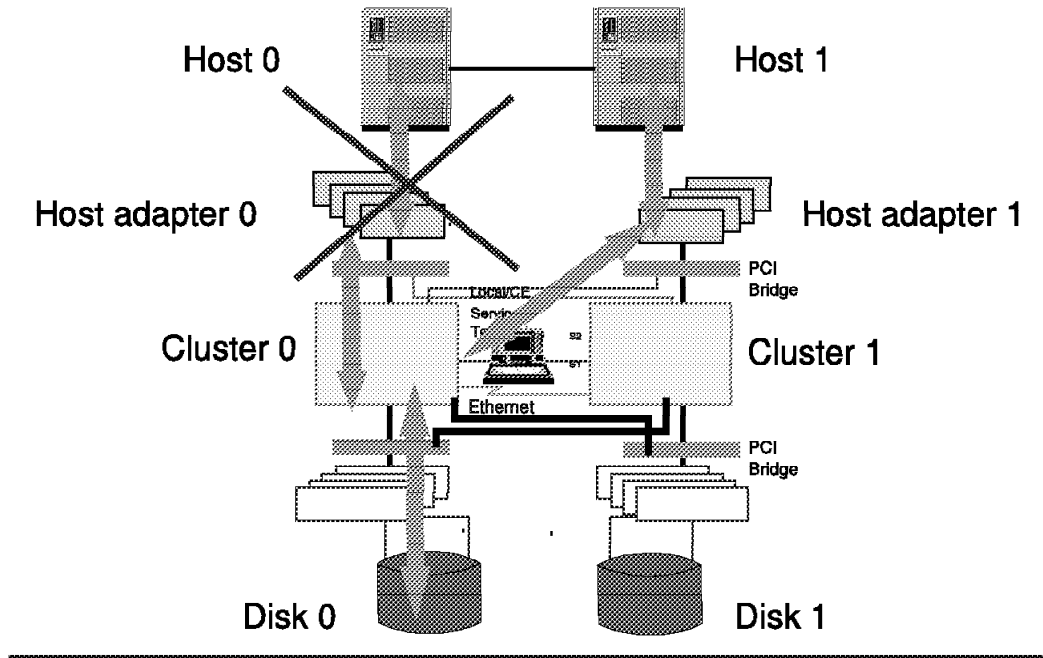


© IBM Corporation 1998

Disk Mirroring

If the host operating system does not support adapter failover, an alternative approach is disk mirroring. In this foil we show two alternative paths to two mirror copies of the data. Data access is maintained in the event of a failure in one part of the host adapter subsystem.

High Availability with Multihost Dual Access

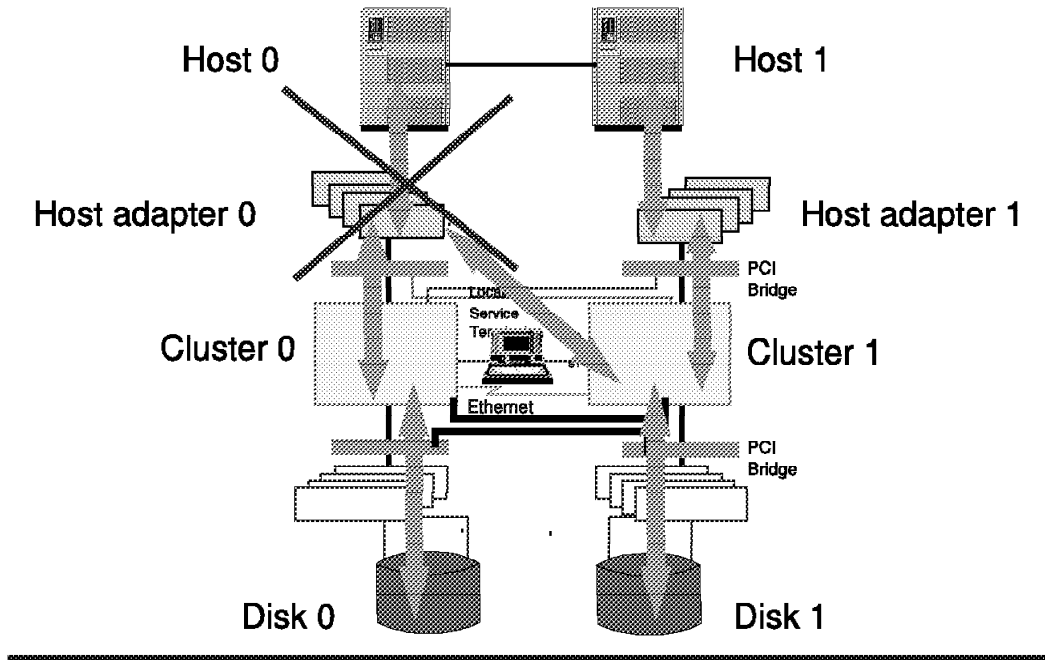


© IBM Corporation 1998

High Availability with Multihost Dual Access

In this foil, we show how we can use VSS to provide high availability. VSS can be configured so that the same data area on a disk can be shared by two different processors. If either Host 0 or Host adapter 0 fails, then Host 1 will be able to process the data on Disk 0. If storage server Cluster 0 fails, then automatic failover to storage server Cluster 1 will occur and both hosts will be able to access Disk 0.

HACMP with Mirroring



© IBM Corporation 1998

HACMP with Mirroring

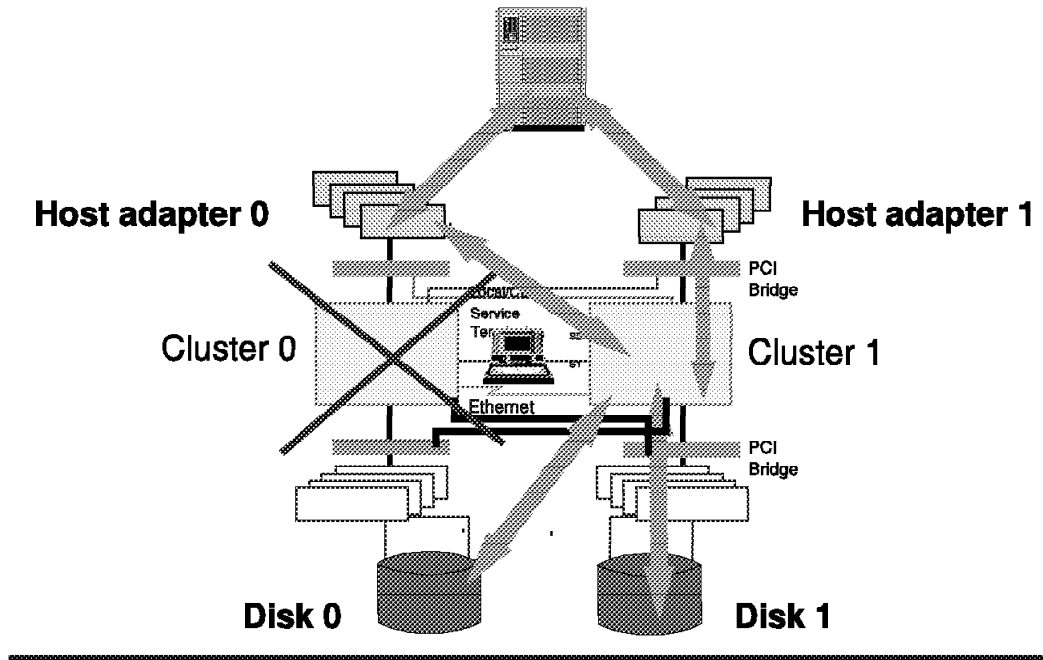
In this foil, we show how we can use VSS to provide high availability. VSS can be configured so that the Host 0 can write data to Disk 0 with a mirror copy on Disk 1. In the event either Host 0 or Host adapter 0 fails, then Host 1 will be able to process the data on Disk 1.

A normal operation scenario is shown with Cluster 0 controlling Host and Disk adapters 0.

Note Any host adapter can be “owned” by either cluster; ownership is set up during configuration. Disk adapters are “tied” to the cluster they are connected to. This is changed only in the event of cluster failover. If Cluster 0 fails, then the adapters will be controlled by Cluster 1.

Storage server Failure

IBM Corporation
1998



© IBM Corporation 1998

Storage Server Failure

In this foil, we show how, in the event of a cluster failure, the normal adapter and cluster relationships change. Cluster 0 has failed. The mirror copy of Cluster 0 microcode stored in Cluster 1 has been activated. Cluster 1 now "owns" all the host and disk adapters. The host is still able to access all the data.

Storage server cache

Any failure within the storage server cache causes the cluster to shift (failover) to the other cluster. Any inflight I/O processing will not complete. Host applications, submitting work through the failed storage server, will not receive return signals from outstanding I/O requests.

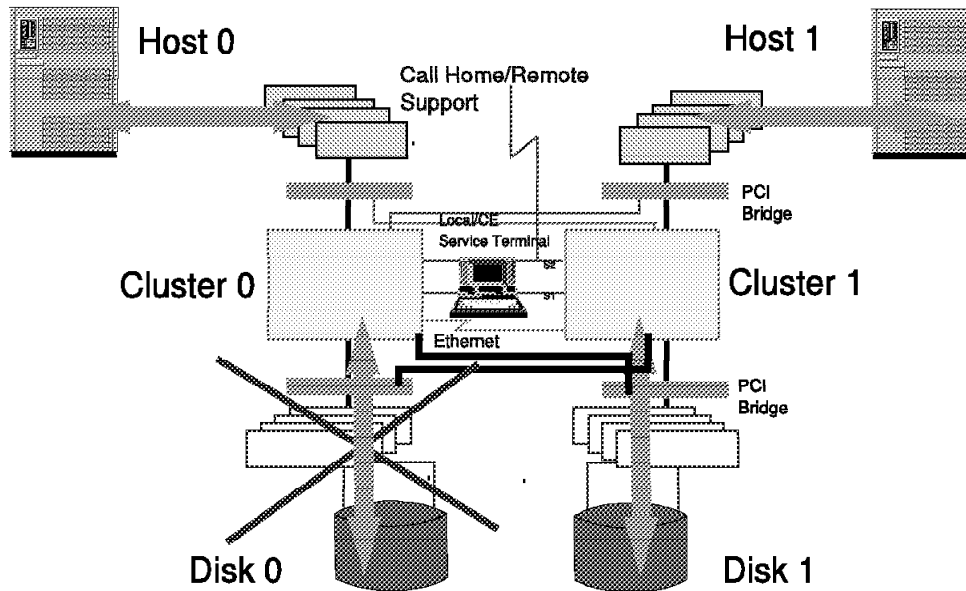
When the I/O request is resubmitted, failover will automatically reroute the requests to the remaining on-line storage server, which will process the request and return a request-complete code to the host.

Storage server processors

Each storage server cluster is of an SMP design, with four-way processors. If any processor in a cluster should fail, the cluster will shut down and all work will failover to the remaining cluster. The failed cluster will then execute an initial program load on itself without restarting the failed processor. Thus, a four-way SMP will restart as a three-way unit. It will then be brought back on-line and resume I/O processing.

Error signals will be generated and dispatched to the service technician or CE. The failed cluster will be replaced without powering the VSS down and access to all the data will be maintained during this operation by the other cluster.

Disk Subsystem Failure



© IBM Corporation 1998

Disk Subsystem Failure

VSS has been designed so that no single failure of the disk subsystem will cause data loss.

Most single failures still allow unrestricted host access to data. The disk subsystem is built on SSA components. Since its introduction in October 1995, over two petabytes of SSA disks have been bought, thus proving the market acceptability and reliability of SSA.

Disk adapter failure

When data is sent from the host server, it is initially stored in the storage server cache and immediately transferred to the cache and Fast Write Cache in the disk adapter. If the adapter fails, then the Fast Write Cache can be removed from the failed adapter and placed in the new one.

On resumption of operation, the Fast Write Cache writes its data to disk; the memory cells within the Fast Write Cache are then marked as available. The write-complete return code is sent back to the host as soon as the data is written to the Fast Write Cache. It is at this point that the integrity of the written data is assured. The disk adapter can be replaced in the VSS without need to power the system down.

The Fast Write Cache in the disk adapter is a mirror copy of the data in the write-through cache in the adapter volatile storage. If the Fast Write Cache

should fail, the adapter will continue to process I/O requests without using the write-through cache. There will be a performance hit in doing this, as every write will suffer the RAID-5 write penalty. An alert and error message that the Fast Write Cache in the adapter has failed will be generated.

If the write-through cache in volatile storage should fail, then the adapter fails and error message and alerts are generated.

Disk drive failure

Data is protected against disk failure through the use of RAID-5 technology. The RAID-5 protection is provided by the dual-loop SSA adapter card. Each RAID-5 array consists of eight drives. There can be up to two arrays per 7133 drawer. The first array in the drawer must be in the form of six drives + parity + spare. The second array can be of this form or can be seven drives + parity. Both of these arrays will be on the same SSA loop and controlled by the same SSA adapter.

All drives in a 7133 should be of the same size. If a disk drive in a RAID array should fail, there is always a hot spare in the loop, so data rebuilding begins immediately after a failure occurs. When the new drive is installed, it becomes the spare, allowing the spare to float among the drives in the array. In fact, the spare will support all the arrays on the loop. The rebuilding process is run as a background task, giving priority to customer requests for data.

SSA cable or connection failure

As the cabling is standard SSA-type loop cabling, then any failure in the loop will cause the adapter to reconfigure the loop into two strings. When the problem has been resolved, the two strings will automatically be reconfigured back into a loop. Thus access to the physical disk drives is always maintained.

SSA disk drawer

The SSA disk drawer is a standard 7133-020 drawer. It comes with redundant power supplies and cooling fans. Because it is designed for on-line maintenance, any problems with these components will not cause any loss of data integrity or availability.

Power System Failure



- Redundant power supplies
- DC power control unit
- Battery backup
- 7133-020 disk draw



© IBM Corporation 1998

Power System Failure

The system is designed so that data corruption will not occur in the event of a power failure. Write-complete return codes are not issued to the host until the data is in Fast Write Cache. If the RAID-5 adapter has not finished its parity write when a power failure occurs, a bit is set in the adapter Fast Write Cache to flag this. When power is resumed, parity is recalculated and the write is completed.

Redundant power

Each VSS rack contains a fully redundant power subsystem. Two 50 or 60 ampere line cords (country and location dependent) provide the power to each rack. Complete system operation can take place on a single line cord. Like the line cords, power distribution units and cooling fans have redundancy, so any single failure will not affect operation. If a power control unit should fail, then the VSS will run from the remaining unit. The failed unit can be replaced without need to power the system down or affect user access to data.

The power subsystem can accept either single-phase or three-phase power. If three-phase power is used, the power control unit will still function even if one of the phases should fail.

DC power control unit

The power control unit provides 350 V DC power to the rack. A DC power bus is fitted to both sides of the rack and all the equipment within the rack is cabled so that it can operate from either bus. In normal operation, both buses are used.

Battery backup

There is an optional battery pack for both the 2105-B09 main rack and the 2105-100 expansion rack. The battery should provide enough power to run the system for 10 minutes after power loss. This battery backup protects the VSS from temporary loss of power in a brownout.

7133-020 disk drawer

The 7133-020 SSA disk drawers have auto-ranging AC/DC power supplies and use the DC power bus. The 7133-020 disk drawer has, as standard, three power supplies and fans. One power supply fan cooling unit is redundant, as the 7133 can function correctly with any two working.

Concurrent maintenance can take place so that users have access to data while a failed unit is being replaced.

Disaster Recovery



- Data recovery
 - Responsibility of each attached host
 - VSS should be treated as a series of SCSI disks
 - Normal backup and restore plans

- Configuration data
 - Small amount
 - stored on diskette
 - diskette drives in each cluster



© IBM Corporation 1998

Disaster Recovery

Many factors have to be considered when planning a complete disaster recovery plan, but most of them are outside the scope of this presentation guide. There are two main considerations when planning for a catastrophic failure of VSS. They are:

- Recovery of data stored in VSS
- Recovery of configuration information of VSS

Data Recovery

It is the responsibility of the host systems that are attached to VSS to back up data that they own. In this respect, VSS should be viewed as a series of standard SCSI disks and all normal operational procedures will apply.

Regular backups should be taken of the data and copies stored off site. Tools such as ADSM are designed to aid this process and they have special modules to cover the possibilities of site failure.

Configuration data

If a replacement VSS must be used before previously backed up data can be restored, it will have to be configured exactly the same as the machine that it is replacing. Configuration data can be backed up on VSS using the diskette drives that are part of each cluster. The amount of data is relatively small as only that data relating to host attachments and logical disk allocation is unique to each system.

Chapter 11. Software Support

Software Support



- SSA RAID management
 - SSA software support
- Device driver support
 - Current Support
 - ▶ *AIX*
 - ▶ *HP/UX*
 - ▶ *Solaris*
 - ▶ *OS/400*
 - ▶ *DG/UX*
 - ▶ *Windows NT*
 - Planned Support



© IBM Corporation 1998

Software Support

This foil shows the topics that we discuss in this chapter.

SSA RAID management

We first explain the SSA RAID management software installed in the VSS.

Device Driver Support

We next discuss the platforms and operating systems supported by the device drivers of the VSS. VSS supports Solaris, HP-UX, AIX, OS/400, DG/UX, and Windows NT platforms. We also indicate the platforms and operating systems expected to be supported in future releases of the Versatile Storage Server.

SSA Software Support



Configuration Methods

- ▶ **SMIT**
 - Configure RAID array
 - SSA disk and RAID array repair actions
 - Show status of SSA RAID arrays and SSA disks

Device Drivers

- ▶ **Four-port SSA RAID adapter device driver**
- ▶ **SSA router**
- ▶ **SSA physical disk device driver**
- ▶ **SSA logical disk device driver**

Adapter+Firmware

- ▶ **RAID-5 function support**



© IBM Corporation 1998

SSA Software Support

The software for the SSA subsystem includes the following components, which must be installed in the Versatile Storage Server.

Configuration methods software

The configuration methods (SMIT) configure the SSA four-port RAID adapters in the VSS subsystem and the drives connected to the ports of these adapters.

Device Drivers

- **SSA RAID adapter device driver**

This device driver supports the SSA RAID adapter, providing access to the adapter for communications and for managing the adapter itself.

- **SSA router**

The SSA router has the function of handling the SSA frames that flow in the SSA loop. The SSA router receives every SSA frame coming through the SSA loop, sending it on to the adjacent target (or initiator) if the frame is not for that router, or keeping it if it is.

- **SSA physical disk device driver**

This driver handles the physical disks.

- **SSA logical disk device driver**

This driver handles the RAID array. The RAID array must be configured as 6+P+S or 7+P, and the logical unit is the group of seven or eight physical disks that store logically related data. This device driver configures the RAID array, or reconfigures it if necessary.

Adapter

The SSA RAID adapter software carries out RAID-5 implementation of functions such as parity processing and data reconstruction.

Current Support



Hewlett Packard

- HP 9000 800
- HP 9000 D,E,G,H,I,K,T Series
- HP 9000 EPS
 - ▶ HP-UX 10.01
 - ▶ HP-UX 10.10
 - ▶ HP-UX 10.20
 - ▶ HP-UX 10.30

Sun Microsystems

- Sun sparc 1000, 1000E, 2000,
- 2000E, 3000, 4000, 5000, 6000
 - ▶ Solaris 2.5.1, 2.6

Compaq

- ProLiant 3000,5000,5500,6500,7000
 - ▶ NT 4.0

Data General

- AViiON 4900, 5000
 - ▶ DG/UX 4.2

IBM

- RS/6000
- SP
 - ▶ AIX 4.1.5
 - ▶ AIX 4.2.x
 - ▶ AIX 4.3
 - ▶ AIX 4.3.1
- AS/400 (9406)
 - ▶ OS/400 V3R1
 - ▶ OS/400 V3R2
 - ▶ OS/400 V3R6
 - ▶ OS/400 V3R7
 - ▶ OS/400 V4R1
 - ▶ OS/400 V4R2
- PC
 - ▶ 325,704,3500,5500,7000
 - ▶ NT 4.0



© IBM Corporation 1998

Current Support

This foil lists platforms and operating systems that the current release of VSS supports. It includes a variety of models from Hewlett Packard, Sun Microsystems, Compaq, Data General, and IBM, using UNIX-related, RISC-based, OS/400, and Windows NT operating systems.

Planned Support



NCR

- 3455, 3555, 5100
 - AT&T System V 3.4 and later

Digital

- Alpha Series
 - Digital UNIX, Windows NT

SGI

- Challenger Series
 - IRIX 5.2 and later

IBM

- PC server 704, Netfinity
 - Novell NetWare 4.1 and later

Sequent

Pyramid

* Other platform support will be available as needed - please refer to <http://w3.ssd.ibm.com/pendisk/vss> for current list and procedures to request support



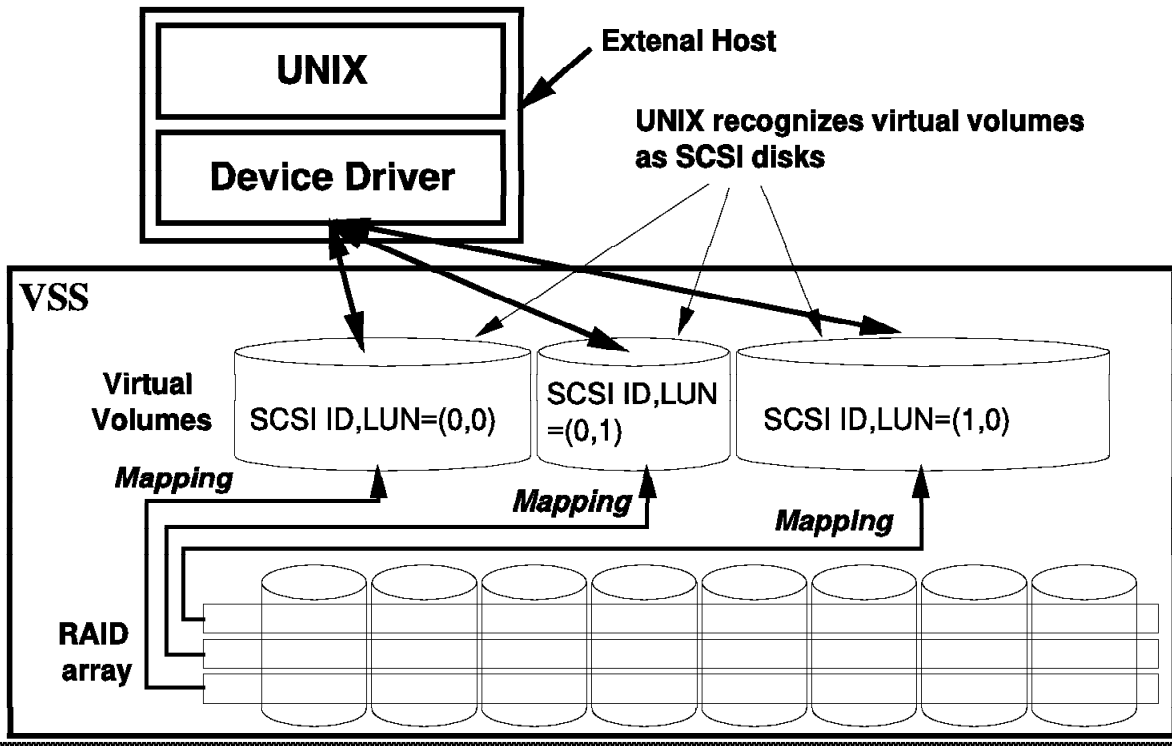
© IBM Corporation 1998

Planned Support

This foil lists platforms and operating systems that may be supported in the future. The order in which they are listed does not indicate the sequence (or even the likelihood) of their eventual release.

Support for other platforms is expected to become available over time as needed. Please check the Web site <http://w3.ssd.ibm.com/pendisk/vss> for the current list and procedures to request support.

Device Driver Function (UNIX)



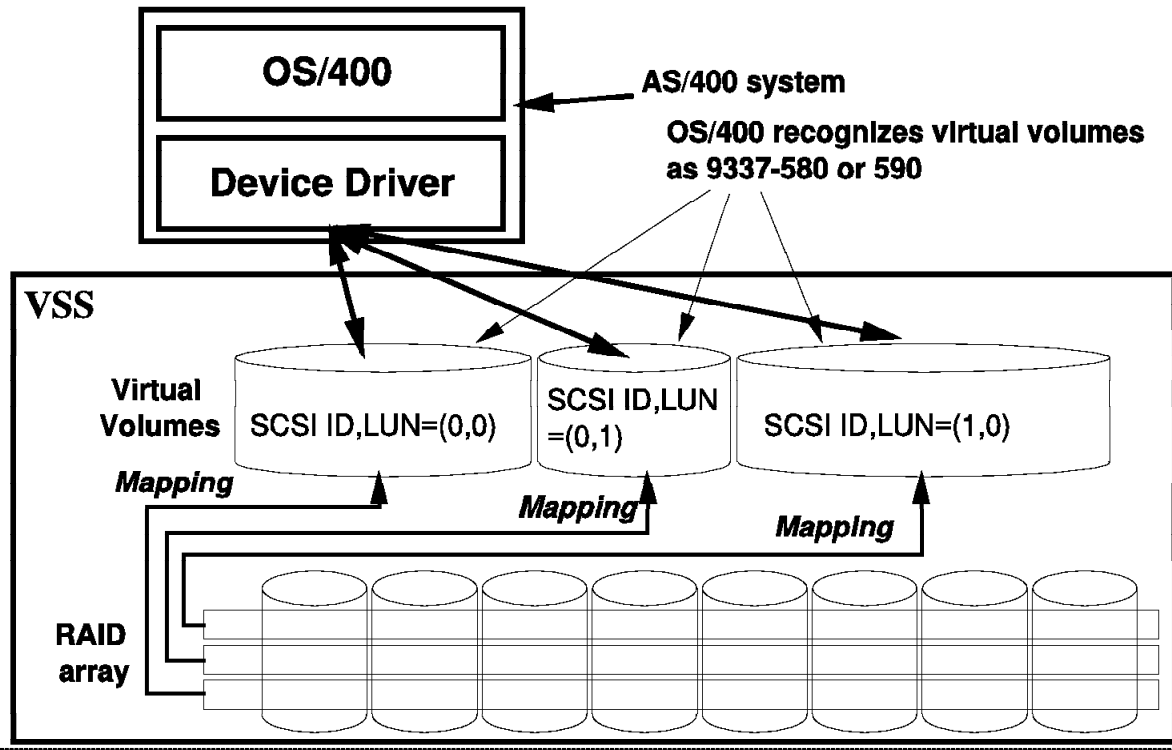
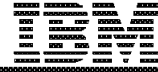
© IBM Corporation 1998

Device Driver Function (UNIX)

This foil shows the device driver function that should be installed in an external UNIX-based host such as HP-UX, Solaris, or AIX. The example shows how the host device driver serves as an interface between VSS and the UNIX operating system. The device driver can send commands and data to the VSS storage server through the SCSI bus and receive requested data.

The UNIX operating system recognizes the virtual volume mapped by the VSS storage server as one physical, generic SCSI disk. The host does not know whether the RAID-5 function is performed on this SCSI disk drive, or on the virtual volume.

Device Driver Function (OS/400)



© IBM Corporation 1998

Device Driver Function (OS/400)

This foil shows the device driver function that should be installed on OS/400 to use the VSS. The OS/400 recognizes the VSS as a 9337-580 or 9337-590 disk subsystem, so it can access or handle the VSS storage as it does the 9337-580 or -590 disk subsystem. The device driver can send commands and data to the VSS and vice versa. The device driver deals with the virtual volumes that are created from the RAID array by the VSS storage server.

Appendix A. Special Notices

This publication is intended to help customer management and technical support staff evaluate and plan for the implementation of IBM's Versatile Storage Server storage subsystem. The information in this publication is not intended as the specification of any programming interfaces that are provided by the IBM Versatile Storage Server.

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 pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

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 terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

AIX

AS/400

BookManager	DB2
ECKD	ES/9000
ESCON	IBM
Magstar	MVS
Netfinity	OS/400
Predictive Failure Analysis	PROFS
RAMAC	RETAIN
RS/6000	S/390
System/390	Ultrastar

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc.

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

Microsoft, Windows, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

Java and HotJava are trademarks of Sun Microsystems, Inc.

Other trademarks are trademarks of their respective companies.

Appendix B. Related Publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

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

Other Publications

These publications are also relevant as further information sources:

- *VSS Executive Overview*, G225-6718
- *Seascope Architecture Presentation*, G325-3347
- *VSS Introduction and Planning Guide*, GC26-7223

These publications are shipped with the product:

- *VSS User's Guide*, SC26-7224
- *VSS Host Systems attachment Guide*, SC26-7225
- *VSS SCSI Command Reference*, SC26-7226

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
1-403-267-4455
(+45) 48 14 2207 (long distance charge)

- **1-800-IBM-4FAX (United States) or (+1)001-408-256-5422 (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
IBM Direct Publications Catalog

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

IBM Redbook Order Form

Please send me the following:

Title	Order Number	Quantity

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

• Invoice to customer number _____

• Credit card number _____

Credit card expiration date Card issued to Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Index

Numerics

- 2105-100 expansion rack 172
- 2105-100 rack 46, 49
 - components 46
 - maximum usable storage size 49
- 2105-100 racks 46
 - placement distance 46
- 2105-B09 components. 172
- 2105-B09 rack 45, 47
 - components 45
 - dimensions 47
 - power supply 47
- 7131 Model 105 storage tower 5
- 7133 disk drawer 310
 - maintenance 310
 - hot replacement of drives and power supplies 310
- 7133 disk drawers 44
- 7133 Model 20 6
- 7133 Model 600 6
- 7133 SSA drive adapter 80
 - dummy modules for slots without drives 80
- 7133-020 77, 78, 349
 - improved cable connector numbering 77
 - improved front cover and on-off switch 77
 - power supply and cooling redundancy 349
 - SSA bypass cards 78
- 7133-020 disk drawer 347
- 7135 RAIDiant array 6
 - dual active controller configuration 6
 - for RS/6000 attachment only 6
- 7137 RAID 6, 7
 - RAID-5 or RAID 0 configuration 6
 - redundant power supplies 7
 - storage for open systems platforms 6
- 7204 external disk drive 5

A

- access skew smoothing 259
- adaptive cache algorithms 169
 - locality of reference 169
- adaptive cache management 139, 140, 170
 - algorithms 140
 - statistical basis 140
 - sequential prediction capability 170
 - use of one caching algorithm for all data in a band 139
- adaptive cache read options in VSS 131
- adaptive caching 296
- adaptive caching algorithm 134, 138
 - dynamic selection of best algorithm 138
- AIX system 266
 - use of memory for host cache 266

- application servers 205
 - view of VSS disk drives 205
 - view of VSS RAID arrays 205
- AS/400 host perception of VSS 231
 - emulation of 9337 subsystem 231
- ASCII terminal 316
 - as maintenance interface 316

B

- bibliography 363
- business data 9
 - full-time availability 9

C

- CE 311
 - maintenance and repairs 311
 - centrally managed common storage pool 22
- character-based browsers 314
 - lack of support for 314
- code EC management 313
- common gateway interface 314
- Compaq Proliant software needs to support VSS 228
- concurrent maintenance 310

D

- data loss and corruption caused by human error 333
- data migration 298
 - RAID-1 to RAID-5 298
- data migration to VSS 222
- data sharing 14
 - disk assigned to more than one host 14
 - locking during updates 14
- database management system 267
 - specification of host cache size 267
- DB2 Parallel Edition 267
- disk drawer power supplies 313
- disk drive buffer 176
- disk drive failure types 179
- disk modules 313
- disk subsystem cache size 269
 - relation to host cache size 269
- DRAM 114

E

- e-mail customer notification of problems 318
- engineering change process 313
 - access to 313

F

- Fast Write Cache 63, 114, 120, 121, 255, 286, 346
 - avoiding data flooding 63
 - battery backup 114
 - battery-powered data retention 120
 - mirrors 4MB of SSA adapter volatile cache 121
 - number 286
 - shared by SSA adapter loops 120
- Fast Write Cache battery 120
 - data retention with battery power 120
 - SRAM compared with DRAM 120
 - lithium battery, not rechargeable 120
- Fast write data protection 118, 119
 - by Fast Write Cache 118, 119
 - by SSA adapter cache 118, 119
 - not by storage server cache 118
- four-way SMP 276

G

- GEM circuit 179
 - items measured 179
- Generalized Parallel File System 267
 - exploits virtual shared disk architecture 267

H

- host adapter 65
 - Ultra SCSI adapter, 32-bit 65
- host bypass circuit 175
- host cache size 269
 - relation to disk subsystem cache size 269
- host caching effectiveness 265
 - reduced in a shared disk environment 265
- host caching effectiveness 264
 - dependence on database management system 264
- host system cache effectiveness 262, 264
- host system software 358, 359
 - OS/400 device driver 359
 - UNIX device driver 358
- hosts supported by Versatile Storage Server 22
- HP-UX 279
 - number of initiators supported 279

I

- input/output operations 258
 - back-end I/O 258
 - front-end I/O from hosts 258

J

- JBOD disks 259

L

- LCD status display 57

- LED status indications 79
- licensed internal code 313
 - concurrent management 313
- locality of reference 296
- Lynx character-based browser 314
 - not supported 314

M

- magnetoresistive head technology 89
- Magstar virtual tape server 25
 - hierarchical storage management system 25
- maintenance 310, 311
- maintenance diagnostics 317
 - run from character-based interface 317
- management information base 322
 - pending faults 322
- management information bases 322
- MESI protocol 53
- microcode upgrade 313
- migrating data 298
 - RAID-1 to RAID-5 298
- migrating data to or from VSS 224
 - carried out by host system 224
 - methods for UNIX hosts 224
 - AIX commands 224
 - software and commands 224
 - requirements 224
- migrating data to VSS 222, 233
 - UNIX commands available 233
- MR 203
 - multihost environment 22
- multivendor environments 9

N

- Netstore 23
 - common parts philosophy 23
- Netstore central control unit 23
 - RS/6000 Model R20 23
- network computing 16
 - data movement as a bottleneck 16
- No-ID sector format 100, 101, 102
 - improved data recovery chances 102
 - increasing capacity of disk drives 102
 - locating the desired sector 101

O

- Oracle Parallel Server 14, 29, 267

P

- parallel query processing 303, 304
 - situation to avoid 304
- PCI local bus 64
- PE 311
 - maintenance and repairs 311

- PFA advantages 178
- PFA measurements 180
 - timing 180
- PFA monitoring schemes 179
- PFA monitors 178
- Power PC 44
- Predictive cache algorithms 133
 - used to manage storage server cache 133
- PRML 203

R

- race condition 298
 - as contention for resources 298
- rack power supplies 313
- RAID adapter 83
 - taking a drive out of service 83
 - drive sparing 83
- RAID adapter disk sparing 313
- RAID Advisory Board 112, 153
- RAID array 83, 112, 113, 285
 - mixed sequential and random workloads 285
 - response to drive failure 83
 - sequential read workloads 285
 - virtual disks 113
- RAID array configurations 81
- RAID configuration 81
 - parity data storage 81
- RAID parity 118
 - resides only in SSA adapters 118
- RAID parity data 119
 - recreated by SSA adapter if necessary 119
- RAID-1 disk subsystem 290
 - read to write ratio 290
- RAID-1 to RAID-5 data migration 298
- RAID-5 adapter 310
 - detection of failed disk drive 310
- RAID-5 array 113, 347
 - data protection 347
 - number of member disks 113
 - parity striping across member disks 113
- RAID-5 arrays 13
 - logical partitions 13
 - dynamic changes in partition size 13
- RAID-5 arrays per SSA adapter 261
 - number recommended 261
- RAID-5 disk subsystem 288, 290
 - read to write ratio 288, 290
 - write penalty for random writes 290
- RAID-5 disk subsystems 259
- RAID-5 write penalty 259, 290
 - avoided by stripe writes 259
 - reduction 290
- redundancy group stripe 113
- remote support interface 318
- repair actions 311
- reuse of existing 7133 and SSA disk drives 225
 - need for added features 225

- RISC 604e CPU 53
 - clock frequency 53
- RISC CPU 54
 - snooping 54
- RISC CPU cache 53
- RISC planar components 52
- RS-232 serial port 318
 - remote support interface 318
- RS-232 serial ports 57, 316
 - as maintenance interface 316
- RS/6000 6
 - 7133 directly attached 6
- RS/6000 AIX 279
 - number of initiators supported 279

S

- SCSI ports 260
 - number per SMP cluster side 260
- SCSI-2 differential fast and wide adapter 6
 - Seascape integrated storage server 21
 - see 'Customer Engineer' CE
 - see 'dynamic random access memory' DRAM
 - See 'EPO' emergency power-off switch
 - See 'Exclusive Or' SSA disk adapter, XOR function
 - see 'generalized error measurement' GEM
 - see 'LBA' VSS sequential access detection, logical block address
 - See 'LCU' logical control units
 - see 'Product Engineer' PE
 - see 'self-monitoring analysis and reporting technology' SMART
 - see 'static random access memory' SRAM
 - see 'VSS' Versatile storage servers
 - see 'CGI' common gateway interface
- sequential I/O 294
 - throughput sensitivity 294
- sequential prediction 62
 - counter checking against preset value 62
- sequential prestage 295
- sequential write 291
 - striping across multiple drives 291
- Serial storage architecture 44
- shared storage 9
- SMART protocol 179
- SMP 52
 - data integrity preservation 52
- SMP cluster 52, 55, 56, 59, 64
 - cache memory 59
 - CD-ROM drive 55
 - diskette drive 56
 - error-correcting code 59
 - Ethernet adapter 56
 - functional parts in RISC planar 52
 - internal SCSI disk drive 55
 - PCI buses, number and speed. 64
- SNMP customer notification of problems 318
- SNMP traps 322

- spatial reuse 107
- SRAM 114
- SSA adapter 114, 115, 122, 123
 - data transfer from disk buffer 123
 - Fast Write Cache, 4 MB 114
 - loop topology 122
 - changes in 122
 - volatile cache 115
 - mirroring Fast Write Cache 115
 - volatile cache in DRAM 115
- SSA adapter cache 122
 - use in RAID-5 write processing 122
- SSA adapter DRAM 121
 - two 16 MB chips 121
- SSA back store 259
 - formula to calculate write I/Os for sequential writes 259
- SSA back store disk 259
 - read operations formula 259
- SSA back store disks 259
 - write I/O load formula 259
- SSA bus 124
 - data transfer from disk to buffer 124
 - interleaved data for different devices 124
- SSA cable 173
- SSA disk 115
 - disk buffer 115
- SSA disk adapter 70
 - RAID array 70
 - amount of data written to one drive 70
- SSA disk adapter 61, 62, 66, 70
 - disk loops supported 66
 - full-stripe destage 61
 - idle-time cache destaging 62
 - parity calculation 70
 - XOR function for calculating parity 70
- SSA disk adapter memory 68, 69
 - DRAM 68
 - SRAM 69
- SSA disk adapters 60
 - cache memory 60
 - Fast Write cache 60
- SSA disk buffer 123
- SSA disks 127
 - formatting by sectors 127
- SSA initiator 67
- SSA node functioning 107
 - three-way router 107
- SSA RAID adapter 184, 186
 - data sparing on disk failure 184
 - data sparing with hot spare disk 186
 - data sparing with operator intervention 186
- storage sharing 13
- storage systems 15
 - platform and vendor independence 15
- StorWatch Versatile storage specialist 18, 205
- strip size 112

- stripe constituents 113
 - data strips 113
 - parity strip 113
- stripe writes 291
 - strips collected in Fast Write Cache 291
- Sun SSA adapter 6
 - 7133 attached 6
- Sun systems running Solaris 279
 - number of initiators supported 279
- synchronous DRAM 59
- synchronous writes 292, 293
 - response-time sensitive 292
 - storage of two copies 293

T

- thin-film disks 87
 - layer constituents 87

U

- Ultrastar 2XP disk drive 176, 181
 - calibration data logs 181
 - disk sweep 181
 - periodic calibration 181
- Ultrastar 2XP disk drives 44, 177
 - disk size effect on performance 177
 - individually addressed 177
- Ultrastar 2XP drives 86, 103
 - embedded servo technology 86
 - features 86
 - predictive failure analysis 86
 - 24 hour notice of pending failure 86
 - self-monitoring and imminent failure warning 103
- Ultrastar disk drive 91, 92, 95
 - overhead reduction 95
 - embedded servo 95
 - read channel as data storage limiter 91
 - Viterbi algorithm in the PRML read channels 92
- Ultrastar disk drives 44, 93, 100, 104
 - areal density across the platter 93
 - innovative features 44
 - maximum data rates across the disk 93
 - measurement-driven self-monitoring 104
 - reducing overhead by using No-ID sector format 100
 - symptom-driven self-monitoring 104
 - zoned bit recording 93
- Ultrastar drive head 90
 - inductive write element and magnetoresistive read element 90
- Ultrastar 2XP drives 86
 - data rate 86
- UNIX 6, 39
 - 7133 attached 6
 - kernel storage in VSS 39
- UNIX file system 143, 258, 259
 - compared with VSS operation 143
 - SCSI writes 258
 - synchronization daemon 258

- UNIX file system (*continued*)
 - sequential processing detection 259
- UNIX file system block size 128
 - defined by system administrator 128
- UNIX file systems 266
 - use of host caching 266
- UNIX host perception of VSS 230
 - emulation of SCSI disk drive 230
- UNIX read request outcomes 263

V

- Versatile storage server 17, 18, 27, 28, 31, 32, 33, 34, 36, 38, 39, 40, 42, 51, 112, 117, 118, 127, 128, 130, 142, 195, 226, 227, 228, 229, 246, 254, 279, 310, 311, 353, 354
 - 7133 disk drives 31
 - 524-byte sectors for AS/400 compatibility 31
 - adaptive caching algorithm 254
 - AS/400 needs to support VSS 227
 - cache management by segment size 128
 - cache not under application control 130
 - cluster read cache 38
 - cache size 38
 - configuration software 354
 - cross-cluster interconnection 34
 - Data General AViiON software needs to support VSS 229
 - detection of internal data corruption 127
 - direct notification of system malfunction 38
 - disk adapter bus 28
 - disk track equivalence to RAID-5 strip size 128
 - emulation of AS/400 system 42
 - emulation of UNIX system 42
 - failover support 34
 - host adapter bus 27
 - PCI bridge 27
 - host hardware needs 226
 - host interface adapter type 33
 - host interface cards 33
 - HP needs to support VSS 228
 - I/O types 142
 - random reads, sequential reads, writes 142
 - individual host storage pools 36
 - influences on storage server cache size 117
 - levels of maintenance and repair actions 311
 - logical control units 38
 - maintenance interface 311
 - ASCII terminal 311
 - remote support interface 311
 - web browser 311
 - maintenance limitations on customers 311
 - maintenance philosophy 310
 - number of adapter bays and adapters 51
 - number of hosts connected 33
 - online cluster microcode updates 34
 - operating system levels for supporting systems 226
 - PCI bus RAID-5 adapter 31

- Versatile storage server (*continued*)
 - platforms supported 353
 - AIX 353
 - DG/UX 353
 - HP-UX 353
 - OS/400 353
 - Solaris 353
 - Windows NT 353
 - platforms supporting VSS 226
 - RAID array subdivision 36
 - RAID-5 array configuration 28
 - removing data from cache 118
 - LRU algorithm 118
 - removing least recently used data 118
 - requirement for homogeneous host types 279
 - reuse of 7133 drawers 246
 - RISC planar boards 31
 - RISC System 6000 needs to support VSS 226
 - Seascape architecture 17
 - SSA disk adapters 40
 - number of SSA loops supported 40
 - SSA RAID adapter 40
 - fast-write cache size 40
 - storage processor bus 27
 - RISC planar with SMP engines 27
 - storage server rack 2105-B09 32
 - strip 112
 - stripe element 112
 - Sun Microsystem needs to support VSS 228
 - Ultra-SCSI adapter 195
 - peak data rates 195
 - realistic operating data rates 195
 - UNIX kernel 39
 - usable storage capacity 18
 - Versatile storage server architecture 21
 - Versatile Storage Server hosts supported 22
 - Versatile storage servers 18
 - supported hosts 18
 - AS/400 18
 - Data General (some models) 18
 - HP 9000 800 series 18
 - RS/6000 and RS/6000 SP 18
 - Sun Ultra series 18
 - Versatile Storage System 248, 257
 - disk drive sizes supported 248
 - improving system performance 257
 - Vicom SLIC 6
 - virtual disks 113
 - virtual shared disk 281
 - limitations on sharing 281
 - virtual shared disk architecture 267
 - ownership of AIX logical volumes 267
 - virtual shared disks 281
 - partitioning indications 281
 - Virtual Storage Server 28, 29, 30
 - file and record level locking 29
 - host attachment definition 30
 - logical disk size definition 30

- Virtual Storage Server (*continued*)
 - number of disk drawers 28
 - disk configuration supported 28
 - reassigning storage pools 30
- VS Specialist 18
- VSS 7133 drawer 171, 173, 174, 175
 - contents 173
 - fault detection 175
 - physical interface to disks 171
 - power supplies and cooling 174
- VSS adaptive cache read options 131
- VSS component failure 333
 - warning and customer notification 333
- VSS component replacement 313
- VSS configuration 192, 193, 194, 195, 196, 197, 199, 200, 201, 203, 205, 206, 207, 209, 210, 211, 212, 213, 215, 216, 217, 218, 219
 - access to the configuration manager 216
 - adding disk drawers 200
 - application server view 205
 - base configuration 199
 - cable choices 194
 - cache memory in SMP clusters 197
 - compromise and its impacts 192
 - connection kit choice 194
 - connection to and from the host 219
 - customer connection 217
 - data sharing control by application code 196
 - disk adapters 199
 - disk array representation 205
 - StorWatch Versatile Storage Specialist 205
 - disk drawer power cords 213
 - disk drawers 210
 - disk drive features 203
 - disk drive reformatting 211
 - disk drive size influences 203
 - disk sizes 203
 - expansion rack 209
 - expansion rack power supply 209
 - Fast Write cache 207
 - first storage server rack, 2105-B09 207
 - host adapter cards 207
 - host interface adapter 193
 - host system support 193
 - AIX 193
 - HP/UX 193
 - Solaris 193
 - influences 192
 - availability 192
 - budget constraints 192
 - performance 192
 - logical disks 192
 - number and size 192
 - LUN choices 205
 - LUN limitation 205
 - main influences on 195
 - maximum configuration 212
 - number of hosts possible 193
- VSS configuration (*continued*)
 - number of 604eCPUs 197
 - number of disk adapters 199
 - number of SMP process boards 197
 - one SSA loop per disk drawer 199
 - physical dimensions 215
 - power control subsystem 207
 - power supplies 213
 - RAID array configuration options 201
 - RAID-5 disk adapters 207
 - RAID-5 protected storage 207
 - rectifying wrong sizing assumptions 218
 - redundancy to protect data availability 195
 - SSA cable lengths 210
 - storage pools per server 205
 - storage server memory 207
 - system backup needs as an influence 195
 - system size selection 192
 - workload influence on RAID array assignment 206
- VSS configuration manager 336
- VSS configuration menus 314
- VSS disks 249
 - formatted in 524-byte sectors 249
- VSS drawer 171, 188
 - design 171
 - loop connection between disk drives 188
- VSS maintenance 318, 320, 321, 322, 323, 325, 327, 328, 329, 330, 331, 332, 334, 338, 339
 - access restrictions 338
 - call-home feature 318, 325
 - concurrent maintenance 334
 - conditions that require service action 321
 - customer repair actions 327
 - customer repair limitations 328
 - disk mirroring to protect data 339
 - engineering change levels of subsystem components 332
 - engineering change management 331
 - error log analysis 318
 - error reporting and error log 320
 - exception symptom codes 323
 - field replaceable units 323
 - management information base, configuration 322
 - online menus for customer engineer use 330
 - primary interface for the CE and PE 329
 - problem record 321
 - remote access 318
 - remote support interface 331
 - emergency code fixes only 331
 - security implications of call-home feature 325
 - service request numbers 323
 - storage server service processor 329
 - access for customer engineer 329
 - storage server system management services 329
- VSS maintenance tools 312
 - call-home function 312
 - e-mail notification of users 312
 - error log 312

- VSS maintenance tools (*continued*)
 - SNMP alerts 312
- VSS maximum usable capacity 172
- VSS microcode 342
 - processed within storage server clusters 342
- VSS nonvolatile storage 114
- VSS operating microcode 317
 - installed from character interface 317
- VSS predictive access management 133, 134
 - adaptive caching algorithm 134
 - anticipatory data staging 133
- VSS racks 45, 172
 - 2105-100 45, 172
 - 2105-B09 45, 172
- VSS random read processing 144, 145, 146
- VSS remote diagnosis and analysis 336
- VSS remote service 335
 - call-home feature 335
 - dedicated phone line recommendation 335
- VSS remote services 336
 - PE support may entail root access 336
- VSS SCSI bus throughput 279
- VSS sequential access detection 255
 - logical block address 255
- VSS sequential read processing 147, 148
- VSS shutdown destaging 162
 - data moved from Fast Write Cache to disk 162
- VSS status screen 315
 - maintenance options shown 315
- VSS storage server 260, 272
 - cache size configuration 260
 - calculating usable capacity of drawers 272
- VSS subsystem 46, 60, 61, 62, 65, 66, 72, 73, 74, 79, 85, 223, 224, 225, 259, 270, 274, 276, 278, 282, 284, 287, 288, 296, 310, 333, 337, 344, 345, 346, 347, 348, 349, 350
 - 7133 SSA disk drawer 72
 - drive configurations 72
 - 7133 SSA disk drawer power supply 74
 - 7133-020 SSA disk drawer 73
 - advantages 73
 - 7133-10 SSA drawers 73
 - access density effect on I/O rate 276
 - configurations to provide needed capacity 276
 - access density for different disk drives 284
 - adaptive cache algorithm 62
 - adaptive caching 296
 - avoiding SCSI bus arbitration overhead 278
 - battery backup for racks 349
 - built-in diagnostic and warning systems 333
 - cache size for less effective host caching 270
 - cache size options 60
 - cache size relative to host cache size 274
 - caching algorithms 61
 - calculating cache size given less effecting host caching 270
 - calculating number of drawers 287
 - component replacement without service disruption 310
- VSS subsystem (*continued*)
 - connection to host 223
 - required for data storage 223
 - data availability using alternative paths 337
 - data transfer to free space 61
 - disk capacity choice 282
 - disk drive module 79
 - disk drive types supported 85
 - Ultrastar 2XP drives 85
 - disk specifications 282
 - effect of raising cache size on I/O performance 270
 - formula to calculate back end I/O load 259
 - full track staging 62
 - host responsibility for data backup 350
 - need for enough power outlets 223
 - need for spare drives 225
 - number of drawers 287
 - number of host adapters included 65
 - number of SSA adapters 287
 - other possible racks 46
 - partial track staging 62
 - predicting sequential access 62
 - read to write ratio 288
 - record mode staging 62
 - response to cache failure 344
 - response to cluster processor failure 345
 - response to disk adapter failure 346
 - response to power failure 348
 - response to SSA cable failure 347
 - reuse of existing disk drives 224
 - sector size of 524 bytes 224
 - SSA disk adapter 66
 - RAID-5 adapter with special firmware 66
 - synchronous writes 288
 - throughput with sequential workloads 288
- VSS subsystem accessed by different hosts 223
 - need for host shut down to make connection 223
- VSS subsystem configuration 223
 - using browser network client 223
 - using CE interface 223
- VSS subsystem failures 334
 - failure types possible 334
- VSS subsystem maintenance 313
 - SSA RAID adapter disk sparing 313
- VSS subsystem recovery 333, 335, 337, 340, 342
 - automatic failover 340
 - data integrity assurance 333
 - failover controlled by microcode 342
 - host connection failures 337
 - object data manager 342
 - remote services 335
- VSS subsystem replacement 351
 - configuration restrictions 351
- VSS subsystem software 354, 355
 - four-port SSA RAID adapter 354
 - SSA logical disk device driver 355
 - SSA physical disk device driver 355

- VSS subsystem software (*continued*)
 - SSA router 354
- VSS subsystem to replace existing storage 224
 - virtual disk configuration needed 224
- VSS subsystems 277
 - reasons to choose four-way SMPs 277
- VSS task-complete signal to host 255
- VSS track blocks 129
 - use of cache segments 129
- VSS UltraSCSI adapters 279
 - speed and compatibility 279
- VSS write options 132
 - update writes compared with sequential writes 132
- VSS write processing 149, 150, 151, 152, 155, 157, 158, 163
 - data removal from Fast Write Cache 163
 - data removal from SSA adapter cache 163
 - ensuring data integrity 150
 - fast write 149
 - masking of RAID-5 write penalty 152
 - not fast write 149
 - RAID-5 penalty and update writes 151
 - sequential writes 158
 - stripe writes 155, 157

W

- Web Cache Manager 24
 - RS/6000 central control unit 24
 - Web Traffic Express Software 24
- workload aspects that can affect VSS configuration 253

ITSO Redbook Evaluation

IBM Versatile Storage Server
SG24-2221-00

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!)**

